STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:2377–2397

Posterior predictive model checks for disease mapping models

Hal S. Stern[1,*,†] and Noel Cressie[2]

[1]*Department of Statistics, Iowa State University, Ames, IA 50011-1210, U.S.A.*
[2]*Department of Statistics, Ohio State University, Columbus, OH 43210-1247, U.S.A.*

SUMMARY

Disease incidence or disease mortality rates for small areas are often displayed on maps. Maps of raw rates, disease counts divided by the total population at risk, have been criticized as unreliable due to non-constant variance associated with heterogeneity in base population size. This has led to the use of model-based Bayes or empirical Bayes point estimates for map creation. Because the maps have important epidemiological and political consequences, for example, they are often used to identify small areas with unusually high or low unexplained risk, it is important that the assumptions of the underlying models be scrutinized. We review the use of posterior predictive model checks, which compare features of the observed data to the same features of replicate data generated under the model, for assessing model fitness. One crucial issue is whether extrema are potentially important epidemiological findings or merely evidence of poor model fit. We propose the use of the cross-validation posterior predictive distribution, obtained by reanalysing the data without a suspect small area, as a method for assessing whether the observed count in the area is consistent with the model. Because it may not be feasible to actually reanalyse the data for each suspect small area in large data sets, two methods for approximating the cross-validation posterior predictive distribution are described. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

Disease incidence or disease mortality rates for a collection of geographic areas are commonly displayed on maps. For example, one might see a map of cancer incidence rates for counties in the United States. Maps of raw rates, defined as disease incidence or mortality counts divided by at-risk population, have been criticized as unreliable due to non-constant variance associated with heterogeneity in at-risk population sizes. This has led to the suggestion that model-based Bayes or empirical Bayes point estimates be used to create maps [1–8], often maps of relative risk. Because such maps have important epidemiological and political consequences, it is important that the

*Correspondence to: Hal S. Stern, Department of Statistics, Iowa State University, Ames, IA 50011-1210, U.S.A.
†E-mail: hstern@iastate.edu

Contract/grant sponsor: Environmental Protection Agency; contract/grant number: CR822919-01-0
Contract/grant sponsor: Office of Naval Research; contract/grant number: N00014-99-1-0214
Contract/grant sponsor: National Cancer Institute; contract/grant number: CA78169-02

assumptions of the underlying models be scrutinized. For example, it is often of interest to identify regions corresponding to extremely high or low unexplained risk of disease, and, consequently, it is important to determine whether observed extrema among the estimated risks are too large (or too small) to have occurred by chance under the model. If the observed value is too large (or too small), then the model is incorrect either because an important (maybe undiscovered) risk factor has been omitted or because one of the statistical assumptions is inappropriate. In the former case we should investigate the site in question, whereas in the latter case respecification of the model is the remedy. There is a potential problem in that we may end up always blaming the statistical assumptions of the model for the existence of extreme values and consequently fail to investigate potentially interesting sites. Here, we ask whether it is possible to use diagnostics for all of the regions, rather than just a single region, to determine whether the model is flawed. In that way, we hope that areas with extreme estimates will still be identified for further study while cutting down on false candidates suggested by invalid models.

We use a common statistical model for disease data in which the observed count of disease cases (incidence or mortality) in a small area is assumed to be a Poisson random variable with mean equal to the product of the expected number of cases (based on known risk factors) and a relative risk parameter. The logarithms of the relative risk parameters are assumed to follow a Gaussian distribution with mean that may incorporate potentially relevant risk factors, and variance matrix that incorporates the possibility of spatial dependence. The spatial dependence allows for the accommodation of correlation induced by unmeasured covariates, for example. Examples of this form of model are those described by Besag *et al.* [4] and Stern and Cressie [9, 10]. These models may also be extended to accommodate repeated observations over time (see, for example, Waller *et al.* [11] and Knorr-Held and Besag [12]), but we do not consider that case here.

The adequacy of such a model can be addressed using posterior predictive model checks as described by Rubin [13] and Gelman *et al.* [14, 15]. The posterior predictive approach to assessing model fit compares features of the observed data to the same features of replicate data generated under the model. Posterior predictive checks are easily carried out given simulations from the posterior distribution of the model parameters; such simulations are often available from a Bayesian analysis of the data under the model. We argue that posterior predictive model checks are useful for assessing the overall fit of a model and for checking specific assumptions. As generally applied, however, they are less useful for assessing whether there exist one or more extreme observations inconsistent with the model. We propose a cross-validation posterior predictive approach to assessing individual observations. The basic idea is to assess the model's fit to the count in a given area by comparing the observed disease count in that area with a predictive distribution obtained by reanalysing the original data without the area in question. One potential disadvantage of this approach is that for large data sets there may be many suspect areas and consequently a large number of additional data analyses. To avoid refitting the model without each small area, we describe the use of importance weighting and importance resampling to approximate the posterior distribution that would be obtained if the analysis were repeated without the small area.

Section 2 reviews notation and statistical models for analysing disease-incidence and disease-mortality data, concentrating on the Poisson-log-Gaussian model. The basic model is applied to a well known data set, the Scotland lip cancer data of Clayton and Kaldor [2]. Section 3 describes posterior predictive model checking and how it can be applied in the disease mapping context. There, the primary focus is on overall questions of fit and model specification. The important question of identifying whether regions with extreme rates are evidence of general model failure is addressed using cross-validation in Section 4. The model checking and cross-validation techniques

are applied to the Scotland lip cancer data in Section 5. We provide some concluding remarks in Section 6.

2. MODELS FOR DISEASE MAPPING

2.1. A Poisson-log-Gaussian model

For purposes of discussion we use the terminology associated with analysing disease incidence data. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ represent the vector of observed cases of a disease from n geographical regions or districts and let $\boldsymbol{E} = (E_1, \ldots, E_n)$ indicate expected counts based on known risk factors. The expected counts $\boldsymbol{E}$ represent a form of standardization of the data. Introduce $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ as a vector of relative risk parameters. Then, conditional on the expected counts and the relative risk parameters, we model the observed cases as independent Poisson random variables

$$Y_i, \mid \lambda_i, E_i \sim \text{Poisson}(\lambda_i E_i); \quad i = 1, \ldots, n \tag{1}$$

Because the expected counts $\boldsymbol{E}$ play an important role in the development of cross-validation diagnostics, we take a brief digression here to explain them further. Suppose that we consider the population at risk of developing the disease as consisting of K demographic groups or strata. These may be defined, for example, based on age and gender. Define q_k to be the proportion of the at-risk population in stratum k expected to develop the disease. If the population at risk in region i is R_i with R_{ik} people in the kth stratum, then the expected count for region i is

$$E_i = \sum_{k=1}^{K} R_{ik} q_k \tag{2}$$

Of course the q_k's need to be estimated. If estimates are available from a source outside of the current data set (perhaps an international data registry), then the data are said to be externally standardized. A common alternative is to use internal standardization whereby the same population that yields the data $\boldsymbol{Y}$ are also used to estimate the rates for the strata. An obvious estimator for q_k is the proportion of the at-risk population in stratum k (totalled over all regions) that developed the disease. To formally define this estimator, let O_{jk} denote the number of observed disease cases from region j in the kth stratum; notice that $Y_j = \sum_k O_{jk}$. Then the sample proportions are

$$\hat{q}_k = \frac{\sum_{j=1}^{n} O_{jk}}{\sum_{j=1}^{n} R_{jk}}; \quad k = 1, \ldots, K \tag{3}$$

It is easy to verify that internal standardization using the sample proportions introduces a form of dependence among the elements of $\boldsymbol{Y}$, in that $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} E_i$. It is common practice to analyse the data conditional on the E_i, ignoring the dependence, and we follow that practice here. In general we do not explicitly include this conditioning in our notation.

The standardized morbidity ratio (SMR) for the ith area is defined as $\text{SMR}_i \equiv Y_i/E_i$; $i = 1, \ldots, n$. The relative risk parameters, $\boldsymbol{\lambda}$, can be thought of as smooth versions of the SMRs. The relative risk parameters are modelled as having a joint population (or prior) distribution. Since the λ_i are positive, we place a prior distribution on the logarithms of the relative risks, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, where $\theta_i = \ln \lambda_i$. Assume

$$\boldsymbol{\theta} \mid \boldsymbol{\beta}, \tau^2, \phi \sim \text{Gau}(X\boldsymbol{\beta}, \Phi\tau^2) \tag{4}$$

where X is an $n \times p$ matrix of predictor variables, $\boldsymbol{\beta}$ is the unknown regression coefficient vector, and $\Phi\tau^2$ is a variance matrix that allows for spatial dependence. The p columns of the covariate matrix X reflect factors thought to be associated with variation in disease incidence rates (after the rates have been adjusted for the known demographic risk factors incorporated in $\boldsymbol{E}$).

