Multiple cancer sites incidence rates estimation using a multivariate Bayesian model

Renato M Assunção¹ and Mónica SM Castro²

Accepted 3 October 2003

Background In Brazil cancer incidence rates have to be estimated from occasional surveys, due to lack of continuous cancer registries. Many estimated rates have very large variances, because only few years of data were collected. When dealing with a single cancer site, it is possible to adopt a Bayesian method which borrows information about the cancer rates from other geographical areas to estimate the cancer rate in a given area. We suggest an additional improvement to this method which explores the correlation between multiple cancer sites rates in a same area and in different areas.

Methods Our method works with a multivariate vector of different cancer sites rates in several areas and it borrows information from both, across geographical areas and across different cancer sites. We applied our method to data from a survey carried out in 18 Brazilian cities in São Paulo State in 1991. We estimated age and sex indirect standardized incidence rates for the six most common cancers in men and women, and calculated the 95% interval estimation for the incidence rates.

Results The usual indirect standardized incidence rates had very large confidence intervals for many cancers and cities due to small expected number of cases. The use of the multivariate Bayesian method led to more precise estimates.

Conclusions More precise age-standardized cancer incidence rates can be calculated using data from other cancers. The method is conceptually simple, easy to perform, has low cost, and can improve substantially the estimation of cancer incidence and other vital rates.

Keywords Cancer incidence rates, cancer registries, epidemiological methods, Bayesian estimation, neoplasms, statistics

Health authorities frequently require different population morbidity and mortality rates, such as different age, sex and ethnic group rates or rates for different areas. If the risk populations are small, the disease is rare, or the observation period is short, the usual rates have little precision, producing unstable rate estimates. It is usual to have the extreme rates values associated with the smaller populations, with no relationship with the underlying risks.¹ If many populations are under investigation, we can observe a great variation in the rates estimates, poorly reflecting genuine geographical heterogeneity of the underlying rates. Therefore, more efficient estimation is essential if accurate information is desired for surveillance and decision making.

One example of these problems appears in developing countries, which have great difficulties in creating and maintaining continuous cancer registries from which incidence rates can be regularly calculated. Occasionally, these countries make large and expensive efforts to collect incidence data for several cancers in one or more geographical locations. However, due to the short period of data collection, many of the estimated rates have large variances, becoming unreliable measures of risk for their respective populations.

To overcome this kind of difficulty, as well for other practical purposes, Bayesian methods have been proposed in the literature.²⁻⁷ Among its main advantages, Bayesian methods allow for the incorporation of prior information, the ease of complex model computation, and the use of posterior probabilities to make inferences about the unknown parameters. The large literature on those methods is due to recent developments in Markov chain Monte Carlo (MCMC) methodology, which

¹ Department of Statistics, Minas Gerais Federal University (UFMG), Belo Horizonte, Minas Gerais, Brazil.
² Spatial Statistics Laboratory—LESTE/UFMG and Belo Horizonte Municipal Health Division (SMSA-BH), Belo Horizonte, Minas Gerais, Brazil.

Correspondence: Dr R Assunção, Department of Statistics, Minas Gerais Federal University (UFMG), Av. Antonio Carlos 662M, Belo Horizonte CEP 31.270-901, Minas Gerais, Brazil. E-mail: assuncao@est.ufmg.br
facilitates the implementation of Bayesian analysis of complex models and datasets.\textsuperscript{8}

The main idea underlying use of Bayesian method to estimate incidence rates of several geographical areas is the recognition that useful information exists in other population’s data to estimate a given population cancer rate.\textsuperscript{9,10} The resulting Bayesian relative risk estimate for each local population is a form of shrinkage of the observed risk (i.e. the indirect standardized incidence ratio (SIR)—observed/expected cases) towards the mean relative risk in the global population. The amount by which each SIR is shrunk towards the global mean is inversely related to the expected count in that local population. Since smaller populations tend to have smaller expected counts, the amount of shrinkage is usually larger for those populations where there is less confidence in their observed risk. Alternatively, the Bayesian estimates can be seen as adjustments due to overdispersion with respect to a Poisson model for the counts. That is, additionally to known risk factors such as the age–sex structure, other unobserved local population characteristics can affect the expected counts, making the observed risk variation larger than that assumed in a Poisson model based solely on the covariates.\textsuperscript{10,11} The main consequence of using Bayesian estimators is that the relative risks of all populations are estimated with larger precision than if naive rates estimators are used and this is the reason for Bayesian techniques being widely used for simultaneous estimation of related parameters.

