# Bayesian Hierarchical Analysis of Esophageal Cancer Risk Factors: Evaluating the Role of Age, Tobacco, and Alcohol Consumption

2023-11-29

## Introduction

Esophageal cancer is cancer in the esophagus – the food pipe connecting the throat and the stomach. In recent years, esophageal cancer has been among the most common types of cancer: in 2020, it ranked 8th in the most diagnosed type of cancer globally, with 604,000 new cases. It is also the 6th most deadly cancer killing over 544,000 people in 2020.

Smoking and alcohol use have long been known as two major risk factors associated with esophageal cancer, and they are among the few risk factors that can be controlled or changed by the subject. The goal of this project is to utilize Bayesian modeling to the data from a case-control study of esophageal cancer in hopes of drawing a relation between a person's level of tobacco and alcohol intake and their chance of developing esophageal cancer. Our modelling will also account for the patients' age as that is an important factor in the development of cancers.

## Data set

Our data set was created by Tuyns AJ, Péquignot G, Jensen OM. in a paper *Esophageal cancer in Ille-et-Vilaine about levels of alcohol and tobacco consumption. Risks are multiplying. (1977).*

It was also used in a statistical book by *Breslow, N. E. and Day, N. E. (1980) Statistical Methods in Cancer Research. Volume 1: The Analysis of Case-Control Studies. IARC Lyon / Oxford University Press.* In the book, the data set was used to showcase statistical methods and techniques. There was also logistic regression performed on the data set. However, no Bayesian statistics was used as the book focuses on multiple regression models.

```
##   agegp     alcgp    tobgp ncases ncontrols total prob
## 1 25-34 0-39g/day 0-9g/day      0        40    40    0
## 2 25-34 0-39g/day    10-19      0        10    10    0
## 3 25-34 0-39g/day    20-29      0         6     6    0
## 4 25-34 0-39g/day      30+      0         5     5    0
## 5 25-34     40-79 0-9g/day      0        27    27    0
## 6 25-34     40-79    10-19      0         7     7    0
```

The data set contains the following columns:

1. **agegp**: Age group
2. **alcgp**: Alcohol consumption group
3. **tobgp**: Tobacco consumption group
4. **ncases**: Number of cases

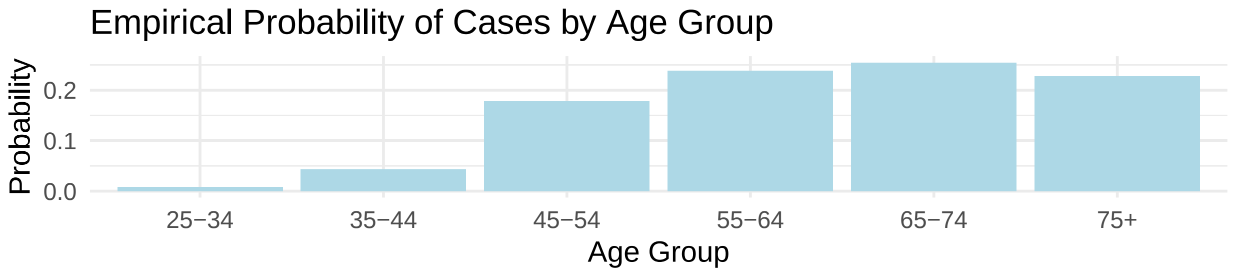
5. **`ncontrols`**: Number of controls
6. **`total`**: Total number of subjects
7. **`prob`**: Empirical probability

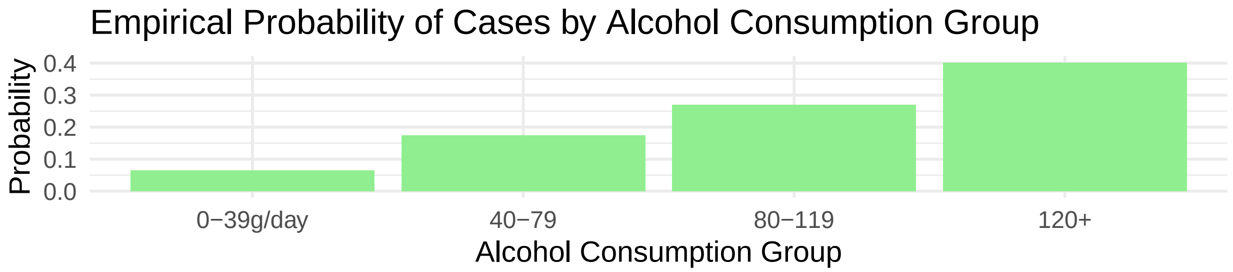
## Visualization

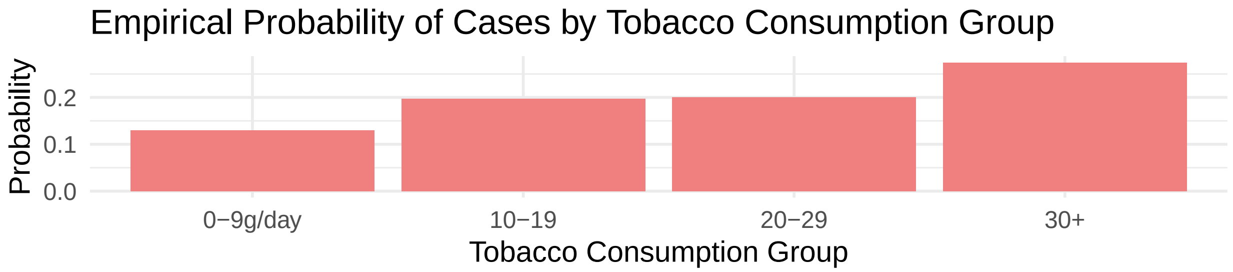
**Empirical Probabilities by Age Group**: This chart shows the empirical probabilities across different age groups.

**Empirical Probabilities by Alcohol Consumption Group**: This chart displays the empirical probabilities in relation to different levels of alcohol consumption.

