Issues in the mapping of two diseases

Alan R Dabney and Jon C Wakefield
Department of Biostatistics, University of Washington, Seattle, WA, USA

Recently, there has been increased interest in the geographical modelling of two or more diseases. In this article, we consider a number of issues relating to such an endeavour including the standardization process and the comparison of univariate and bivariate disease mapping models. A principle motivation for the examination of two or more diseases is to discover similarities or dissimilarities in the geographical distribution of risk. In this article, we propose a proportionality mortality approach to give clues to areas of similarity and dissimilarity. A secondary aim of bivariate modelling is to ‘borrow strength’ between diseases in order to provide better estimates of risk in each area. We will illustrate various modelling strategies using incidence data from 1996 to 2000 on lung and bladder cancer in Washington state.

1 Introduction

Disease mapping has a long history in epidemiology and has been used to: provide a description of geographical variation in risk, suggest possible risk factors that may explain variation (hypothesis generation), and provide estimates of risk in specific areas in order to provide a means of allocating public health services.1–3 An important extension to standard univariate disease mapping methodologies is the joint analysis of two (or more) diseases. Benefits of such bivariate disease mapping include the ability to explore shared and divergent trends in risk as well as increased precision for the estimation of each collection of disease risks. In the latter case, when interest is in a relatively rare disease, a joint model that incorporates data from a more common, and related, disease may be used to gain precision in the estimates for the rare disease.

It has been suggested that finding similar geographical patterns of risk for multiple diseases might provide stronger evidence of common risk factors that have spatial patterns than that obtained by separate univariate analyses.4 As an example, risk factors for one disease may be well established (but unmeasured at the area level), and so similarity of rates with a second set of disease rates may suggest that the risk factors, or a subset, are common to both diseases. Likewise, dissimilarity may suggest that the risk factor is not a cause of the second disease. In the preceding discussion, ‘positively correlated’ may replace ‘similar’ when the effects of the common risk factor are of the same sign, and ‘negatively correlated’ if of the opposite sign. Likewise, ‘uncorrelated’ may be equated with ‘dissimilar’.

More speculatively, it has also been suggested that at least partial adjustment for unobserved covariates may be carried out by considering two similar diseases and

Address for correspondence: Alan R Dabney, Department of Biostatistics University of Washington F-600 Health Sciences Building, Box 357232 Seattle, WA 98195-7232, USA. E-mail: adabney@u.washington.edu

© Arnold 2005 10.1191/0962280205sm340oa
picking out shared spatial trends that represent the unobserved covariates. In this spirit, it has been suggested that disease incidence rates from another disease may act as a surrogate for unobserved exposures, with the obvious example being lung cancer incidence as a surrogate for smoking. As another example, it has been hypothesized that primary Epstein-Barr virus (EBV) infection occurring during adolescence or adulthood is associated with elevated breast cancer risk. Data on EBV are unavailable routinely, but EBV is thought to be a causal factor for Hodgkin’s disease, and so the latter may be used as a surrogate. Hence, a Poisson regression analysis was carried out in which breast cancer incidence was regressed on measures of Hodgkin’s disease, and a positive association was found to be present.

Area level covariates may be added to the linear predictor in order to obtain better estimates of area level risk in mapping studies, whereas the area level association between covariates and risk may be examined via ecological correlation models. The former aim is prediction, whereas the latter is to gain clues to causal relationships. Prediction is a far more straightforward task, as the model does not have to be causal. Even at the level of the individual, attempting to make causal statements in an observational study is fraught with difficulties, and at the group (area) level the task is even more difficult owing to so-called ecological bias. There is a huge literature on this topic. It has been argued that modelling spatial dependence in an ecological correlation setting is of secondary importance to dealing with confounding and within area variability in exposures and confounders. The modelling of spatial dependence in residual risks in a regression setting has been considered, but when such dependence is introduced, regression coefficients often change in magnitude (as the exposures of interest often have spatial structure), making interpretation difficult. By analogy with longitudinal data analysis, care must be taken in a nonlinear setting when estimating regression coefficients in the presence of correlation, because different models for the correlation will lead to different estimates (as well as measures of uncertainty). With this in mind, it would seem that modelling a pair of diseases in order to predict area level risks is less risky than attempting to estimate regression coefficients.

We focus on the problem of exploring shared and divergent trends between two diseases in this article. A so-called shared component model was recently proposed, in which the risk surfaces for two diseases was decomposed into disease-specific components and a shared component. We briefly comment on a number of other approaches that we do not consider further in this article. Empirical Bayes methods have been used to estimate relative risks for two diseases simultaneously by modelling the pair of relative risks as arising from a bivariate lognormal distribution. These authors did not consider spatial dependence between the random effects; this extension was considered elsewhere. Other work generalized a conditional autoregressive (CAR) distribution to two diseases or worked directly from multivariate CAR distributions.

In this article, we propose the use of a proportional mortality model as a tool for bivariate disease mapping. This model has elements in common with the shared component model, in that it highlights trends of similarity and dissimilarity. We explore this new model and the shared component model using data on bladder and lung cancer incidence in the 1318 census tracts of Washington state during the years 1996–2000.
This article is organized as follows. In Section 2, we describe the data in more detail, and provide exploratory analyses. In Section 3, we review some univariate spatial distributions and in Section 4, we describe the shared component and proportional mortality models. In Section 5, we return to the illustrative example; we first carry out separate univariate analyses of each cancer before moving to joint analyses. Section 6 contains a concluding discussion.

2 Illustrative example

Suppose that a study region $A_i$ is partitioned into subregions $A_{ij}$, with this partition usually defined by data availability, $i = 1, 2, \ldots, n$. Further, assume that we a priori identify confounders that we would like to control for in the analysis; we assume there are $J$ confounder strata. For a generic cancer we let $Y_{ij}$ and $N_{ij}$ denote, respectively, the number of incident cases and the population, in area $i$, stratum $j$, $i = 1, \ldots, n$; $j = 1, \ldots, J$. In our study, the demographic confounders correspond to race, gender, and age, and we also control for year of incidence. We specifically considered two race groups, white and black, with all other race groups being excluded owing to sparsity of data. This corresponds to losing 2.2 and 3.1% of the incidence counts for bladder and lung cancer, respectively. For age, 18 five-year age bands $[0,5), [5, 9), \ldots, [85, \ldots$ were considered. Hence, we have $2 \times 2 \times 18 \times 5 = 360$ strata. Populations were available from the 2000 US Census, so that the same populations were used for each of the years 1996–2000. These data are available at the level of the census tract, of which there are $n = 1318$ in Washington state. The cancer incidence data were provided by the Washington State Cancer Registry (WSCR). Table 1 displays a number of descriptive statistics for the population and disease data. The total population is $\sim 5$ million, and lung cancer incidence is around three times greater than bladder cancer, with median number of cases per area four and 12 for bladder and lung, respectively.