The Bayesian method has been intensely developed in the context of mapping cancer incidence rates. Besag \textit{et al.} proposed a model that generated much interest.\textsuperscript{11} They imposed a plausible spatial relationship structure among the geographical areas in a map and modelled the relative risks with a spatial Markov distribution. As a consequence, the information on neighbouring areas was used to improve the estimation in a given area. Their model has been extended in several directions to deal, for example, with space and time simultaneously, with missing observations, with mismatched geographical data structures, and with errors in covariates.\textsuperscript{12–14}

In these methods, only information from a single cancer site rate is used. In contrast with the large literature dealing with univariate rates, there has been little work done on methods for vectors of different cancer rates. There have been a few exceptions such as the empirical Bayes method proposed by Longford\textsuperscript{15} and the fertility rates pattern in small areas\textsuperscript{16} not been continuous in time. In 1995, there were only six registries operating, all of them covering only large city geographical areas and with recent and different starting dates. A recent study used simple methods to geographically interpolate from those sparsely dispersed registries, but it is likely to produce biased results because there were large variations in risk among the few centres used in the study, which are in general geographically far from each other.\textsuperscript{20}

In an attempt to collect more geographically refined data, a cancer incidence survey was conducted in 18 cities spread in São Paulo state.\textsuperscript{21} Incidence data for the six most common cancers for each sex in those cities provided estimated rates that showed large variations, especially in smaller cities. These variations were attributed to a variety of factors, including small population sizes and the short reference time period. We applied our method to these data, illustrating a number of practical issues in the analysis.

**Methods**

We used data from a field survey done in 18 cities in São Paulo state in 1991. Incidence information on the six most common cancers in those cities in men (lung, stomach, prostate, oesophagus, colon/rectum, larynx) and women (lung, stomach, breast, cervix/uteri, colon/rectum, ovary) were collected by field workers and were used to calculate standardized rates.\textsuperscript{21}

Initially statistical evidence of correlation between different cancer sites was sought through an analysis of the 12 cancer rates set. In each area \(i\), we have the number \(C_{ij}\) of cases from cancer site \(j\) which we assume to be a Poisson random variable with expected value equal to \(\theta_j E_{ij}\) where \(E_{ij}\) is the expected number of cases under some incidence rates schedule and \(\theta_j\) is the area-specific relative risk for cancer site \(j\). The counts \(C_{ij}\) are supposed to be conditionally independent given the underlying relative risks. Usually, the \(E_{ij}\) are calculated by the sum

\[
E_{ij} = \sum_m \pi_{im} r_{jm}q
\]

where \(\pi_{im}\) is the area \(i\) risk population in age–sex class \(m\) and \(r_{jm}\) is the known cancer \(j\) incidence rate from some reference population in age–sex class \(m\). In this paper, we used the age–sex rates from the pooled 18 cities’ populations as the reference rates set.

The unknown parameter \(\theta_j\) is the area \(i\) relative risk for cancer site \(j\). Hence, each geographical area has an unknown vector \(\theta_j = (\theta_{j1}, \ldots, \theta_{jk})\) of true underlying incidence relative risks where \(k\) is the number of cancer sites in the study. The usual estimator of \(\theta_j\) is the standardized incidence ratio (SIR) \(C_{ij}/E_{ij}\), which is the maximum likelihood estimator under the above distributional assumptions.

There can be a large positive correlation between the true underlying incidence relative risks from two different cancer sites, as we will demonstrate in the next section. If this is the case, when incidence of cancer site \(A\) in a given area is greater (or smaller) than the global cancer site \(A\) incidence, cancer site \(B\) incidence in that same area also tends to be larger.
(or smaller) than global cancer site B incidence. Not all pairs of cancers are highly correlated but when some of them are so, the Bayesian method explained next explores this correlation between multiple cancers to better estimate the incidence rates.

Rather than modelling the \( \theta_i \) vectors directly, it is usual to work with the logarithm of the relative risks, represented by

\[
\phi_i = (\phi_{i1}, \ldots, \phi_{ik}) = \log(\theta_i) = (\log(\theta_{i1}), \ldots, \log(\theta_{ik})).
\]

Working with restricted parameter spaces, such as positive real numbers in the case of \( \theta_i \), is mathematically inconvenient and is the main reason for the transformation of the relative risks.