**Empirical Probabilities by Tobacco Consumption Group**: Similar to the alcohol consumption chart, this one shows the empirical probabilities across various levels of tobacco consumption.







## Data pre-processing

Transform the original data set into a new data set suitable for Bernoulli fit:

1. **`ID`**: ID of the patient
2. **`agegp`**: Age group

3. **alcgp**: Alcohol consumption group

4. **tobgp**: Tobacco consumption group

5. **cancer**: Binary indicator whether there is presence in the patient

```
##   ID   age       tob      alc cancer
## 1  1 25-34 0-39g/day 0-9g/day      1
## 2  2 25-34 0-39g/day 0-9g/day      1
## 3  3 25-34 0-39g/day 0-9g/day      0
## 4  4 25-34 0-39g/day 0-9g/day      0
## 5  5 25-34 0-39g/day 0-9g/day      0
## 6  6 25-34 0-39g/day 0-9g/day      0
```

# Split the Data

We will split the Data into Training and Test Sets using the function split(). The `SplitRatio = 0.7` argument of the function indicates that approximately 70% of the data should be allocated to the training set, and the remaining 30% in the test set.

# Motivation

The motivation of the project is to model the probability of having cancer based on several explanatory variables: age, tobacco usage, and alcohol consumption. The response variable (cancer occurrence) is binary, where '1' indicates the presence of cancer and '0' indicates its absence.

### Mathematical Notation:

The logistic regression model can be expressed as follows:

$$\text{logit}(P(Y=1)) = \log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Where: - $P(Y=1)$ is the probability of having cancer. - $\beta_0$ is the intercept. - $\beta_1, \beta_2, \cdots, \beta_k$ are the coefficients for the explanatory variables $X_1, X_2, \cdots, X_k$ (age, tobacco, and alcohol consumption in your case). These parameter estimate will show the change in the log-odds of the outcome when the variable changes from its reference level (usually 0) to the other level. In practical terms, this parameter tells us the difference in the log-odds of the probability of having cancer.

# Pooled ordinal logistic regression model

### Priors and Likelihood:

1. **Priors**:

- For the $\beta$ coefficients, a common choice is to use normal distributions as priors due to their mathematical convenience and the central limit theorem. For example, $\beta_i \sim \mathcal{N}(0, \sigma^2)$, where $\sigma^2$ is the variance. However, we can also use the flat priors suggested by BRMS and then later test the prior sensitivity of the model.

- According to American Cancer Society, Age, Tobacco, and Alcohol are ones of the risk factors [2]. A risk factor is ‘anything that increases your chance of getting the esophageal cancer’. But having a risk factor, or even many, does not mean that you will get the cancer. Therefore, one possible choice for the priors of the coefficients is $\beta_i \sim \mathcal{N}(0.2, 10)$, with a close-to-zero positive mean and a very big standard deviation for it to be a flat priors, resulting in a posterior mainly influenced by the data.

```
#Examine the default priors suggested
get_prior(cancer ~ 1 + tob + alc + age,
             data = train_data,
             family = bernoulli("logit"))
```

```
##                 prior     class  coef group resp dpar nlpar lb ub       source
##                (flat)         b                                        default
##                (flat)         b age.C                             (vectorized)
##                (flat)         b age.L                             (vectorized)
##                (flat)         b age.Q                             (vectorized)
##                (flat)         b ageE4                             (vectorized)
##                (flat)         b ageE5                             (vectorized)
##                (flat)         b alc.C                             (vectorized)
##                (flat)         b alc.L                             (vectorized)
##                (flat)         b alc.Q                             (vectorized)
##                (flat)         b tob.C                             (vectorized)
##                (flat)         b tob.L                             (vectorized)
##                (flat)         b tob.Q                             (vectorized)
##  student_t(3, 0, 2.5) Intercept                                        default
```

2. **Likelihood**: The likelihood of observing the data given the parameters is modeled by a binomial distribution (for binary outcomes), which in the case of logistic regression simplifies to a product of Bernoulli trials.

### BRMS code

```
# Defining the ordinal logistic model
model <- brm(cancer ~ 1 + tob + alc + age,
             data = train_data,
             family = bernoulli("logit"),
             prior = c(prior(normal(0, 10), class = "b"),
                       prior(normal(0, 10), class = "Intercept")),
             chains=4,
             file="model1",
             cores=4)
```

The model uses 4 chains, each with 2000 iterations. The warm-up length is 1000. This are mostly default settings but, as we show below, the MCMC chains converge well.

### Convergence diagnostic

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: cancer ~ 1 + tob + alc + age
##    Data: train_data (Number of observations: 863)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept    -0.96      0.12    -1.19    -0.74 1.00     7224     3105
## tob.L         1.37      0.21     0.97     1.78 1.00     7140     3502
## tob.Q        -0.20      0.19    -0.57     0.16 1.00     7517     3354
## tob.C        -0.02      0.17    -0.36     0.31 1.00     7453     2954
## alc.L         0.97      0.19     0.60     1.33 1.00     6926     3006
## alc.Q         0.21      0.19    -0.16     0.57 1.00     6710     3391
## alc.C         0.12      0.20    -0.25     0.51 1.00     7502     3104
## age.L         0.71      0.32     0.06     1.34 1.00     5234     3361
## age.Q         0.20      0.30    -0.41     0.77 1.00     6033     3111
## age.C        -0.33      0.26    -0.85     0.18 1.00     5270     3275
## ageE4        -0.00      0.23    -0.47     0.46 1.00     7083     3233
## ageE5        -0.12      0.19    -0.50     0.26 1.00     7483     3419
##
## Draws were sampled using sample(hmc). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

