Tongue cancer rates by DHB

Patrick Graham

09/09/2021

Contents

1 Tongue cancer rates by DHB

We consider rates of tongue cancer by DHB, based on cancers registered in 2018.

1.1 Setup the data

```r
cancerdata <- read.csv("data/tongue_cancer.csv",header=TRUE)
str(cancerdata)
```

```
## 'data.frame':    20 obs. of  6 variables:
##  $ Year      : int  2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...
##  $ Desc      : chr  "Tongue - C01-C02" "Tongue - C01-C02" "Tongue - C01-C02" "Tongue - C01-C02" ...
##  $ Sex       : chr  "AllSex" "AllSex" "AllSex" "AllSex" ...
##  $ Demography: chr  "Northland" "Waitemata" "Auckland" "Counties Manukau" ...
##  $ Cases     : int  7 23 14 12 19 3 10 4 6 3 ...
##  $ Population: int  185800 615100 493300 567000 421000 113400 249700 49500 172300 121300 ...
```

```r
#only a small dataset so why not just print it to have a look

cancerdata
```

```
##    Year             Desc    Sex        Demography Cases Population
## 1  2018 Tongue - C01-C02 AllSex         Northland     7     185800
## 2  2018 Tongue - C01-C02 AllSex         Waitemata    23     615100
## 3  2018 Tongue - C01-C02 AllSex          Auckland    14     493300
## 4  2018 Tongue - C01-C02 AllSex  Counties Manukau    12     567000
## 5  2018 Tongue - C01-C02 AllSex           Waikato    19     421000
## 6  2018 Tongue - C01-C02 AllSex             Lakes     3     113400
## 7  2018 Tongue - C01-C02 AllSex     Bay of Plenty    10     249700
## 8  2018 Tongue - C01-C02 AllSex        Tairawhiti     4      49500
## 9  2018 Tongue - C01-C02 AllSex        Hawke's Bay     6     172300
## 10 2018 Tongue - C01-C02 AllSex          Taranaki     3     121300
## 11 2018 Tongue - C01-C02 AllSex        MidCentral     9      66700
```

```
## 12 2018 Tongue - C01-C02 AllSex          Whanganui     4     181700
## 13 2018 Tongue - C01-C02 AllSex    Capital & Coast    12     153900
## 14 2018 Tongue - C01-C02 AllSex        Hutt Valley     7     315900
## 15 2018 Tongue - C01-C02 AllSex          Wairarapa     1      46800
## 16 2018 Tongue - C01-C02 AllSex Nelson Marlborough     4     155500
## 17 2018 Tongue - C01-C02 AllSex         West Coast     3      32400
## 18 2018 Tongue - C01-C02 AllSex         Canterbury    20     560800
## 19 2018 Tongue - C01-C02 AllSex   South Canterbury     1      60900
## 20 2018 Tongue - C01-C02 AllSex           Southern    15     307400
```

```
rates <- cancerdata$Cases / cancerdata$Population
summary(rates)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 1.642e-05 2.409e-05 3.524e-05 4.371e-05 4.605e-05 1.349e-04
```

```
#may make more  sense if expressed as rate per 100000
rates100000 <- rates*100000
summary(rates100000)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.642   2.409   3.524   4.371   4.605  13.493
```

```
rates.df <-  data.frame(cancerdata$Demography,rates100000,rates,
                        cancerdata$Cases,cancerdata$Population)
names(rates.df) <-  c("Dhb","rate_100000","rate","Cases", "Population")
rates.df
```

```
##                   Dhb rate_100000         rate Cases Population
## 1           Northland    3.767492 3.767492e-05     7     185800
## 2           Waitemata    3.739229 3.739229e-05    23     615100
## 3             Auckland    2.838030 2.838030e-05    14     493300
## 4     Counties Manukau    2.116402 2.116402e-05    12     567000
## 5              Waikato    4.513064 4.513064e-05    19     421000
## 6                Lakes    2.645503 2.645503e-05     3     113400
## 7        Bay of Plenty    4.004806 4.004806e-05    10     249700
## 8           Tairawhiti    8.080808 8.080808e-05     4      49500
## 9          Hawke's Bay    3.482298 3.482298e-05     6     172300
## 10            Taranaki    2.473207 2.473207e-05     3     121300
## 11          MidCentral   13.493253 1.349325e-04     9      66700
## 12           Whanganui    2.201431 2.201431e-05     4     181700
## 13     Capital & Coast    7.797271 7.797271e-05    12     153900
## 14         Hutt Valley    2.215891 2.215891e-05     7     315900
## 15           Wairarapa    2.136752 2.136752e-05     1      46800
## 16 Nelson Marlborough    2.572347 2.572347e-05     4     155500
## 17          West Coast    9.259259 9.259259e-05     3      32400
## 18          Canterbury    3.566334 3.566334e-05    20     560800
## 19    South Canterbury    1.642036 1.642036e-05     1      60900
## 20            Southern    4.879636 4.879636e-05    15     307400
```

```
#What is the overall rate
rawrate <- sum(cancerdata$Cases) /sum(cancerdata$Population)
rawrate*100000 #3.634198
```

```
## [1] 3.634198
```