The matrix Φ in the variance of the prior distribution (4) can be parameterized according to a spatial-dependence model, namely $\Phi = (I - \phi C)^{-1}M$, where $C = (c_{ij})$ is a spatial-association matrix with zeros on the diagonal, ϕ is a parameter measuring spatial dependence, and M is a known diagonal matrix chosen so that Φ is positive-definite. Examination of Φ^{-1} indicates that Φ is symmetric as long as $m_{jj}c_{ij} = m_{ii}c_{ji}$. Note that the variance matrix can be expressed as $\Phi = M^{1/2}(I - \phi M^{-1/2}CM^{1/2})^{-1}M^{1/2}$. Then the variance matrix is positive-definite for $\phi \in (\phi_{\min}, \phi_{\max})$, where the upper and lower limits are determined by the eigenvalues of $M^{-1/2}CM^{1/2}$ (Section 7.6 of Cressie [16]). There is considerable flexibility in the choice of M and C for defining the variance matrix in the model for $\boldsymbol{\theta}$. In the remainder of this paper we take $c_{ij} = (E_j/E_i)^{1/2}$ if area j is a neighbour of area i and zero otherwise, and $m_{ii} = E_i^{-1}$ for $i = 1, \ldots, n$. This leaves $\tau^2 > 0$ and ϕ as free parameters, as was done in Stern and Cressie [9, 10]. The Gaussian model on $\boldsymbol{\theta}$ is an example of the conditional autoregressive model (see, for example, Besag [17] and Section 6.6 of Cressie [16]). Let $N_i \equiv \{j : c_{ij} \neq 0\}$ represent the 'neighbours' of i and $\boldsymbol{\theta}_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)'$; then

$$\theta_i \mid \boldsymbol{\theta}_{-i} \sim \mathrm{N}\left((X\boldsymbol{\beta})_i + \phi \sum_{j \in N_i} c_{ij}(\theta_j - (X\boldsymbol{\beta})_j), \tau^2 m_{ii} \right); \quad i = 1, \ldots, n \tag{5}$$

where $(X\boldsymbol{\beta})_i$ is the ith element of the vector $X\boldsymbol{\beta}$. The conditional autoregressive model assumes there is a linear association between the logarithm of the relative risk in region i and the logarithms of the relative risks in neighbouring regions. The parameter ϕ and the matrix C determine the degree of association. In most disease-mapping applications, negative values of ϕ seem unlikely so that we restrict $\phi > 0$.

To complete the Bayesian model specification, we require a prior distribution on the remaining parameters $(\boldsymbol{\beta}, \tau^2, \phi)$. We take the non-informative prior distribution corresponding to independent flat prior distributions for $\boldsymbol{\beta}$ and ϕ and a nearly flat prior distribution for τ:

$$p(\boldsymbol{\beta}, \tau^2, \phi) \propto \mathrm{e}^{-\varepsilon/\tau^2} \quad \text{for} \quad \phi \in (0, \phi_{\max}), \tau^2 > 0 \tag{6}$$

with $\varepsilon = 0.01$. This is not a proper distribution but it does lead to a proper posterior distribution.

The model we consider here is a special case of a more general class of models that assumes $\mathrm{var}(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \tau^2, \sigma^2, \phi) = D\sigma^2 + \Phi\tau^2$, where $D\sigma^2$ is a diagonal matrix accounting for inhomogeneities in risk not presumed to be spatially correlated. This broader class also includes the popular model of Besag *et al.* [4].

2.2. Posterior inference

In the remainder of this paper, we take the model to be that given by (1), (4), (6). We can obtain posterior inferences for the parameters of the model by simulating from the posterior distribution using Markov chain Monte Carlo. For an overview of Bayesian modelling and computation, see Gelman *et al.* [14], Carlin and Louis [18], and Gilks *et al.* [19]. We use a Gibbs sampling

algorithm, wherein the conditional posterior distributions of each parameter (or set of parameters) given all of the others are used in succession to construct a Markov chain in the parameter space that has the joint posterior distribution as its stationary distribution. After the chain has been run for a sufficiently long time, the simulated draws may be taken as being representative of the joint posterior distribution. In this case, the conditional distributions of $\boldsymbol{\beta}$ and τ^2 are easily recognized as normal and inverse chi-squared distributions, respectively. The remaining conditional distributions are not standard distributions so that alternative algorithms are required. For the example, we used univariate Metropolis–Hastings steps for the θ_i. The conditional distribution of ϕ is quite complex because of the way ϕ appears in the covariance matrix. It too is sampled within the Gibbs sampling algorithm using a Metropolis–Hastings step. Convergence was assessed from independent Markov chain simulations using the approach of Gelman and Rubin [20].

2.3. Example: Scotland lip cancer data

The Poisson-log-Gaussian model is used to analyse data representing male lip cancer rates (over the period 1975–1980) in the $n = 56$ districts of Scotland. These are the districts prior to the 1995 reorganization of local government. Table I repeats the lip cancer data (originally analysed by Clayton and Kaldor [2]) from Breslow and Clayton [21] with district names provided in Cressie [16]. The table includes district names and identifying numbers, and for district i with $i = 1, \ldots, 56$: the number of observed cases Y_i; the number of expected cases E_i; the standardized morbidity rate $\mathrm{SMR}_i = Y_i/E_i$; the value of a single covariate (per cent of population employed in agriculture, fishing and forestry) X_i; and a list of the neighbouring regions. The expected counts are computed using a form of internal standardization (the method of Mantel and Stark [22]) to adjust for the age distribution in the districts. We apply the Poisson-log-Gaussian model of Section 2.1 with the matrix X consisting of a column of ones corresponding to an intercept and the single covariate in Table I. For the neighbourhood structure indicated in Table I and the choices of C and M described in Section 2.1 we find that $\phi_{\max} = 0.175$.

One thousand simulations from the posterior distribution were obtained using the algorithm described in Section 2.2; the results are summarized in Table II. In addition, histograms showing the posterior distribution of λ_i for four districts of special interest are given in Figure 1.

These correspond to the districts with largest and smallest values of SMR_i, Skye-Lochalsh (district 1) and Annandale (district 55), respectively, and the districts with largest and smallest values of E_i, Glasgow (district 49) and Badenoch (district 17), respectively. The former are included as obvious candidates for values that might be considered extreme. The latter are of interest because under our model E_i is a key factor in determining the posterior variability of λ_i. Thus the posterior distribution for λ_{49} has a very narrow range; the 95 per cent central posterior interval is $(0.30, 0.52)$. District 17 and district 1 have much wider posterior intervals because they have smaller expected counts.

The results in Table II indicate that Skye-Lochalsh (district 1) has an extremely high relative risk. The central 95 per cent posterior interval for the relative risk parameter is (3.0, 11.1). Part of this relative risk is associated with the covariate, however the posterior distribution of $e^{-(X\beta)_1}\lambda_1$, the relative risk unrelated to the covariate, is also quite high, with central 95 per cent posterior interval (1.8,8.4). It is natural to wonder if the extreme risk is an indication of model failure. In the remaining sections we develop methods for assessing the fit of the model, especially for determining whether observed extreme values are too large to have occurred by chance under the model.