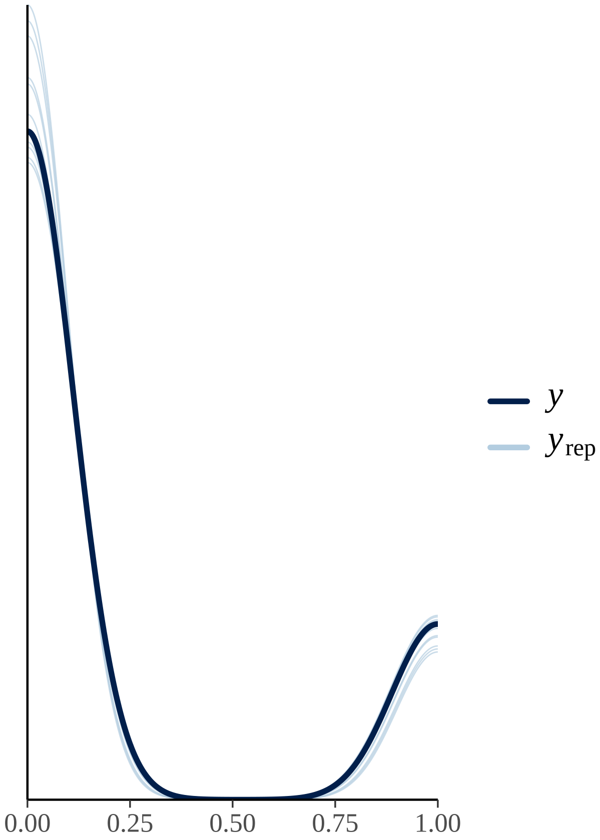
From the summary we can learn that our model estimation converged well: $\hat{R} \sim 1.0$ and effective sample size greater than 100 times the number of chains (=4) for each parameter (Vehtari et al. 2021) [1].

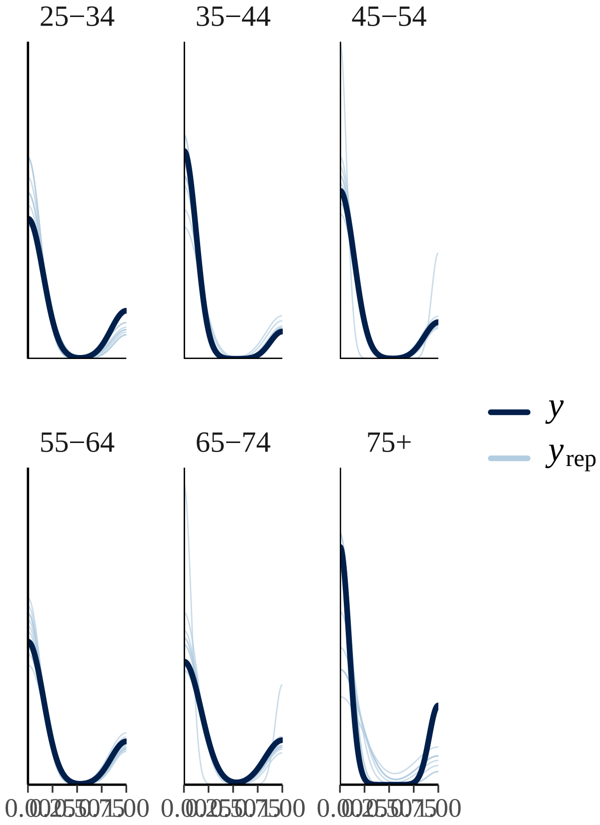
### Posterior predictive checks

These first two plots are used for visualizing the overlay of posterior predictive densities on top of the observed data.

```
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
```

```
## Using 10 posterior draws for ppc type 'dens_overlay_grouped' by default.
```

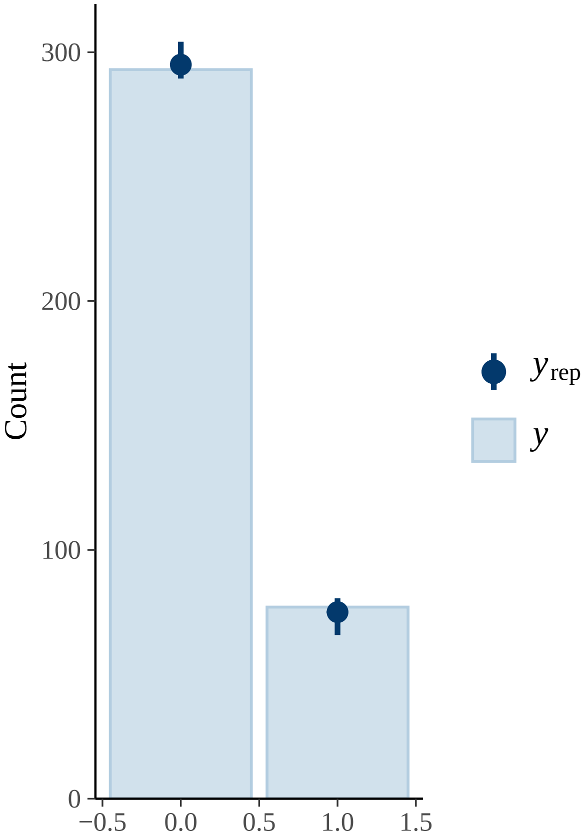


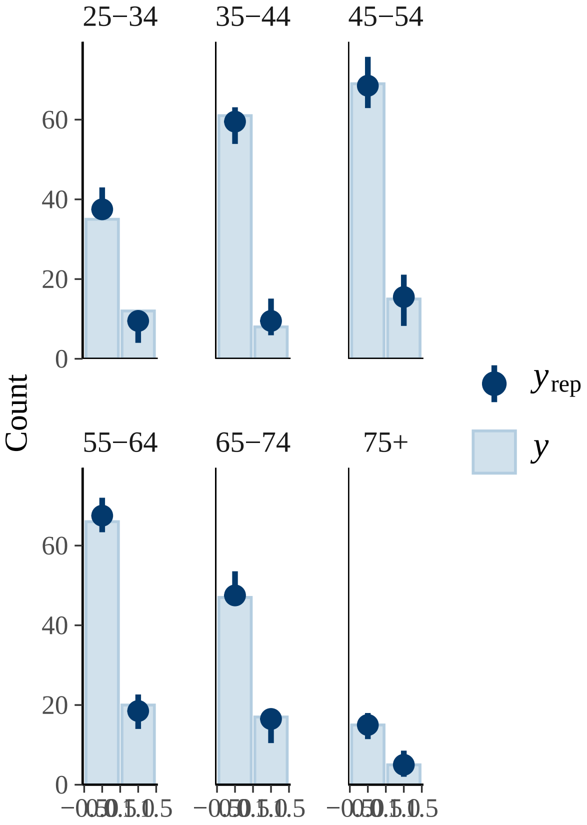


The next two plots displays two bars representing the observed counts of the two outcomes, with overlaid bars from the simulated data.

```
## Using 10 posterior draws for ppc type 'bars' by default.
```

```
## Using 10 posterior draws for ppc type 'bars_grouped' by default.
```



