Appendix

Part A: Statistical models

Incidence model

Given a set of n areas, the statistical model for area i ($i = 1, \dots, n$) can be written as follows,

$$y_i \sim \text{Poisson}(e_i\theta_i),$$
$$log(\theta_i) = \alpha + x\beta_k + u_i + v_i,$$

where y_i are the observed counts of area i, e_i are the expected counts of area i, and θ_i is the SIR of area i. Here α is the intercept term, x is the predictor variable, and β_k is the coefficient of the predictor variable. The component that accounts for spatial correlation between neighbouring areas is denoted by u_i, and v_i accounts for the unstructured (non-spatial) variation in the model.

Relative survival model

The statistical model can be written as

$$d_{ijk} \sim Poisson(\mu_{ijk}),$$
$$\log(\mu_{ijk} - {d^*}_{ijk}) = \log(y_{ijk}) + \alpha_j + x\beta_k + u_i + v_i,$$

where for age group k, follow-up interval j and area i; d_{ijk} is the number of deaths and μ_{ijk} is the expected number of deaths. Here ${d^*}_{ijk}$ is the expected number of deaths due to causes other than the disease of interest and y_{ijk} is the person-time at risk. The intercept varied by follow-up year and is denoted by α_j, x is the predictor variable, and β_k is the coefficient of the predictor variable. Also, u_i accounts for spatial dependence between neighbouring areas, and v_i denotes the unstructured (non-spatial) random effects in the model.

Part B: WinBUGS code

WinBUGS code for the incidence model

```
Model
{
for (i in 1 : N) {
# Likelihood
O[i] ~ dpois(mu[i])
Opred[i] ~ dpois(mu[i])
log(mu[i]) <- log(E[i]) + alpha + u[i] + v[i]
# Area-specific relative risk (for maps)
RR[i] <- exp(alpha + u[i] + v[i])
# Prior distribution for the uncorrelated heterogeneity
v[i] ~ dnorm(0, tauv)
}
# CAR prior distribution for spatial random effects
u[1 : N] ~ car.normal(adj[], weights[], num[], tauu)
for(k in 1:sumNumNeigh) {
weights[k] <- 1
}
# Other priors:
alpha ~ dflat()
# Hyperpriors on precisions
tauu ~ dgamma(0.1, 0.1)
tauv ~ dgamma(0.001, 0.001)
sigmau <- sqrt(1 / tauu)
sigmav <- sqrt(1 / tauv)
#Standard deviations
sdv <- sd(v[]) #marginal SD of heterogeneity
sdu <- sd(u[]) #marginal SD of clustering
}
```

WinBUGS code for the relative survival model

```
Model
{
# Likelihood
for (i in 1 : datarows) {
d[i]  ~ dpois(mu[i])
mu[i]<-d_star[i] + excessd[i]
log(excessd[i]) <-  log(y[i])+ alpha[RiskYear[i]] + beta[1]*agegp2[i]
+ beta[2]*agegp3[i]+ u[slaNo[i]] + v[slaNo[i]]
for (j in 1:N_RiskYear){
alpha[j] ~ dnorm (0, 0.001)
}
}
# CAR prior for spatial effects
u[1:Nsla] ~ car.normal(adj[], weights[], num[], tauu)
for (k in 1:sumNumNeigh) {weights[k] <- 1 }
for (i in 1:Nsla) {
# Prior distribution for the uncorrelated heterogeneity
v[i] ~ dnorm(0, tauv)
logRER[i]<-u[i]+v[i]
RER[i]<-exp(logRER[i])
}
# Other priors
tauu ~ dgamma(0.5, 0.001)
tauv ~ dgamma(0.5, 0.001)
varv <- 1/tauv
varu_con <-1/tauu
varu_marg<-sd(u[])*sd(u[])
}
```

