Practical 2: Disease mapping study of lung cancer incidence in Greater London, 2001-2005

25 February, 2019

Before starting the practical

Organise your work:

- Create a separate subdirectory in your home directory to save your files created during this practical (e.g. "C:/Users/yourusername/Practical2").
- Copy all the files from the blackboard to your subdirectory (created just above). A description of them is given below.
- You will use the following R packages: *sp, maptools, spdep, rgdal, tmap, RColorBrewer, R2OpenBUGS, mcmcplots, coda.*
- To check where the packages are saved type *.libPaths()*
- To check if already installed type *.packages(all.available = TRUE)*

If the packages are not installed you need to do it through:

```
install.packages(c("sp", "maptools", "spdep","rgdal","tmap","RColorBrewer",
                   "R2OpenBUGS", "mcmcplots", "coda"), type = "both",
                 dependencies = TRUE, repos = c(CRAN = "http://cran.r-project.org"))
```

Then you need to load these through:

```
library(sp)           #tools for spatial data
library(maptools)     #tools for reading and handling spatial objects
library(spdep)        #to create the adjacency matrix from the shape file
library(R2OpenBUGS)   #tools to call OpenBUGS from R
library(coda)         #tools for convergence diagnostics
library(mcmcplots)    #tools for mcmc plots
library(rgdal)        #tools for geospatial data
library(RColorBrewer) #provides colour palettes for maps and plots
library(tmap)         #tools for plotting maps
```

PART A

In Part A of this pactical you will be using the R software Statistical Package (www.r-project.org), as well as OpenBUGS to carry out a disease mapping study as discussed in Lecture 2 using the BYM model specification. You will use the same data as in Practical 1, i.e. lung cancer incidence in males in Greater London, at ward level (628 areas).

1. In folder 'Part A' you will find the following files:

- *newLung.csv*: file containing lung cancer data: Observed (O), Expected (E), SMR, non spatially smoothed RR (RR.HET), non spatially residual RR (resRR.HET), posterior probability that the residual RR is above 1 (probaresRR.HET)
- *GreaterLondonWard*: folder containing the shapefile of Greater London without the river as described during lectures 1 and 2
- *model_HET.txt*: BUGS code for the Poisson logNormal model with unstructured random effects (HET model)

2. Health data

1. Check the current working directory (This should be the location where the .Rmd file is saved)
2. Import the health data
3. Print the dimensions of the data and look at the first few rows

3. Creating the adjacency matrix

To run the BYM model (HET+CAR), the adjacency matrix needs to be provided. As seen this morning at the end of lecture 2, we can use the shapefile to produce it. As seen during the lecture, the Thames river leads to unconnected wards. Therefore, we will use the shapefiles without the Thames river only for this purpose.

1. Import the shapefile of Greater London and plot it

```
london <- readOGR(dsn = "PartA/GreaterLondonWard", layer = "GreaterLondonWard")
```

```
## OGR data source with driver: ESRI Shapefile
## Source: "C:\Users\aboulier\Dropbox\SpatialAnalysis2019\Practicals\Practical2_AB\PartA\Greate
## with 628 features
## It has 5 fields
## Integer64 fields read as strings:  POLY_ID
```

```
proj4string(london) <- CRS("+init=epsg:32630")  # we assign a projection manually
## or
# london <-
# readShapeSpatial("PartA/GreaterLondonWard/GreaterLondonWard.shp", IDvar = "STwardcode")
plot(london)
```

2. Adjacency matrix in R

(a) Convert the polygons to a list of neighbours using the function *poly2nb*

```
## Neighbour list object:
## Number of regions: 628
## Number of nonzero links: 3696
## Percentage nonzero weights: 0.9371577
## Average number of links: 5.88535
## Link number distribution:
##
##   1   2   3   4   5   6   7   8   9  10  11  12  13
##   1   4  16  75 162 179 113  56  13   4   2   2   1
## 1 least connected region:
## 383 with 1 link
## 1 most connected region:
## 2 with 13 links
```

(b) Convert this list to OpenBUGS format (i.e. a list of 3 components adj, num and weights) using the function *nb2WB* amd print a summary of the object

```
## [1] "adj"     "weights" "num"
```

```
##         Length Class  Mode
## adj     3696   -none- numeric
## weights 3696   -none- numeric
## num      628   -none- numeric
```