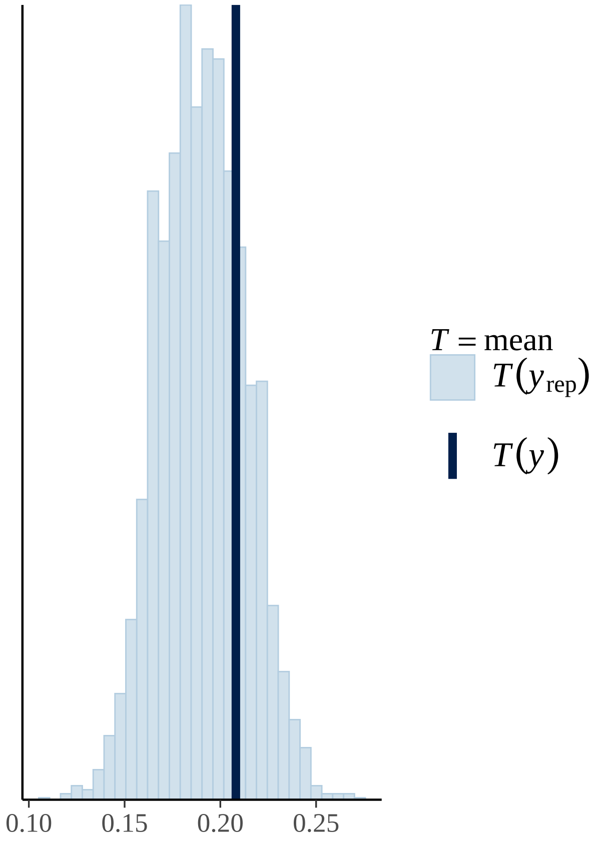


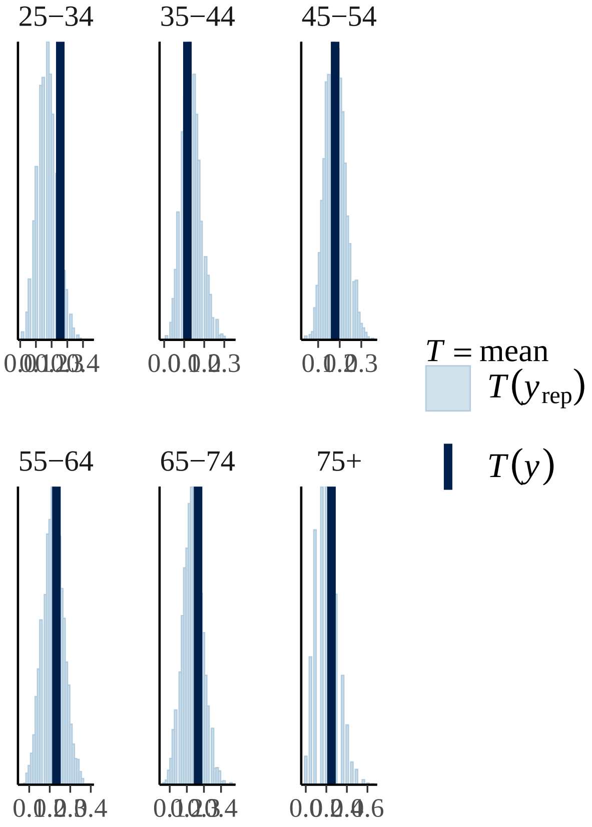
The final two plots show the observed mean statistic and a distribution of the statistic computed from the simulated data set.

```
## Using all posterior draws for ppc type 'stat' by default.
```

```
## Using all posterior draws for ppc type 'stat_grouped' by default.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





**Predictive performance assessments**

Now we will make predictions on the Test Set and then construct a confusion matrix by comparing predictions to the actual outcomes in the test set.

**The Confusion Matrix:**

```
##           Actual
## Predicted   0   1
##         0 287  71
##         1   6   6
```

```
## Accuracy:    0.8
## Precision:   0.5
## Recall:      0.1
## F1 Score:    0.1
```

# Hierarchical logistic regression model

In the hierarchical model, we will assume that each patient will inherently have a distinct tendency to have cancer, accounting for patient-level variation. For example, depending on the patient's genes, the patient is either vulnerable or more resistant to cancer.

### Priors and Likelihood:

1. **Priors**: We can reuse the priors in the Pooled Model as well as priors suggested by BRMS, then test the Model's sensitivity to priors.