Part C: R-INLA code

R-INLA code for the incidence model Assume that data are available for a set of areas as $\{y_i, e_i, x_{1i}, x_{2i}\}$ for $i = 1,...,n$, where y_i is a count, e_i is an expected count, and x_{1i} and x_{2i} are two predictors/covariates. These data should be read into R as vectors and can be held in a list. In the code below, n represents the number of areas, obs represents disease count, expe represents expected count, cov1 and cov2 represent the covariates, u represents the spatial random effects, and v represents the unstructured (non-spatial) random effects.

```
u=seq(1:n)
v=seq(1:n)
data.incid = list(obs=obs, expe=expe, cov1=cov1, cov2=cov2, u=u, v=v)
formula1 = obs ~ cov1 + cov2
 + f(u, model="besag", graph="queensland.graph", param=c(0.1, 0.1))
 + f(v, model="iid", param=c(0.001, 0.001))
result1 = inla(formula1, family="poisson", data=data.incid,
control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE), E=expe)
summary(result1)
```

R-INLA code for the relative survival model

In the code below, n represents the number of areas, d represents the number of deaths (d_{ijk}), d_star represents the expected number of deaths due to causes other than the disease of interest ($d^{*}{}_{ijk}$), y represents the person-time at risk (y_{ijk}), cov1 and cov2 represent the covariates, u represents the spatial random effects, and v represents the unstructured (non-spatial) random effects.

```
u=seq(1:n)
v=seq(1:n)
data.surv = list(d=d, d_star=d_star, y=y, cov1=cov1, cov2=cov2, u=u, v=v)
formula2 = d ~ offset(d_star) + cov1 + cov2
 + f(u, model="besag", graph="queensland.graph", param=c(0.5, 0.001))
 + f(v, model="iid", param=c(0.5, 0.001))
result2 = inla(formula2, family="poisson", data=data.surv,
   control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE), E=y)
summary(result2)
```

Part D: Glossary

Box plot	A visual display that summarizes data using a “box and whiskers” format to show the minimum and maximum values (ends of the whiskers), interquartile range (length of the box), and median (line through the box).
Case-control study	A type of observational analytic study. Enrollment into the study is based on presence (“case”) or absence (“control”) of disease. Characteristics such as previous exposure are then compared between cases and controls.
Covariate	In statistics, a covariate is a variable that is possibly predictive of the outcome under study. A covariate may be of direct interest or it may be a confounding or interacting variable.
Credible interval	An interval in the domain of a posterior probability distribution used for interval estimation. A 95% credible interval is interpreted as “a 95% probability the true estimate lies in this range”
Direct method of standardisation	Apply stratum-specific rates observed in the populations of interest to a standard population. The ratio of two directly standardised rates is called the comparative incidence ratio.
Excess mortality	A measure of the deaths which occur over and above those that would be expected for a given population. These are deaths considered to result from the disease of interest
Hierarchical model	A model written in a hierarchical form or in terms of sub-models
Hierarchical structure	A hierarchy of parameters which are related to one another in a model
Hyperparameter	A parameter in a hyperprior distribution
Hyperprior distribution	A prior distribution on a hyperparameter, i.e., on a parameter of a prior distribution
Incidence	A measure of the risk of developing a disease within a specified period of time