We assume that each geographical area withdraws \( \phi_i \) independently from a multivariate probability distribution with mean \( \mu = (\mu_1, \ldots, \mu_k) \) and a \( k \times k \) covariance matrix \( \Sigma \). If the internal standardization is used, the value \( \mu_i \) should be 0 or very close to 0. The covariance matrix \( \Sigma \) will have usually, but not always, non-negative elements reflecting the fact that, if a given location \( i \) has the relative risk \( \theta_j \) of cancer \( j \) larger than the mean \( \mu_j \), then we would expect to have the cancer sites \( \theta_j \) also above their mean rates \( \mu_j \). The correlation between different cancer sites commented on previously is the empirical justification for this model.

The Bayesian approach adopts a prior distribution to express our uncertainty about the vector \( \mu \) and the matrix \( \Sigma \). This can also be interpreted as introducing an overdispersion in the counts above that accounted for by the Poisson variability to accommodate the presence of unobserved risk factors. For the vector \( \mu_j \), we assumed that each of its entries was independently distributed as an uniform random variable in the interval \([-2, 2]\). The type of the covariance matrix \( \Sigma \) will have usually, but not always, non-negative elements reflecting the fact that, if a given location \( i \) has the relative risk \( \theta_j \) of cancer \( j \) larger than the mean \( \mu_j \), then we would expect to have the cancer sites \( \theta_j \) also above their mean rates \( \mu_j \). The correlation between different cancer sites commented on previously is the empirical justification for this model.

The Bayesian approach adopts a prior distribution to express our uncertainty about the vector \( \mu_j \) and the matrix \( \Sigma \). This can also be interpreted as introducing an overdispersion in the counts above that accounted for by the Poisson variability to accommodate the presence of unobserved risk factors. For the vector \( \mu_j \), we assumed that each of its entries was independently distributed as a uniform random variable in the interval \([-2, 2]\), which is large enough to cover any reasonable deviation from 0 due to mismatching of the reference rates and the areas under study. As a sensitivity analysis on the choice of \( u[-2, 2] \) as the prior distribution of the SIR, we also ran the MCMC procedure assuming a normal distribution with mean zero and large standard deviation (1000) for each entry of the vector \( \mu_j \). For the matrix \( \Sigma \), a useful and flexible choice is the inverse Wishart distribution with parameters \( h \gg k \) and \( R \) where \( R \) is a non-singular matrix. In this paper, we used \( R \) as diagonal matrix with entries equal to 0.1 and \( h = 12 \). Since we know little about the correlation structure between the cancer rates, we chose the minimum possible value (12) for the \( h \) parameter of the Wishart distribution in order to allow the maximum variability in the random matrix probability distribution. This model is easily implemented in a freely available software for Bayesian data analysis called WinBUGS.\(^{21}\)

We provide the source code to run this model in the Appendix. We ran the MCMC chain for 100 000 iterations as the burn-in period and for additional 400 000 iterations saving every 100-th. We ended with a sample size of 4000 vectors \( \theta_i \) from the posterior distribution and inferences are based on these values. The Bayesian estimate is taken as the posterior mean of the parameter and 95% CI are obtained from the posterior distribution quantiles. Convergence to the posterior distribution was checked in WinBUGS by means of the Gelman-Rubin diagnostic,\(^{22}\) using two chains with widely different initial values.

For comparative purposes, we also ran the Bayesian model assuming no correlation between different cancer sites. That is, the matrix \( \Sigma \) was considered a diagonal matrix with \( j \)-th diagonal element equal to \( \sigma_j^2 \) and following an inverse Gamma prior distribution with parameters 0.001 and 0.001. This second model is equivalent to fitting a univariate Bayesian model for each cancer site separately. The comparison between the two fitted models, with the full and with the diagonal \( \Sigma \) matrix, is made by means of the Deviance Information Criterion (DIC) proposed by Spiegelhalter \textit{et al.}\(^{23}\) This measure incorporates two aspects to verify the model adequacy, the deviance from predicted to observed values and the number of effective model parameters. It is a sum of two components, one representing the fitness of predicted to observed values and another representing a penalty of increasing model complexity. Smaller values of DIC indicate a better fitting model.