```
##                   prior     class      coef group resp dpar nlpar lb ub
##                  (flat)         b
##                  (flat)         b     age.C
##                  (flat)         b     age.L
##                  (flat)         b     age.Q
##                  (flat)         b     ageE4
##                  (flat)         b     ageE5
##                  (flat)         b     alc.C
##                  (flat)         b     alc.L
##                  (flat)         b     alc.Q
##                  (flat)         b     tob.C
##                  (flat)         b     tob.L
##                  (flat)         b     tob.Q
##  student_t(3, 0, 2.5) Intercept
##  student_t(3, 0, 2.5)        sd                                  0
##  student_t(3, 0, 2.5)        sd              ID                  0
##  student_t(3, 0, 2.5)        sd Intercept    ID                  0
##        source
##       default
##  (vectorized)
##  (vectorized)
##  (vectorized)
##  (vectorized)
##  (vectorized)
##  (vectorized)
##  (vectorized)
##  (vectorized)
##  (vectorized)
##  (vectorized)
##  (vectorized)
##       default
##       default
##  (vectorized)
##  (vectorized)
```

2. **Likelihood**: We will reuse the Bernoulli family as in the Pooled Model.

### BRMS code

```
# Defining the ordinal logistic model
model2 <- brm(cancer ~ 1 + tob + alc + age + (1 | ID),
             data = train_data,
             family = bernoulli("logit"),
             prior = c(prior(normal(0, 10), class = "b"),
                       prior(normal(0, 10), class = "Intercept")),
             chains=4,
             file="model2",
             cores=4)
```

The model uses 4 chains, each with 2000 iterations. The warm-up length is 1000. This are mostly default settings but, as we show below, the MCMC chains converge well.

### Convergence diagnostic

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: cancer ~ 1 + tob + alc + age + (1 | ID)
##    Data: train_data (Number of observations: 863)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~ID (Number of levels: 863)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)    20.95      5.31    11.79    32.07 1.00     1273     2094
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept   -11.99      3.35   -19.38    -6.38 1.00     1373     2151
## tob.L        15.69      4.24     8.50    24.88 1.00     1483     2094
## tob.Q        -2.09      2.31    -6.98     2.29 1.00     1818     2117
## tob.C        -0.52      2.13    -4.77     3.65 1.00     1684     2079
## alc.L        11.25      3.32     5.64    18.33 1.00     1569     1930
## alc.Q         2.10      2.33    -2.50     7.00 1.00     2114     2342
## alc.C         1.51      2.40    -2.95     6.65 1.00     1848     2338
## age.L         7.39      3.84     0.57    15.67 1.00     1779     2359
## age.Q         1.99      3.33    -4.32     8.72 1.00     1842     2396
## age.C        -4.17      3.09   -10.80     1.58 1.00     2052     2598
## ageE4        -0.11      2.67    -5.52     5.18 1.00     1876     2267
## ageE5        -1.55      2.31    -6.35     2.89 1.00     1915     2433
##
## Draws were sampled using sample(hmc). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

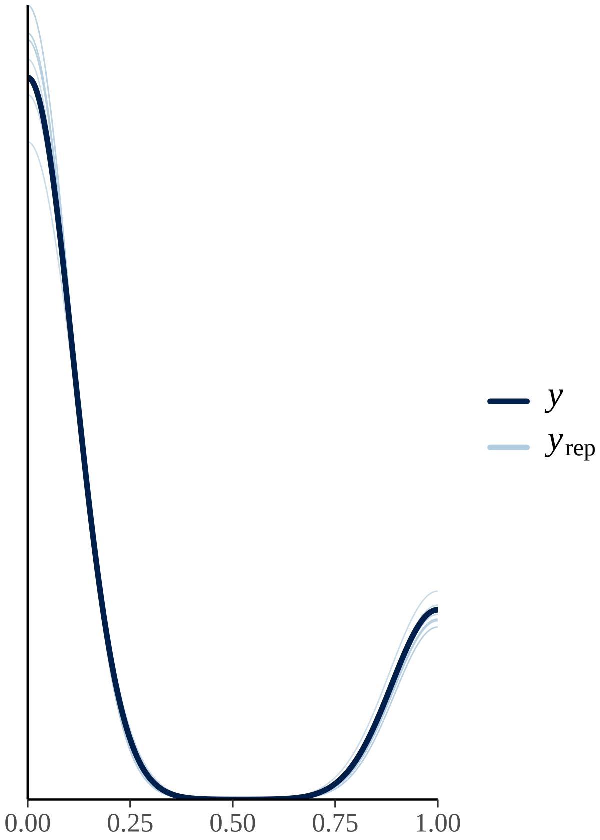
From the summary we can learn that our model estimation converged well: $\hat{R}$ ~ 1.00 and effective sample size greater than 100 times the number of chains (=4) for each parameter (Vehtari et al. 2021) [1].

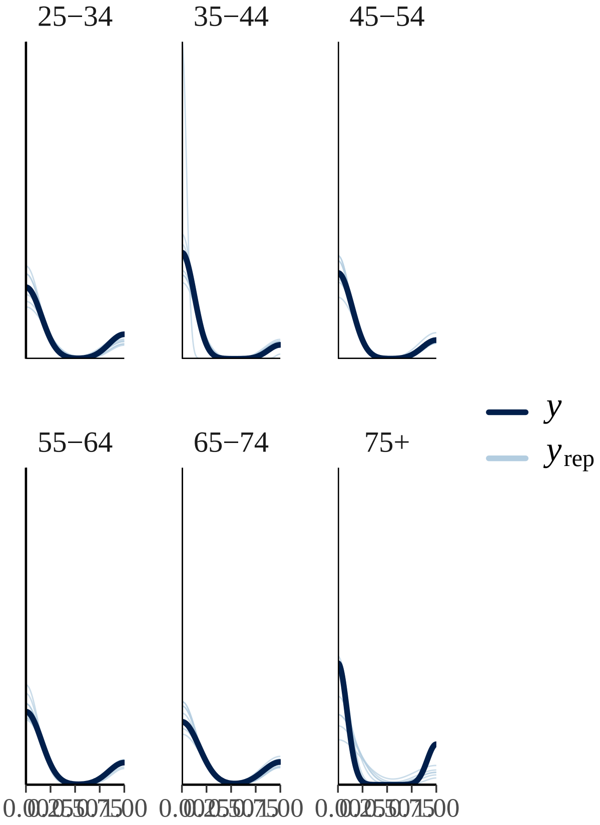
### Posterior predictive checks

These first two plots are used for visualizing the overlay of posterior predictive densities on top of the observed data.