4. Analyse the data using the BYM model

1. First, write the model. DO NOT CLOSE R and open the file *model_HET.txt*. Alternatively, you may open the file within RStudio. Amend the model and save as *model_BYM.txt*:
   (a) Add the spatially structured random effects U and their distribution (CAR model)
   (b) Amend the codes for the RR and residual RR (look back at the slides to see how to write this). REMEMBER THAT ANY PARAMETER THAT IS AREA-SPECIFIC NEEDS TO STAY IN A LOOP *(for i in 1: N)*
   (c) Add any prior distribution for the hyperparameters
   (d) Add any statement (if not specified) to calculate the parameters of interest:
      - posterior probabilities that the residual RR is above 1
      - ratio between the 5% top and 5% bottom of the distribution of the residual RR
      - variance of unstructured effects, conditional variance of the spatial effects, empirical variance of the spatial effects
      - spatial fraction
   (e) Save the file as *model_BYM.txt* and close it.

BACK TO the R file:

2. Order the data so that they match the order of the shapefile.

```
# Order the data according to POLY_ID in order to match the order of the shapefile data
lung <- lung[match(london@data[, 'POLY_ID'], lung[, 'POLY_ID']), ]
```

3. Format the data. Do not forget to include the health data and the adjacency matrix

```
data_A <- list(N = length(lung[, 1]), # nb of areas/wards
               O = lung$O, # observed nb of cases
               E = lung$E, # expected nb of cases
               adj = nbWB_A$adj, weights = nbWB_A$weights,
               num = nbWB_A$num)
# adjacency matrix
```

4. Create the initial values for ALL the unknown parameters. For the spatially structured RE, DO NOT FORGET THE SUM TO ZERO CONSTRAINT. As usual, create two different chains.

5. Which model parameters do you want to monitor? Set these before running OpenBUGS

```
parameters_A <- c("sigma2.v", "sigma2.u","overallRR", "QR90", "RR",
                  "resRR", "frac.spatial", "sigma2.u.marginal", "proba.resRR")
```

6. Specify the MCMC setting

```
ni <- 5000  #nb iterations
nt <- 5     #thinning interval
```

```
nb <- 3000  #nb iterations as burnin
nc <- 2     #nb chains
```

7. Run the MCMC simulations calling OpenBUGS from R using the function *bugs()*.

```
modelBYM.sim <- bugs(data = data_A, parameters = parameters_A,
                     inits = inits_A,
                     working.directory = "PartA",
                     model.file = "model_BYM.txt",
                     n.chains = nc, n.iter = ni, n.burnin = nb,
                     n.thin = nt, debug = FALSE,
                     DIC = TRUE, codaPkg = FALSE, summary.only = FALSE,
                     bugs.seed = 9)
```

8. Check the convergence of non area-specific monitored parameters through density, trace and autocorrelation plots. Remember: if you want to use the package mcmcplots as we did last week you need to say codaPkg = FALSE within the bugs() function.

```
mcmcplot(modelBYM.sim, c("sigma2.v", "sigma2.u", "overallRR",
                         "QR90", "frac.spatial", "sigma2.u.marginal"))
```

```
attach.bugs(modelBYM.sim)
```

9. Compare the DIC with the one from last week practical (Non spatial model)

```
modelBYM.sim$DIC
```

10. Produce summary statistics of the non area-specific monitored parameters. What is the proportion of variability explained by the spatial effect?

11. Now we want to map the smoothed RRs and the posterior probability that the smoothed residual RRs are above 1 in R using *spplot*.

(a) Extract the residual RR (i.e. exp(V)) and the posterior probability that resRR is higher than 1 as

```
RR_BYM <- data.frame(RR_BYM = apply(resRR, 2, mean),
                     pp_BYM = apply(proba.resRR, 2, mean))
```

(b) Create an ID for each ward, from 1 to 628

```
RR_BYM$POLY_ID <- 1:628
```

(c) Change the name of the column corresponding to the posterior mean to BYM

```
colnames(RR_BYM) <- c("BYM", "pp_resRR", "POLY_ID")
```

(d) Merge RR_BYM with lung using POLY_ID as column for merging

```
lung <- merge(lung, RR_BYM, by = "POLY_ID")
```

(e) Merge shapefile with lung data

```
data.London <- attr(london, "data")
attr(london, "data") <- merge(data.London, lung, by = "STwardcode")
```