Our method allows for the introduction of covariates, or additional confounding factors. Our previous model is changed simply by making \( \phi_i \) with a non-constant mean \( \mu_i = (\mu_{i1}, \ldots, \mu_{ik}) \) where each entry \( \mu_{ij} \) is a linear combination of the covariates and unknown regression parameters. These parameters would receive a vague prior distribution such as uniform with a large range and the MCMC procedure would be run as usual. It is straightforward to implement this extension in WinBUGS.

Results

The 1991 total population in the 18 cities ranged from 69 000 to 442 000 inhabitants with the first, second, and third quartiles approximately equal to 94 000, 160 000, and 278 000 inhabitants, respectively. The cases from cancers ranged from zero to 217 and, for male cancers, the median number of cases among the areas varied from 7 (larynx) to 24 (stomach), while for female cancers, it varied from 5 (lung) to 38 (breast). The SIR estimates varied substantially depending on the geographical area and the cancer site. Figure 1 shows the SIR estimates grouped by area and with 95% CI calculated using the Poisson distribution for the SIR numerator.\(^{25}\) The horizontal axis is an area/site order number and, to avoid cluttering, the estimates are plotted for every other area. For some areas, the CI are wide reflecting their small risk population.

Tables 1 and 2 show the correlation between the usual SIR of the 18 cities for male and female cancers separately. For men, incidence rates for larynx and lung, colon and larynx, lung and prostate, prostate and larynx are highly correlated. For women, incidence rates for breast and colon, stomach and colon, stomach and ovary are highly correlated. Other pairs of cancers are just moderately correlated, while others show very low correlation. Table 3 shows the correlation between pairs of male and female cancers. Considering pairs formed by the same cancer site in men and women, we observe some high and moderately high correlations, such as for stomach, colon, and lung cancers, 87.0, 76.5, and 42.6, respectively. However, even for quite different sites we still find highly correlated pairs, such as male lung and female breast with correlation equal to 73.7 and male colon and female ovary with correlation equal to 60.8.

Applying the multivariate method to borrow information across different cancer sites, we can improve the incidence rates estimation. Figure 2 shows the multivariate Bayesian estimates of relative risks and associated 95% credibility intervals calculated from the posterior distribution. Figure 2 uses the same horizontal and vertical scale and the same cities as Figure 1 for comparative purposes. Due to the information borrowed from
MULTIPLE CANCER SITES INCIDENCE RATES ESTIMATION

Other areas and cancers, there is a substantial reduction in the uncertainty associated with the risk estimates. This reduction is a strong argument in favour of Bayesian methods in situations where the local area observations have large variance.

The influence of sample size is clear: there are eight cities/sites where the absolute difference between the multivariate Bayesian and the usual SIR estimates is larger than 0.5 and only one of those occurs in a city with more than 100,000 inhabitants (108,980).

In general, univariate Bayesian estimates of relative risks produced risk relative estimates with larger 95% credibility intervals. In fact, from the 216 estimates from the 18 cities and 12 cancer sites, only seven had multivariate-based intervals larger than the corresponding univariate-based intervals (Figure 3). The larger univariate interval is due to the extra information contained in correlated other cancer sites data in the same area used by the multivariate method. One additional argument favouring the multivariate model is its smaller DIC value, 1338.14, as compared with that of the univariate model, 1372.51.

Table 1 Correlation coefficient \((\times 100)\) between male cancers, 18 cities, São Paulo state, 1991

<table>
<thead>
<tr>
<th>Lung</th>
<th>Stomach</th>
<th>Prostate</th>
<th>Oesophagus</th>
<th>Colon</th>
<th>Larynx</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stomach</td>
<td>24.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prostate</td>
<td>59.4</td>
<td>28.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oesophagus</td>
<td>45.6</td>
<td>29.1</td>
<td>43.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colon</td>
<td>26.9</td>
<td>29.6</td>
<td>28.2</td>
<td>30.9</td>
<td></td>
</tr>
<tr>
<td>Larynx</td>
<td>65.7</td>
<td>31.2</td>
<td>58.7</td>
<td>48.6</td>
<td>60.6</td>
</tr>
</tbody>
</table>

Table 2 Correlation coefficient \((\times 100)\) between female cancers, 18 cities, São Paulo state, 1991