1.2 Model 1: Common underlying rate across all DHBs

Let Y_i and N_i be the number of cancer cases and population at risk for the i^{th} DHB. Initially we consider a simple model that assumes the underlying rate is the same for all DHBS:

$$[Y_i|\lambda] \sim \text{Poisson}(Y_i|N_i\lambda),\ i = 1\ldots,n \tag{1}$$

so, for $i \in 1,\ldots,n$, $E(Y_i|\lambda) = N_i\lambda$ and $E(Y_i/N_i = \lambda)$. That is, λ is the expected tongue cancer rate, often referred to as the underlying rate, in contrast to the observable rate $r_i = Y_i/N_i$. The populations at risk, or "exposures," $\{N_i, i \in 1,\ldots,n\}$ are regarded as known constants. It is arguable whether we should make the conditioning on these values explicit by including them in the model definition and write

$$[Y_i|N_i,\lambda] \sim \text{Poisson}(N_i\lambda), i = 1\ldots,n.$$

To simplify notation a little we will just regard the N_i as extra background information we know in advance of observing the cancer counts and so we won't explicitly condition on them.

Model (1) assumes the expected rate is the same for all DHBs. This means the observable rates $r_i = Y_i/N_i$ vary over DHBs only because of random variation and not because of variation in the underlying rates. The model is a bit simplistic but can be a base to compare other models against. Here we are just using it as a simple example to explore some aspects of Monte Carlo simulation of posterior distributions.

Let $\mathbf{Y} = (Y_1,\ldots,Y_n)$. We assume conditional independence of the Y_i given θ so the likelihood is

$$p(\mathbf{Y}|\lambda) = \prod_{i=1}^{n} \text{Poisson}(Y_i|N_i\lambda).$$

The conjugate prior for the rate parameter of a Poisson model is Gamma(a,b). We set $a = 3/100,000$, $b = 1$ and our prior expectation is therefore $E(\theta) = a/b = 3/100,000$ Thus, our prior guess at the underlying rate is 3 per 100,000, but the weight we put on that prior guess is equivalent to the information that would be provided by learning the number of cases for a DHB with a population of size 1 (impossible of course, because the rate for such a DHB would be 0 or 1, but its a useful analogy.)

The posterior for λ is

$$\begin{aligned}
p(\lambda|\mathbf{Y}) &= \frac{\prod_{i=1}^{n} \text{Poisson}(Y_i|N_i\lambda)\text{Gamma}(\lambda|3/100,000,1)}{\int \prod_{i=1}^{n} \text{Poisson}(Y_i|N_i\lambda)\text{Gamma}(\lambda|3/100,000,1)\,\mathrm{d}\lambda} \\
&\propto \left[\prod_{i=1}^{n}(N_i\lambda)^{Y_i}\exp(-N_i\lambda)\right]\left[\lambda^{((3/100,000)-1)}\exp(-\lambda)\right] \\
&= \left[\prod_{i=1}^{n} N_i^{Y_i}\right]\left[\prod_{i=1}^{n}\lambda^{Y_i}\right]\left[\exp(\lambda\sum_{i=1}^{n} N_i)\right]\left[\lambda^{((3/100,000)-1))}\exp(-\lambda)\right] \\
&\propto \lambda^{\left(\sum_{i=1}^{n} Y_i + ((3/100,000)-1)\right)}\exp -\left(\lambda\left(\sum_{i=1}^{n} N_i + 1\right)\right) \\
&\propto \text{Gamma}\left(\lambda|\left(\sum_{i=1}^{n} Y_i + (3/100,000)\right), (N_i+1)\right)
\end{aligned} \tag{2}$$