```
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
```

```
## Using 10 posterior draws for ppc type 'dens_overlay_grouped' by default.
```

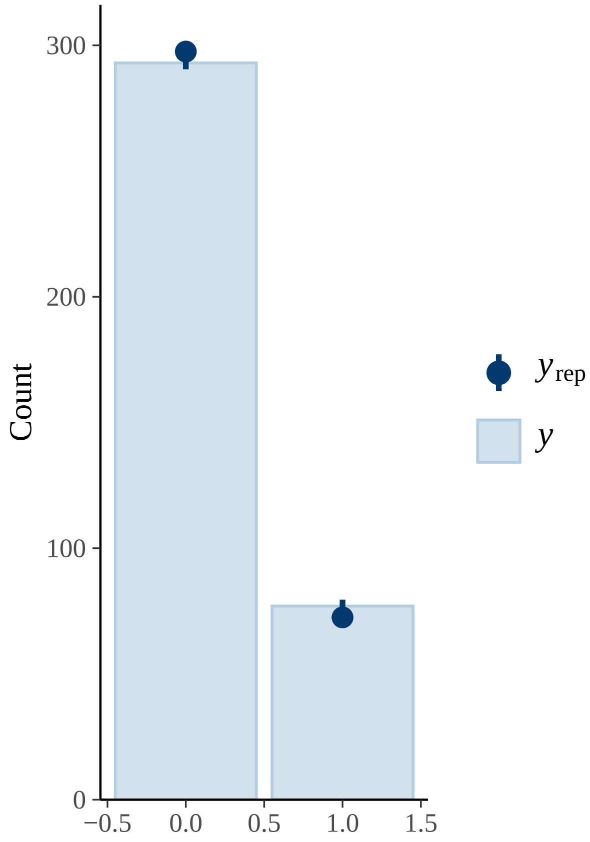


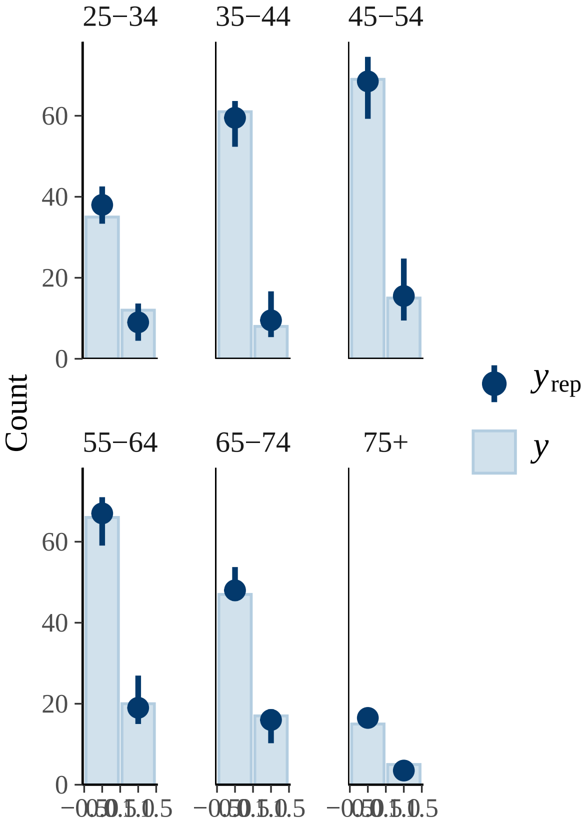


The next two plots displays two bars representing the observed counts of the two outcomes, with overlaid bars from the simulated data.

```
## Using 10 posterior draws for ppc type 'bars' by default.
```

```
## Using 10 posterior draws for ppc type 'bars_grouped' by default.
```



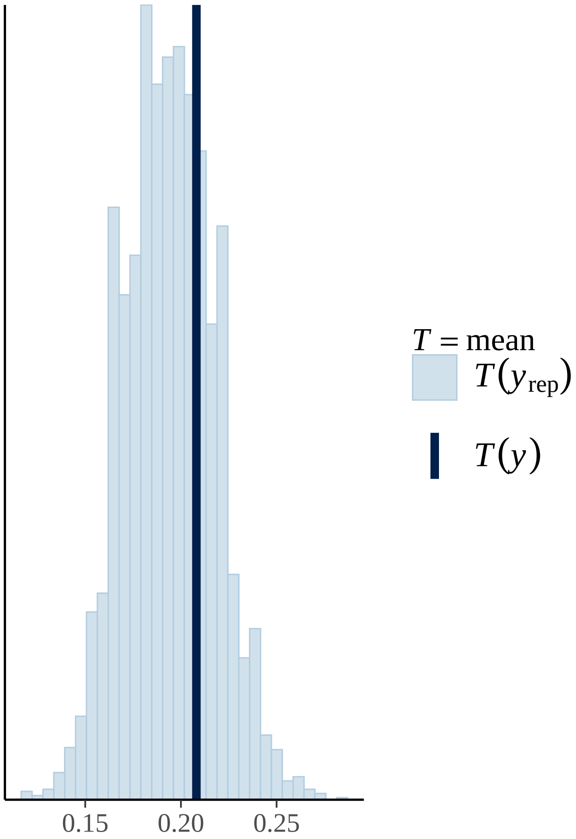


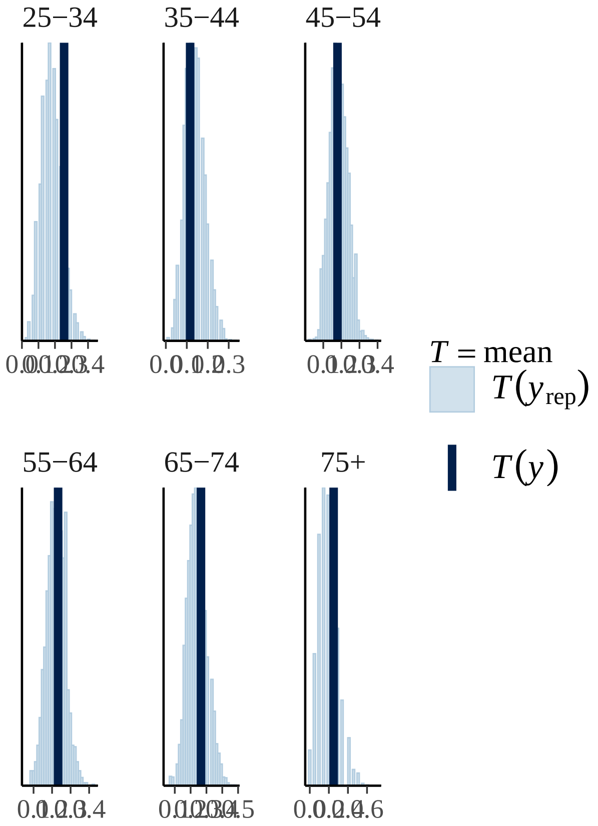
The final two plots show the observed mean statistic and a distribution of the statistic computed from the simulated data set.