<table>
<thead>
<tr>
<th>Lung</th>
<th>Stomach</th>
<th>Breast</th>
<th>Cervix</th>
<th>Colon</th>
<th>Ovary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stomach</td>
<td>22.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast</td>
<td>46.4</td>
<td>40.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cervix</td>
<td>6.8</td>
<td>21.0</td>
<td>48.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colon</td>
<td>-4.0</td>
<td>52.5</td>
<td>58.7</td>
<td>6.6</td>
<td></td>
</tr>
<tr>
<td>Ovary</td>
<td>9.5</td>
<td>51.9</td>
<td>52.8</td>
<td>9.0</td>
<td>39.4</td>
</tr>
</tbody>
</table>
Figure 4 illustrates the effect of the Bayesian multivariate estimation method compared with the Bayesian univariate and the usual SIR estimates for ovary and female lung cancer sites. These cancers had the lowest expected number of cases among the 12 cancers and the Bayesian estimation effect is large in this case. The three pairs of estimates for an area are represented by three points joined by the full lines. The univariate Bayesian estimates are marked by circles and the multivariate Bayesian estimates by triangles. The multivariate estimates are pulled, quite radically, towards the solid line, even more than the univariate Bayesian estimates. Note that some relative risks estimated as above the reference level 1 in the SIR and univariate Bayesian estimates are reversed to be below that level in the case of the multivariate estimation.

Table 3  Correlation coefficient ($\times 100$), between female (F) and male (M) cancers, 18 cities, São Paulo state, 1991

<table>
<thead>
<tr>
<th></th>
<th>MLung</th>
<th>MStomach</th>
<th>MProstate</th>
<th>MOesophagus</th>
<th>MColon</th>
<th>MLarynx</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLung</td>
<td>42.6</td>
<td>39.9</td>
<td>50.2</td>
<td>13.0</td>
<td>-16.3</td>
<td>15.7</td>
</tr>
<tr>
<td>FSthomach</td>
<td>12.4</td>
<td>87.0</td>
<td>26.4</td>
<td>31.5</td>
<td>50.4</td>
<td>24.5</td>
</tr>
<tr>
<td>FBreast</td>
<td>73.7</td>
<td>49.5</td>
<td>43.2</td>
<td>35.8</td>
<td>36.3</td>
<td>57.2</td>
</tr>
<tr>
<td>FCervix</td>
<td>51.3</td>
<td>30.6</td>
<td>20.9</td>
<td>27.4</td>
<td>-0.5</td>
<td>7.0</td>
</tr>
<tr>
<td>FOvary</td>
<td>33.6</td>
<td>43.7</td>
<td>11.5</td>
<td>11.8</td>
<td>76.5</td>
<td>55.6</td>
</tr>
<tr>
<td>FOvary</td>
<td>29.1</td>
<td>30.2</td>
<td>33.2</td>
<td>14.2</td>
<td>60.8</td>
<td>35.2</td>
</tr>
</tbody>
</table>

**Figure 2**  Multivariate Bayesian standardized incidence ratio (SIR) estimates of 12 cancer sites in 1991 in 9 Brazilian cities. The cancers are numbered in the horizontal axis, the first six being male cancers: 1-oesophagus, 2-stomach, 3-colon/rectum, 4-larynx, 5-lung, 6-prostate, 7-stomach, 8-colon/rectum, 9-lung, 10-breast, 11-cervix/uteri, 12-ovary. Vertical scale is the same as in Figure 1 for comparison purposes.
Figure 3 95% credibility interval length for the 18 cities and 12 cancer sites. The horizontal axis shows the multivariate Bayesian relative risk estimates and the vertical axis, the univariate Bayesian estimates. The solid line is the \( y = x \) straight line and a point on this line indicates no change under the two different models while points above (below) the line represents wider (shorter) credibility intervals under the univariate model.

Figure 4 Effect of the Bayesian multivariate estimation method (triangles) compared with the Bayesian univariate (circles) and the usual standardized incidence ratio (SIR) (crosses) estimates for the female ovary and lung cancer sites. The unity reference is indicated by vertical and horizontal broken lines. The solid line is the \( y = x \) line indicating no change under different models.
Bayesian estimates. This is a consequence of the high correlation in $\Sigma$.

