Title: Inferring lung cancer risk factor patterns through joint Bayesian spatio-temporal analysis

Running title: Lung cancer risk factor spatio-temporal patterns

Authors: Susanna M Cramb[a,b], Peter D Baade[a,c,d], Nicole M White[b,e], Louise M Ryan[f], Kerrie L Mengersen[b,e]

[a] Cancer Council Queensland, Brisbane, Australia.
[b] Mathematical Sciences School, Queensland University of Technology, Brisbane, Australia.
[c] School of Public Health and Social Work, Queensland University of Technology, Brisbane, Australia.
[d] Griffith Health Institute, Griffith University, Gold Coast, Australia.
[e] Cooperative Research Centre for Spatial Information, Australia.
[f] School of Mathematical Sciences, University of Technology, Sydney, Australia.

Corresponding author: Ms Susanna Cramb, Cancer Council Queensland, PO Box 201, Spring Hill, QLD 4004, Australia; E-mail: susannacramb@cancerqld.org.au; Telephone: +61-7-36345350; Facsimile: +61-7-32598527.

Article category: Original research

Sources of support: Peter Baade was supported by an Australian National Health and Medical Research Council Career Development Fellowship (#1005334). Kerrie Mengersen and Nicole White acknowledge support from the Cooperative Research Centre for Spatial Information, whose activities are funded by the Australian Commonwealth's Cooperative Research Centres Programme. Louise Ryan and Kerrie Mengersen acknowledge support from the ARC Centre of Excellence in Mathematical and Statistical Frontiers in Big Data, Big Models and New Insights. The views expressed in this paper are those of the authors and not of any funding body.

Conflict of Interest Statement
Nil. All funding sources have been acknowledged.

Abstract: 294 words
Body of text: 4,010 words
Figures: 6
Tables: 2
References: 30

Abstract
Background: Preventing risk factor exposure is vital to reduce the high burden from lung cancer. The leading risk factor for developing lung cancer is tobacco smoking. In Australia, despite apparent success in reducing smoking prevalence, there is limited information on small area patterns and small area temporal trends. We sought to estimate spatio-temporal patterns for lung cancer risk factors using routinely collected population-based cancer data.
Methods: The analysis used a Bayesian shared component spatio-temporal model, with male and female lung cancer included separately. The shared component reflected lung cancer risk factors, and was modelled over 477 statistical local areas (SLAs) and 15 years in Queensland, Australia. Analyses were also run adjusting for area-level socioeconomic disadvantage, Indigenous population composition, or remoteness.
Results: Strong spatial patterns were observed in the underlying risk factor estimates for both males (median Relative Risk (RR) across SLAs compared to the Queensland average ranged from 0.48-2.00) and females (median RR range across SLAs 0.53-1.80), with high risks observed in many remote areas. Strong temporal trends were also observed. Males showed a decrease in the underlying risk across time, while females showed an increase followed by a decrease in the final two years. These patterns were largely consistent across each SLA. The high underlying risk estimates observed among disadvantaged, remote and indigenous areas decreased after adjustment, particularly among females.
Conclusion: The modelled underlying risks appeared to reflect previous smoking prevalence, with a lag period of around 30 years, consistent with the time taken to develop lung cancer. The consistent temporal trends in lung cancer risk factors across small areas support the hypothesis that past interventions have been equally effective across the state. However, this also means that spatial inequalities have remained unaddressed, highlighting the potential for future interventions, particularly among remote areas.

Keywords
Lung cancer, risk factor, tobacco smoking, Bayesian methods, spatio-temporal analysis, shared component model

1. Introduction

Due to its high incidence and low survival, lung cancer is the leading cause of cancer-related death in Australia.[1] More males are affected by this disease than females.[1] Most lung cancers are caused by cigarette smoking, accounting for around 65% of lung cancers among females and 90% among males.[2] Other modifiable risk factors include exposure to air pollution, radon, asbestos and certain heavy metals.[3]

In the absence of effective early diagnostic tools or treatments for advanced lung cancer,[4] preventing the initiation of lung cancer by reducing exposure to risk factors is vital. In particular, there has been much progress in reducing the prevalence of tobacco smoking in many developed countries.[5] Between 1964 and 2010, the percentage of Australians who smoked cigarettes decreased from 43% to 15%,[6] although the prevalence of smoking among females increased until around the late 1970s, when it started to decline.[7] Yet this smoking prevalence varies geographically, with people living in rural and disadvantaged areas more likely to smoke.[1] However these geographical data are often compromised by small numbers and a reliance on self-reported surveys. This limits the ability to understand small area patterns of smoking prevalence, particularly over time.

Given the lack of data on most risk factors at the spatial level, recent work has sought to model selected cancers jointly to extract spatial or spatio-temporal estimates of the common underlying risk factor components. Where high quality, population-based cancer registry data are available, this can be used to obtain objective risk factor estimates. When a cancer has similar risk factors for both sexes, but a differential impact across space and/or time, there may be benefit in jointly modelling one cancer type and dividing into sex-specific components, e.g. male and female lung cancer. This joint modelling is often conducted using a shared component model.

The premise of the shared component model, as first proposed by Knorr-Held and Best,[8] was to jointly model the relative risk by dividing into separate components, including one common to both diseases (e.g. representing the underlying risk factor exposure), as well as residual variation components (one for each disease). This enables information to be borrowed between diseases. In this model the shared component acts as a surrogate for spatially structured unobserved risk factors common to both diseases.[8] The model has been extended by incorporating covariates,[9] adjusting the number of components,[9] increasing the number of diseases,[10] and including temporal trends.[11, 12] The joint modelling of multiple cancers at the spatial or spatio-temporal level has commonly been applied within a Bayesian context.[8, 11]

When there is only one shared component in these models, this component provides an estimate of all the risk factors common to the included diseases. However, when a particular risk factor is prominent in developing disease, such as tobacco smoking with lung cancer, underlying risk estimates are likely to reflect the most prominent risk factor.

Our aims were to apply Bayesian spatio-temporal shared component models to routinely collected, population-based male and female lung cancer data to:

1. Infer the spatio-temporal patterns of underlying lung cancer risk factors in Queensland.

2. Determine how known influences (socioeconomic, remoteness and Indigenous status) impact on the modelled underlying risk factor patterns.

3. Identify geographical areas where the temporal underlying risk factor pattern differed from the pattern for total Queensland.

2. Methods

2.1 Data

Lung cancers diagnosed among Queensland residents between 1997 and 2011 were sourced from the Queensland Cancer Registry,[13] a population-based cancer registry with high-quality data covering the entire state of Queensland. Australian legislation requires this Registry to be notified of every invasive cancer diagnosed in a Queensland resident, excluding only keratinocytic skin cancers. Ethical approval was obtained from Queensland Health (approval number: HREC/09/QHC/25).

Details about patients' usual residence at diagnosis was provided at the statistical local area (SLA) level. Geocoding was used to match the residence at diagnosis to the 2006 SLA

definition, thus overcoming limitations of changing geographical boundaries over time. In 2006 there were 478 SLAs, with a median population of 5,723.