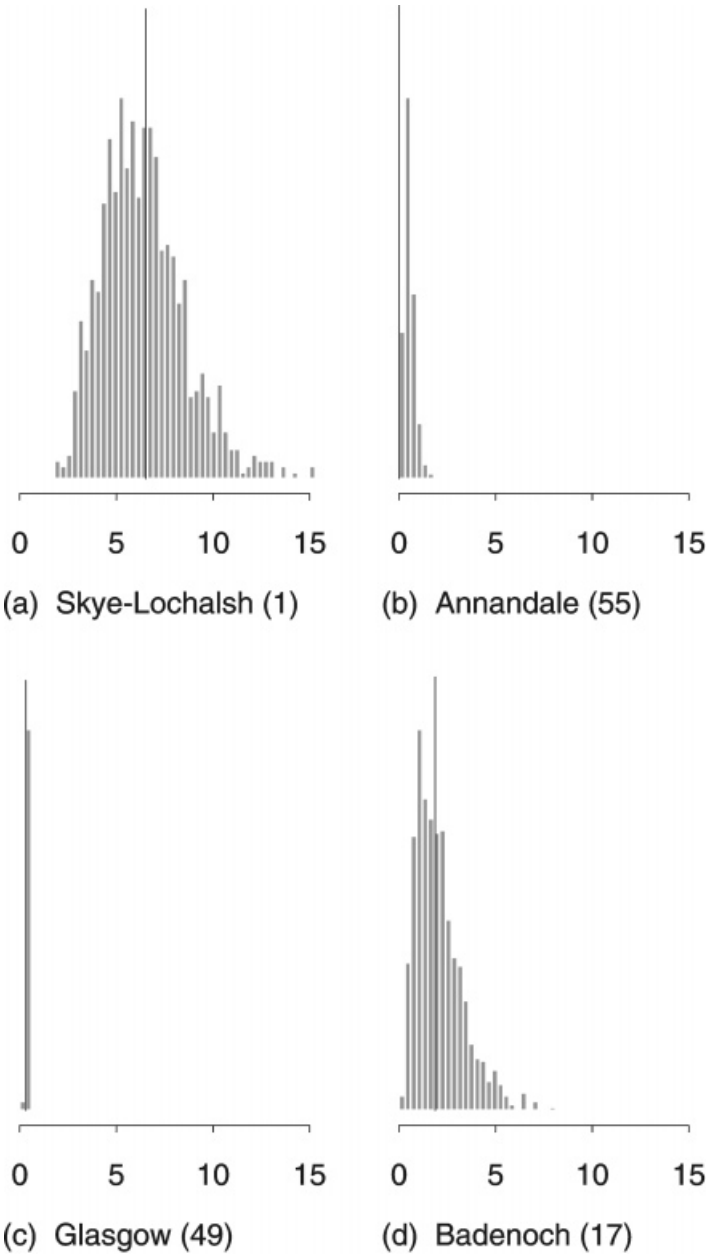

Figure 1. Posterior distributions of relative risk parameters, λ_i, for four districts estimated using 1000 realizations simulated from the posterior distribution. Distributions are presented on a common horizontal scale. Observed values of the standardized mortality rate, $\mathrm{SMR}_i = Y_i/E_i$, are indicated by a solid vertical line on each plot. (*a*) District 1, Skye-Lochalsh, has the largest SMR_i (and a fairly small E_i). (*b*) District 55, Annandale, has the smallest SMR_i. (*c*) District 49, Glasgow, has the largest E_i. (*d*) District 17, Badenoch, has the smallest E_i.

Table I. Scotland lip cancer data.

ID	District name	Y	E	SMR	X	Neighbours
1	Skye-Lochalsh	9	1.38	6.52	16	5,9,11,19
2	Banff-Buchan	39	8.66	4.50	16	7,10
3	Caithness	11	3.04	3.62	10	6,12
4	Berwickshire	9	2.53	3.56	24	18,20,28
5	Ross-Cromarty	15	4.26	3.52	10	1,11,12,13,19
6	Orkney	8	2.40	3.33	24	3,8
7	Moray	26	8.11	3.21	10	2,10,13,16,17
8	Shetland	7	2.30	3.04	7	6
9	Lochaber	6	1.98	3.03	7	1,11,17,19,23,29
10	Gorden	20	6.63	3.02	16	2,7,16,22
11	Western Isles	13	4.40	2.95	7	1,5,9,12
12	Sutherland	5	1.79	2.79	16	3,5,11
13	Nairn	3	1.08	2.78	10	5,7,17,19
14	Wigtown	8	3.31	2.42	24	31,32,35
15	NE Fife	17	7.84	2.17	7	25,29,50
16	Kincardine	9	4.55	1.98	16	7,10,17,21,22,29
17	Badenoch	2	1.07	1.87	10	7,9,13,16,19,29
18	Ettrick	7	4.18	1.67	7	4,20,28,33,55,56
19	Inverness	9	5.53	1.63	7	1,5,9,13,17
20	Roxburgh	7	4.44	1.58	10	4,18,55
21	Angus	16	10.46	1.53	7	16,29,50
22	Aberdeen	31	22.67	1.37	16	10,16
23	Argyll-Bute	11	8.77	1.25	10	9,29,34,36,37,39
24	Clydesdale	7	5.62	1.25	7	27,30,31,44,47,48,55,56
25	Kirkcaldy	19	15.47	1.23	1	15,26,29
26	Dunfermline	15	12.49	1.20	1	25,29,42,43
27	Nithsdale	7	6.04	1.16	7	24,31,32,55
28	East Lothian	10	8.96	1.12	7	4,18,33,45
29	Perth-Kinross	16	14.37	1.11	10	9,15,16,17,21,23,25,26,34,43,50
30	West Lothian	11	10.20	1.08	10	24,38,42,44,45,56
31	Cumnock-Doon	5	4.75	1.05	7	14,24,27,32,35,46,47
32	Stewartry	3	2.88	1.04	24	14,27,31,35
33	Midlothian	7	7.03	1.00	10	18,28,45,56
34	Stirling	8	8.53	0.94	7	23,29,39,40,42,43,51,52,54
35	Kyle-Carrick	11	12.32	0.89	7	14,31,32,37,46
36	Inverclyde	9	10.10	0.89	0	23,37,39,41
37	Cunninghame	11	12.68	0.87	10	23,35,36,41,46
38	Monklands	8	9.35	0.86	1	30,42,44,49,51,54
39	Dumbarton	6	7.20	0.83	16	23,34,36,40,41
40	Clydebank	4	5.27	0.76	0	34,39,41,49,52
41	Renfrew	10	18.76	0.53	1	36,37,39,40,46,49,53
42	Falkirk	8	15.78	0.51	16	26,30,34,38,43,51
43	Clackmannan	2	4.32	0.46	16	26,29,34,42
44	Motherwell	6	14.63	0.41	0	24,30,38,48,49
45	Edinburgh	19	50.72	0.37	1	28,30,33,56
46	Kilmarnock	3	8.20	0.37	7	31,35,37,41,47,53
47	East Kilbride	2	5.59	0.36	1	24,31,46,48,49,53
48	Hamilton	3	9.34	0.32	1	24,44,47,49
49	Glasgow	28	88.66	0.32	0	38,40,41,44,47,48,52,53,54
50	Dundee	6	19.62	0.31	1	15,21,29

Table I. Continued.

ID	District name	Y	E	SMR	X	Neighbours
51	Cumbernauld	1	3.44	0.29	1	34,38,42,54
52	Bearsden	1	3.62	0.28	0	34,40,49,54
53	Eastwood	1	5.74	0.17	1	41,46,47,49
54	Strathkelvin	1	7.03	0.14	1	34,38,49,51,52
55	Annandale	0	4.16	0.00	16	18,20,24,27,56
56	Tweeddale	0	1.76	0.00	10	18,24,30,33,45,55

Table II. Posterior inference for $\boldsymbol{\lambda}$ and other parameters of the Poisson-log-Gaussian model for the Scotland lip cancer data.

Parameter	Posterior distribution			Parameter	Posterior distribution		
	2.5%	Median	97.5%		2.5%	Median	97.5%
λ_1	3.01	6.14	11.14	λ_{29}	0.76	1.16	1.70
λ_2	2.93	4.08	5.44	λ_{30}	0.54	0.97	1.61
λ_3	1.55	3.07	5.40	λ_{31}	0.38	0.89	1.78
λ_4	1.76	3.42	5.94	λ_{32}	0.47	1.26	2.86
λ_5	1.82	3.15	4.98	λ_{33}	0.47	0.93	1.64
λ_6	1.61	3.25	5.85	λ_{34}	0.37	0.76	1.42
λ_7	1.93	2.81	4.04	λ_{35}	0.48	0.83	1.34
λ_8	1.09	2.50	4.90	λ_{36}	0.39	0.71	1.22
λ_9	1.03	2.70	5.29	λ_{37}	0.53	0.87	1.45
λ_{10}	1.83	2.89	4.18	λ_{38}	0.30	0.61	1.13
λ_{11}	1.41	2.58	4.18	λ_{39}	0.55	0.99	1.71
λ_{12}	0.95	2.69	5.44	λ_{40}	0.22	0.54	1.21
λ_{13}	0.68	2.52	6.28	λ_{41}	0.29	0.50	0.77
λ_{14}	1.13	2.30	4.26	λ_{42}	0.47	0.77	1.20
λ_{15}	1.11	1.79	2.83	λ_{43}	0.23	0.72	1.60
λ_{16}	1.05	1.99	3.30	λ_{44}	0.23	0.43	0.73
λ_{17}	0.46	1.83	5.13	λ_{45}	0.33	0.47	0.65
λ_{18}	0.59	1.29	2.68	λ_{46}	0.22	0.48	0.97
λ_{19}	0.83	1.52	2.71	λ_{47}	0.13	0.35	0.84
λ_{20}	0.68	1.36	2.58	λ_{48}	0.18	0.39	0.74
λ_{21}	0.84	1.32	2.07	λ_{49}	0.30	0.40	0.52
λ_{22}	1.05	1.43	1.96	λ_{50}	0.27	0.47	0.74
λ_{23}	0.65	1.15	1.88	λ_{51}	0.09	0.34	0.98
λ_{24}	0.44	0.94	1.81	λ_{52}	0.09	0.31	0.92
λ_{25}	0.68	1.03	1.51	λ_{53}	0.10	0.29	0.72
λ_{26}	0.59	0.95	1.50	λ_{54}	0.11	0.30	0.66
λ_{27}	0.48	0.99	1.79	λ_{55}	0.13	0.48	1.20
λ_{28}	0.55	0.99	1.65	λ_{56}	0.04	0.31	1.38
τ^2	1.23	2.23	4.18	β_1	−0.899	−0.566	−0.209
ϕ	0.040	0.146	0.174	β_2	0.036	0.062	0.090

3. POSTERIOR PREDICTIVE MODEL CHECKING

3.1. Preliminary remarks

Model checking is a broad term that encompasses a large number of ideas for determining if a model provides an adequate fit to a particular data set. One particularly powerful approach

for checking the fit of a model is to embed the model inside a larger model by introducing one or more additional parameters. It is then possible to fit the larger model and examine the posterior distribution of the added parameters to determine if the existing model is adequate. If these parameters do not differ from their null value, often zero, then we might be content to use the smaller model. In the disease mapping context, this approach could for example be used to assess which of several potential covariates should be included in the final model.