The estimates presented so far were obtained with a uniform prior distribution for each entry of the vector $\mu$. We verified the sensitivity of these results to this uniform distribution choice by running the MCMC procedure with a normal distribution with mean zero and standard deviation equal to 1000. We found virtually the same results. For example, the arithmetic mean of the absolute differences between the $\theta_{ij}$ posterior means from the two models was 0.004835 while the mean of the relative absolute differences was 0.005756. The maximum absolute and relative differences were 0.024 and 0.030.

Concerning convergence, we ran the MCMC procedure with one additional set of suitably overdispersed starting values for the vector $\mu$ (uniform values varying from $-5$ to $5$) and the $\phi_{ij}$ elements (uniform random values varying from $-5$ to $5$) and calculated the Gelman-Rubin convergence statistic $R$, as modified by Brooks and Gelman. To run this procedure with dispersed initial values beyond the ($-2$, $2$) interval limits, we adopted the normal discussed above for the vector $\mu$. We monitored the convergence of the statistic $R$ for all $\phi_{ij}$ parameters and found all of them within 0.01 of the target value of 1. Therefore, we assumed the chains converged adequately.

Discussion
The method we used is conceptually simple, easy to perform, has low cost, and can improve substantially the estimation of cancer incidence and other vital rates. The Bayesian method eliminates an important part of the variability unrelated to the true underlying cancer incidence rate by considering the information contained in other cancer sites and areas incidence rates.

It is important to emphasize that the method is not proposed as an alternative to good quality routine data. Even where this kind of data exists, more reliable estimation of incidence rates can be obtained with the Bayesian approach if the expected counts of cases are small as is usual in small geographical areas or rare disease studies. In the situation examined in this paper, the method has the additional advantage of making better use of rarely available good quality incidence information, collected at great cost for a sample of cities in a given year.

The idea that other cancer sites incidence can provide information to better estimate a given cancer site incidence is not usual. In fact, as far as we know, our paper is the first attempt to implement this idea with many cancers simultaneously, including cancers not obviously related, such as female breast and male lung cancers, for example. Due to the highly specialized nature of cancer research, studies generally address only one specific cancer type at a time. Exceptions to this are epidemiological studies analysing incidence or mortality of several cancers, sometimes comparing different regions. However, these studies did not explore multiple correlations among different cancers as we have here. Additional research is merited to explore the possibility of routinely using other cancers incidence to estimate a given cancer incidence when reliable rates are difficult to obtain due to the small number of expected cases.

We want to emphasize the ecological nature of the statistical correlations we found. In fact, at the individual level, there is no reason to believe that one type of cancer will be associated with another one in the same person, with the obvious exceptions of primary cancer metastases. However, at the aggregate level, we believe that it is justifiable to use other cancer sites information to better estimate one site incidence or mortality, as we argue next.

Besides the empirical correlations we found in our analysis, some aspects of cancer epidemiology corroborate our hypothesis that one cancer site rate gives information about other cancer sites rates. The variation in cancer rates between countries, the differing rates among migrants from one place to another, and the trends over time suggest that most cancers can be environmentally induced. Furthermore, there is evidence that many of these determinants are risk factors for several cancer sites simultaneously. It is estimated that approximately 75–80% of all cancers could have environmental determinants and these include lifestyle habits, like tobacco and alcohol consumption, food and medicine intake, water, soil and air contaminants, and occupational exposures. Tobacco consumption is responsible for nearly one-third of all cancer cases in the US and it is the major cause of lung, larynx, oral cavity, pharynx and oesophagus cancers. Alcohol interacts with tobacco to cause oral cavity, pharynx, oesophagus and larynx cancers. It is estimated that one-third of all cancers could be related to dietary and nutritional practices. There is an inverse association between risk of certain epithelial cancers, particularly oral, oesophageal, stomach and lung cancers, and intake of fresh fruits and vegetables. High-fat, low-fibre diets are positively associated with the risk of colon, breast and prostate cancers.

Since the aetiology of many cancers is not completely understood, the association of tobacco and alcohol consumption, inadequate diet and other common environmental factors could justify why many cancers were highly correlated in our analysis. Also, similarities in the cities’ health care systems can contribute to the correlation, since incidence rates are highly dependent on the health system’s early detection capacity.