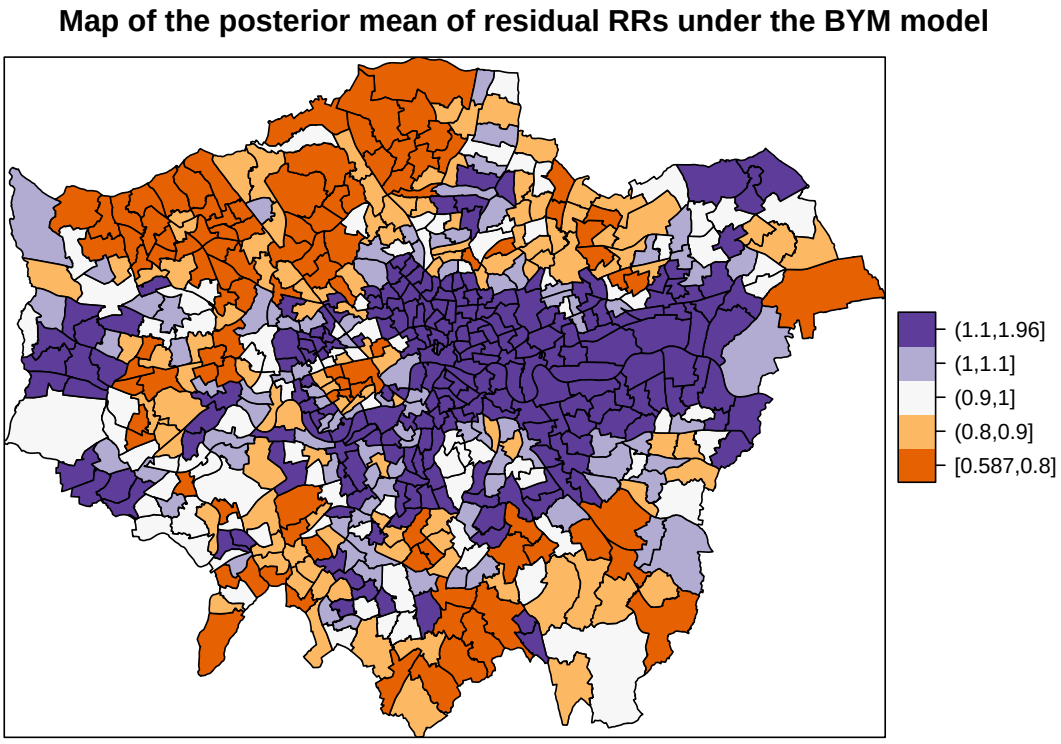
(f) Map the smoothed residual RR (resRR)

```
# Create classes for the column "BYM" using sensible cut off values
summary(london$BYM)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5872  0.8474  1.0029  1.0366  1.1865  1.9570
```

```
# We use cut off values 0.8, 0.9, 1, 1.1
london$BYM_classes <- cut(london$BYM,
                      breaks = c(min(london$BYM), 0.8, 0.9, 1, 1.1, max(london$BYM)),
                      include.lowest = T)

# display.brewer.all()
my.palette1 <- brewer.pal(n = 5, name = "PuOr")
spplot(obj = london, zcol = "BYM_classes", col.regions = my.palette1,
       main = "Map of the posterior mean of residual RRs under the BYM model")
```