We assume that in area $i$ and stratum $j$, $Y_{ij} \sim \text{Poisson}(N_{ij} p_{ij})$, where $p_{ij}$ is an area-and stratum-specific incidence rate. In this model, there are $n \times J = 1318 \times 360 = 474480$ area and stratum cells in our study, which far exceed the number of disease events

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Mean</th>
<th>Median</th>
<th>Standard deviation</th>
<th>Minimum</th>
<th>Maximum</th>
<th>No. (%) of areas with zero counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>5 012 090</td>
<td>3802.80</td>
<td>3663.50</td>
<td>1573.45</td>
<td>34</td>
<td>13 523</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Bladder observed</td>
<td>5811</td>
<td>4.41</td>
<td>4.0</td>
<td>3.21</td>
<td>0</td>
<td>19</td>
<td>87 (7)</td>
</tr>
<tr>
<td>Bladder expected</td>
<td>5770</td>
<td>4.38</td>
<td>4.0</td>
<td>2.26</td>
<td>0.02</td>
<td>15.1</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Bladder SMRs</td>
<td>–</td>
<td>1.06</td>
<td>0.95</td>
<td>1.35</td>
<td>0</td>
<td>42.9</td>
<td>87 (7)</td>
</tr>
<tr>
<td>Lung observed</td>
<td>17 824</td>
<td>13.5</td>
<td>12.0</td>
<td>8</td>
<td>0</td>
<td>48</td>
<td>13 (1)</td>
</tr>
<tr>
<td>Lung expected</td>
<td>16 697</td>
<td>12.7</td>
<td>11.6</td>
<td>6.46</td>
<td>0.05</td>
<td>44.8</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Lung SMRs</td>
<td>–</td>
<td>1.14</td>
<td>1.04</td>
<td>1.15</td>
<td>0</td>
<td>37.4</td>
<td>13 (1)</td>
</tr>
</tbody>
</table>

Note: The population data are from the 2000 census, and the cancer data are from the Washington State Cancer Registry. The Standardized Morbidity Ratios (SMRs) are given by $Y_{ij}/E_{ij}$, $i = 1, \ldots, 1318$, where the $E_{ij}$ denote expected counts, controlling for race, gender, age and year.
for either cancer. To reduce dimensionality, it is common to make an assumption of proportionality,

\[ p_{ij} = \theta_i \times p_j, \]

with \( \theta_i = \frac{p_{ij}}{p_j} \) interpreted as a relative risk associated with area \( i \). Hence, we are assuming that, for each stratum, the effect of living in area \( i \) is to multiply all of the stratum specific risks by \( \theta_i \). It is important to emphasize that this relative risk is not at the level of the individual. To make such an assumption would leave one open to the ecological fallacy, occurring when inference at the area level is incorrectly assumed to hold at the level of the individual.

To summarize the general approach: we have risk factors area and confounders (which are predictive of disease outcome and have difference distributions by area). Equation (1) assumes that these act independently (i.e., have no interaction). In a mapping study, we are interested in geographical variation in risk. Hence, the \( \theta_i \) parameters assume a special status, and effort is placed in modelling these parameters, while a priori control of confounders is carried out.

Before continuing, we consider the proportionality assumption (1). This model choice is convenient, as it allows us to sum over many strata and consider only a single parameter for each area. An examination of the literature reveals that checking whether the proportionality assumption is valid seems to be rarely reported, though it is clearly a crucial assumption. If it is inaccurate, then important relationships within particular strata may be missed, and it will not be appropriate to report a single measure for each area. Some checks have been suggested, including a goodness-of-fit test and a simple graphical method. The suggested graphical check involves plotting log \( p_{ij} \) against the strata \( j \) in each area \( i \). We now report a variation of this, where log \( p_{ij} \) is plotted against area \( i \) for each of a simplified collection of strata. As the rates within census tracts are very unstable, we work at the county-level (there are 39 counties within Washington state). Examining all 360 strata in each county is clearly impractical given the rarity of disease, and so we examine eight strata only:

2) Males aged <70 years, contracting disease during 1999–2000.
4) Males aged ≥70 years, contracting disease during 1999–2000.

Race was not included due to sparse counts for blacks. Figures 1 and 2 show the resulting graphical check. Note that we have arbitrarily ordered all rates according to the rates in stratum 1, for each cancer, in order to impose some structure to the plots. In the plots, numbers correspond to the strata numbers above.
Figure 1. Assessment of proportionality assumption, $P_{ij} = \theta_i \times P_j$, for area $i$, stratum $j$ in bladder cancer. The plots are of $\log P_{ij}$ against county index. The numbers indicate the eight coarsened strata described in Section 2. All rates have been ordered according to that for stratum 1.
Figure 2  Assessment of proportionality assumption for lung cancer.
If the proportionality assumption (1) holds, then we expect \( \log p_{i1} - \log p_{ij} = \log p_1 - \log p_j \) to be constant across county \( i \), for each \( j \neq 1 \). As ordering was done by the rates for stratum 1, we expect the pattern for strata 2–8 to resemble the pattern for stratum 1, with vertical displacements representing the difference in log stratum-specific rates, comparing the stratum of interest with stratum 1. As an example, consider the plot for stratum 3 and bladder cancer in Figure 1. The area- and stratum-3-specific rates appear to be centered on a line which is parallel to that for stratum 1. That it is not perfectly parallel may well be due to variation of the estimates from small disease counts. The above mentioned is roughly true for the rest of the strata in Figure 1, giving informal support to the proportionality assumption (1). However, in the plots for lung cancer in Figure 2, it is less plausible that the area- and stratum-specific rates for strata 2–8 are parallel to strata 1.

We briefly note another diagnostic plot for checking (1). In this approach, we estimate \( \log \theta_i \) for all strata by \( \log (p_{ij}/p_j) \), \( i = 1, 2, \ldots, n \), \( j = 1, 2, \ldots, 8 \) and plot these estimates by county index, together with the log SMR’s. If Equation (1) holds, then the points should be tightly clustered for each area. Figure 3 shows such plots for both cancers. There is a moderate amount of variability in the rates about the SMR’s. Hence, on the basis of this and the previous diagnostic approach, we do not have strong evidence to defend the proportionality assumption.

**Figure 3** An alternative assessment of proportionality assumption. The plots are of \( \log(p_{ij}/p_j) \) against county index. The numbers indicate the eight coarsened strata described in Section 2. The filled circles are the SMR’s.
If assumption (1) is appropriate, standardization may be carried out to reduce the dimensionality of the data. There are both direct and indirect versions of standardization. The latter is preferred in mapping of small area data, because the former requires rates by stratum to be calculated in each area, and this is not possible with the sparse amount of data available in small areas. Indirect standardization in area \( i \) corresponds to a counterfactual argument in which we ask, ‘What would be the expected number of cases if standard rates were applied to the population of area \( i \)? Summing over strata, making the proportionality assumption (1), and letting \( Y_i = \sum_{j} Y_{ij} \), we have \( Y_i \sim \text{ind Poisson} (\theta_i \sum_j N_{ij} p_j) \), or

\[
Y_i \sim \text{ind Poisson}(E_i, \theta_i),
\]

where the expected count of area \( i \) is given by

\[
E_i = \sum_{j=1}^{J} N_{ij} p_j,
\]

and \( p_j \) represent a set of standard rates for stratum \( j \). If these rates are obtained from another region then we have external standardization, whereas, if the data of the study are used, we have internal standardization; we use the latter here. The obvious method of calculating standard rates is via

\[
\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_{ij}}{N_{ij}}, \quad j = 1, 2, \ldots, J,
\]