Population estimates based on the 2006 SLA boundaries were obtained from the Australian Bureau of Statistics, for each SLA, year and 5-year age groups up to 85+ years. Due to zero population counts in one SLA for several years during the time period of interest, only 477 SLAs were used in our analyses (population range in 2006: 78 to 74,804).

Each SLA was assigned a value for area-level socioeconomic disadvantage (3 categories (Advantaged: top 20%, Middle SES: middle 60%, Disadvantaged: lowest 20%), defined using the Index of socioeconomic advantage and disadvantage (IRSAD) from the Australian Bureau of Statistics Socioeconomic Indexes For Areas (SEIFA), remoteness (Urban (Major city), Regional (Inner/Outer regional) and Remote (Remote/Very remote) based on the Accessibility/Remoteness Index of Australia+), and Indigenous population (2 categories based on 2006 census data: <10% or ≥10%).

2.2 Model

Most shared component models use a standard Poisson likelihood, as is appropriate for rare and non-contagious diseases. However, when area-specific count data are particularly sparse, an alternative formulation allowing for excess zero counts may be preferred. Therefore, we extended previous approaches by incorporating and comparing alternate distributions for the counts within the shared component framework. Specifically, we compared four alternative variants of the Poisson count distribution:[14, 15]

1. Poisson $O_{dij} \sim \text{Poisson}(\rho_{dij} E_{dij})$
2. Negative binomial $O_{dij} \sim \text{Poisson}(x_d \rho_{dij} E_{dij})$ where $x_d \sim \text{Gamma}(r_d, r_d)$
3. Zero-inflated Poisson (ZIP) $O_{dij} \sim \text{Poisson}\left((1 - u_{dij})\rho_{dij} E_{dij}\right)$ if $O_{dij} > 0$
4. Poisson hurdle $O_{dij} \sim \text{Poisson}\left(\frac{(1-u_{dij})}{1-\exp(-\rho_{dij} E_{dij})} \rho_{dij} E_{dij}\right)$ if $O_{dij} > 0$

where O_{dij} are the observed lung cancer counts for each sex d=1,2 (representing males and females, respectively), i=1,2…477 areas and j=1,2…15 years, ρ_{dij} is commonly referred to as the relative risk[16] and E_{dij} represents expected counts. To enable comparisons over time, the expected counts were calculated using the sex- and age-specific Queensland lung cancer incidence rates in 1997-99. In the negative binomial model, here expressed as a Poisson-gamma mixture for comparability, r_d is the sex-specific overdispersion parameter, which forms the shape and scale parameters in the gamma distribution, while the u_{dij} is the probability of zero in the ZIP and hurdle models.

The Poisson hurdle model separates the zeros from anything above zero, modelling counts under a truncated Poisson distribution. The ZIP model can be considered a special type of hurdle model. Here the zero counts are separated into excess (those above what is expected under a Poisson distribution) and non-excess zeros (those expected under a Poisson distribution).

Using a modified version of the shared component model from Richardson et al,[11] the log relative risk for each of these models was then expressed as:

$$\log(\rho_{dij}) = \alpha_d + \mu_{dij}$$

The sex-specific intercept is given by α_d, while the space-time structure is modelled through μ_{dij}, which represents exposure to the risk factors for lung cancer, here referred to as the underlying risk factor component.

The underlying risk component is separated into several components so that spatial clustering and temporal trends can be presented separately. Log RRs of the underlying risk factor component for males (μ_{1ij}) are constrained to capture the shared spatial and temporal trends, while the log RRs of the underlying risk factor estimates for females (μ_{2ij}) include additional terms providing the sex differential, as follows:

$$\mu_{1ij} = \lambda_i\delta + \xi_j\kappa + \phi_{1ij}$$

$$\mu_{2ij} = \frac{\lambda_i}{\delta} + \frac{\xi_j}{\kappa} + \beta_i + \gamma_j + \phi_{2ij}$$

where $\lambda_i$represents the common spatial pattern for SLA_i, β_i gives the female spatial difference (the sex-space interaction) for SLA_i, ξ_j is the shared time trend for calendar $year_j$, and γ_j the female temporal difference (the sex-time interaction) for calendar $year_j$. A sex-specific residual term ϕ_{dij} was also included for the i^{th} SLA and j^{th} year combination. The terms δ and κ are scaling parameters, enabling different risk gradients between sexes.[11]

Prior distributions were assigned to each parameter as follows: the spatial components (λ_i, β_i) had a conditional autoregressive (CAR) prior with neighbours based on adjacent SLAs, while temporal parameters had a one-dimensional CAR prior (ξ_j, γ_j) with neighbours consisting of the immediately previous and subsequent time periods. Because the CAR prior smooths the log RRs, spatio-temporal patterns can be recovered even when data are sparse. A zero-mean multivariate normal distribution with covariance matrix Σ was assigned to ϕ_{dij}. This term captures additional spatio-temporal variation in each disease that is not explained by the other terms. To improve convergence, a centered parameterisation was used with the distribution specified on μ_{dij}, rather than directly on ϕ_{dij}. For identifiability, and as there were only 15 time periods, κ was fixed at a value of one.[11] Finally, log δ was described by a normal distribution, and α_d by a normal distribution with large variance. Refer to the Appendix for further details on priors.

To explore the impact of factors known to be associated with the prevalence of the main risk factor for lung cancer, tobacco smoking, covariates for area-level socioeconomic disadvantage, remoteness, and Indigenous population were added to the linear predictor.

All models were run with single chain Markov Chain Monte Carlo (MCMC) using Stata version 13.1 (StataCorp LP, College Station, Texas, USA) interfaced with WinBUGS 1.4 (Imperial College and Medical Research Council, UK). The first 300,000 iterations were discarded as burn-in, and a further 50,000 iterations monitored (with every tenth iteration kept).

2.3 Sensitivity analyses

We compared three commonly used versions of the hyperparameter distributions on the variance component of the spatial and temporal parameters for each of the four count distributions to check the influence exerted by priors on the results:
Version 1: Gamma on the inverse variance (ie. precision), $\tau \sim \Gamma(0.5, 2000)$
Version 2: Uniform on the standard deviation, $\sigma \sim \mathrm{U}(0.1,100)$
Version 3: Uniform on the standard deviation, $\sigma \sim \mathrm{U}(0.1,20)$

These equate to means and variances on the precision of (10,1000) for the gamma distribution (version 1), and on the standard deviation of (50,4990) for version 2 and (10,198) for version 3. These distributions deliberately aim to be non-informative to minimise the risk of substantive influence on the estimates produced.

All gave similar estimates and uncertainty measures for most parameters, so after examining deviance cumulative distribution functions, convergence trace and density plots, we selected version three. Uniform distributions on the standard deviation have been recommended as more robust than gamma distributions on the precision,[17] and the tighter boundaries minimised convergence issues.

One concern when examining diseases such as cancer is the potential influence of patient migration. People may have been exposed to an environmental or personal risk factor in one location, but then moved residence prior to diagnosis. Since information on residential history was not available before diagnosis, a sensitivity analysis was conducted to estimate the impact of changing location. Three alternatives were compared assuming up to a 10%, 20% and 40% population movement between SLAs. This internal migration was approximated by randomly increasing or decreasing the expected incidence count in each SLA up to the desired percentage, while constraining the overall count to match the Queensland total. The adjusted risk estimates for each scenario were categorised as low (<0.909), average (0.909-<1.10), and high (1.10+), and then these categories compared to those from the original scenario (0% migration).