```
## Using all posterior draws for ppc type 'stat' by default.
```

```
## Using all posterior draws for ppc type 'stat_grouped' by default.
```

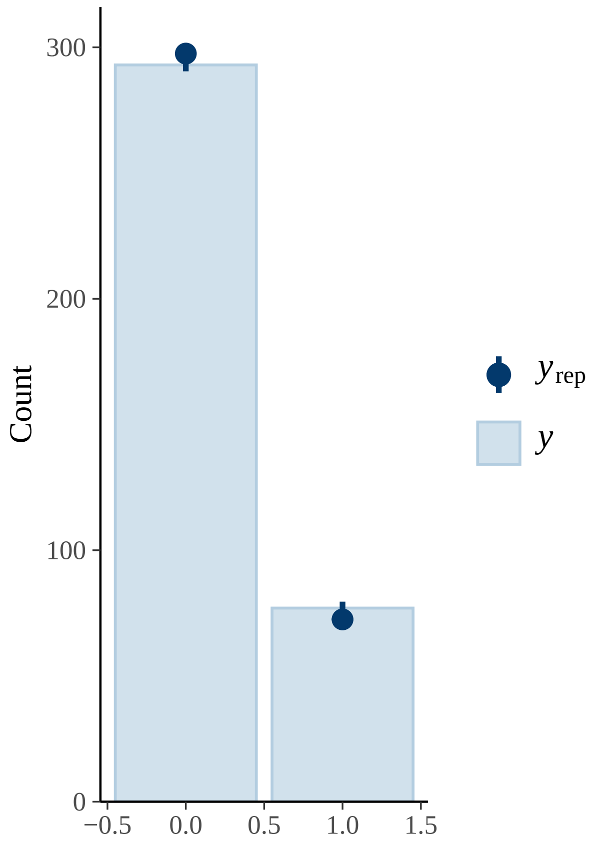
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

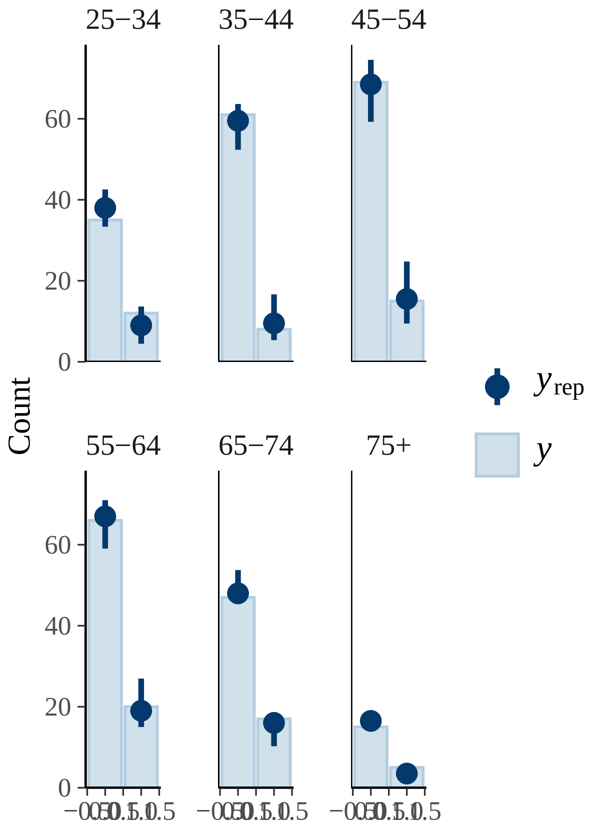


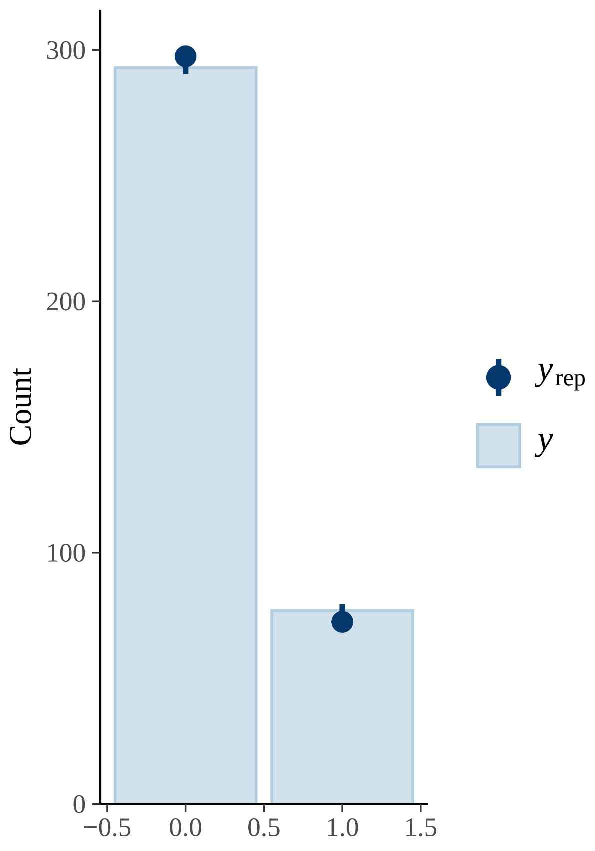


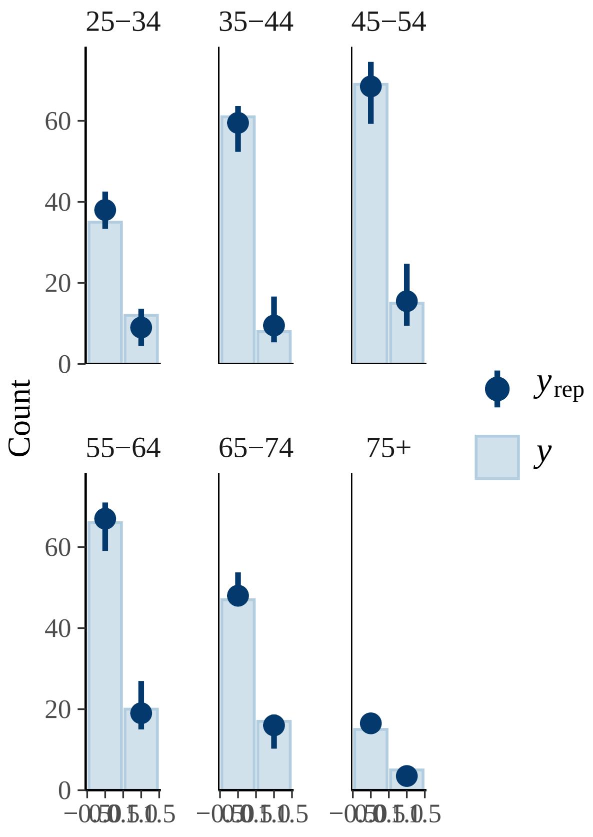
**Predictive performance assessments**

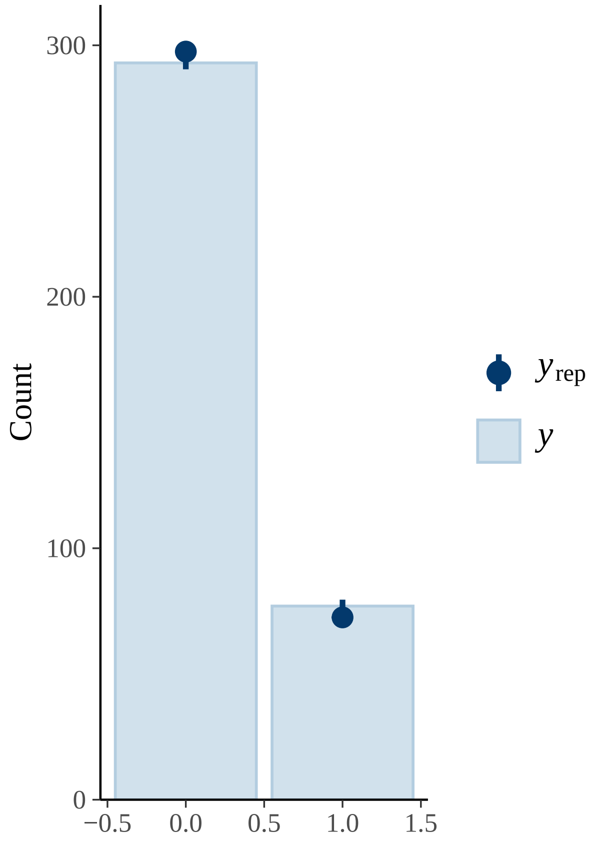
Now we will make predictions on the Test Set and then construct a confusion matrix by comparing predictions to the actual outcomes in the test set.