where, again, we have \( J = 360 \) stratum. Unfortunately, for blacks in particular, these estimated rates exhibited high variability owing to the sparsity of population data. To alleviate this problem, we carry out log-linear modelling of the data marginalized across areas, in order to obtain a set of rates that both reflect the patterns across strata, and are stable. We let \( Y_j = \sum_i Y_{ij} \) and \( N_j = \sum_i N_{ij} \) denote the number of cases and the populations in stratum \( j \) and then assume that \( Y_j \sim \text{ind Poisson}(N_j \times p_j) \) with log-linear models of the form

\[
\log p_j = x_j \beta,
\]

where \( x_j \) is a \( 1 \times q \) vector of zeroes and ones and \( \beta \) a \( q \times 1 \) vector. Hence, we are modelling the counts of the \( J \) strata as arising from a contingency table which, in our example, is of dimension \( 2 \times 2 \times 18 \times 5 \). The simplest model that is of interest epidemiologically includes main effects only for year, race, sex and age, to give \( q = 24 \) terms. At the other extreme, the saturated model has \( q = 360 \) terms. The interest is in choosing a parsimonious model which differs little from the saturated model in terms of fit. Once a model is selected we can take \( \hat{p}_j = \exp(x_j \beta) \), where \( j \) implies a cell of the contingency table, and \( x_j \) reflects the terms of the model that are used to model the cell mean.
On the basis of biological plausibility, we felt that it was necessary to at least allow for interactions between race and sex, race and age, and sex and age, and so this was our minimal model. This model has a deviance of 84.5 on 131 degrees of freedom for bladder cancer and 174 on 181 degrees of freedom for lung cancer. The model with all second order interactions gives a residual deviance of 41.7 on 58 degrees of freedom and 64.8 on 58 degrees of freedom for bladder and lung, respectively, indicating that there is little gain in fit, at least in terms of statistical significance, in adding the rest of the second order interactions; however, the use of the asymptotic distribution may be dubious here. Raw and modelled rates were examined graphically and this provided an informal reassurance that little was lost with the simplified model, which was subsequently used to evaluate expected numbers. Other possibilities are available for modelling rates in this and related age – cohort models. For example, previous work has used conditional autoregressive models to semi-parametrically smooth rates (such models could be used to model the age and year rates, and interactions, in a disease mapping context).

Now we believe that the proportionality assumption (1) is appropriate and have obtained a set of stable rates, we can return to the model (2). Viewing each area in isolation, the maximum likelihood estimator of \( \theta_i \) is given by \( \hat{\theta}_i = Y_i / E_i \), and is called the Standardized Morbidity Ratio (SMR) for area \( i \). As \( \text{var}(\theta_i) \propto E_i^{-1} \), the SMRs for sparsely populated areas are highly variable, and these areas often produce the most extreme SMRs. This was one of the original motivations for smoothing of disease maps. We calculated the SMRs at the level of the census tract and indeed found that four out of the five highest SMRs for bladder cancer occurred in counties with expected counts <0.50 (Table 1 gives summary statistics for both the expected numbers and the SMRs and shows that such expected numbers are in the lower tail of the distribution). For lung cancer, the highest SMRs are more stable, because the rates of lung cancer are greater than for bladder cancer. The exception is census tract 9501 of Skamania County for which the SMR is 37.4, corresponding to 10 observed and 0.27 expected counts. Interestingly, the SMR for bladder cancer in this county is also high, 42.9, corresponding to five observed and 0.12 expected counts; this is the county which contains Mount St Helens, which erupted in 1980. As an aside, there is evidence that exposure to dust is a risk factor for lung cancer, and so it is plausible that there would be an increase in cancer risk following the eruption of a volcano. Figures 4 and 5 show maps of the county level SMRs; all maps were produced with the Maptitude Geographical Information Systems (GIS) program (1994–2002 Caliper Corporation). Such maps are notoriously difficult to interpret and in particular depend on the colour scheme used and the cut points of the categories chosen. However, it appears that for both diseases, the rates are higher in the west. Note also that the high census tract rates in tract 9501 have been ‘smoothed away’ by aggregating to the level of the county, showing a disadvantage of examining maps at coarser geographies, particularly if we wish to detect high area specific relative risks. We could have presented maps at the census tract level but such maps are very hard to read (with 1318 tracts), particularly in densely populated areas.

As smoking is a major risk factor for both cancers, we chose to include a measure of smoking prevalence in our models. Our smoking information is based on market segmentation data, provided by Claritas (© 2004 Claritas Inc.). The prevalence measure used is annual per capita spending on cigarettes in 2003, in hundreds of dollars.
It is well documented that most cancer risks show marked gradients when compared with census based measures of socio economic status; many studies in the United Kingdom have used the so called Carstairs index, whereas in the United States income has been commonly used (it is convenient because it is available from the census). The interpretation of such gradients is difficult as we do not know what risk factors
represented by income are being reflected, but there is clearly an element of such measures acting as surrogates for the lifestyle characteristics of individuals within the area (for example known risk factors such as smoking, alcohol and diet). Smoking behaviour can be predicted by area level measures of socioeconomic status. As an attempt to capture information beyond that provided by the smoking data, we also include median family income from the 2000 United States Census, in tens of thousands of dollars; the correlation between the smoking and income variables is $-0.43$. Figure 6 shows scatterplots of the log SMRs for bladder and lung cancer against the smoking and income data, at the level of the census tract. To ease inspection of these plots, the largest SMRs (corresponding to census tract 9501) were removed, as were any areas in which no cases were observed (this is not recommended in general, but here only removed a small fraction of the data, Table 1). As expected, an increase in smoking appears to be associated with an increased risk for both cancers, especially for lung cancer, and an increase in income appears to be associated with a decreased risk, again more markedly for lung cancer.

Figure 6  log SMRs for bladder and lung cancer plotted against smoking and income at the census tract level, local smoothers superimposed. Smoking is measured by annual per-capita spending on cigarettes in 2003 (in hundreds of dollars), and income is median family income (in tens of thousands of dollars) from the 2000 census.
3 Univariate disease mapping

We briefly review some of the details of univariate disease mapping. More details may be found elsewhere.29,30

When spatial smoothing of relative risks is carried out, it has usually been done via the introduction of random effects into the modelling of the area-level relative risks \( \theta_i \) in Equation (2). The convolution model29 is a common choice and takes the form

\[
\log \theta_i = \beta_0 + U_i + S_i, \tag{6}
\]

where \( U_i \) and \( S_i \) represent unstructured and spatially structured random effects, respectively. Let \( S = (S_1, \ldots, S_n) \) be a vector of length \( n \) representing the spatial random effects. There are two general approaches to specifying a spatially structured distribution for \( S \). The first is based on conditional distributions, \( S_i | S_{-i} \), where \( S_{-i} \) denotes the collection \( S \) excluding \( S_i \). The second is through the specification of a multivariate distribution. Note that it is the structured prior for \( S \) that allows us to separate \( U_i \) from \( S_i \) in Equation (6).