As we mentioned in Methods, covariates can be introduced to account for known or suspected risk factors. For example, if information on tobacco consumption is available, it should be introduced in the model for lung, larynx, and oesophagus cancers. Similarly, if information on alcohol consumption is available, it should be introduced in the model for oesophagus and larynx cancers. In general, if we suspect that the $j$-th cancer relative risk in area $i$ is associated with $p$ variables $x_{i1}, \ldots, x_{ip}$, then our model can be easily extended by making $\mu_{ij} = \beta_0 + \beta_1 x_{ij} + \cdots + \beta_p x_{ip}$. As prior distributions, the parameters $\beta_0, \ldots, \beta_p$ have independent normal distribution with mean zero and large variance. If these risk factors are available, which was not the case for the data we used, the model can be easily implemented in WinBUGS. One clear advantage of this model is that the risk factors can be the same for some cancer sites, although this is not necessary.

We believe that our methodological contribution is useful for improving the risk assessment of such an important health problem in modern societies as cancer, although it can also be used to improve rates estimation for other health problems.
Besides that, we believe that our paper contributes to reinforcing the notion that environmental factors present in many modern populations can act as common risk factors for many types of cancers and, because of this, that cancer can be a largely preventable disease.

Acknowledgements

To Donaldo Botelho Veneziano, from Fundação Oncocentro, São Paulo State, who kindly provided the data and information about their collection.

KEY MESSAGES

- Cancer incidence rates can be unreliable if the expected counts are small.
- Cancer incidence rates from other geographical areas can be used to obtain better estimates of a given area rate.
- We explored the correlation among different cancer sites by using other cancer sites incidence rates to improve the estimation of a given cancer site rate which leads to more precise estimates.
- The method we propose is conceptually simple, easy to perform, has low cost, and can improve substantially the estimation of cancer incidence and other vital rates.

References

Appendix

WinBUGS code to fit the multivariate Bayesian relative risk model:

```winbugs
model{
  for( i in 1 : N ) {
    for( j in 1 : K ) {
      Y[i, j] ~ dpois(lambda[i, j]) # distribution of observations
      lambda[i, j] <- E[i, j] * theta[i, j]
      theta[i, j] <- exp(phi[i, j]) # log parametrization
    }
    phi[i, 1:K ] ~ dmnorm(mu[ ], Omega[ , ])
  }
  for(j in 1:K){
    mu[j] ~ dunif(-2,2)
  }
  Omega[1:K, 1:K ] ~ dwish(R[ ], 12) # Wishart on prec. matrix
}
```

Commentary: Robust estimation of population parameters with sparse data

B Rachet

Assunção and Castro present a Bayesian approach based on Markov chain Monte Carlo (MCMC) methodology designed to estimate cancer incidence rates simultaneously for a number of cancers. The methodology is appropriate for providing reliable estimates when the available data are geographically and temporally sparse.

Health authorities have demanded incidence, mortality, and survival rates and other economic, socio-demographic, and health-related parameters for many years. Knowledge of geographical and temporal incidence and survival data at local level is naturally of interest to health authorities, especially if covariates such as socioeconomic status can be included in the analyses. Such statistics may allow authorities to evaluate health care system performance, anticipate shortages or crisis situations, and as a consequence, adjust their policies. Such data may also help construct or confirm research hypotheses, such as the presence of new latent risk factors for lung cancer. In both situations—the first related to public health, the second to aetiological epidemiology—the aim is to obtain parameter estimates that are as accurate as possible. However, even though the accuracy of data collection at population level has dramatically improved during the last century, there are many situations where either the accuracy of the population data or the reliability of parameters derived from accurate data cannot be readily assured.

Many countries, mostly the less-developed countries, have only recently begun collecting population data. In some cases, data are collected during occasional surveys. Even when registration procedures have taken place, completeness still varies widely. In summary, we are often still faced with spatially and temporally sparse data.

Health authorities and epidemiologists increasingly often want to analyse temporo-spatial variations in parameters at small geographical level or within specific sub-groups of the population defined by socioeconomic status or ethnicity. Even in countries with experience in collecting population data and where completeness is high, the number of events in a cell of a multi-dimensional matrix may be small and subject to random instability, leading to unreliable estimates.

For the last 15 years, much research has been devoted to the accurate estimation of such parameters. The most promising, and most recent, have been based on Bayesian methods, particularly those using Markov field methodology. MCMC