An idea closely related to model checking is model selection, in which the goal is to select the best of a set of models. The most common approach to model selection under the Bayesian approach to inference relies on Bayes factors (reviewed for example by Kass and Raftery [23]). Again, one can easily imagine taking this approach to select the best subset from among a given set of covariates. Bayes factors can be used to compare the 2^p models corresponding to the inclusion or exclusion of each of p covariates.

In the present context, we assume that only a single model is being fit and thus do not consider the use of Bayes factors. Instead we ask whether the proposed model fits the observed data. It is similar to a traditional significance-testing approach in the sense that a specific alternative model is not specified. If test statistics or measures are constructed carefully, then a failure of the proposed model may suggest some way of extending the model, but we assume that there are no specific alternative models under consideration. The remainder of this section briefly reviews posterior predictive model checks and describes their application to disease mapping.

3.2. Posterior predictive model checks

The goal in model checking is to determine whether the observed data are representative of the type of data we might expect under the model. Model fit can be assessed using draws from the posterior predictive distribution [13, 15] to represent what we can expect under the model. Let $\boldsymbol{Y}^{\text{rep}}$ denote a replication of the data with the same (unknown) values of the parameters that produced the data $\boldsymbol{Y}$. The posterior predictive distribution of $\boldsymbol{Y}^{\text{rep}}$ is defined as

$$p(\boldsymbol{Y}^{\text{rep}} \mid \boldsymbol{Y}) = \int p(\boldsymbol{Y}^{\text{rep}} \mid \boldsymbol{\eta}, \boldsymbol{Y})\, p(\boldsymbol{\eta} \mid \boldsymbol{Y})\, \mathrm{d}\boldsymbol{\eta} = \int p(\boldsymbol{Y}^{\text{rep}} \mid \boldsymbol{\eta})\, p(\boldsymbol{\eta} \mid \boldsymbol{Y})\, \mathrm{d}\boldsymbol{\eta}, \tag{7}$$

where $\boldsymbol{\eta}$ is used as generic notation for all of the model parameters and the second equality reflects assumed conditional independence of $\boldsymbol{Y}^{\text{rep}}$ and $\boldsymbol{Y}$ given the parameters. In practice, we study the posterior predictive distribution via simulation. Draws from the posterior distribution of $\boldsymbol{\eta}$ are typically available from the Markov chain Monte Carlo procedure described in Section 2.2. Then a replicate data set is obtained from each draw of $\boldsymbol{\eta}$ using $p(\boldsymbol{Y}^{\text{rep}} \mid \boldsymbol{\eta})$.

To assess the fit of the model we introduce $T(\boldsymbol{Y}; \boldsymbol{\eta})$ as a discrepancy measure that is intended to measure the fit of the model to the data. For example, T may be an overall measure of fit or a measure designed to tell whether a particular source of variability is adequately addressed by the model. Note that we do not restrict attention to test statistics in the formal sense; our notation allows T to depend on the parameters and the data. The fit of the model with respect to the discrepancy T is judged by comparing the posterior distribution of $T(\boldsymbol{Y}; \boldsymbol{\eta})$ (recall that T is a function of the parameters so it has a posterior distribution) to the posterior predictive reference distribution $T(\boldsymbol{Y}^{\text{rep}}; \boldsymbol{\eta})$. The joint posterior distribution of $T(\boldsymbol{Y}^{\text{rep}}; \boldsymbol{\eta})$ and $T(\boldsymbol{Y}; \boldsymbol{\eta})$ can be studied empirically using simulations, with the simulations displayed in a scatter plot. If the points in the scatter plot are far removed from the 45 degree line, then the data generated by the model do not

resemble the observed data as regards the measure T. One summary of that joint distribution is the posterior predictive p-value

$$p_{\text{pp}} = \Pr(T(\boldsymbol{Y}^{\text{rep}}; \boldsymbol{\eta}) \geqslant T(\boldsymbol{Y}; \boldsymbol{\eta}) \mid \boldsymbol{Y}) \tag{8}$$

where the probability is calculated over the posterior distribution of $(\boldsymbol{\eta}, \boldsymbol{Y}^{\text{rep}})$. Extremely small posterior predictive p-values indicate a clear rejection of the current model. More moderate values of p_{pp} may cause us to question the model but whether the model is rejected or not may depend on the ultimate purpose for which it will be used.

There is some simplification in the model-checking procedure if the discrepancy measure T does not depend on the model parameters, that is, if T is a test statistic in the traditional sense. Then the observed value of the test statistic $T(\boldsymbol{Y})$ is compared to the posterior predictive reference distribution of $T(\boldsymbol{Y}^{\text{rep}})$, and the two-dimensional scatter plot used for comparison can be collapsed to a one-dimensional histogram.

There are a number of alternative ways to define replications for use in model checking; the definition (7) is common but certainly not the only possibility. The key choices to be made in defining replications are reviewed by Gelman *et al*. [15]. Here we briefly review one alternative, the model-checking approach described by Box [24], which relies on the prior predictive or marginal distribution of the data $\boldsymbol{Y}$. Under that approach, we compare test statistics $T(\boldsymbol{Y})$ to the reference distribution obtained by averaging over the prior distribution of the model parameters

$$p(\boldsymbol{Y}^{\text{rep}}) = \int p(\boldsymbol{Y}^{\text{rep}} \mid \boldsymbol{\eta}) p(\boldsymbol{\eta}) \, \mathrm{d}\boldsymbol{\eta}$$

A difficulty with this approach is that it requires proper prior distributions for all parameters. Moreover, the common approach of using flat prior distributions over wide ranges to approximate improper prior distributions does not lead to useful prior predictive model checking. Since improper prior distributions are common in statistical practice, including in the disease-mapping context, we do not consider the prior predictive approach here.

3.3. Posterior predictive model checks for disease mapping

To carry out the posterior predictive approach for the disease mapping model of Section 2.1, we need to define suitable discrepancy measures. We mention a few possibilities here and demonstrate them in the context of a real data set in Section 5. One class of discrepancy measures are omnibus measures of fit. An example from this class is the discrepancy based on the usual chi-squared goodness-of-fit measure

$$T(\boldsymbol{Y}; \boldsymbol{\eta}) = \sum_i \frac{(Y_i - E(Y_i|\boldsymbol{\eta}))^2}{\text{var}(Y_i|\boldsymbol{\eta})} \tag{9}$$

Classical goodness-of-fit tests for the null hypothesis that the data $\boldsymbol{Y}$ come from the given model (for some $\boldsymbol{\eta}$) are based on the test statistic obtained by replacing $\boldsymbol{\eta}$ in (9) with its maximum likelihood or minimum chi-squared estimate. For these classical test statistics, there are analytic results establishing the asymptotic distributions as chi-squared under the null hypothesis. The posterior predictive distribution provides a suitable reference distribution for any sample size.

Omnibus discrepancy measures are useful but provide less power than measures designed to test specific features of the data. It is difficult to describe how such measures are constructed in general because that will depend on the specifics of the application and the model [25–27]. To

illustrate the idea for disease mapping models, we consider one hypothetical scenario. If it were hypothesized that a covariate $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ omitted from the model were in fact important, then the discrepancy, $T(\boldsymbol{Y}; \boldsymbol{\eta}) = \text{corr}(\log Y_i - \log E_i - (X\boldsymbol{\beta})_i, Z_i)$, could be used to assess the correlation of the unexplained variation in $\boldsymbol{Y}$ and $\boldsymbol{Z}$. This is analagous to the use of residuals for assessing the importance of a covariate in ordinary linear regression.

3.4. Discussion of the posterior predictive approach

Posterior predictive model checks have been criticized on several grounds. First, it has been pointed out that the choice of discrepancy measures very often depends (at least implicitly) on the specification of alternative models and in those cases it might be better to actually fit the alternative models [28]. In the hypothetical scenario of a missing covariate considered above for the disease mapping model, the obvious alternative is a model that includes the covariate Z along with an additional parameter (the regression coefficient of Z). This alternative can be fit and the importance of the new parameter assessed by examining its posterior distribution. We have found that posterior predictive checks are of most use when the computational cost of refitting the alternative models (additional programming etc.) are prohibitive. A second aspect of posterior predictive model checks is that they, or more precisely the posterior predictive p-values used to summarize the checks, are quite conservative [15, 29, 30]. Finally, posterior predictive p-values do not generally have a uniform distribution under the null hypothesis (that the data were generated by the model in question), not even asymptotically [30, 31]. Recently Bayarri and Berger [30] have developed a related approach based on posterior distributions that condition on only part of the information in the data rather than using the full posterior distribution to define the reference distribution. Their p-values are uniformly distributed under the null and are not as conservative as the posterior predictive p-values. However, their approach requires more calculation than the posterior predictive approach described here and can be quite difficult to apply for the kinds of complex models that are most challenging to check in practice. Thus, despite their limitations, posterior predictive model checks are practical and quite informative. They are at their best when evaluating the fit of a single model developed for a particular application; see examples in Gelman *et al.* [14, 15] and Glickman and Stern [25].