We begin by assuming the joint distribution

\[
S \sim N_n(0, \sigma_s^2 \Sigma), \tag{7}
\]

where \( \Sigma \) is an \( n \times n \) positive definite correlation matrix, and \( \sigma_s^2 = \text{var}(S_i) \) is the common variance of the \( S_i, i = 1, \ldots, n \). As we now demonstrate, it is convenient to consider the precision of the correlation matrix \( Q = \Sigma^{-1} \). From standard properties of the normal distribution,

\[
S_i | S_{-i} \sim N \left( -\sum_{j=1}^{n} \frac{Q_{ij}}{Q_{ii}} S_j, \frac{\sigma_s^2}{Q_{ii}} \right).
\]

The two ways of proceeding, via the covariance or its reciprocal, can now be discussed. One way is to specify the joint distribution (7) and assume specific forms for \( \Sigma \). One such choice is the powered exponential form with \( S \sim N(0, \Sigma) \), where \( \Sigma \) has element \( (i, j) \):

\[
\Sigma_{ij} = \sigma_s^2 \exp \{- (\phi d_{ij})^\kappa \}, \tag{8}
\]

where \( d_{ij} \) is the distance between areas \( i \) and \( j \), \( \phi > 0 \) is a measure of the scale of spatial dependence in the sense that large values for \( \phi \) indicate that spatial dependence drops off rapidly with distance, whereas small values indicate a less rapid decline, and \( \kappa \) is a parameter constrained to be in the interval \([0, 2)\) which controls the shape of the smoothing. We take \( \kappa = 1 \) in the analyses of Section 5. Note that this model is stationary and isotropic and so quite restrictive. Other forms are available, such as the Matérn class.31 The multivariate model is computationally expensive to use in practice, as it is necessary to invert an \( n \times n \) matrix in order to obtain the conditional distributions that are required for implementation via Markov Chain Monte Carlo (MCMC) sampling.
This lead to a discretisation of the parameters indexing the correlation matrix, so that the required matrix inversions could be carried out in advance.32 The restriction to a finite number of values is in practice not too restrictive, because there is often very little information in the data to inform on the scale of spatial dependence (this is borne out in the examples of Section 5).

As an alternative to modelling the joint distribution, we may model conditionally through the Q matrix by writing $\rho W_{ij} = -Q_{ij}/Q_{ii}$ and $D_{ii} = 1/Q_{ii}$ to give

$$S_j|S_{-i} \sim N\left(\rho \sum_{j=1}^{n} W_{ij}S_j, \sigma_s^2D_{ii}\right),$$

with $W_{ii} = 0$. We may now define the areas that we would like to contribute to the distribution of $S_j|S_{-i}$, the ‘neighbours’, by specifying $W_{ij} \neq 0$ for those areas and taking care to ensure that the resultant choices are probabilistically legal. For example, symmetry of Q requires

$$W_{ij}D_{jj} = W_{ji}D_{ii}. \quad (9)$$

Let $\partial i$ denote the set, and $n_i$ the number, of neighbours of area $i$. It seems natural to take $D_{ii} = 1/n_i$, so that the more neighbours an area has, the greater the precision in the prior for that area. To satisfy Equation (9), we then take $W_{ij} = n_i^{-1}$ for areas $j \in \partial i$ to give the conditional autoregression (CAR) model

$$S_j|S_{-i} \sim N\left(\frac{\rho}{n_i} \sum_{j \in \partial i} S_j, \frac{\sigma_s^2}{n_i}\right),$$

where constraints are required on $\rho$ to ensure a proper joint distribution.18 It can be shown30 that

$$\text{corr}(S_i, S_j|S_k, k \neq i,j) = \rho(n_i n_j)^{-1/2},$$

so that the conditional correlation is constrained from being high (since $n_i \times n_j$ is usually much greater than 1). It has been further illustrated that it is not possible to model strong spatial correlation using this form.33 As an alternative, the intrinsic CAR (ICAR) model29 has been used which is the limiting form

$$S_j|S_{-i} \sim N\left(\tilde{S}_j, \frac{\sigma_c^2}{n_i}\right),$$

where $\tilde{S}_j = 1/n_i \sum_{j \in \partial i} S_j$. This form does not produce a joint distribution that is proper, and so $\sigma_c^2$ no longer has an interpretation as a marginal variance (hence our change of notation) but rather as a component of the conditional variance (which also depends on the number of neighbours); this makes prior specification for $\sigma_c^2$ more difficult than for
\( \sigma^2_s \). In the proper CAR form, the parameter \( \rho \) is estimated, and so if there is little spatial dependence in the data, this parameter will be close to zero. In the ICAR model, however, this parameter assumes its limiting form, and so to ensure that spatial dependence is not induced, it is recommended that an unstructured component always be included in the convolution model.\(^29\)

The high dimensionality and complicated form for the random effect distributions makes classical inference challenging, although approximate solutions may be possible.\(^24\) Instead, a Bayesian formulation is usually used. A complete Bayesian model requires hyperprior distributions to be assigned to the parameters of the spatial models; these are \((\sigma^2_s, \phi)\) for the joint model and \(\sigma^2_c\) for the conditional model.

In a serious mapping study, one needs to carefully consider artefacts in the maps that may be introduced owing to data anomalies. For example, there may be differential reporting of cancer cases across the region, problems with the population counts due, for example, to migration and census undercount, and errors in the exposure measures used (such as income and smoking in our example). We do not comment further on these aspects (a more complete discussion can be found elsewhere), but note that cancer registration in Washington state is one of the most accurate in the United States. For example, the Fred Hutchinson Cancer Research Center (FHCRC) Cancer Surveillance System (CSS) registry covers the 13 northwest counties of Washington state. The registry has won many recent awards, scoring either first or second in Data Quality Profile among the 10 SEER registries in the United States. The Department of Health data we use in this article share this registry’s data for the 13 northwest counties.

4 Bivariate disease mapping

4.1 Shared component model

In this section, we suppose that we wish to model counts \( Y_{1i}, Y_{2i} \) from two diseases in area \( i \). We assume that these counts arise from two Poisson distributions, \( Y_{ki} \sim \text{Poisson}(E_{ki}/\theta_{ki}) \), \( k = 1, 2, i = 1, 2, \ldots, n \), where \( E_{1i}, E_{2i} \) are the pair of expected numbers for the two diseases.

The key element of the shared component model\(^4\) is a random effect that is common to both log linear relative risks, in addition to disease specific components. Suppose that the true model is of the form

\[
\log \theta_{ki} = \gamma_{k0} + \gamma_{k1} x_{ki} + \gamma_{k2} z_i, \quad (10)
\]

where \( x_{ki} \) is a risk factor specific to disease \( k \), and \( z_i \) is a risk factor for both diseases; suppose further that each of these risk factors are unmeasured. It is natural to model the relative risks as

\[
\log \theta_{1i} = \beta_{10} + \eta_{1i} + \delta \phi_i, \\
\log \theta_{2i} = \beta_{20} + \eta_{2i} + \phi_i / \delta \quad (11)
\]
where $\eta_{1i}$ and $\eta_{2i}$ are disease specific components and $\phi_i$ is a shared component, all of which are assigned distributions (otherwise they would not be identifiable). The reason for the introduction of the $\delta$ parameter in Equation (11) is that there is an additional identifiability associated with this model that arises because the $z_i$ are unmeasured. In particular, consider the idealized situation in which there is only a shared covariate $Z$; that is, $\gamma_{k1} = 0$, $k = 1, 2$. If we fit the model $\log \theta_{1i} = \beta_{10} + \delta \phi_i$, $\theta_{2i} = \beta_{20} + \phi_i/\delta$, then it is straightforward to show that $\delta^2 = \gamma_{12}/\gamma_{22}$. In general, if the shared covariate accounts for most variability, $\eta_1$ and $\eta_2$ will be approximately zero, and the above relation will be approximately true.