Indirect method of standardisation	Apply stratum-specific reference rates to the populations of interest. The ratio of two indirectly standardised rates is called the SIR.
Inference, statistical	In statistics, the development of generalizations from sample data, usually with calculated degrees of uncertainty.
Likelihood	Likelihood is a tool for summarizing the data's evidence about unknown parameters. It is the probability of a given sample being randomly drawn regarded as a function of the parameters of the population.
Markov chain	A mechanism for generating plausible parameter value, whereby the value to be drawn depends on the previously drawn value in some way
Markov chain Monte Carlo (MCMC)	A class of algorithms for sampling from probability distributions by constructing a Markov chain that has the desired distribution as its equilibrium distribution
Parameter	A value used to represent a certain population characteristic which is usually unknown and therefore has to be estimated
Percentile	The set of numbers from 0 to 100 that divide a distribution into 100 parts of equal area, or divide a set of ranked data into 100 class intervals with each interval containing 1/100 of the observations. A particular percentile, say the 5th percentile, is a cut point with 5 percent of the observations below it and the remaining 95% of the observations above it.
Posterior distribution	A probability distribution on the values of an unknown parameter that combines prior information about the parameter contained in the observed data to give a composite picture of the final judgements about the values of the parameter
Predictor	A predictor variable is also known as an independent variable
Prevalence	The number or proportion of cases or events or conditions in a given population.
Prior distribution	A probability distribution that represents the uncertainty about the parameter before the current data are examined
Random effects	Effects that account for differences among the individual observational units in the sample, which are randomly sampled from the population. These effects usually conform to a specified distribution (typically a Normal distribution) and have a mean of 0
Regression	A statistical technique for estimating the relationships among variables.
Relative excess risk (RER)	A measure that informs the relative survival of a disease, by reporting the risk of death within a certain number of years of diagnosis after adjusting for broad age groups, compared to the average
Relative risk	A comparison of the risk of some health-related event such as disease or death in two groups.
Relative survival	A standard measure of excess mortality due to a disease in population-based disease survival studies
Risk factors	An aspect of personal behavior or lifestyle, an environmental exposure, or an inborn or inherited characteristic that is associated with an increased occurrence of disease or other health-related event or condition.

Sensitivity analysis	A sensitivity analysis is the study of how the uncertainty in the output of a mathematical model or system (numerical or otherwise) can be apportioned to different sources of uncertainty in its inputs.
Standardised incidence ratio (SIR)	An estimate of relative risk within each area based on the population size, that compares the observed incidence against the expected incidence

Part E: Boxes

Box 1: Bayesian model

Given Bayes' theorem (Gelman *et al*., 2014),

$$P(A \mid B) \propto P(A)P(B|A)$$

The posterior distribution ($P(A \mid B)$) is proportional to the prior distribution for parameters ($P(A)$) multiplied by the data-based distribution given parameters (also known as the likelihood (Appendix Part D)) ($P(B|A)$).

- Posterior estimates (model output) are a combination of the prior information and the data
- Parameter Parameters in the model are assigned prior distributions
- A prior distribution is the probability distribution that represents the uncertainty about the parameter before the current data are examined
- Parameters in the prior distribution can also be assigned distributions
- Parameters in the prior distribution (called 'hyperparameters') can also be assigned distributions

Box 2: Normal distribution

A distribution contains information on every possible observation and its associated probability. For instance, a Normal distribution is a continuous distribution that is "bell-shaped", at which data are most likely to be distributed around the mean and are less likely to be farther away from the mean. A Normal distribution is often specified in terms of its mean (μ) and variance (σ^2) and can be written in the form of Normal(μ, σ^2). A parameter can be assigned a Normal distribution with mean 0 and variance 100 which can be denoted as parameter~Normal(0, 100). Alternatively, instead of specifying the values (0, 100), uncertainty about these parameters can also be described probabilistically. For example, instead of specifying '100' for the variance, the prior distribution could be written as Normal(0, σ_0^2) and then σ_0^2 is described by another probability distribution. Here σ_0^2 is termed a hyperparameter (Appendix Part D) and the distribution on σ_0^2 a hyperprior distribution (Appendix Part D).

Box 3: Selecting regional scale
Important questions to consider when deciding on an appropriate area scale to conduct the analysis include:
1. Is there a risk of patient confidentiality being compromised?
2. Are population data available at the same scale as disease occurrences?
3. Will boundaries change over time? If so, what options are possible for keeping your data consistent?
4. Is there a digital boundary file available?
5. Will areas have a practical and relevant interpretation?
6. How does the size of the areas compare relative to the spatial pattern of the variation? If there is a lot of variation in an environmental effect within areas, this will limit the scope to measure the effect.
7. How many areas will there be? This affects computational time.
8. Are some areas likely to have zero population? This is likely to cause difficulties in modelling and estimation, e.g., zero denominator causes difficulties when using a Poisson distribution.
9. What scale have other similar studies used?
10. What spatial scale is available for covariate data? If spatial variation that takes fixed effects into account is of interest, it is not necessary to have a spatial scale finer than the available covariate data.