Code to compuet the posterior for the assumed common rate for all DHBs is straightforward.

```
##Gamma prior
a <- 3/100000
b <- 1      #Like saying prior  evidence is  equivalent  to
            #one extra tiny DHB of size 1
```

```r
totcases <- sum( cancerdata$Cases)
totpop <- sum(cancerdata$Population)

##update to get  parameters of the posterior, using conjugacy

apost <- a + totcases
bpost <- b + totpop

##Compute posterior  summaries
postmean <- apost/bpost
post_median <-  qgamma(0.5,shape=apost,rate = bpost)

postmean * 100000
```

```
## [1] 3.634198
```

```r
post_median * 100000
```

```
## [1] 3.627357
```

```r
q025 <- qgamma(0.025,shape=apost,rate=bpost)
q975 <-  qgamma(0.975,shape=apost,rate=bpost)

exact_quantiles <- 100000*c(q025,post_median,q975)
exact_quantiles
```

```
## [1] 3.118513 3.627357 4.188762
```

```r
# Simulation approach - first look at modest size Monte Carlo  sample
post_lambda100 <- rgamma(n=100,shape=apost,rate=bpost)

##check quantiles
post_quantiles100 <-
  quantile(post_lambda100,probs=c(0.025,0.5,0.975))

post_quantiles100
```

```
##         2.5%          50%        97.5%
## 3.107372e-05 3.636848e-05 4.154266e-05
```

```r
exact_quantiles
```

```
## [1] 3.118513 3.627357 4.188762
```

```r
#check_mean
post_mean100 <- mean(post_lambda100)
100000*post_mean100  #3.662318
```

```
## [1] 3.65006
```

```r
exact_mean <- apost / bpost
100000*exact_mean   #3.634198
```

```
## [1] 3.634198
```

```r
#check standard deviation
post_sd100 <-  sd(post_lambda100)
```

The Monte Carlo error for the posterior mean is

```r
MCerror <- post_sd100 / sqrt(100)
MCerror
```

```
## [1] 3.020686e-07
```

which is tiny. It makes more sense when multiplied by 100,000, as per the rates themselves. Recall

$$\mathrm{Var}(C\theta) = C^2\mathrm{Var}(\theta)$$

so

$$\mathrm{sd}(C\theta) = C\mathrm{sd}(\theta)$$

```r
100000 * MCerror #0.02695065,
```

```
## [1] 0.03020686
```

which is still pretty small. The MC mean is about 1 MC standard error from the exact mean.

Now, see what happens for a bigger posterior sample.

```r
post_lambda1000 <- rgamma(n=1000,shape=apost,rate=bpost)

post_quantiles1000 <- quantile(post_lambda1000,probs=c(0.025,0.5,0.975))
##Compare true and simulation results

exact_quantiles
```

```
## [1] 3.118513 3.627357 4.188762
```

```r
100000*post_quantiles100
```

```
##     2.5%      50%    97.5%
## 3.107372 3.636848 4.154266
```

```r
100000*post_quantiles1000
```

```
##     2.5%      50%    97.5%
## 3.119403 3.618105 4.210571
```

```r
#tail quantiles looking pretty  good by the time Monte Carlo
#simulation size reaches 1000

#MC error for nsim=1000

post_sd1000 <-  100000*sd(post_lambda1000)
post_sd1000
```

```
## [1] 0.2733795
```

```r
MC_error1000 <- post_sd1000/sqrt(1000)
MC_error1000
```

```
## [1] 0.00864502
```

```r
post_mean1000 <- 100000*mean(post_lambda1000)
post_mean1000
```

```
## [1] 3.636131
```

```r
100000*exact_mean
```

```
## [1] 3.634198
```