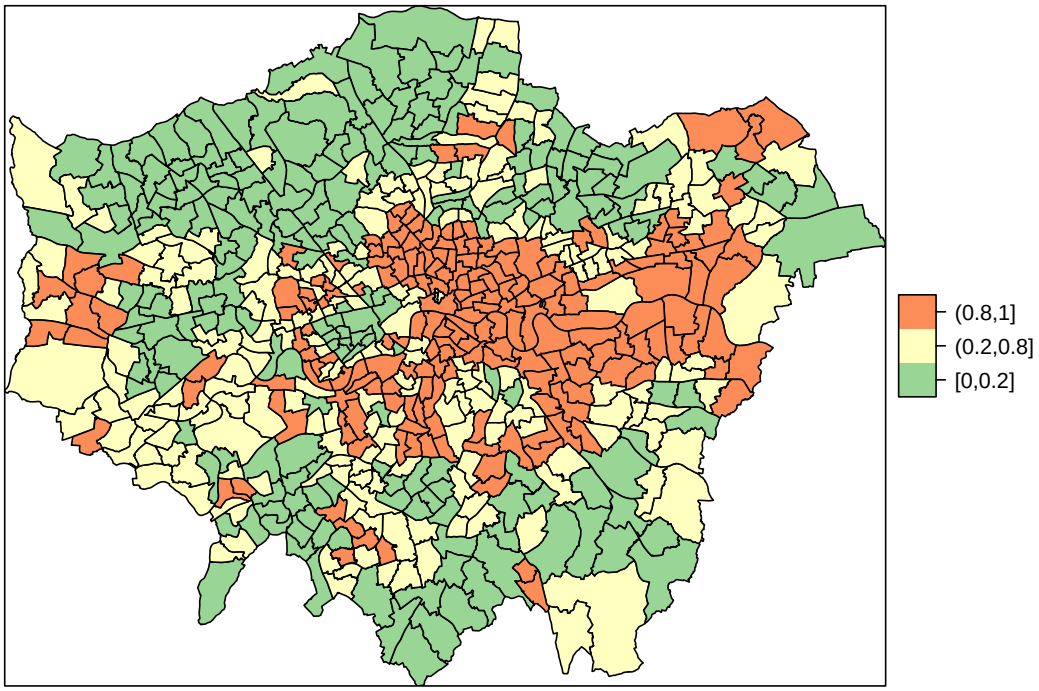


(g) Map the posterior probablity that resRR is above 1 (pp_resRR)

```
# create classes for the column "pp_resRR" using cut off values 0.2 and 0.8
london$pp_resRR_classes <- cut(london$pp_resRR,
                           breaks = c(min(london$pp_resRR), c(0.2, 0.8),
                                      max(london$pp_resRR)), include.lowest = T)

my.palette2 <- brewer.pal(n = 3, name = "Spectral")
spplot(obj = london, zcol = "pp_resRR_classes", col.regions = rev(my.palette2),
       main = "Map of the posterior probablity that resRR is above 1 under the BYM model")
```

Map of the posterior probablity that resRR is above 1 under the BYM model



16. Compare the results with what you obtained last week using the model with heterogeneity only.

PART B

In Part B of this practical, you will be using the R software Statistical Package (www.r-project.org), and OpenBUGS to carry out an ecological regression analysis following lecture 2. You will use larynx cancer data in 148 areas of West Yorkshire. You will estimate the association between larynx cancer and smoking at area level.

1. In folder 'Part B' you will find the following files:

- *larynx2.txt*: file containing larynx data with observed cases (O), expected cases (E) and total number of areas (N)
- *smoke2.txt*: file containing area-level smoke data
- *WestYorkshire*: folder containing the shapefile of West Yorkshire
- *model-BYM.txt*: BYGS code for the Poisson logNormal model with the unstructured and structured spatial effects (BYM model)

2. Health data

1. Import larynx and smoke data

```
larynx <- dget("PartB/larynx2.txt")
smoke <-  dget("PartB/smoke2.txt")
```

2. Print the first few rows of both datasets

3. Creating the adjacency matrix

1. Import the shapefile of West Yorkshire and plot it

```
## OGR data source with driver: ESRI Shapefile
## Source: "C:\Users\aboulier\Dropbox\SpatialAnalysis2019\Practicals\Practical2_AB\PartB\WestY
## with 148 features
## It has 3 fields
## Integer64 fields read as strings:  POLY_ID
```

2. Adjacency matrix in R

(a) Convert the polygons to a list of neighbours using the function *poly2nb*

```
## Neighbour list object:
## Number of regions: 148
## Number of nonzero links: 754
## Percentage nonzero weights: 3.442294
```

```
## Average number of links: 5.094595
## Link number distribution:
##
##  1  2  3  4  5  6  7  8  9
##  1  2 21 29 35 39  9 10  2
## 1 least connected region:
## 67 with 1 link
## 2 most connected regions:
## 5 87 with 9 links
```

(b) Convert this list to OpenBUGS format (i.e. a list of 3 components adj, num and weights) using the function *nb2WB* amd print a summary of the object

```
## [1] "adj"     "weights" "num"
```

```
##         Length Class  Mode
## adj     754    -none- numeric
## weights 754    -none- numeric
## num     148    -none- numeric
```