Heuristically, the shared component can be thought of as a weighted average of spatial random effects from individual models. Similarly, the disease specific components can be thought of as differences between the individual level random effects and their average; these comments are explored further in Section 5.2. The shared component model thus allows us to investigate shared and divergent spatial trends in two diseases. Maps of these trends can be used to gain clues to risk factors that are common to the two diseases or are disease specific.

The shared component model was originally specified using a cluster model as the prior distribution for the random effects. However, owing to the fact that the cluster model is not available in standard software packages and hence requires specialist software to be written, we do not use it here.

A natural alternative to using the cluster model is to use conditional or joint models as discussed in Section 3. The convolution model is a standard tool in univariate disease mapping, and so we might consider extending it to the bivariate setting. In its most obvious formulation, this would involve having two shared components, one unstructured and one spatial, and two disease specific components per disease, giving six random effects for each area in total. Such a model seems overly complex and unidentifiable and so here we will use model (11) with $\eta_{1i}$, $\eta_{2i}$, and $\phi_i$ each assigned spatial priors with the powered exponential model. Examination of the disease specific and shared components across space will then reveal the extent to which the risks are similar or dissimilar.

### 4.2 Proportional mortality model

The proportional mortality model has been used when counts are available for deaths only, and not for the population at risk. In this situation, one may derive the distribution of the number of deaths for the cause of interest, relative to all deaths. We now describe how this model may be used in the two disease situation.

Let us assume $Y_{ki} \sim \text{Poisson}(\lambda_{ki})$, $k = 1, 2$, $i = 1, 2, \ldots, n$, where we begin by assuming $\log \lambda_{ki} = x_{k0} + x_{k1} x_i$, so that there are no random effects, but $x_i$ represents an area level covariate associated with both diseases through the area level relative risks $\exp(x_{1i})$ and $\exp(x_{2i})$. If we define $M_i = Y_{1i} + Y_{2i}$ as the sum of the two disease counts in area $i$, then

$$Y_{1i}|M_i = m_i \sim \text{Binomial}(m_i, \gamma_i), \quad (12)$$
where
\[
\gamma_i = \frac{\exp \{ (\alpha_{10} - \alpha_{20}) + (\alpha_{11} - \alpha_{21}) x_i \}}{1 + \exp \{ (\alpha_{10} - \alpha_{20}) + (\alpha_{11} - \alpha_{21}) x_i \}}.
\] (13)

Hence, we can model the data with a logistic model of the form
\[
\text{logit}(\gamma_i) = \beta_0 + \beta_1 x_i,
\]
where the parameter \(\beta_1\) can be interpreted as the difference \(\alpha_{11} - \alpha_{21}\). In this way, inference can be made on the differences in log relative risks without knowledge of the population counts of those at risk (which, recall from Section 3, may be subject to data anomalies). Notice also that, when the second disease is not related to the covariate \(x\) (i.e., \(\alpha_{21} = 0\)), \(\beta_1 = \alpha_{11}\).

We now apply the proportional mortality model to the problem of mapping two diseases by introducing random effects. In this setting, we have expected counts available, and we use the proportional mortality model as a method for jointly modeling the two diseases in a way that highlights their differences. Consider first the following individual disease mapping models for each disease: \(\log \theta_{ki} = \alpha_{k0} + U_{ki} + S_{ki}\), where the \(U_{ki}\)-values are spatially unstructured, and the \(S_{ki}\)-values are spatially structured random effects for disease \(k\), \(k = 1, 2\). We can then consider the Binomial distribution that results from conditioning on the total number of deaths in each area, identical to Equation (12) with
\[
\gamma_i = \frac{\exp \{ \log (E_{1i}/E_{2i}) + (\alpha_{10} - \alpha_{20}) + (U_{1i} - U_{2i}) + (S_{1i} - S_{2i}) \}}{1 + \exp \{ \log (E_{1i}/E_{2i}) + (\alpha_{10} - \alpha_{20}) + (U_{1i} - U_{2i}) + (S_{1i} - S_{2i}) \}},
\] (14)
giving the logistic model
\[
\text{logit}(\gamma_i) = \log \frac{E_{1i}}{E_{2i}} + \beta_0 + U_i^* + S_i^*.
\] (15)

The log ratio of expected counts is included as an offset term to account for any differential spatial trends in expected numbers for the two diseases. Now \(S_i^*\) can be thought of as the difference \(S_{1i} - S_{2i}\), so that the \(S_i^*\)-values capture similarity and dissimilarity between the spatial random effects of each disease. That is, a large absolute value of \(S_i^*\) indicates that the two diseases differ significantly in risk in area \(i\) in terms of the spatial component, whereas a value close to zero indicates the two disease risks are similar in this area. A map of the spatial random effects hence highlights shared and divergent trends between the two diseases. Such a map could be used in conjunction with maps of various risk factors to generate hypotheses about risk patterns that distinguish the two diseases. Note that the \(U_i^*\) can similarly be thought of as the difference \(U_{1i} - U_{2i}\), and a map of these quantities would point out areas in which the effects of unobserved nonspatial covariates are similar or different between the two diseases.

To summarize this model, we assume a binomial model for \(Y_{1i}|Y_{1i} + Y_{2i} \sim \text{ind} \text{Binomial}(Y_{1i} + Y_{2i}, \gamma_i)\), with the logit of \(\gamma_i\) given by Equation (15). For both the shared component and the proportional mortality models, area level covariates may be added to the linear predictor.
5 Analyses of lung and bladder cancer incidence

5.1 Univariate analyses

We first analyse the bladder and lung cancer data via individual disease mapping models. We fit a series of models and compare them with shared component and proportional mortality models in subsequent sections. The comparison will clarify some of the features of these models. We will analyse the data at the census tract level, with random effects introduced at the county level to alleviate the computational burden of the spatial models. Letting $i$ index county and $l$ index census tract, we assume $Y_{i,l} \sim \text{Poisson}(E_{i,l}, \theta_{i,l})$ for each disease and fit the models:

- **Model I**: $\log \theta_{i,l} = \beta_0 + U_i$, \\
- **Model II**: $\log \theta_{i,l} = \beta_0 + U_i + S_i$, \\
- **Model III**: $\log \theta_{i,l} = \beta_0 + \beta_{\text{SMK}} \times \text{Smoke}_l + U_i + S_i$, \\
- **Model IV**: $\log \theta_{i,l} = \beta_0 + \beta_{\text{INC}} \times \text{Income}_l + U_i + S_i$, \\
- **Model V**: $\log \theta_{i,l} = \beta_0 + \beta_{\text{SMK}} \times \text{Smoke}_l + \beta_{\text{INC}} \times \text{Income}_l + U_i + S_i$.

Comparison of random effect variances between Models I and II will inform on the amount of spatial dependence between counties. In addition, recalling that we interpret the random effects as unobserved covariates, comparison of Models II, III, IV, and V will reveal how much information about the unobserved covariates is recovered by including smoking and/or income in the models. We note that we have used a priori standardization, and so some of the effects of smoking and income may already be accounted for in the expected numbers. For example, it may be that the elderly are more likely to smoke, and so some of the age effect may be due to smoking. This is closely related to the issue of mutual standardisation.36 Here, we are not interested in obtaining accurate measures of the relative risks associated with the area level covariates but rather in obtaining better area level predictions and accounting for variability in the observed risks.