Box 4: Data required to produce incidence estimates
Given a disease of interest, the information required to produce incidence estimates includes
• Number of disease cases among people within a certain time period for each small area
• Estimated population counts by age group, sex, year and small area of residence − this is used as the denominator for calculating rates and for age-standardisation (see Appendix Part D direct and indirect methods of standardisation)
• Geographical boundaries − this is used to compute the adjacency matrix required for spatial smoothing
• *Optional*: any desired small area level covariates (if available) such as rurality and socioeconomic status

Box 5: Data required to produce survival estimates
To produce relative survival estimates of a disease of interest, the input data required include
• From the patients with the disease of interest (if not available for each individual then aggregated over each small area, any covariates and follow-up time intervals): − The observed number of deaths (from any cause) within a certain time period − Person-time at risk (the length of time between diagnosis and either death or censoring)
• General population mortality data used to calculate the expected number of deaths, which represents deaths due to causes other than the disease of interest for each small area, sex and

broad age group

- Geographical boundaries − this is used to compute the adjacency matrix required for spatial smoothing
- *Optional*: individual or area-level covariates, including age, tumour stage, or area rurality and socioeconomic status

Box 6: Probability distributions used in epidemiology

- For common diseases, the Binomial distribution models the number of disease occurrences in a sample size n from a population size N. The Binomial distribution is also commonly used in the analysis of disease prevalence data and case-control studies (see Glossary) (Thomas, 2014).
- When the disease is rare or less common (i.e., the probability of a disease is small), the Poisson distribution is used as an approximation to a Binomial distribution (Wakefield, 2003, 2004). A Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time and/or space.
- For over-dispersed count distributions (where the data admit more variability than expected under the assumed distribution), a Negative Binomial distribution may be appropriate (Gardner *et al*., 1995).
- For empirical data that show more zeroes than would be expected, zero-inflated models may be employed (Gardner *et al*., 1995)

Box 7: The incidence model

Given a set of *n* areas, the model for area *i* (*i* = 1,…, *n*) can be written as follows,

Observed counts in area *i* □Poisson(expected counts of area *i* × SIR of area *i*),

log(SIR of area *i*) = intercept term + coefficient × predictor variable vector for area *i* + spatial random effect of area *i* + unstructured random effect of area *i*.

- Apply stratum-specific reference rates to the populations of interest.
- The ratio of two indirectly standardised rates is called the SIR.

Box 8: The relative survival model

The model can be written as below, where for area *i*, follow-up interval *j*, and age group *k*,

Number of deaths$_{(ijk)}$ □Poisson(expected number of deaths$_{(ijk)}$),

log(expected number of deaths$_{(ijk)}$ − expected number of deaths due to causes other than disease of interest$_{(ijk)}$) = log(person-time at risk$_{(ijk)}$) + intercept varied by follow-up year *j* +

coefficient$_{(k)}$ × predictor variable vector + spatial random effect of area *i* + unstructured random effect of area *i*.

Box 9: Prior distributions for the random effects

Unstructured

The unstructured random effects are assumed to follow a Normal distribution with mean zero and a hyperparameter for variance.

Unstructured random effect of area *i* □Normal(0, variance hyperparameter).

Spatial

The spatial random effects are assumed to follow a conditional autoregressive (CAR) prior (Besag *et al.,* 1991) with some hyperparameters, as follows

Spatial random effect of area *i* □ Normal (average of spatial effects of neighbours of area *i,* variance hyperparameter / number of neighbours of area *i*).

References

Gelman A, Carlin JB, Stern HS, Rubin DB, 2014. Bayesian Data Analysis (Vol. 2). London: Chapman & Hall/CRC.