4. EVALUATING EXTREME OBSERVATIONS

4.1. Cross-validation

A key question in the analysis of disease incidence data is whether extreme relative risks indicate a model failure. A posterior predictive model check using the obvious test statistic, $T(\boldsymbol{Y}) = \max_i(Y_i/E_i)$, the maximum standardized morbidity ratio, is of limited use in addressing this question. This is because a truly unusual value of Y_i/E_i will likely inflate the estimated value of variance parameters in the model such that the posterior predictive distribution will generate simulated extrema that are approximately as large as the observed maximum. Thus, a district with truly exceptional relative risk might be missed by posterior predictive checks in the same way that an extremely influential point in a traditional regression analysis might not be uncovered just by looking at residuals. The solution proposed here is similar to the solution in regression – we can apply a form of cross-validation by leaving out the ith region while assessing whether Y_i is unusual.

Let $\boldsymbol{Y}_{-i}$ denote the data vector without the count for the ith region and let $p(\boldsymbol{\eta} \mid \boldsymbol{Y}_{-i})$ denote the posterior distribution of $\boldsymbol{\eta}$ computed without the ith region. Define $Y_{i,-i}^{\text{rep}}$ as a predicted value for the number of disease cases in region i based on the given model and data $\boldsymbol{Y}_{-i}$. Then we can define a cross-validation or leave-one-out posterior predictive distribution of $Y_{i,-i}^{\text{rep}}$ as

$$p(Y_{i,-i}^{\text{rep}} \mid \boldsymbol{Y}_{-i}) = \int p(Y_{i,-i}^{\text{rep}} \mid \boldsymbol{\eta})\ p(\boldsymbol{\eta} \mid \boldsymbol{Y}_{-i})\, \mathrm{d}\boldsymbol{\eta} \tag{10}$$

The position of the observed value Y_i within the leave-one-out posterior predictive distribution can be used to assess the fit of the model. The leave-one-out posterior predictive distribution (10), evaluated at the observed value Y_i, is called the conditional predictive ordinate (CPO) by Geisser [32]. A small CPO suggests an observed value that is unlikely under the model fit without the observation in question. Of course, the calculation of the distribution (10) requires refitting the model to $\boldsymbol{Y}_{-i}$.

There is one issue that must be resolved before the definition (10) can be implemented for the disease mapping model. Recall that the data $\boldsymbol{Y}$ are modelled conditional on the expected counts $\boldsymbol{E}$. If the data are internally standardized (recall from Section 2.1 that this means the data being analysed were also used to create $\boldsymbol{E}$), then leaving out one small area creates a situation in which the expected counts are no longer the appropriate standardizations. It is necessary to recalculate expected counts $\boldsymbol{E}_{-i}$ when deleting area i; we cannot just delete the expected count for the ith region because the remaining expected counts will not represent an internal standardization of $\boldsymbol{Y}_{-i}$. If the raw data used to compute the original standardizations (2) in Section 2.1 are available, then it is straightforward to recalculate the expected counts by developing new estimates for the risk within each stratum without the data from area i. If the raw data are not available, then a simple approximation is to multiply all the expected counts in the original vector $\boldsymbol{E}$ by a constant such that the sum of the elements of $\boldsymbol{Y}_{-i}$ is equal to the sum of the elements of $\boldsymbol{E}_{-i}$. The required constant is

$$c_i = \frac{\sum_{j \neq i} Y_j}{\sum_{j \neq i} E_j} \tag{11}$$

This means that the expected count in area j when area i is deleted is $E_{j,-i} = c_i E_j$. Then the leave-one-out posterior distribution $p(\boldsymbol{\eta} \mid \boldsymbol{Y}_{-i})$ is the posterior distribution under the model of Section 2.1 except that the ith area is omitted and the expected counts $\boldsymbol{E}_{-i}$ are used. The adjustment of the expected counts also affects the first term in the integral (10) that defines the leave-one-out posterior predictive distribution. The predictive distribution $p(Y_{i,-i}^{\text{rep}} \mid \boldsymbol{\eta})$ is a Poisson distribution with mean $\lambda_i c_i E_i$, because $c_i E_i$ represents the best estimate of the expected count in region i when conditioning on $\boldsymbol{Y}_{-i}$. It should be pointed out that if the data are externally standardized (recall that this means $\boldsymbol{E}$ is constructed from information outside the current data set), then the expected counts do not need to be recalculated when a region is dropped from the analysis.

In practice we might like to apply this approach to a large number of suspect regions. If fitting the model requires a sophisticated MCMC simulation, then it may not be feasible to consider many suspect regions. Carlin and Louis [18] describe one approach for calculating the CPO without refitting the model. They suggest using the approximation, $p(\boldsymbol{\eta} \mid \boldsymbol{Y}) \approx p(\boldsymbol{\eta} \mid \boldsymbol{Y}_{-i})$, unless Y_i is an extreme outlier. Of course this approximation is not appropriate here since we are hoping to identify extreme outliers. There is an additional difficulty with this approximation. In models like the disease mapping model, where there is one parameter (λ_i) corresponding directly to each

observed count (Y_i), the posterior mean for a given λ_i will be a compromise between the observed count and the expected count under the spatial regression model. Consequently, assessing the fit of the model based on the approximation will tend to be quite conservative in that all CPOs are reasonably large. In the remainder of this section, we describe two approximations that perform better while still not requiring us to reanalyse the data without region i. Both approaches reweight the posterior simulations from the complete data analysis to approximate the posterior distribution without the ith region.

4.2. Importance weighting

Let $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_M$ denote a collection of M posterior simulations from the complete data posterior distribution, $p(\boldsymbol{\eta} \mid \boldsymbol{Y})$. The importance ratio for the jth posterior simulation, when we are approximating the posterior distribution $p(\boldsymbol{\eta} \mid \boldsymbol{Y}_{-i})$, is defined as

$$w_{-i}(\boldsymbol{\eta}_j) = \frac{p(\boldsymbol{\eta}_j \mid \boldsymbol{Y}_{-i})}{p(\boldsymbol{\eta}_j \mid \boldsymbol{Y})} \propto \frac{\prod_{k=1,k\neq i}^{n} p(Y_k \mid c_i E_k)}{\prod_{k=1}^{n} p(Y_k \mid E_k)}; \quad j = 1, \ldots, M \tag{12}$$

where the numerator and the denominator of the final ratio are products of Poisson densities. The prior distribution is the same in numerator and denominator and cancels in forming the importance ratio. The importance ratios are only determined up to a multiplicative constant because the numerator and denominator have different normalizing constants. The importance ratios can be used to compute estimates of various summaries of the posterior distribution without small area i. For example, the posterior mean of $\boldsymbol{\eta}$ conditional on Y_{-i} would be estimated as the weighted average of the posterior draws with importance ratios used as weights, $(\sum_j \boldsymbol{\eta}_j w_{-i}(\boldsymbol{\eta}_j))/(\sum_j w_{-i}(\boldsymbol{\eta}_j))$. Similarly, the importance-weighted CPO estimate is

$$\frac{\sum_{j=1}^{M} \Pr(Y_{i,-i}^{\text{rep}} = Y_i \mid \boldsymbol{\eta}_j) w_{-i}(\boldsymbol{\eta}_j)}{\sum_{j=1}^{M} w_{-i}(\boldsymbol{\eta}_j)}. \tag{13}$$

The conditional probability in the CPO estimate is easily calculated using the fact that $Y_{i,-i}^{\text{rep}}$ has a Poisson distribution with mean equal to $c_i E_i$ multiplied by the 'λ_i' component of $\boldsymbol{\eta}_j$. We can also compute estimates of $\Pr(Y_{i,-i}^{\text{rep}} < Y_i \mid \boldsymbol{Y}_{-i})$ and $\Pr(Y_{i,-i}^{\text{rep}} > Y_i \mid \boldsymbol{Y}_{-i})$ to assess whether replicate data tend to be larger or smaller than the observed value.

4.3. Importance resampling

Importance weighting can be used to approximate the posterior expected value of any function of the parameters conditional on all of the data except the count from region i. It is sometimes convenient to have a set of simulations to approximate a posterior distribution, which allows one to obtain all manner of posterior summaries, not just expectations. The importance resampling algorithm of Rubin [33, 34] uses the importance weights $\{w_{-i}(\boldsymbol{\eta}_j)\colon j = 1, \ldots, M\}$, to generate such simulations. The algorithm can be expressed simply in two steps:

1. Define $\pi_j = w_{-i}(\boldsymbol{\eta}_j)/\sum_{j=1}^{M} w_{-i}(\boldsymbol{\eta}_j)$, for $j = 1, \ldots, M$.
2. From the orginal sample, $\{\boldsymbol{\eta}_j\colon j = 1, \ldots, M\}$, select a subsample of size L without replacement using probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$.