In Model I, we assign $\beta_0$ a $N(0, 1.531^2)$ prior distribution, corresponding to a prior belief that the intercept is between $-3$ and $3$ with probability $0.95$. In Models II–V, the inclusion of the ICAR random effect in the WinBUGS software package37 requires a flat prior on the intercepts. It is expected that smoking has a greater impact on risk than income. We then assume independent $N(0, 2.551^2)$ and $N(0, 1.531^2)$ priors for $\beta_{\text{SMK}}$ and $\beta_{\text{INC}}$ parameters, respectively. The first assumes that the ecological smoking relative risk associated with a one unit change in annual per capita spending (in hundreds of dollars) is between $\exp (\pm 5)$ with probability $0.95$, whereas the second assumes that the ecological income relative risk is between $\exp (\pm 3)$ with probability $0.95$. The unstructured random effects are assumed to have the distribution $N(0, \sigma_u^2)$, whereas the structured random effects $S_i$ are assumed to follow the ICAR model with conditional variance $\sigma_c^2$. Independent InvGamma $(0.5, 0.0005)$ priors are used for the hyperparameters $\sigma_u^2$ and $\sigma_c^2$ (more details on this model can be found elsewhere38).
All analyses in the article were carried out using WinBUGS which implements Markov Chain Monte Carlo (MCMC) in order to simulate dependent samples from the posterior distributions (all of which are analytically intractable for the models considered). In the MCMC runs, a single chain was used. We used the first 20 000 draws as the burn-in period and then drew 10 000 more samples. After thinning by 10, we were left with 1000 samples to base posterior summaries upon. Tables 2 and 3 show summary statistics from Models I–V for bladder and lung cancer, respectively; the intercepts were not reported but were close to zero in all models. Recall that $\sigma^2_c$ is difficult to interpret. To allow comparison with $\sigma_u$, we empirically evaluated the standard deviation, which we call $\sigma_s$, of the collection $S$. Point estimates are posterior medians, and interval estimates are 95% posterior credible intervals.

Comparing Models I and II, we see that when the spatial random effects are added, they account for the majority of the variability, for bladder cancer in particular. The proportion of variance attributed to the structured random effect is $0.13^2/(0.04^2 + 0.13^2) = 0.91$ for bladder cancer and 0.77 for lung cancer. This is an indication that missing risk factors (or data anomalies) are mostly spatially structured.

Ideally, adding information on smoking and income would substantially reduce the residual spatial and unstructured variability, as measured through the random effects. Comparing Models III–V with Model II, we see that this is not the case in this example. This could be explained by the fact that the smoking and income information available to us does not explain any of the variability in the incidence rates. However,

<table>
<thead>
<tr>
<th>Model</th>
<th>$\sigma_u$</th>
<th>$\sigma_s$</th>
<th>$\exp(\beta_{SMK})$</th>
<th>$\exp(\beta_{INC})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I</td>
<td>0.13</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Model II</td>
<td>0.04</td>
<td>0.13</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Model III</td>
<td>0.04</td>
<td>0.12</td>
<td>1.18 (1.12, 1.25)</td>
<td>–</td>
</tr>
<tr>
<td>Model IV</td>
<td>0.03</td>
<td>0.13</td>
<td>–</td>
<td>0.9953 (0.9935, 0.9970)</td>
</tr>
<tr>
<td>Model V</td>
<td>0.04</td>
<td>0.11</td>
<td>1.19 (1.12, 1.25)</td>
<td>0.9979 (0.9959, 1.0001)</td>
</tr>
</tbody>
</table>

**Table 2** Summaries for univariate disease mapping models for bladder cancer: posterior medians and 95% credible intervals

*Note:* The relative risks for smoking are the multiplicative changes in risk for two areas that differ in their per capita cigarette sales by 100 dollars; those for income are associated with two areas that differ in their median income by 10 000 dollars.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\sigma_u$</th>
<th>$\sigma_s$</th>
<th>$\exp(\beta_{SMK})$</th>
<th>$\exp(\beta_{INC})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I</td>
<td>0.13</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Model II</td>
<td>0.06</td>
<td>0.11</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Model III</td>
<td>0.08</td>
<td>0.10</td>
<td>1.36 (1.32, 1.40)</td>
<td>–</td>
</tr>
<tr>
<td>Model IV</td>
<td>0.06</td>
<td>0.10</td>
<td>–</td>
<td>0.9872 (0.9862, 0.9883)</td>
</tr>
<tr>
<td>Model V</td>
<td>0.07</td>
<td>0.10</td>
<td>1.36 (1.32, 1.40)</td>
<td>0.9896 (0.9883, 0.9901)</td>
</tr>
</tbody>
</table>

**Table 3** Summaries for univariate disease mapping models for lung cancer: posterior medians and 95% credible intervals

*Note:* The relative risks for smoking are the multiplicative changes in risk for two area that differ in their per capita cigarette sales by 100 dollars; those for income are associated with two area that differ in their median income by 10,000 dollars.
the large coefficient for the smoking variable suggests that its inclusion should reduce the model’s variability. It has been noted in previous work\(^\text{39}\) that it is possible for random effect variability to actually increase with the inclusion of predictor variables. A similar situation holds here. For both King and Snohomish counties, the average of the component census tract level log SMRs is very close to the global mean of all the census tract level log SMRs. This means that the intercept term in a model without any covariates provides a good estimate of the county level SMR. The random effects in these counties will then have values close to zero. However, these two counties have the highest incomes out of all the Washington counties. As the regression coefficient for income is slightly less than one, including income in the model results in county level estimates of log SMRs which are smaller than the global mean. The random effects will therefore need to be large to provide estimates which correspond to the observed values. This increase in magnitude of random effects prevents the random effect variance from decreasing. In our example, the spatial random effects for bladder cancer are 0.003 and 0.024 for Snohomish and King counties, respectively, in Model II. In comparison, the same random effects are 0.198 and 0.178 in Model IV, the model that includes income.

The point estimate of the ecological relative risk associated with smoking from Model III indicates an 18% increase in risk between areas that differ by 100 dollars in their per capita cigarette sales. With income only in the model (Model IV) there is a 0.47% decrease in risk for two areas whose median income differs by 100 000 dollars. With smoking and income in the model, the income point estimate moves towards 1 (as expected given the negative correlation between the variables) and the income coefficient is no longer statistically significant. For lung cancer the conclusions are similar, though income remains statistically significant when smoking is added to the model.

Figure 7 compares the raw SMRs with the smoothed relative risks, at both the county and census tract level. It is evident that the smoothed relative risks shrink the SMRs towards 1. There is greater shrinkage for bladder cancer, as that is less prevalent than lung cancer. Figures 8 and 9 show maps of the smoothed relative risks. Note that the scaling used is the same as that in Figures 4 and 5 (all maps in this article are on this scale). For bladder cancer, as we would anticipate from the shrinkage plots, the map is much ‘flatter’.

For both cancers, there appears to be a trend of increasing risk from east to west. At a finer level of detail there appears to be an increasing trend from southeast to northwest for bladder cancer and a, perhaps less distinct, increasing trend from northeast to southwest for lung cancer. We would thus expect a shared component model to reflect the east–west trend in its shared component and the separate southeast–northwest and northeast–southwest trends in its disease specific components. Similarly, we might expect a proportional mortality model to highlight the regions of dissimilarity, namely in the northwest and southwest.