So the true mean is about on MC standard error from the exact mean. The MC error is fairly trivial though and represents 1/sqrt(1000) = 3.2% of the posterior standard deviation

1.3 Model 2: DHB specific underlying rates and inference for functions of these rates

We now turn our attention to Monte Carlo simulation for more challenging estimands (things to be estimated). We let the underlying rates vary by DHB and address questions such as the rank of each DHB and the probability that the underlying rate in each DHB is the largest of all DHBs. In reality, any such comparisons should take account of other factors that vary by DHB and affect cancer rates, such as age structure. However, to keep the illustration relatively simple we ignore such factors here. Consequently, the comparisons by DHB presented below should not be taken too seriously. Further analysis would be needed to rigorously compare tongue cancer rates by DHB.

Our model is now

$$[Y_i|\lambda_i] \sim \text{Poisson}(N_i\lambda_i),\; i = 1, \ldots, n \tag{3}$$

which differs from (1) by allowing each DHB to have its own underlying rate parameter λ_i in contrast to the common underlying rate assumed in (1). Assuming conditional independence over DHBs and letting $\boldsymbol{\lambda} = (\lambda_1, \ldots \lambda_n)^{'})$, the likelihood is now

$$p(\mathbf{Y}|\boldsymbol{\lambda}) = \prod_{i=1}^{n} \text{Poisson}(Y_i|N_i\lambda_i). \tag{4}$$

If we assume *a priori* independence,then $p(\boldsymbol{\lambda}) = \prod_{i=1}^{n} p(\lambda_i)$. We will assume independent Gamma priors for the underlying tongue cancer rates, and, in particular, assume that each DHB parameter has the same $\text{Gamma}((3/100,000), 1)$ prior as adopted for the assumed common λ in model (1). Therefore the prior for the underlying rates is

$$p(\boldsymbol{\lambda}) = \prod_{i=1}^{n} \text{Gamma}(\lambda_i|(3/100,000), 1). \tag{5}$$

From (4) and (5) the joint posterior for the underlying rate parameters is

$$\begin{aligned} p(\boldsymbol{\lambda}|\mathbf{Y}) &= \frac{\prod_{i=1}^{n} \text{Poisson}(Y_i|N_i\lambda_i) \prod_{i=1}^{n} \text{Gamma}(\lambda_i|(3/100,000), 1)}{\int \prod_{i=1}^{n} \text{Poisson}(Y_i|N_i\lambda_i) \prod_{i=1}^{n} \text{Gamma}(\lambda_i|(3/100,000), 1)\, \mathrm{d}\boldsymbol{\lambda}} \\ &\propto \prod_{i=1}^{n} \text{Poisson}(Y_i|N_i\lambda_i) \times \text{Gamma}(\lambda_i|(3/100,000), 1) \\ &\propto \prod_{i=1}^{n} \text{Gamma}\left(\lambda_i|(Y_i + (3/100,000)), (N_i + 1)\right) \end{aligned} \tag{6}$$

where the last line (6) follows fron the conjugacy of the Poisson and Gamma distributions.