Given the data sparseness, a sensitivity analysis was conducted to ascertain if the modelled small area temporal trends were likely to reflect only the average trend due to inadequate data for individual SLAs. A modified version of the model was run with data aggregated by five broad remoteness groupings. No local spatial smoothing was performed between these areas, and the Poisson count distribution was used. All other model details remained the same.

2.4 Model comparison
The deviance information criterion (DIC) is widely used in comparing Bayesian hierarchical models. However, it has a tendency to under-penalise complex models unless the effective number of parameters is much smaller than the number of independent observations, which may not occur in disease mapping.[18]

In light of this, we considered a collection of criteria representing different features of model fit – the overall goodness of fit (via the median squared predicted error (MSPE), Bayesian predictive p-value and L-criterion,[19] all of which compare model estimates against the data), the effective number of parameters (model complexity, defined as $p_{D,}$ which is the mean deviance minus the deviance at the mean of the posteriors and a component of DIC),[20] and the predictive distribution (via the conditional predictive ordinance (CPO)).[21] Lower values generally indicate better model fit, apart from Bayesian p-values (ideal is 0.5), or CPO, where many very low values suggest poor fit.[21]

3. Results

The median number of observed lung cancer cases by SLA in 2011 was 2 for males (range: 0 to 29) and 1 for females (range: 0 to 25). The proportion of SLAs with no lung cancer cases diagnosed ranged from 33% (males) and 56% (females) in 1997 to 28% and 39% in 2011 among males and females, respectively. Further details on the study cohort are available in Table 1.

The different model distributions produced similar results for the majority of parameters, although the shared underlying risk factor estimates occasionally differed in very sparsely populated areas. There was minimal difference in model goodness of fit between the models (Table 2), but a slight preference for the negative binomial formulation based on p_D. Results presented are from the negative binomial model.

Mapping the underlying risk factor component showed strong spatial variation throughout Queensland (Figure 1). The median SLA-specific underlying relative risks ranged across the State from 0.48 to 2.00 for males ($\exp(\lambda)$), and 0.53 to 1.80 for females ($\exp(\lambda + \beta)$). When females were compared to males ($\exp(\beta)$), many regions had similar risks (Figure 1). However, there were higher risk factor estimates among females in some urban South East areas, and lower risks among females in selected remote areas (Figure 1).

There was also strong evidence of trends across time in the underlying risk component (Figures 2-3). Males ($\exp(\xi)$) showed a decrease in the underlying risk across time, while females ($\exp(\xi + \gamma)$) showed an increase followed by a decrease in the final two years (Figure 4). These patterns were practically universally consistent across each SLA. When data were aggregated by remoteness groupings, the same broad trend ($\exp(\mu)$) was observed across each remoteness group (Figure 4).

The high underlying risk factor estimates observed among disadvantaged, remote and indigenous areas decreased after adjustment, particularly among females (Figure 5). Specifically, areas with a high Indigenous proportion largely explained the increased risk among disadvantaged and remote areas for females.

Risk factor estimates remained quite similar even after allowing for hypothetical migration patterns (Figure 6). As the proportion of migration increased, greater differences from the initial estimates were observed. However, even allowing for up to 40% migration, few spatial underlying risk factor estimates changed between the broad categories of low, average or high risk. Both males and females had 6% of SLAs change from a higher to a lower category, while for males, 5% of SLAs moved from a lower to a higher category, and 9% of SLAs among females

4. Discussion

This population-based study found strong evidence for differences by region of residence and across time in the shared underlying lung cancer risk factors. These patterns and trends are consistent with known trends in tobacco smoking prevalence.[22] Almost all areas followed a similar trend to that observed in the underlying risk factors overall (males decreasing and females increasing before recently decreasing).

Tobacco smoking is the leading risk factor for developing lung cancer in Australia,[22] and the detected underlying risk factor component is likely to strongly reflect past smoking

patterns. Other key risk factors such as radon and/or air pollution exposure have very low levels in Australia.[1] However, some caution is required since 10-15% of lung cancers are diagnosed among non-smokers.[23]

Patterns in the underlying risk factors by remoteness, socioeconomic disadvantage or areas with a high Indigenous population are also consistent with that reported for tobacco smoking.[22] Surveys have suggested around 50% of Indigenous Australians smoke cigarettes.[24] Our results showed the increased risk in many disadvantaged or remote areas was diminished or annulled after adjusting for the Indigenous composition. This contrasted with the more minor changes observed after adjusting for remoteness or socioeconomic disadvantage.

This methodology also allows trends over time for each region to be obtained. When data are very sparse, region-specific trends may simply reflect the overall average, so our consistent trends should be interpreted with caution. However, our sensitivity analysis using five broad remoteness groupings also obtained consistent temporal trends across these regions, supporting the hypothesis that the temporal patterns were consistent across most areas of Queensland.

This is the first time the consistency in trends over time for lung cancer risk factors has been demonstrated at the small-area level within the Australian context. Tobacco smoking information reported in early surveys was not able to be analysed by small-area geographic regions, and it has been unclear how trends across time varied across small regions. The similar trends across time obtained for these small areas is consistent with the suggestion that smoking-related interventions were equally effective across the different regions of Queensland. Given that smoking-related interventions have incorporated state- or nation-wide price increases, public awareness of the health risks and advertising restrictions,[7] this consistency in trends is not surprising.

Recently, small-area estimates of smoking prevalence were released for 2007-2008 across Australia based on modelled self-reported survey data.[25] A current smoker was defined as smoking cigarettes, cigars or pipes at least once a week. Although Queensland estimates were not available for many rural and remote SLAs, or sometimes only provided by aggregated SLAs, the results showed that smoking prevalence was generally higher outside of Brisbane. This also agrees with 2011-2012 self-reported daily smoking estimates across 73 larger Queensland regions, with results released for 43 regions.[26] Of the remote areas with available estimates, most showed higher smoking prevalence, while many urban areas tended to have lower estimates.

The similarity of our geographical patterns to these recent results suggests past geographic differentials haven't changed. Despite the overall decrease across time for males and more recently, females, many remote and rural areas are likely to continue to have higher smoking prevalence for many years into the future, unless preventive and remedial efforts are targeted at these areas. Given that population-wide intervention programs have been shown to be more cost effective in tobacco control,[27] potential interventions should aim to address the prevailing social norms and practices, rather than an individualistic approach.[28] This may include addressing the higher density of tobacco outlets and lower cigarette prices in socioeconomically disadvantaged areas.[29]

Although there are advantages in using population-based cancer data to understand risk factor patterns, there are also limitations. Only past estimates can be obtained, as the lag period between exposure and development of cancer suggests our underlying risk estimates are likely to reflect smoking prevalence up to 30 years previously.[3] Although theoretically all cancers with smoking as a risk factor could be included to obtain estimates, we found including a less common cancer (oesophageal), despite a strong link to smoking, decreased the precision of the estimates. This might reflect the influence of other key risk factors on oesophageal cancer, such as alcohol intake,[30] or the sparseness of our data. In addition to tobacco smoke exposure, it is possible that these patterns reflect the impact of other risk factors (despite their rarity in the Australian context) that may also be captured in the underlying exposure component.