The resulting subsample represents an approximation of the desired posterior distribution, $p(\boldsymbol{\eta} \mid \boldsymbol{Y}_{-i})$. The samples are taken without replacement so that each of the posterior simulations appears at most once in the final subsample. This affords a measure of protection in case one of the original posterior draws ends up with very large proportion of the total importance weight.

4.4. Discussion of cross-validation

Cross-validation is an old and valuable idea. The preceding sections demonstrate how to carry out cross-validation in the disease mapping context. Separate reanalysis of the data after deleting a single case is possible but requires care when the data have been internally standardized. In situations where data analysis is time-consuming or the number of small areas is large, it may be desirable to avoid the complete reanalysis of the data. Importance weighting and importance resampling provide approximations to the desired posterior distribution. The accuracy of importance weighting and importance resampling approximations depend on how great the change in the posterior distribution is when region i is eliminated. The distribution of the importance weights provides some, but not perfect information, about this. A heavily skewed distribution of importance weights (dominated perhaps by a small number of extreme values) will tend to produce unreliable results. In that case, the importance-weighted analysis serves primarily to identify a subset of potentially unusual observations – a complete reanalysis would be required to fully assess these regions. Importance resampling provides a bit of protection in cases where the distribution of importance ratios is extremely skewed, because individual posterior simulations appear at most once in the importance-weighted subsample and thus one large ratio will not dominate the calculation. Of course, in that case, the importance-weighted subsample represents a distribution that is intermediate between the complete-data posterior distribution $p(\boldsymbol{\eta}|\boldsymbol{Y})$ and the target leave-one-out posterior distribution $p(\boldsymbol{\eta} \mid \boldsymbol{Y}_{-i})$.

5. MODEL CHECKING FOR THE SCOTLAND LIP CANCER DATA

5.1. Overall measure of fit

The overall measure of fit (9) from Section 3.3 was computed for 1000 simulations from the posterior distribution of $\boldsymbol{\eta}$ and the posterior predictive distribution of $\boldsymbol{Y}^{\text{rep}}$. The joint distribution of $T(\boldsymbol{Y};\boldsymbol{\eta})$ and $T(\boldsymbol{Y}^{\text{rep}};\boldsymbol{\eta})$ is displayed in Figure 2(a). The points appear to be well scattered about the 45 degree line suggesting no evidence of lack of fit. The upper tail area probability, $\Pr(T(\boldsymbol{Y}^{\text{rep}};\boldsymbol{\eta}) \geqslant T(\boldsymbol{Y};\boldsymbol{\eta})|\boldsymbol{Y})$, is estimated as 0.38.

For purposes of illustration, we have also fit the clearly inferior simple Poisson regression model to the Scotland lip cancer data; that model corresponds to taking $\tau^2 = 0$ and $\boldsymbol{\lambda} = \mathrm{e}^{X\boldsymbol{\beta}}$. The same discrepancy measure is used to evaluate the fit of this model in Figure 2(b). There the observed values (horizontal axis) are extremely large relative to what we would expect to see under the model. The tail area probability estimate would be 0.00 suggesting the model could not have generated the type of data we have observed. The solution of course is to introduce extra-Poisson variation, and this is precisely what the Poisson-log-Gaussian model does. Note that this diagnostic is quite similar to traditional methods used to check for overdispersion in Poisson regression models. We also note that the vertical scale in both diagnostic plots is centred on values near 56 which corresponds roughly to the expected value of the asymptotic reference distribution for the classical chi-squared goodness-of-fit test.

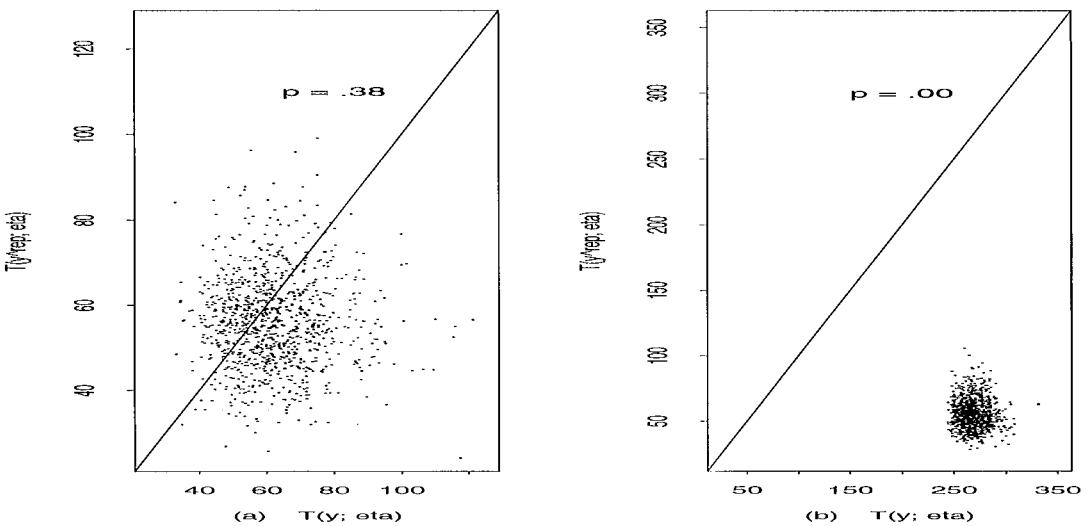

Figure 2. Scatter plots of the joint posterior distribution of the chi-squared discrepancy measure evaluated at the posterior predictive replicate $\boldsymbol{Y}^{\text{rep}}$ and the observed data $\boldsymbol{Y}$: (a) for the Poisson-log Gaussian model; (b) for a Poisson regression model ($\tau^2 = 0$).

5.2. *Extreme values*

Figure 3 provides a histogram of the posterior predictive distribution for the maximum SMR_i based on the analysis of the entire data set. The distribution of the maximum SMR_i has a number of peaks because it is the maximum among a set of ratios Y_i/E_i, where the E_i are considered fixed and the Y_i are restricted to be integers. Note that the observed maximum is not terribly unusual; 54 per cent of the posterior predictive replications have maximum SMR_i greater than the observed maximum and another 7 per cent have maximum SMR_i equal to the observed maximum. This result was anticipated by our discussion at the beginning of Section 4.1. A similar check based on the minimum SMR_i indicates that the observed minimum SMR_i (zero) is not at all unexpected; a minimum SMR_i of zero occurs in 96 per cent of the posterior predictive replications. This is not surprising given the large number of regions with small expected counts.

We now consider our cross-validation posterior predictive approach to assessing the fit of the model to individual observations. For several quantities of interest we present results based on the original analysis of all 56 districts along with cross-validation results leaving out individual districts. Three different computational approaches described in Section 4 are used to obtain the cross-validation posterior distribution: a complete reanalysis of the data without the district in question (Section 4.1); an importance weighting approximation (Section 4.2); and an importance resampling approximation (Section 4.3). The most accurate approach is to obtain the cross-validation posterior distribution without a given district by repeating the Bayesian analysis of the Scotland lip cancer data described in Section 2, except that we leave out the district in question and adjust the expected counts as described in Section 4.1. The importance weighting estimates for the quantities of interest are obtained by reweighting the 1000 posterior draws from the complete data posterior, $p(\boldsymbol{\eta}|\boldsymbol{Y})$, as described in Section 4.2. For importance resampling, the algorithm of Section 4.3 was applied

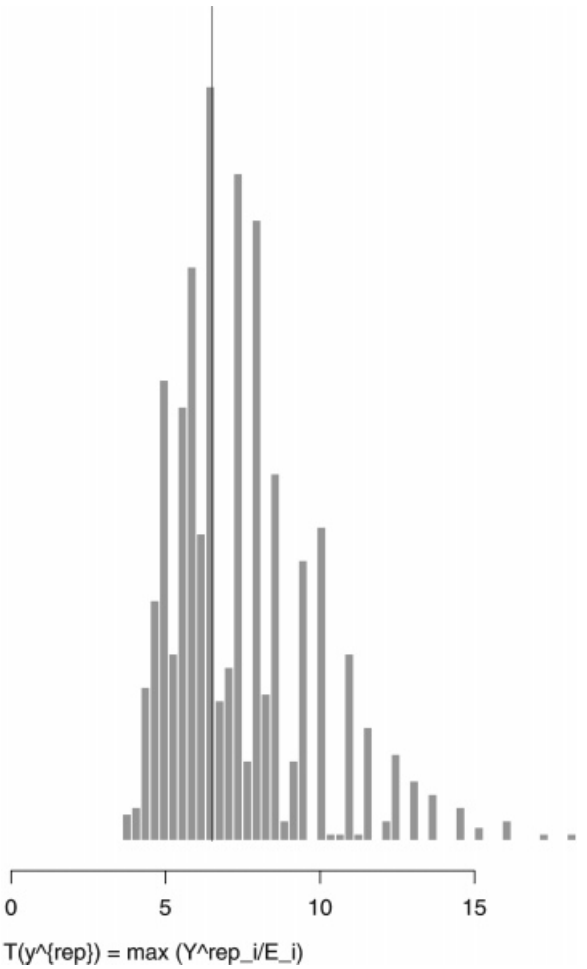

Figure 3. Posterior predictive distribution of the maximum SMR_i with a vertical line indicating the observed maximum.

with an initial sample of 5000 posterior draws (we supplemented the original simulation which contained only 1000 posterior draws) and a subsample of size 500.