So, we can generate the joint posterior distribution for the underlying rates by drawing from independent Gamma distributions. As well as summarising the posterior for the underlying rates for each DHB, we can compare rates acroos DHBs and compute such this as the posterior ditribuion of each DHBs rank and $\Pr(\lambda_i > \lambda_j, \forall j \neq i|\mathbf{Y})$, i.e the posterior probability that the underlying rate for DHB i is greater than the underlying rate for all other DHBs.

```
fulla_post <-  a + cancerdata$Cases        #vector
fullb_post <- b +  cancerdata$Population  #vector

fulla_post
```

```
##  [1]  7.00003 23.00003 14.00003 12.00003 19.00003  3.00003 10.00003  4.00003  6.00003
## [10]  3.00003  9.00003  4.00003 12.00003  7.00003  1.00003  4.00003  3.00003 20.00003
## [19]  1.00003 15.00003
```

```
fullb_post
```

```
##  [1] 185801 615101 493301 567001 421001 113401 249701  49501 172301 121301  66701 181701
## [13] 153901 315901  46801 155501  32401 560801  60901 307401
```

```
## rgamma is partially vectorised; Easiest to  loop
## over simulations and on each iteration generate   the vector of lambda
##lambda values for the 20 DHBs
## also  need to  work out the maximum and rank for  each  set  of  lambdas
##generated

M <- 1000 ##number of  draws from  the posterior
n <- length(rates)  #number of groups - DHBs in this case

##Set-up structures  for  storing output

post_fullambda <- matrix(nrow=M,ncol=n )

post_max <- matrix(nrow=M,ncol=n)

post_rank  <- matrix(nrow=M,ncol=n)

for (i in 1:M ) {

  fullambda <- rgamma(n,shape=fulla_post,rate=fullb_post)

  ranks <- rank(fullambda)
  ismax <- (ranks == max(ranks) )

  post_fullambda[i,] <- fullambda
  post_rank[i,]  <- ranks
  post_max[i,]  <- ismax

  }

##check results

##posterior  quantiles for  each DHB

fullpost_quantiles <-  apply(post_fullambda,MARGIN=2,FUN=quantile,
                             probs=c(0.025,0.5,0.975))

fullpost_quantiles.df <-  data.frame(rates.df$Dhb,t(100000*fullpost_quantiles))


fullpost_quantiles.df <-
cbind(fullpost_quantiles.df,rates.df$Cases)

names(fullpost_quantiles.df) <- c("DHB","q025","q50","q975","cases")
fullpost_quantiles.df
```

```
##                    DHB       q025       q50       q975 cases
```

```
## 1           Northland 1.65444626  3.517522  6.920025     7
## 2           Waitemata 2.42320310  3.676191  5.359995    23
## 3            Auckland 1.51521882  2.796301  4.408050    14
## 4    Counties Manukau 1.06150977  2.096230  3.403149    12
## 5             Waikato 2.70014759  4.426788  6.961326    19
## 6               Lakes 0.63930037  2.367682  6.586044     3
## 7        Bay of Plenty 1.95621243  3.798187  6.804720    10
## 8          Tairawhiti 2.34942450  7.542662 17.904782     4
## 9          Hawke's Bay 1.23543386  3.275601  6.657502     6
## 10           Taranaki 0.48299559  2.211004  6.215087     3
## 11          MidCentral 6.21342678 12.854694 24.599970     9
## 12           Whanganui 0.62397582  2.005451  4.756890     4
## 13     Capital & Coast 3.72194130  7.517346 12.664989    12
## 14         Hutt Valley 0.87653985  2.104554  4.044580     7
## 15           Wairarapa 0.05356417  1.520950  7.653331     1
## 16 Nelson Marlborough 0.68726132  2.390353  5.757225     4
## 17          West Coast 1.88841422  8.491304 21.066409     3
## 18          Canterbury 2.18543470  3.519093  5.370977    20
## 19    South Canterbury 0.04188125  1.075816  5.969569     1
## 20            Southern 2.79727183  4.788989  7.546623    15
```

```
##posterior quantiles for  each DHB's rank
fullpost_ranks_quantiles.df <-
  apply(post_rank,MARGIN=2,FUN=quantile,
                                   probs=c(0.025,0.5,0.975))

fullpost_ranks_quantiles.df <-
 data.frame( t(apply(post_rank,MARGIN=2,FUN=quantile,
       probs=c(0.025,0.5,0.975)) ) )

fullpost_ranks_quantiles.df$Dhb <- rates.df$Dhb
fullpost_ranks_quantiles.df
```

```