The similarity of smoking patterns and lung cancer patterns raises the question of whether our latent component is simply detecting lung cancer, rather than an estimate of underlying exposure. In our model, the lung cancer relative risk (ρ_{dij}) is dependent on the sex specific intercept, any included covariates, and our latent component. The intercept terms consistently differed from one, with resultant differences in the relative risk estimates of lung cancer and latent components.

The novelty of our approach is two-fold. Firstly, a review of the literature found no published comparisons of these four alternate count distributions in a fully Bayesian spatio-temporal shared component model. We are also not aware of these models being applied to such sparse data before, resulting both from the Australian context with its relatively small population across a large land area, and examining annual time periods.

The small difference in model fit between the Poisson model and the three models allowing for excess zeros (negative binomial, ZIP and hurdle Poisson) at first seemed counter-intuitive given the large proportion of zeros. This is likely to be influenced by both the random effects included in the model, which allow for overdispersion, and the comparatively large number of cases observed in some SLAs.

Both the ZIP and Poisson hurdle models can be considered mixture models, and we found that some underlying exposure estimates did not converge as well as under Poisson or negative binomial. The assumptions of a hurdle model (that zeros represent an inability to have a positive result) is questionable, although in our modelling of μ we assumed all areas and time periods had the ability to have lung cancer diagnosed, equating to assuming that all areas/time periods had a positive count. The negative binomial distribution was slightly preferred, but was the most computationally intensive model, taking twice as long as the Poisson model to run. If time had been an issue, the Poisson distribution could have been used instead in this study.

Many variations on this model are possible, either by adjusting the included components (inserting and/or removing terms), or using alternative priors. For instance, we explored using a first-order autoregressive (AR(1)) prior instead of a CAR prior on the temporal components, which would have only smoothed based on the previous time period. This prior is useful when the aim is to extrapolate into the future. However, our aim was to identify smoothed patterns, and the larger uncertainty around estimates and less smoothing under the AR(1) resulted in preferring the CAR prior. We also considered including a shared spatio-temporal interaction term ν_{ij}. However, estimates were all close to 1, so the additional model complexity was not justified.

Similar methodology could be used to explore spatio-temporal variation in other disease risk factors. For instance, trends and patterns in diet-related influence were examined in Greece using a factor analysis model containing six cancers with particular dietary factors as recognised risk factors.[12] Obesity has strong links to several cancers, and this may be a useful approach to obtain temporal and small-area estimates of obesity, which can be poorly self-reported.

In conclusion, these shared component models have provided evidence supporting the similarity of temporal trends in lung cancer risk factors across small geographical areas, consistent with the hypothesis that past interventions designed to reduce lung cancer risk factors have been equally effective across the state. However, this consistency in temporal trends also means that current inequalities in these risk factors between areas have remained unaddressed, highlighting the potential for future interventions targeting the social norms and practices of people living in rural and remote areas.

Appendix

Prior distributions (expressed as mean, precision):

$$\alpha_d \sim \text{Normal}(0, 0.001)$$

$$\begin{bmatrix}\mu_{1ij}\\ \mu_{2ij}\end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix}\eta_{1ij}\\ \eta_{2ij}\end{bmatrix}, \Sigma^{-1}\right)$$

$$\boldsymbol{\lambda} \sim \text{CARNormal}(W, \tau_\lambda)$$

$$\boldsymbol{\beta} \sim \text{CARNormal}(W, \tau_\beta)$$

$$\boldsymbol{\xi} \sim \text{CARNormal}(Q, \tau_\xi)$$

$$\boldsymbol{\gamma} \sim \text{CARNormal}(Q, \tau_\gamma)$$

$$\log\delta \sim \text{Normal}(0,0.2)$$

MVN=Multivariate Normal, CARNormal=Conditional Autoregressive Normal.

Note that by centering $\log\delta$ around 0, we are assuming that any value of δ is as likely as any value of $1/\delta$.[8] This would allow the indices for the sexes to be switched and still the same posterior distributions to be obtained for each sex, as the posterior distribution on δ would change to the reciprocal.

Hyperprior distributions:

$$\Sigma^{-1} \sim \text{Wishart}\left(\begin{bmatrix}0.01 & 0\\ 0 & 0.01\end{bmatrix}, 2\right)$$

$$\frac{1}{\sqrt{\tau_\lambda}} \sim \text{Uniform}(0.1,20)$$

$$\frac{1}{\sqrt{\tau_\beta}} \sim \text{Uniform}(0.1,20)$$

$$\frac{1}{\sqrt{\tau_\xi}} \sim \text{Uniform}(0.1,20)$$

$$\frac{1}{\sqrt{\tau_\gamma}} \sim \text{Uniform}(0.1,20)$$

The Wishart distribution is the conjugate for the precision parameter of the multivariate normal distribution, and is treated as a multivariate chi-squared distribution.

References

[1] Australian Institute of Health and Welfare & Cancer Australia. Lung cancer in Australia: an overview. Canberra: AIHW, 2011.

[2] Ridolfo B, Stevenson C. The quantification of drug-caused mortality and morbidity in Australia, 1998. Canberra: AIHW, 2001.

[3] Youlden DR, Cramb SM, Baade PD. The International Epidemiology of Lung Cancer: geographical distribution and secular trends. *J Thorac Oncol* 2008;3(8):819-831.

[4] Kathuria H, Gesthalter Y, Spira A, Brody JS, Steiling K. Updates and controversies in the rapidly evolving field of lung cancer screening, early detection, and chemoprevention. *Cancers (Basel)* 6(2):1157-1179.

[5] Jha P, Peto R. Global effects of smoking, of quitting, and of taxing tobacco. *N Engl J Med* 2014;370(1):60-68.

[6] Organisation for Economic Cooperation and Development. OECD Health Data 2013. Paris: OECD, 2013.

[7] Scollo MM, Winstanley MH. Tobacco in Australia: Facts and issues. Melbourne: Cancer Council Victoria, 2012.

[8] Knorr-Held L, Best NG. A shared component model for detecting joint and selective clustering of two diseases. *J R Stat Soc Ser A Stat Soc* 2001;164(1):73-85.

[9] Dabney AR, Wakefield JC. Issues in the mapping of two diseases. *Stat Methods Med Res* 2005;14(1):83-112.

[10] Held L, Natario I, Fenton SE, Rue H, Becker N. Towards joint disease mapping. *Stat Methods Med Res* 2005;14(1):61-82.

[11] Richardson S, Abellan JJ, Best N. Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in Yorkshire (UK). *Stat Methods Med Res* 2006;15(4):385-407.

[12] Tzala E, Best N. Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. *Stat Methods Med Res* 2008;17(1):97-118.

[13] Queensland Cancer Registry. Cancer in Queensland: Incidence, Mortality, Survival and Prevalence,1982 to 2010. Brisbane: QCR, Cancer Council Queensland and Queensland Health, 2013.