Estimates of the posterior mean of the relative risks for a subset of the regions are provided in Table III. In the first two columns, posterior means are reported for the complete data analysis and for the leave-one-out reanalysis of the data. The 15 districts in Table III include the four districts identified earlier (see Figure 1) as being of interest, along with others for which large changes were observed. There are a number of large differences between the complete data analysis and the leave-one-out reanalysis. Districts 2, 3, 11 and 15 exhibit much lower mean relative risks when the model is refit without the district in question, and districts 17, 42, 55 and 56 have much higher mean relative risks when the model is refit without the district in question. Especially noteworthy is the large increase for the posterior mean of λ_{17} when the data are reanalysed without district

Table III. Posterior mean of relative risk for selected districts under the complete data analysis, and estimated posterior means for the leave-one-out analyses using importance weights and importance resampling.

ID	District name	Posterior mean of relative risk λ_i			
		Complete data	Leave-one-out		
			Reanalysis	Importance weights	Importance resample
1	Skye-Lochalsh	6.37	8.80	6.45	6.10
2	Banff-Buchan	4.10	2.14	3.19	3.74
3	Caithness	3.18	1.70	1.93	2.54
11	Western Isles	2.63	1.66	0.85	2.28
15	NE Fife	1.83	1.07	1.26	1.52
17	Badenoch	2.07	9.44	2.42	2.39
26	Dunfermline	0.97	0.59	0.63	0.80
38	Monklands	0.64	0.42	0.40	0.49
42	Falkirk	0.78	1.78	1.16	1.03
45	Edinburgh	0.47	0.63	0.65	0.54
49	Glasgow	0.41	0.49	0.48	0.43
50	Dundee	0.48	0.77	0.66	0.60
54	Strathkelvin	0.32	0.55	0.44	0.44
55	Annandale	0.53	2.12	1.24	0.90
56	Tweeddale	0.41	2.02	0.79	0.70

17; one possible explanation is that this district has the lowest expected count and thus its relative risk parameter has the greatest variability according to our model. With the exception of district 17, the results in Table III look about as they should; the posterior mean of the relative risk based on the complete data analysis is heavily influenced by the data from that region (as it should be!), with the model providing some smoothing of risks over nearby regions, whereas the leave-one-out reanalysis ignores the information from the region under consideration. The predictive distributions for the disease counts in the regions are much more informative than the posterior means; we examine these below.

One additional use of Table III is that it allows us to compare the results obtained using the approximations to that obtained by actually analysing the data without the region in question. The importance weighting and importance resampling results are generally intermediate between the complete data analysis and the results obtained by reanalysing the data without a given region. This suggests that the approximations may be useful for identifying those regions worthy of further study, but may not be useful for providing accurate estimates of what happens to the posterior distribution of individual parameters when a region is removed. The distribution of the logarithms of the importance weights for four of the leave-one-out analyses are shown in Figure 4. Of these, none seems very highly skewed. In the case of district 49, Glasgow, one of the 1000 posterior samples, has importance weight equal to 27 per cent of the total; however, the importance-weighted estimates are accurate in that case because the posterior mean does not change much.

Table IV gives summaries for leave-one-out predictive distributions of $Y^{\text{rep}}_{i,-i}$, the count in district i conditional on the data excluding that district, for the same 15 districts. In particular, we calculate the probability that $Y^{\text{rep}}_{i,-i}$ is less than, equal to, or greater than the observed count Y_i. We briefly review how these quantities are estimated under the different approaches to cross-validation using the 'equal to' case, that is, the CPO, as an illustration. For the complete reanalysis without district i, we have available a sample from the posterior distribution of the relative risk parameter

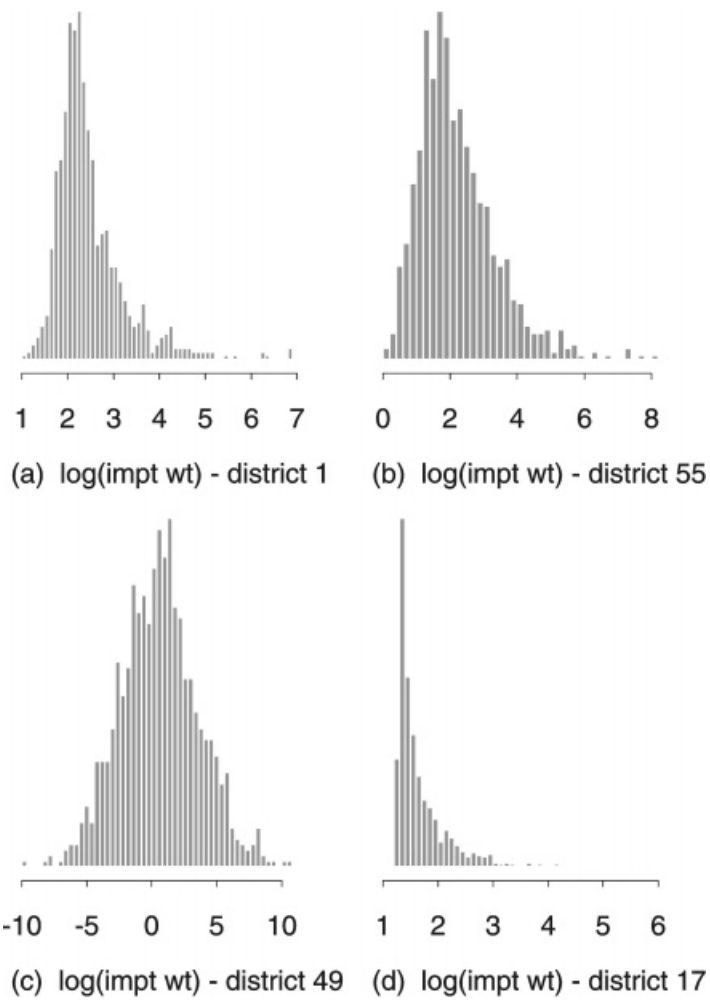

Figure 4. Histogram of the logarithms of the importance weights for leave-one-out analyses. (*a*) District 1, Skye-Lochalsh, has the largest SMR_i. (*b*) District 55, Annandale, has the smallest SMR_i. (*c*) District 49, Glasgow, has the largest E_i. (*d*) District 17, Badenoch, has the smallest E_i.

λ_i conditional on $\boldsymbol{Y}_{-i}$. If we denote this sample as $\lambda_i^{(1)}, \ldots, \lambda_i^{(M)}$, then the CPO is estimated as $\sum_{j=1}^{M} \Pr(Y_{i,-i}^{\mathrm{rep}} = Y_i \mid \lambda_i^{(j)})/M$ with the probabilities computed from the assumed Poisson distribution. The importance-weighted estimate of the CPO is computed via (13) of Section 4.2. That formula uses the posterior draws of λ_i from the complete data analysis, and computes a weighted average of Poisson probabilities. The importance-resampled estimate of the CPO is a simple average of Poisson probabilities, using the 500 resampled posterior draws of λ_i.

The leave-one-out estimates of the posterior distribution for $Y_{i,-i}^{\mathrm{rep}}$, especially at the bottom of the table, give us reason to be concerned about the fit of the model. There are a number of districts for which the leave-one-out analyses suggest observed values are smaller than would be expected

Table IV. Posterior predictive probability distributions for replicate data $\boldsymbol{Y}^{\text{rep}}$.