### 5.2 Shared component model

As in Section 5.1, we fit a collection of models to explore various characteristics of the shared component model. Given that the spatial random effects were dominant for both cancers, we only consider such effects in the shared component model, and suppose that they arise from the exponential model. We chose this model because the ICAR alternative requires an unstructured component also, and we wish to include as small a number of random effects as possible. For a generic model, we have a correlation between areas }
Figure 7  Shrinkage of SMRs for bladder and lung cancer. The plots show line segments connecting: 1) the SMR with the smoothed relative risk from an individual analysis (ID) and 2) the smoothed relative risk from an individual analysis with the smoothed relative risk from a shared component analysis (SC). The first two plots show county level estimates, whereas the second two are at the census tract level.

Figure 8  Smoothed relative risks for bladder cancer.
and \( j \) of \( \exp(-\phi d_{ij}) \) where \( d_{ij} \) is the distance between the centroids of areas \( i \) and \( j \). Letting \( \Sigma(\phi) \) denote the correlation matrix as a function of \( \phi \) we have

\[
\begin{align*}
S_{BLA} & \sim N(0, \sigma^2_{BLA} \Sigma(\phi_{BLA})), \\
S_{LUN} & \sim N(0, \sigma^2_{LUN} \Sigma(\phi_{BLA})), \\
S & \sim N(0, \sigma^2_{Shared} \Sigma(\phi_{Shared}))
\end{align*}
\]

for the three sets of spatial random effects, with an obvious notation.

We then consider the following models:

**Model I:**

\[
\begin{align*}
\log \theta_{BLA,i,l} &= \beta_{BLA,0} + S_{BLA,i} + \delta \times S_i \\
\log \theta_{LUN,i,l} &= \beta_{LUN,0} + S_{LUN,i} + S_i/\delta,
\end{align*}
\]

**Model II:**

\[
\begin{align*}
\log \theta_{BLA,i,l} &= \beta_{BLA,0} + \beta_{BLA,SMK} \times \text{Smoke}_i + S_{BLA,i} + \delta \times S_i \\
\log \theta_{LUN,i,l} &= \beta_{LUN,0} + \beta_{LUN,SMK} \times \text{Smoke}_i + S_{LUN,i} + S_i/\delta,
\end{align*}
\]

**Model III:**

\[
\begin{align*}
\log \theta_{BLA,i,l} &= \beta_{BLA,0} + \beta_{BLA,INC} \times \text{Income}_i + S_{BLA,i} + \delta \times S_i \\
\log \theta_{LUN,i,l} &= \beta_{LUN,0} + \beta_{LUN,INC} \times \text{Income}_i + S_{LUN,i} + S_i/\delta,
\end{align*}
\]

**Model IV:**

\[
\begin{align*}
\log \theta_{BLA,i,l} &= \beta_{BLA,0} + \beta_{BLA,SMK} \times \text{Smoke}_i + \beta_{BLA,INC} \times \text{Income}_i + S_{BLA,i} + \delta \times S_i \\
\log \theta_{LUN,i,l} &= \beta_{LUN,0} + \beta_{LUN,SMK} \times \text{Smoke}_i + \beta_{LUN,INC} \times \text{Income}_i + S_{LUN,i} + S_i/\delta.
\end{align*}
\]
Hence, Model I has no covariates, II includes smoking for both cancers, III includes income for both cancers and IV includes smoking and income for both cancers. Comparison of the variability of the shared component $S$ in Model I with that in Models II–IV will inform on the extent to which the shared component is picking up on smoking and income information. If, for example, the shared component is flattened out when smoking and/or income information is added to the model, it would be an indication that these variables are part of the missing information that is shared by both diseases. We will also be able to compare the shared components with the random effects from the individual models fit in Section 5.1 to gain intuition into what the shared component represents.

We used priors for the intercepts and slope parameters that were identical to their counterparts in Section 5.1. The log of the $\delta$ parameter was assigned a $N(0, 0.4106^2)$ prior, corresponding to the prior belief that the gradient $\delta^2$ is within $(1/5, 5)$ with probability 0.95. Each of the variance components were given independent InvGamma$(0.5, 0.0005)$ priors. For the correlation parameters $\phi_{BLA}$, $\phi_{LUN}$ and $\phi_{Shared}$, we chose uniform prior distributions with lower bound such that there is correlation 0.01 at the maximum distance possible and upper bound such that there is correlation 0.01 at the minimum distance possible. The minimum and maximum distances possible between Washington state county centroids are 19.06 and 334.37 miles, respectively. We partitioned the map of Washington state into units of 100 miles. Thus, for correlation 0.01 at distance 334.37 miles, we require $\phi = -\log (0.01)/3.3437 = 1.3773$ (this value gives correlation of 0.5 at 50 miles). Similarly, we desire an upper bound of 24.1656, and so we assign independent U$(1.3773, 24.1656)$ priors for the correlation parameters. The MCMC algorithm was run with the same choices of number of chains, burn-in, and number of posterior draws as in Section 5.1.

Table 4 shows results for the variance components. The summary measure reported for the correlation parameters is the estimated distance at which correlation drops to 0.5, in miles, $d_{1/2} = \log 2/\phi$ where $\phi$ is the posterior median. Table 5 shows results for the regression parameters. Except as noted for the correlation parameters, point estimates are posterior medians and interval estimates are 95% posterior credible intervals.

The shared component has greater variability than the disease specific components. In Model I, for example, the shared component accounts for 77% of the variability for bladder cancer and 71% of the variability for lung cancer. Not only is the magnitude of the shared spatial components greater, but the correlations of the shared components

| Table 4 Summary statistics for the variances of the random effects distributions from the shared components model, and the distances (in miles) at which the correlations fall to 0.5 |
|---------------------------------|-----|-----|-----|-----|-----|
| Model I            | $\sigma_{BLA}$ | $\sigma_{LUN}$ | $\sigma_{Shared}$ | $d_{BLA,1/2}$ | $d_{LUN,1/2}$ | $d_{Shared,1/2}$ |
| Model II           | 0.05 | 0.08 | 0.11 | 6.2  | 6.1  | 7.4  |
| Model III          | 0.06 | 0.07 | 0.11 | 6.6  | 5.5  | 16.2 |
| Model IV           | 0.06 | 0.06 | 0.11 | 7.0  | 7.0  | 7.0  |
die down at a slower rate than in the disease specific components, as seen in the
distances at which correlations drop to 0.5, and as shown in Figure 10.

As smoking and income information is added to the models, the variance components
remain largely unchanged. This is similar to the results of Section 5.1.

The relative risks associated with smoking and income in Models II–IV are nearly
identical to those reported in the individual disease mapping analyses. The gradient
parameter \( \delta \) is centered at one for each model. Any residual variation in risk, then, is
assumed to be due to missing covariates whose associations with the two diseases are
similar in magnitude.