[14] Ntzoufras I. Bayesian modeling using WinBUGS. New Jersey, USA: John Wiley & Sons, 2009.

[15] Neelon B, Ghosh P, Loebs PF. A Spatial Poisson Hurdle Model for Exploring Geographic Variation in Emergency Department Visits. *J R Stat Soc Ser A Stat Soc* 2013;176(2):389-413.

[16] Lawson AB. Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. Boca Raton: CRC Press, 2013.

[17] Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 2006;1(3):515-533.

[18] Plummer M. Penalized loss functions for Bayesian model comparison. *Biostatistics* 2008;9(3):523-539.
[19] Laud PW, Ibrahim JG. Predictive model selection. *J R Statist Soc Ser B Method* 1995;57(247-262.
[20] Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *J R Statist Soc Ser B Method* 2002;64(4):583-639.
[21] Congdon P. Chapter 2: Bayesian Model Choice, Comparison and Checking. In Bayesian Statistical Modelling, 2nd edition, Edition. West Sussex, England: Wiley, 2006.
[22] Australian Institute of Health and Welfare. Drugs in Australia 2010: tobacco, alcohol and other drugs. Canberra: AIHW, 2011.
[23] McCarthy WJ, Meza R, Jeon J, Moolgavkar SH. Chapter 6: lung cancer in never smokers: epidemiology and risk prediction models. *Risk Anal* 2012;32 Suppl 1(S69-84.
[24] Australian Bureau of Statistics. National Aboriginal and Torres Strait Islander Social Survey, 2008. Canberra: ABS, 2009.
[25] Population Health Information Development Unit. Social Atlas of Australia. Adelaide: University of Adelaide, 2012.
[26] Department of Health. Self reported health status 2011-12. Health indicators: chronic disease and behavioural risk factors, local government areas. Brisbane: Department of Health, Queensland Government, 2013.
[27] World Health Organization. WHO Report on the Global Tobacco Epidemic, 2013: Enforcing bans on tobacco advertising, promotion and sponsorship. Geneva: WHO, 2013.
[28] Blue S, Shove E, Carmona C, Kelly MP. Theories of practice and public health: understanding (un)healthy practices. *Crit Public Health* 2014;1-15.
[29] Dalglish E, McLaughlin D, Dobson A, Gartner C. Cigarette availability and price in low and high socioeconomic areas. *Aust N Z J Public Health* 2013;37(4):371-376.
[30] International Agency for Research on Cancer. Alcohol consumption and ethyl carbamate. Lyon, France: IARC, 2010.

Table 1: Study cohort and population characteristics, 1997-2011

	Population	Lung cancer cases	Median IRSAD percentile	N SLAs	N SLAs with high indigenous population
Total Queensland	57,990,293	26,664	50.5	477	55
Sex					
Male	28,937,540	17,313			
Female	29,052,753	9,351			
Age structure					
0-49 years	41,254,096	1,334			
50-64 years	9,781,744	7,402			
65-79 years	5,226,054	13,111			
80+ years	1,728,399	4,817			
Years					
1997-99	10,215,429	4,491			
2000-02	10,734,471	4,810			
2003-05	11,491,585	5,152			
2006-08	12,338,515	5,898			
2009-11	13,210,293	6,313			
Socioeconomic (IRSAD)					
Advantaged (top 20%)	9,164,720	2,952	90	95	0
Middle SES (middle 60%)	41,412,012	19,383	50.5	286	10
Disadvantaged (lowest 20%)	7,413,561	4,329	11	96	45
Remoteness (ARIA+)					
Urban	33,456,103	15,034	71	252	0
Regional	21,443,351	10,239	35.5	144	6
Remote	3,090,839	1,391	14.5	81	49
Indigenous population					
High (10%+)	2,416,775	1,185	6	55	55
Other (<10%)	55,573,518	25,479	56	422	0

ARIA+=Accessibility/Remoteness Index of Australia plus; IRSAD=Index of Socioeconomic Advantage and Disadvantage; SLA=Statistical Local Area

Notes: IRSAD percentiles are Queensland-specific, and high values indicate socioeconomic advantage. IRSAD, ARIA+ and Indigenous population are defined based on SLA characteristics in 2006.

Table 2: Comparison of model fit measures under the final prior choice

	MSPE		Bayesian PPV		L-criterion	Effective number of parameters	% CPO <0.01
	Males	**Females**	**Males**	**Females**			
Poisson	2.38	1.25	0.60	0.63	12888.4	574.2	0.8
Negative binomial	2.45	1.25	0.60	0.64	13101.2	388.2	0.9
ZIP	2.83	1.40	0.60	0.65	13753.2	518.8	1.1
Poisson hurdle	2.51	1.32	0.60	0.64	13159.8	n.a.	1.0

ZIP=Zero-inflated Poisson; n.a.=not available

MSPE= Median squared predicted error, ie. $(O-m)^2$.

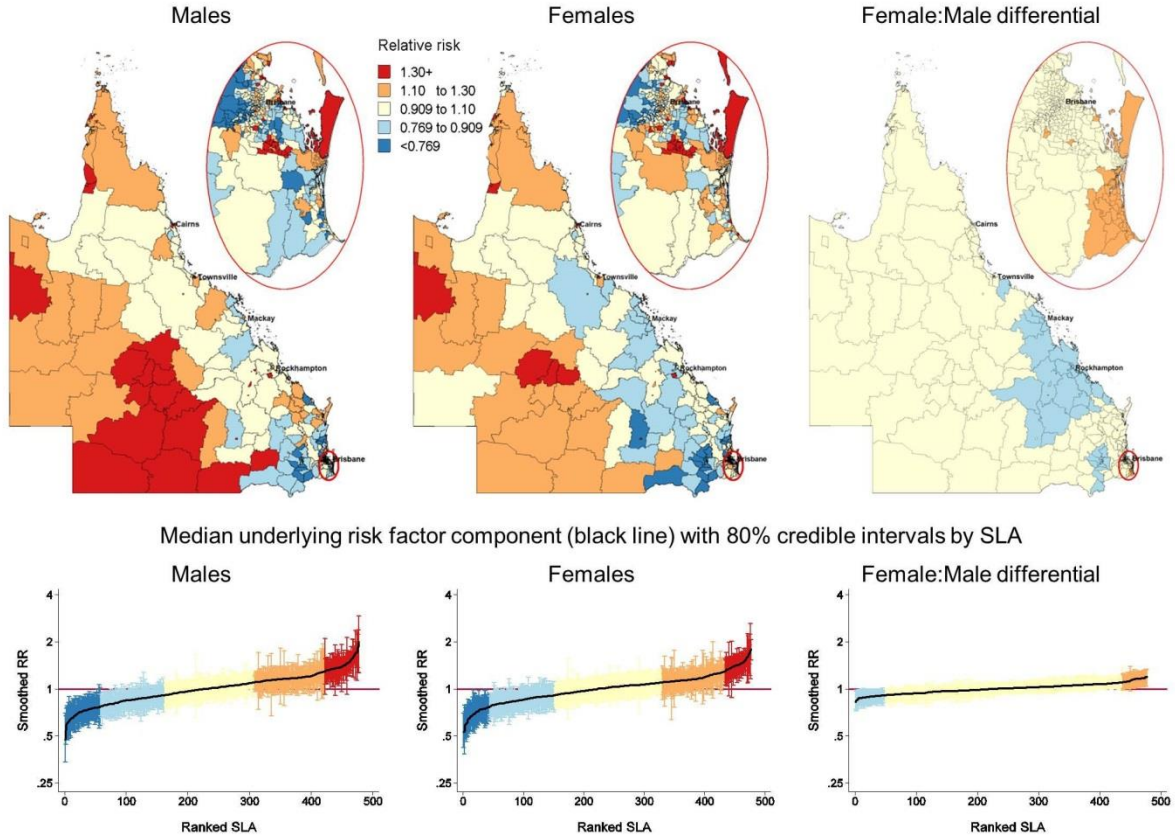
Bayesian PPV=predictive p-value, calculated as the probability of m>O, and ideally equal to 0.5.