District ID	Posterior predictive probabilities for Y_i^{rep} (complete data) and $Y_{i,-i}^{\text{rep}}$ (leave-one-out)											
	Complete data			Leave-one-out analysis								
				Reanalysis			Importance weights			Importance resample		
	$<Y_i$	$=Y_i$	$>Y_i$	$<Y_i$	$=Y_i$	$>Y_i$	$<Y_i$	$=Y_i$	$>Y_i$	$<Y_i$	$=Y_i$	$>Y_i$
1	0.52	0.09	0.39	0.66	0.03	0.31	0.59	0.05	0.36	0.60	0.06	0.34
2	0.66	0.04	0.30	0.96	0.00	0.04	0.92	0.01	0.07	0.83	0.02	0.14
3	0.62	0.08	0.30	0.89	0.02	0.10	0.87	0.02	0.11	0.78	0.04	0.17
11	0.61	0.07	0.31	0.84	0.02	0.14	0.98	0.00	0.02	0.74	0.05	0.21
15	0.69	0.06	0.25	0.93	0.01	0.06	0.90	0.02	0.08	0.84	0.03	0.13
17	0.44	0.20	0.36	0.34	0.11	0.55	0.46	0.15	0.40	0.45	0.16	0.39
26	0.72	0.06	0.21	0.94	0.02	0.04	0.92	0.02	0.06	0.85	0.03	0.11
38	0.72	0.08	0.20	0.89	0.03	0.07	0.91	0.03	0.06	0.85	0.05	0.10
42	0.14	0.07	0.79	0.01	0.01	0.99	0.02	0.01	0.96	0.03	0.02	0.95
45	0.19	0.05	0.75	0.04	0.01	0.94	0.04	0.01	0.95	0.07	0.02	0.90
49	0.14	0.03	0.83	0.03	0.01	0.96	0.01	0.00	0.99	0.03	0.01	0.96
50	0.15	0.08	0.77	0.02	0.01	0.97	0.04	0.03	0.92	0.07	0.04	0.89
54	0.15	0.24	0.61	0.08	0.15	0.77	0.09	0.17	0.74	0.09	0.17	0.74
55	0.00	0.18	0.82	0.00	0.03	0.97	0.00	0.03	0.97	0.00	0.05	0.95
56	0.00	0.56	0.44	0.00	0.32	0.68	0.00	0.37	0.63	0.00	0.41	0.59

if the district followed the same model as the other districts. This includes the districts 42, 45, 49 and 50 which have four of the six largest expected counts; districts 45 and 49 correspond to the two biggest urban centres in Scotland. Hence a modification of the model that incorporates the degree of urbanization of a district might be warranted. The approximations (importance weighting and importance resampling) to the leave-one-out reanalysis seem to perform quite well in Table IV even though they were not terribly reliable for the posterior means of the relative risk parameters.

Although not directly comparable to the leave-one-out reanalyses and approximations, we also present results for the complete data analysis. There we are actually computing the posterior predictive distribution of Y_i^{rep} based on analysis of all of the data. As one would expect, the complete-data estimates of the posterior predictive distribution for Y_i^{rep} suggest a better fit than the leave-one-out estimates of the posterior predictive distribution of $Y_{i,-i}^{\text{rep}}$. As explained earlier, the estimated relative risk parameters λ_i, and hence the predictive distribution of Y_i^{rep}, are heavily influenced by the observed Y_i and thus the predictive distribution will tend to generate values like the observed Y_i. The leave-one-out analyses represent a more reasonable attempt to assess whether a district is unusual. Incidentally, the complete data results can also be viewed as the leave-one-out results that would be obtained if we accept the crude approximation $p(\boldsymbol{\eta}|\boldsymbol{Y}) \approx p(\boldsymbol{\eta}|\boldsymbol{Y}_{-i})$. The results in Table IV suggest that this is a poor idea.

6. CONCLUSIONS

Hierarchical models are commonly used to provide smoothed estimates of small area disease risks. Given the political and epidemiological importance of the estimated risks, it is important to check the fit of the model to the data. This paper describes the posterior predictive approach for model

checking in the context of disease mapping models, where the data are disease counts from n small areas. Extreme values are of great interest in disease mapping because the associated areas can then be examined for factors that may be associated with increased (or decreased) risk. Just examining the posterior distribution of the relative risk parameters (or the posterior predictive distribution of replicate data) can be misleading in these cases because the posterior distribution tries to fit the observed data and hence will not easily identify outliers. We propose that a cross-validation posterior predictive distribution, conditional on all of the data except small area i, be used to determine whether the observed value in small area i is consistent with the model; $i = 1, \ldots, n$. We also describe two approaches for constructing approximations to the cross-validation posterior predictive distribution. Both approaches apply importance weights to the simulations from the complete data posterior distribution. The approaches are demonstrated on the Scotland lip cancer data where it does in fact seem that the model is not adequate; the high-population low-risk areas do not appear to fit the model very well.

ACKNOWLEDGEMENTS

The authors are grateful to Deanne Reber for computing assistance. Cressie's research was partially supported by the Environmental Protection Agency (CR822919-01-0) and the Office of Naval Research (N00014-99-1-0214). Cressie and Stern's research was supported by the National Cancer Institute (CA 78169-02). The computing for this research was performed on equipment purchased with funds provided by an NSF SCREMS grant award DMS-9707740.

REFERENCES

1. Tsutakawa RK, Shoop GL, Marienfeld CJ. Empirical Bayes estimation of cancer mortality rates. *Statistics in Medicine* 1985; **4**:201–212.
2. Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 1987; **43**:671–681.
3. Manton KG, Woodbury MA, Stallard E, Riggan WB, Creason, JP, Pellom AC. Empirical Bayes procedures for stabilizing maps of U.S. cancer mortality rates. *Journal of the American Statistical Association* 1989; **84**:637–650.
4. Besag J, York JC, Mollie A. Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* 1991; **43**:1–59.
5. Mollie A, Richardson S. Empirical Bayes estimates of cancer mortality rates using spatial models. *Statistics in Medicine* 1991; **10**:95–112.
6. Cressie N. Smoothing regional maps using empirical Bayes predictors. *Geographical Analysis* 1992; **24**:75–95.
7. Clayton, D, Bernardinelli L. Bayesian methods for mapping disease risk. In *Geographical and Environmental Epidemiology*: *Methods for Small-Area Studies*, Elliot P, Cuzick J, English D, Stern R (eds). Oxford University Press: London, 1992; 205–220.
8. Bernardinelli L, Clayton D, Montomoli C. Bayesian estimates of disease maps: How important are priors? *Statistics in Medicine* 1995; **14**:2411–2431.
9. Stern HS, Cressie N. Bayesian and constrained Bayesian inference for extremes in epidemiology. In *1995 Proceedings of the Section on Epidemiology*. American Statistical Association: Alexandria, 1995; 11–20.
10. Stern HS, Cressie N. Inference for extremes in disease mapping. In *Disease Mapping and Risk Assessment for Public Health*, Lawson A, Biggeri A, Böhning D, Lesaffre E, Viel J-F, Bertollini R (eds). Wiley: Chichester, 1999; 63–84.
11. Waller LA, Carlin BP, Xia H, Gelfand AE. Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association* 1997; **92**:607–617.
12. Knorr-Held L, Besag J. Modelling risk from a disease in time and space. *Statistics in Medicine* 1998; **17**: 2045–2060.
13. Rubin DB. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* 1984; **12**:1151–1172.
14. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. Chapman and Hall: London, 1995.
15. Gelman A, Meng X-L, Stern HS. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica* 1996; **6**:733–807.
16. Cressie NAC. *Statistics for Spatial Data*, revised edition. Wiley: New York, 1993.
17. Besag J. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1974; **36**:192–236.

18. Carlin BP, Louis TA. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall: London, 1996.
19. Gilks WR, Richardson S, Spiegelhalter DJ (eds). *Markov Chain Monte Carlo in Practice*. Chapman and Hall: London, 1996.
20. Gelman A, Rubin D. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* 1992; **7**:457–511.
21. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; **88**:9–25.
22. Mantel N, Stark CR. Computation of indirect-adjusted rates in the presence of confounding. *Biometrics* 1968; **24**:997–1005.
23. Kass RE, Raftery AE. Bayes factors. *Journal of the American Statistical Association* 1995; **90**:773–795.
24. Box GEP. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society*, *Series A* 1980; **143**:383–430.
25. Glickman ME, Stern HS. A state-space model for National Football League (NFL) scores. *Journal of the American Statistical Association* 1998; **93**:25–35.
26. Boscardin WJ, Gelman A. Bayesian regression with parametric models for heteroscedasticity. *Advances in Econometrics* 1996; **11A**:87–109.
27. Gelman A, Meng X-L. Model checking and model improvement. In *Practical Markov Chain Monte Carlo*, Gilks WR, Richardson S, Spiegelhalter DJ (eds). Chapman and Hall: New York, 1996, 189–201.
28. Draper D. Comment: Utility, sensitivity analysis, and cross-validation in Bayesian model-checking, discussion of A. Gelman, X.-L. Meng, H. Stern, Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* 1996; **6**:760–767.
29. Rubin DB, Comment: On posterior predictive *p*-values, discussion of A. Gelman, X.-L. Meng, H. Stern, Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* 1996; **6**:787–792.
30. Bayarri MJ, Berger JO. P-values for composite null models. Technical Report, Institute of Statistics and Decision Sciences, Duke University, 1999.
31. Robins JM, van der Vaart A, Ventura V. The asymptotic distribution of P-values in composite null models. Technical Report, Department of Epidemiology, Harvard School of Public Health, 1999.
32. Geisser S. *Predictive Inference: An Introduction*. Chapman and Hall: London, 1993.
33. Rubin DB, A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when the fractions of missing information are modest: the SIR algorithm, discussion of M.A. Tanner and W.H. Wong, The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 1987; **82**:543–546.
34. Rubin DB. Using the SIR algorithm to simulate posterior distributions. In *Bayesian Statistics 3*, Bernardo JM, DeGroot MH, Lindley DV, Smith AFM (eds). Oxford: New York, 1988; 395–402.