3. Carry out an ecologicial regression analysis

1. Modify the BYM model *model-BYM.txt* to include the covariate SMOKE as a contextual effect and save it as *model-BYM-reg.txt*
2. Format the data. Don't forget to include the smoke data and the adjacency matrix.
3. Create the initial values for ALL the unknown parameters. For the spatially structured RE, DO NOT FORGET THE SUM TO ZERO CONSTRAINT. As usual, create two different chains.
4. Specify the prameters to monitor

```
parameters_B = c("resRR", "proba.resRR","sigma2.v", "QR90", "RR","RR.smoke",
                 "RR.smoke10", "overallRR", "sigma2.u", "frac.spatial",
                 "sigma2.u.marginal")
```

6. Specify the MCMC setting

```
ni <- 5000  #nb iterations
nt <- 10     #thinning interval
nb <- 3000  #nb iterations as burnin
nc <- 2     #nb chains
```

7. Run the MCMC simulations calling OpenBUGS from R using the function *bugs()*

```
modelREG.sim  <- bugs(data = data_B, inits = inits_B,
                      working.directory = "PartB",
                      model.file = "model-BYM-reg.txt",
                      n.chains = nc, n.iter = ni, n.burnin = nb,
                      parameters = parameters_B, debug = FALSE, DIC = TRUE,
```

```
                        codaPkg = FALSE, summary.only = FALSE,
                        bugs.seed = 9)
```

8. Check convergence by using the function *mcmcplot()* and print summary statistics

```
mcmcplot(modelREG.sim, c("sigma2.v", "sigma2.u", "overallRR", "QR90", "frac.spatial",
                         "sigma2.u.marginal", "RR.smoke"))
```

9. Print summary statistics for the parameters of interest

```
summary(modelREG.sim$sims.list$QR90)
summary(modelREG.sim$sims.list$RR.smoke)
quantile(modelREG.sim$sims.list$RR.smoke, probs=c(0.05, 0.95))
```

10. What is the relative risk (and 95% uncertainty interval) of larynx cancer associated with living in a ward with 100% smoking prevalence versus a ward with zero smoking prevalence?

Answer: The relative risk of larynx cancer in areas with 100% smoking prevalence compared to areas with 0% smoking prevalence is 4.48 (95% interval 3.67 to 5.53). The 95% CI excludes 1 so larynx cancer risk is significantly associated with smoking prevalence of the area.

11. How would you calculate the relative risk of larynx cancer for wards with a 10% difference in smoking prevalence?

Answer: To calculate the relative risk of larynx cancer associated with a 10% increase in area smoking prevalence, we need to include an additional statement in the BUGS model code: *RR.smoke10 <- exp(beta1 * 0.1)* and then monitor the variable *RR.smoke10*. Note, it is not correct to simply plug in the log of the posterior mean of *RR.smoke* into the formula, i.e. *exp([mean of RR.smoke] * 0.1)* since the posterior mean of a non-linear function of a parameter does not equal the value of that function evaluated at the posterior mean of the parameter itself.

```
summary(modelREG.sim$sims.list$RR.smoke10)
```

12. Report and interpret the posterior mean (and 95% uncertainty interval) of the 90% RR ratio (QR90).

Answer: The 90% RR ratio indicates that there is around 1.1-fold variation (95% CI 1.0-fold to 1.3-fold) in the residual relative risk of larynx cancer across 90% of areas in the study region, after adjusting for the effects of area-level smoking.

13. Provide the following maps:

    - the smoking covariate
    - posterior mean relative risk
    - posterior mean residual relative risk
    - posterior probabilities that relative risks and residual relative risks exceed 1

To create the maps requested, we re-read the shapefile in order to have a new clean dataframe

```
westyork = readOGR("PartB/WestYorkshire","WestYorkshire")
```

Then we create a dataframe with the variables to plot and we merge it to the *westyork* spatial object.

```
results  <- data.frame(resRR = apply(modelREG.sim$sims.list$resRR, 2, mean),
                       pp_resRR = apply(modelREG.sim$sims.list$proba.resRR, 2, mean),
                       RR = apply(modelREG.sim$sims.list$RR, 2, mean),
                       POLY_ID = westyork$POLY_ID,
                       smoke = smoke$SMOKE)

westyork@data <- data.frame(westyork@data,
                 results[match(westyork@data[, 'POLY_ID'], results[, 'POLY_ID']), ])
```