L-criterion=(sum of square root of (variance(m) + difference from observed value(i.e. O-m)^2)).

Effective number of parameters calculated as the posterior mean of the deviance minus the deviance of the posterior means (a component of Deviance Information Criterion (DIC). DIC is not calculated for hurdle models).

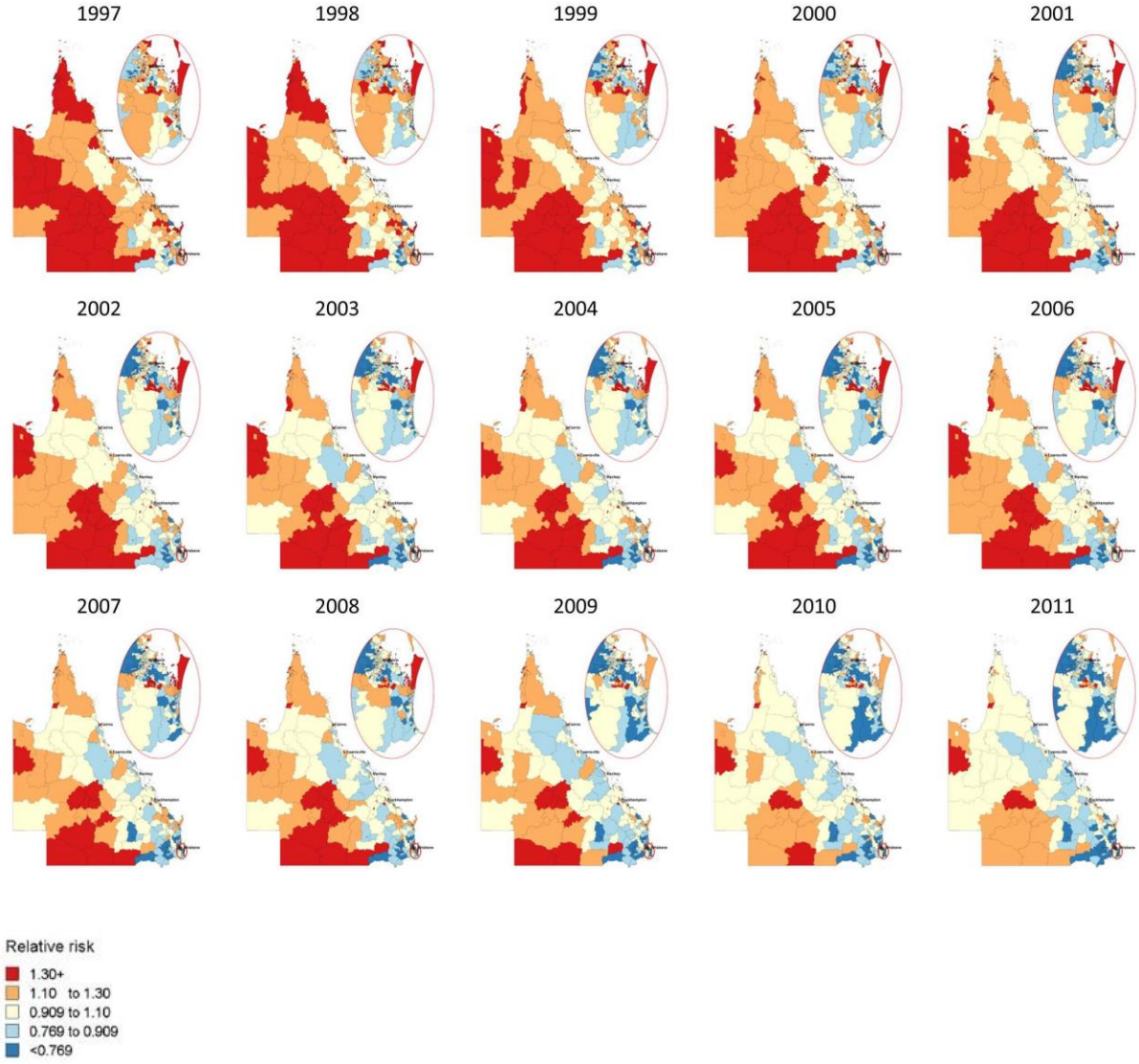
CPO=Conditional predictive ordinate, also known as the leave-one-out predictive density as it represents the posterior probability of observing the value of O_i when the model is fitted to all data except O_i. Approximated here using the harmonic mean of the inverse likelihood of O_i. Very low values may represent outliers/influential observations. Model fit is considered adequate if few values are <0.01.

Figure 1: Spatial variation in the underlying risk factor component by sex.