##     X2.5. X50. X97.5.                Dhb
## 1    3.000   11     18          Northland
## 2    5.975   12     16          Waitemata
## 3    3.000    8     15           Auckland
## 4    1.000    5     11   Counties Manukau
## 5    8.000   14     18            Waikato
## 6    1.000    6     17              Lakes
## 7    4.000   12     18      Bay of Plenty
## 8    6.000   18     20         Tairawhiti
## 9    2.000   10     17        Hawke's Bay
## 10   1.000    6     17           Taranaki
## 11  17.000   20     20         MidCentral
## 12   1.000    5     15          Whanganui
## 13  12.000   18     20    Capital & Coast
## 14   1.000    5     13        Hutt Valley
## 15   1.000    3     18          Wairarapa
## 16   1.000    6     16 Nelson Marlborough
## 17   5.000   18     20         West Coast
## 18   6.000   11     16         Canterbury
## 19   1.000    2     17   South Canterbury
## 20   8.000   15     18           Southern
```

```r
names(fullpost_ranks_quantiles.df)[1:3] <- c("q025","q50","q975")

fullpost_ranks_quantiles.df
```

```
##      q025 q50 q975                 Dhb
## 1   3.000  11   18           Northland
## 2   5.975  12   16           Waitemata
## 3   3.000   8   15            Auckland
## 4   1.000   5   11    Counties Manukau
## 5   8.000  14   18             Waikato
## 6   1.000   6   17               Lakes
## 7   4.000  12   18       Bay of Plenty
## 8   6.000  18   20          Tairawhiti
## 9   2.000  10   17         Hawke's Bay
## 10  1.000   6   17            Taranaki
## 11 17.000  20   20          MidCentral
## 12  1.000   5   15           Whanganui
## 13 12.000  18   20     Capital & Coast
## 14  1.000   5   13         Hutt Valley
## 15  1.000   3   18           Wairarapa
## 16  1.000   6   16 Nelson Marlborough
## 17  5.000  18   20          West Coast
## 18  6.000  11   16          Canterbury
## 19  1.000   2   17    South Canterbury
## 20  8.000  15   18            Southern
```

```r
##posterior probability that rate in each DHB is the maximum

fullpost_max <- colMeans(post_max)
fullpost_max.df <- data.frame(rates.df$Dhb,fullpost_max)
names(fullpost_max.df) <- c("Dhb","prob")
fullpost_max.df
```

```
##                   Dhb  prob
## 1           Northland 0.003
## 2           Waitemata 0.000
## 3            Auckland 0.000
## 4    Counties Manukau 0.000
## 5             Waikato 0.000
## 6               Lakes 0.000
## 7       Bay of Plenty 0.001
## 8          Tairawhiti 0.113
## 9         Hawke's Bay 0.001
## 10           Taranaki 0.000
## 11         MidCentral 0.630
## 12          Whanganui 0.000
## 13    Capital & Coast 0.053
## 14        Hutt Valley 0.000
## 15          Wairarapa 0.003
## 16 Nelson Marlborough 0.001
## 17         West Coast 0.191
## 18         Canterbury 0.000
## 19   South Canterbury 0.003
## 20           Southern 0.001
```

```r
##What about probability in the top 5

intop5 <- (post_rank >= 16)

Prtop5 <- colMeans((intop5) )

Prtop5.df <- data.frame(rates.df$Dhb,Prtop5)
Prtop5.df
```

```
##          rates.df.Dhb Prtop5
## 1           Northland  0.148
## 2           Waitemata  0.075
## 3            Auckland  0.003
## 4     Counties Manukau  0.000
## 5             Waikato  0.269
## 6               Lakes  0.067
## 7        Bay of Plenty  0.178
## 8          Tairawhiti  0.768
## 9          Hawke's Bay  0.123
## 10            Taranaki  0.051
## 11          MidCentral  0.993
## 12           Whanganui  0.021
## 13     Capital & Coast  0.887
## 14         Hutt Valley  0.003
## 15           Wairarapa  0.088
## 16 Nelson Marlborough  0.041
## 17          West Coast  0.787
## 18          Canterbury  0.060
## 19   South Canterbury  0.037
## 20            Southern  0.401
```

We can see that very few of the DHBS are definitively in the top 5. We would be most confident about Mid-central, West Coast and Taranaki ranking in the top 5.

The fact that we could only confidently assert that three of the 20 DHBS are in the top 5 of all DHBs, illustrates the uncertainty in the estimation and the difficulties that would be faced if some treatment or screening programme was to be targeted at the DHBs with the highest rates. However, the data analysed here pertain to a single year, 2018. Analysing more years of data would give more stable picture of variation by DHB in tongue cancer rates, though adjustment for differences in population structure between DHBs would be necessary before any firm conclusions could be drawn from an analysis of this sort.

Later in the course we will consider models that are a compromise between model (1) and (3).