And then plot using *tmap* functions

```
tm_shape(westyork) + tm_polygons("resRR", style = "pretty", palette = "PuOr",
                                 title = "Residual RR")

tm_shape(westyork) + tm_polygons("pp_resRR", style = "pretty", palette = "Blues",
                                 title = "Probability")

tm_shape(westyork) + tm_polygons("RR", style = "pretty", palette = "Greens",
                                 title = "Relative Risk")

tm_shape(westyork) + tm_polygons("smoke", style = "pretty", palette = "Purples",
                                 title = "Smoke")
```

PART C - Optional exercise

In folder Part C you will find a pdf of the paper 'Joint disease mapping using six cancers in the Yorkshire region of England' (Downing et al, 2008, International Journal of Health Geographics)

The aims of this paper were to model jointly the incidence rates of six smoking related cancers in the Yorkshire region of England (532 wards), between 1983 and 2003, to explore the patterns of spatial correlation amongst them. The six cancer sites included in this study were oesophagus (d1), stomach (d2), pancreas (d3), lung (d4), kidney (d5), and bladder (d6). The main known risk factors for each cancer are as follows:

- oesophagus: smoking, bodyweight/obesity, diet and alcohol consumption
- stomach: smoking, diet and alcohol consumption
- pancreas: smoking, bodyweight/obesity
- lung: smoking
- kidney: smoking, bodyweight/obesity
- bladder: smoking

The authors carried out a shared component analysis. The log relative risks were modelled as:

$$
\begin{aligned}
\log R_{1i} &= \alpha_1 + \phi_{1i}\kappa_{1,1} + \phi_{2i}\kappa_{2,1} + \phi_{3i}\kappa_{3,1} + \epsilon_{1i} \\
\log R_{2i} &= \alpha_2 + \phi_{1i}\kappa_{1,2} + \phi_{3i}\kappa_{3,2} + \epsilon_{2i} \\
\log R_{3i} &= \alpha_3 + \phi_{1i}\kappa_{1,3} + \phi_{2i}\kappa_{2,2} + \epsilon_{3i} \\
\log R_{4i} &= \alpha_4 + \phi_{1i}\kappa_{1,4} + \epsilon_{4i} \\
\log R_{5i} &= \alpha_5 + \phi_{1i}\kappa_{1,5} + \phi_{2i}\kappa_{2,3} + \epsilon_{5i} \\
\log R_{6i} &= \alpha_6 + \phi_{1i}\kappa_{1,6} + \epsilon_{6i}
\end{aligned}
$$

1. There are 3 shared components. Identify them according to the common risk factors
2. What do the random effects ϵ represent?
3. What do the parameters κ represent? Their estimates are in the table below. How would you interpret them?

Table 3: Posterior median (95% CI) relative weights of each cancer in the shared components analysis (unadjusted for socioeconomic background)

		Oesophagus	Stomach	Pancreas	Lung	Kidney	Bladder
Oesophagus	1	1.00					
	2	1.00					
	3	1.00					
Stomach	1	0.53 (0.26–0.92)	1.00				
	2	-	1.00				
	3	2.05 (0.56–6.15)	-				
Pancreas	1	0.89 (0.46–1.62)	1.67 (1.04–3.02)	1.00			
	2	1.72 (0.67–3.59)	-	1.00			
	3	-	-	1.00			
Lung	1	0.48 (0.24–0.80)	0.90 (0.67–1.17)	0.54 (0.30–0.84)	1.00		
	2	-	-	-	-		
	3	-	-	-	-		
Kidney	1	1.10 (0.52–2.22)	2.06 (1.19–3.95)	1.23 (0.65–2.42)	2.30 (1.34–4.47)	1.00	
	2	1.81 (0.66–4.09)	-	1.05 (0.49–2.32)	-	1.00	
	3	-	-	-	-	-	
Bladder	1	0.81 (0.41–1.45)	1.51 (0.96–2.73)	0.91 (0.52–1.58)	1.69 (1.07–3.05)	0.73 (0.38–1.40)	1.00
	2	-	-	-	-	-	-
	3	-	-	-	-	-	-

1 – The shared component representing smoking; 2 – The shared component representing bodyweight/obesity; 3 – The shared component representing diet/alcohol consumption.
The figures in the main body of the table represent the weight of the cancer listed along the top row relative to the cancers listed along the left hand side (with 95% confidence intervals). If the RR is > 1.00 the cancer along the top row has more weight, if the RR is < 1.00 the cancer along the left hand side has more weight.