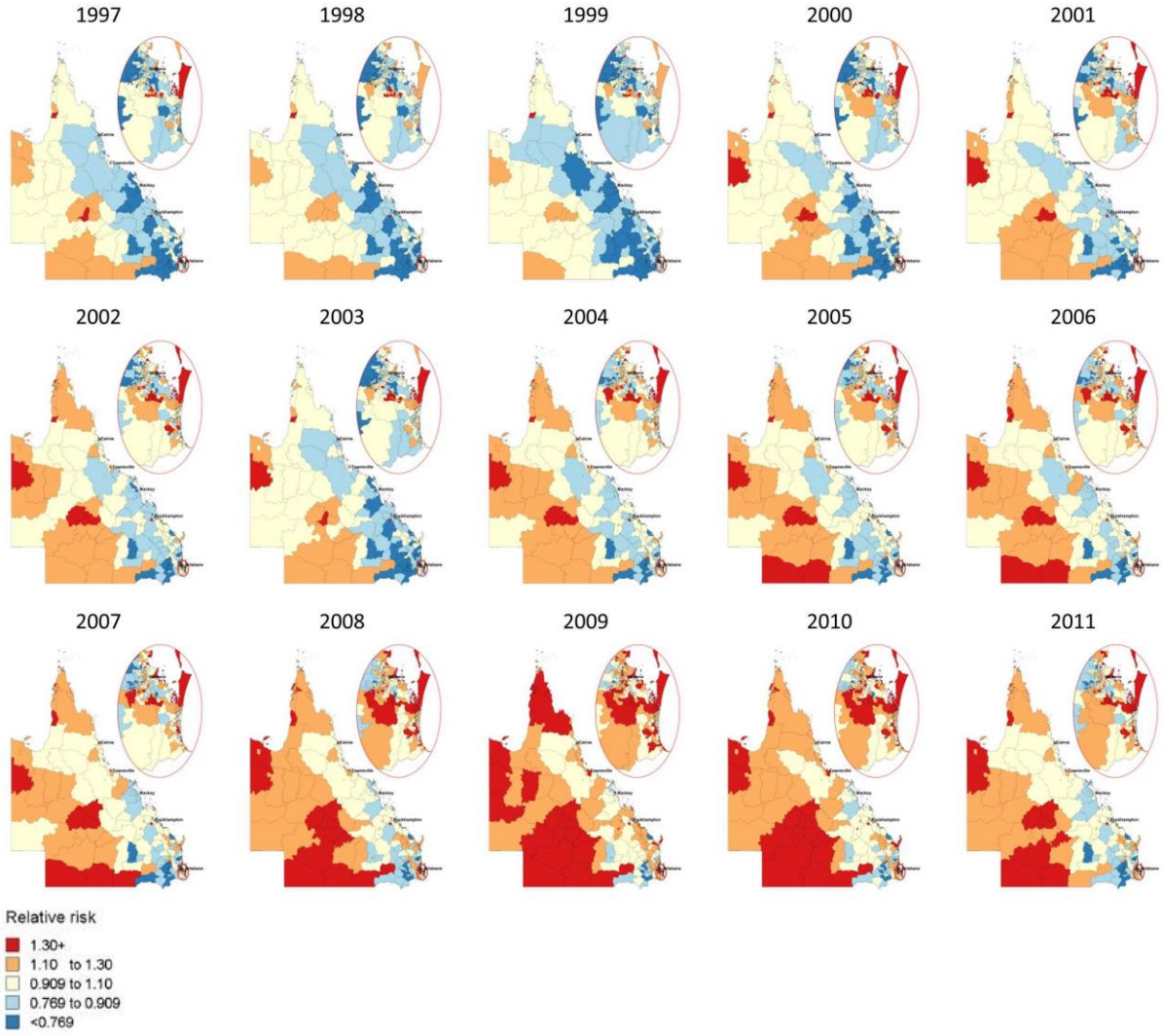
Note: Relative risk=1 corresponds to the specified Queensland average risk in 1997-99 (ie. males, females and the female: male differential, respectively).

Figure 2: Median relative risk of the male underlying risk component across time ($\exp(\mu_{1ij})$)



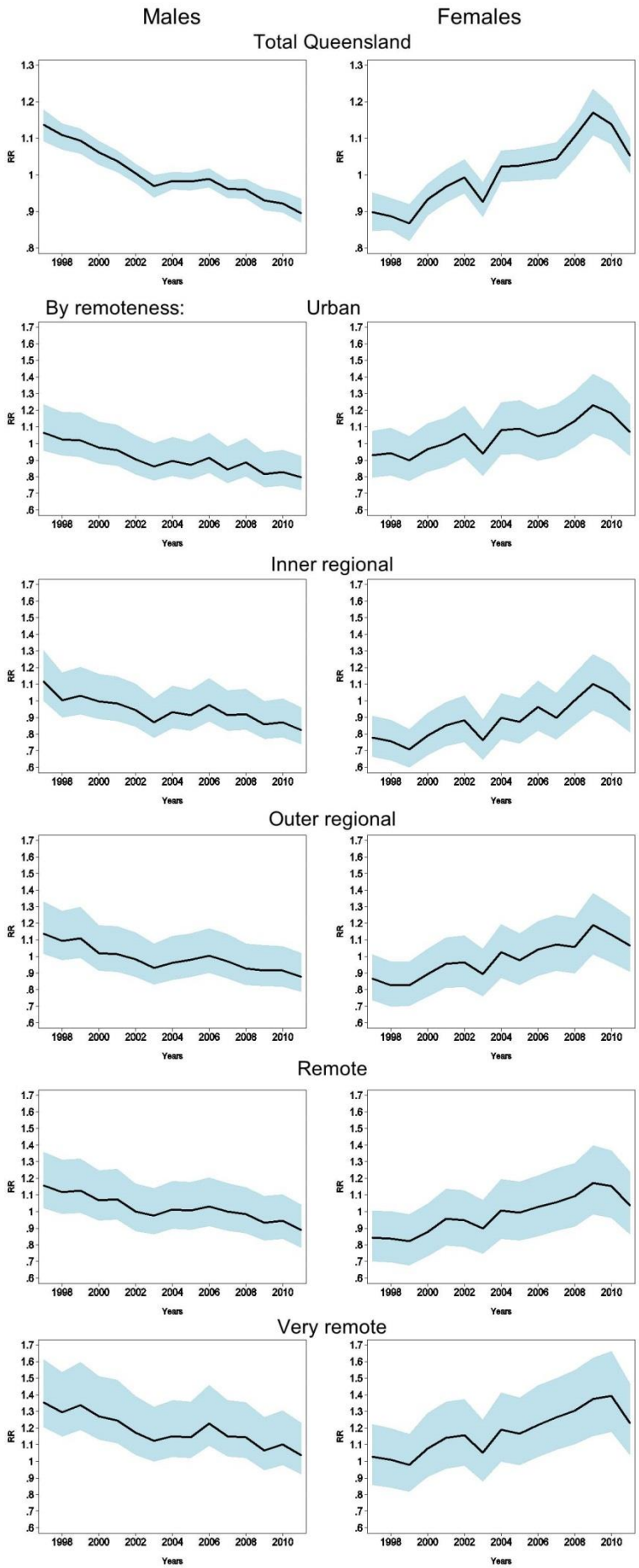
Note: Relative risk=1 corresponds to the Queensland male average risk in 1997-99.

Figure 3: Median relative risk of the female underlying risk component across time ($\exp(\mu_{2ij})$)



Note: Relative risk=1 corresponds to the Queensland female average risk in 1997-99

Figure 4: Time trends in the underlying risk by sex.



Notes: RR=Relative Risk

Black line is the median, blue shading represents the 80% credible interval.

Total Queensland results produced by the model based on statistical local areas (SLAs).
Results by remoteness produced by the model with broad remoteness groups replacing SLAs.

Figure 5: Relative risk (RR) of the underlying risk factor component before and after model adjustment, by remoteness, socioeconomic position and Indigenous population composition.