Figures 11 and 12 show the exponentiated disease specific random effects for bladder
and lung cancer, respectively. Figure 13 shows the exponentiated shared components.
There is some evidence of an increasing southeast–northwest trend that is specific to
bladder cancer and of an increasing northeast–southwest trend that is specific to lung

<table>
<thead>
<tr>
<th></th>
<th>( \exp(\beta_{\text{BLA,SMK}}) )</th>
<th>( \exp(\beta_{\text{LUN,SMK}}) )</th>
<th>( \exp(\beta_{\text{BLA,INC}}) )</th>
<th>( \exp(\beta_{\text{LUN,INC}}) )</th>
<th>( \delta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>1.1</td>
</tr>
<tr>
<td>Model II</td>
<td>1.20 (1.13, 1.27)</td>
<td>1.36 (1.31, 1.41)</td>
<td>–</td>
<td>–</td>
<td>1.1</td>
</tr>
<tr>
<td>Model III</td>
<td>–</td>
<td>–</td>
<td>0.9950 (0.9932, 0.9968)</td>
<td>0.9872 (0.9862, 0.9883)</td>
<td>1.1</td>
</tr>
<tr>
<td>Model IV</td>
<td>1.20 (1.13, 1.26)</td>
<td>1.36 (1.32, 1.40)</td>
<td>0.9976 (0.9955, 0.9997)</td>
<td>0.9896 (0.9884, 0.9909)</td>
<td>1.1</td>
</tr>
</tbody>
</table>

*Note:* Posterior medians and 95% credible intervals are reported.
There is a clear increasing east–west trend that is common to the two cancers. Recall from Section 5.1 and Figures 8 and 9 that this is expected. Notice also that the disease specific random effects are flatter than the shared components. This agrees with the results in Table 4 which showed that the shared component has greater variability than the disease specific components.

The shared component model can then be used to explore trends which are shared between two diseases. It also provides some information about trends that are specific.
to each disease, but the majority of the information is on the shared component. In Figure 7, we also show the shrinkage associated with the shared component model, and we see very little difference from that of the univariate models alone. The shared component represents an average of the two disease risks. Figure 14 illustrates this by showing a scatterplot of the shared components from Model I against the average of structured random effects from Models I in Section 5.1. Given how strongly we can predict the shared component from the individual components, it is not clear that a formal shared model provides much greater information (and requires strong assumptions in order to be identifiable).

5.3 Proportional mortality model

We fit a similar collection of models as in Section 5.1. Specifically, with $\gamma$ defined as in Equation (14),

Model I: \[ \text{logit } \gamma_{i,j} = \beta_0 + U_i, \]
Model II: \[ \text{logit } \gamma_{i,j} = \beta_0 + U_i + S_j, \]
Model III: \[ \text{logit } \gamma_{i,j} = \beta_0 + \beta_{\text{SMK}} \times \text{Smoke}_i + U_i + S_j, \]
Model IV: \[ \text{logit } \gamma_{i,j} = \beta_0 + \beta_{\text{INC}} \times \text{Income}_i + U_i + S_j, \]
Model V: \[ \text{logit } \gamma_{i,j} = \beta_0 + \beta_{\text{SMK}} \times \text{Smoke}_i + \beta_{\text{INC}} \times \text{Income}_i + U_i + S_j. \]
Note that we have formulated the proportional mortality model such that we are modelling bladder cancer as a function of the sum of bladder and lung cancer; hence regression parameters and random effects are of the form $\text{bladder} - \text{lung}$. Prior distributions and MCMC details were chosen precisely as in Section 5.1. Table 6 shows posterior summaries of parameters of interest.

The proportion of random effect variance accounted for by the spatial random effect in Model II is 0.8. In interpreting the odds ratio parameters, it is useful to reference Tables 2 and 3. The odds ratio associated with smoking is $<1$. Recall that this parameter corresponds to a ratio of smoking relative risks, comparing bladder cancer to lung cancer. Thus, the indication is that the relationship between smoking and lung cancer is stronger than that between smoking and bladder cancer. Similarly, the odds ratio associated with income is slightly $>1$, indicating that the relationship between income and bladder cancer is slightly stronger than that between income and lung cancer.

Table 6. Proportional mortality summary statistics: posterior medians and 95% credible intervals

<table>
<thead>
<tr>
<th>Model</th>
<th>$\sigma_u$</th>
<th>$\sigma_s$</th>
<th>exp ($\beta_{\text{SMK}}$)</th>
<th>exp($\beta_{\text{INC}}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I</td>
<td>0.09</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Model II</td>
<td>0.04</td>
<td>0.08</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Model III</td>
<td>0.04</td>
<td>0.10</td>
<td>0.87 (0.82, 0.93)</td>
<td>–</td>
</tr>
<tr>
<td>Model IV</td>
<td>0.05</td>
<td>0.08</td>
<td>–</td>
<td>1.0066 (1.0045, 1.0086)</td>
</tr>
<tr>
<td>Model V</td>
<td>0.04</td>
<td>0.10</td>
<td>0.87 (0.82, 0.93)</td>
<td>1.0065 (1.0039, 1.0093)</td>
</tr>
</tbody>
</table>
Figure 15  Exponentiated spatial random effect from proportional mortality model.

Figure 16  Proportional mortality spatial random effect as difference of individual random effects. The line of equality is dashed, and the solid line is a smoother.
Figure 15 is a map of the exponentiated structured random effects. The dominant feature of the map is an increasing trend from the southeast to the northwest. This is as was expected, because we saw in Section 5.1 that the main distinguishing characteristic of the two diseases was an increasing southeast to northwest trend for bladder cancer and an increasing northeast to southwest trend for lung cancer. It is then clear that the proportional mortality model can be used in a bivariate disease mapping setting to highlight trends of similarity and dissimilarity. This point is reiterated by a plot of the structured random effect from the proportional mortality model against the difference of structured random effects from the individual disease mapping analyses in Section 5.1, shown in Figure 16. Note that the critique of the shared component model at the end of Section 5.2 applies equally here.

6 Discussion

In all models used, we performed sensitivity analyses to assess the degree to which our results were dependent on choices of prior distributions. There was generally little change in result when priors were changed. We do note, however, that the shared component models were sensitive to the choice of prior for the spatial dependence parameter \( \phi \). In addition, it may be appropriate in future studies to include an unstructured disease specific random effect to represent heterogeneity in the shared component models. The large increase in the magnitude of the spatial dependence parameter for the shared component relative to those for the disease specific components observed in Table 4 does not occur when such a heterogeneity term is included in the models.

In this article, we have discussed a number of issues relating to the modelling of a pair of diseases. As in all disease mapping exercises, the quality of the data will determine the extent to which useful information can be obtained. In terms of ecological bias, data at the lowest geographical level is the most desirable, but such data are likely to be the most unreliable, as they have a greater susceptibility to data anomalies such as migration and inaccurate geocoding of addresses.

We have proposed a proportional mortality approach to detecting areas with similar or dissimilar patterns of residual disease risk. For the example of the paper, the mapping of bladder and lung cancer in the state of Washington, the model proved useful, when combined with univariate analyses.

We found that the degree of spatial variability in risk is not of great magnitude (when compared to age effects, for example), but we saw significant increasing east–west trends in residual ecological relative risk for both cancers. This merits future investigation.

Acknowledgements

The authors would like to thank Richard Hoskins of the Washington State Department of Health for many useful discussions and for help in obtaining the incidence data used in the article. We would also like to thank Tom Vaughan of the Fred Hutchinson Cancer Research Center for help in obtaining the smoking information used in the analysis.
References


24 Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models.


The following queries have arisen during the typesetting of your manuscript. Please answer these queries.

<table>
<thead>
<tr>
<th>Q1</th>
<th>Please provide complete author list for Ref. 6 and 20.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q2</td>
<td>Please provide publisher location for Refs. 9, 10 and 30.</td>
</tr>
<tr>
<td>Q3</td>
<td>Please provide the publisher details for Ref. 37.</td>
</tr>
</tbody>
</table>