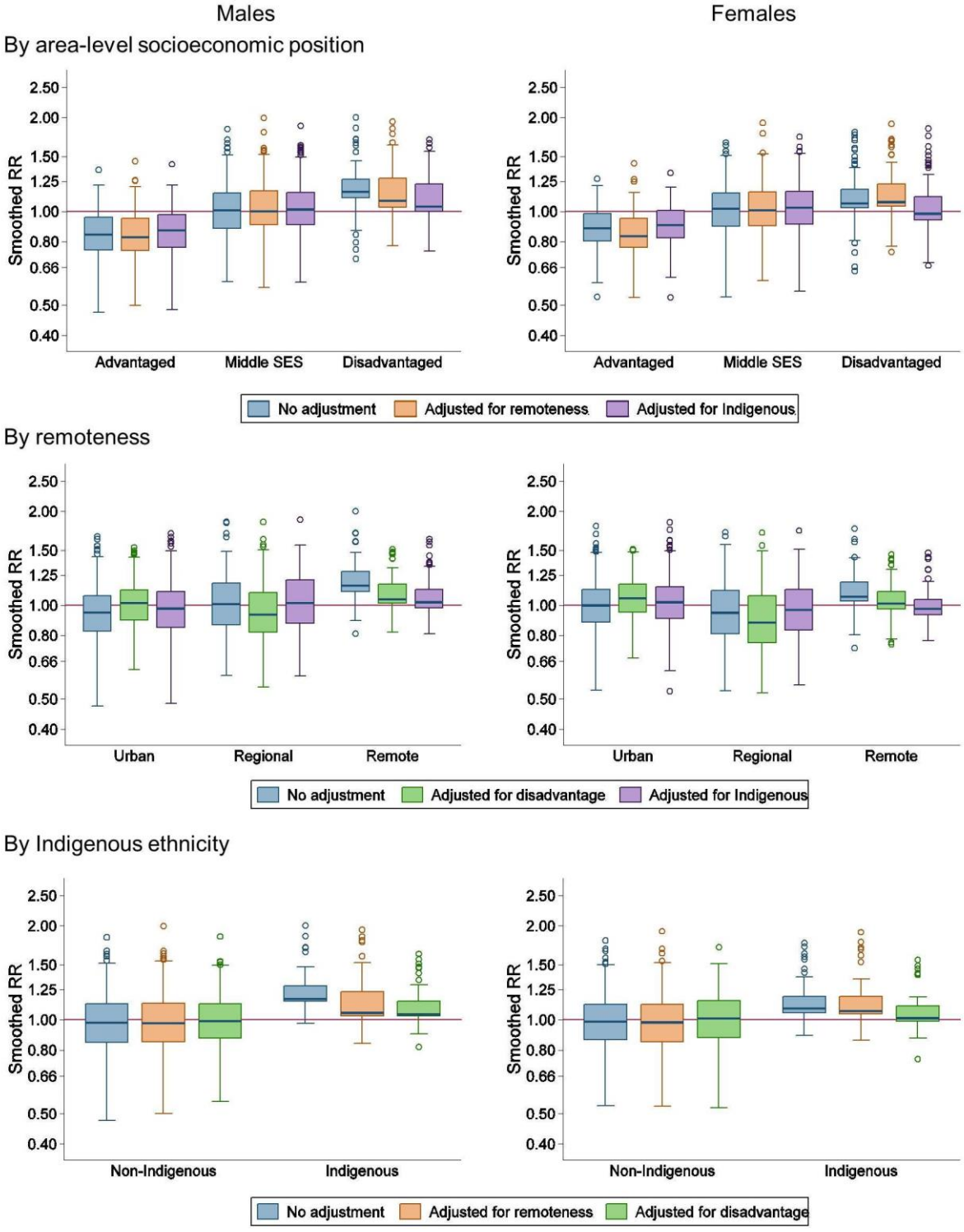
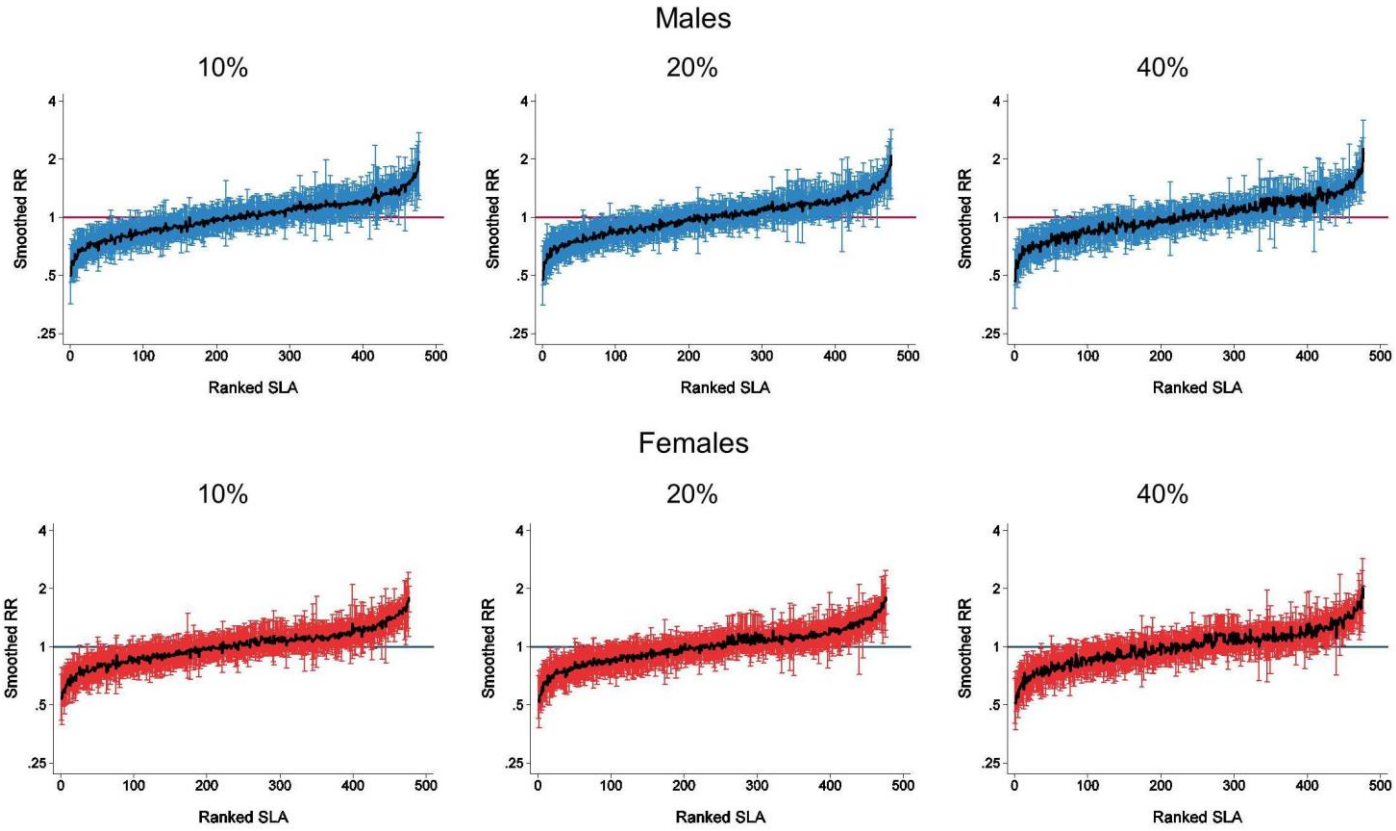


Figure 6: The median underlying risk with 80% credible intervals assuming up to x% migration by sex.



RR=Relative Risk; SLA=Statistical Local Area
Note: SLAs ranked by the order in Figure 1 (0% migration) to enable comparison.