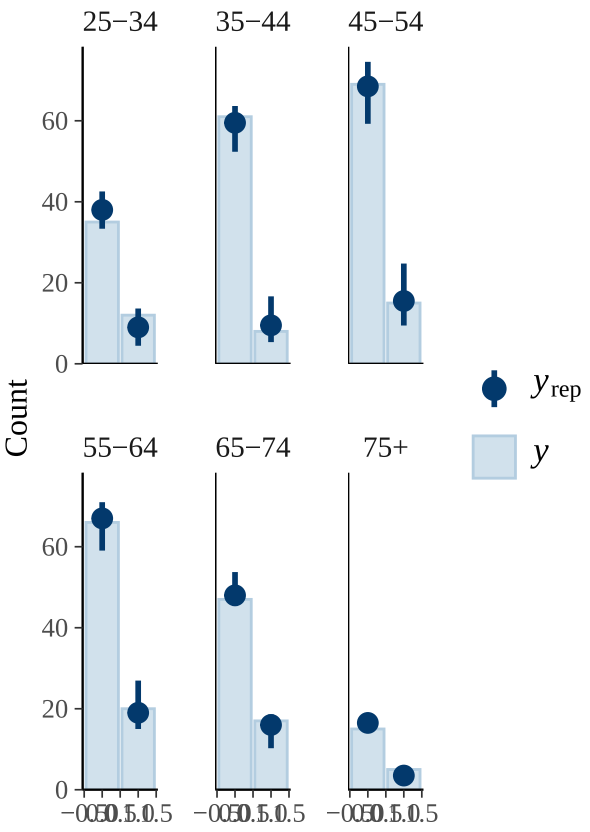

Count
300
200
100
0
−0.5
0.0
0.5
1.0
1.5
$y_{\text{rep}}$
$y$


25−34
35−44
45−54
55−64
65−74
75+
Count
60
40
20
0
−0.5 0.0 0.5 1.0 1.5
$y_{\text{rep}}$
$y$


Count
300
200
100
0
−0.5
0.0
0.5
1.0
1.5
$y_{\mathrm{rep}}$
$y$


25−34
35−44
45−54
55−64
65−74
75+
Count
60
40
20
0
−0.5 0.0 0.5 1.0 1.5
$y_{\mathrm{rep}}$
$y$





**The Confusion Matrix:**

```
##           Actual
## Predicted   0   1
##         0 288  72
##         1   5   5
```

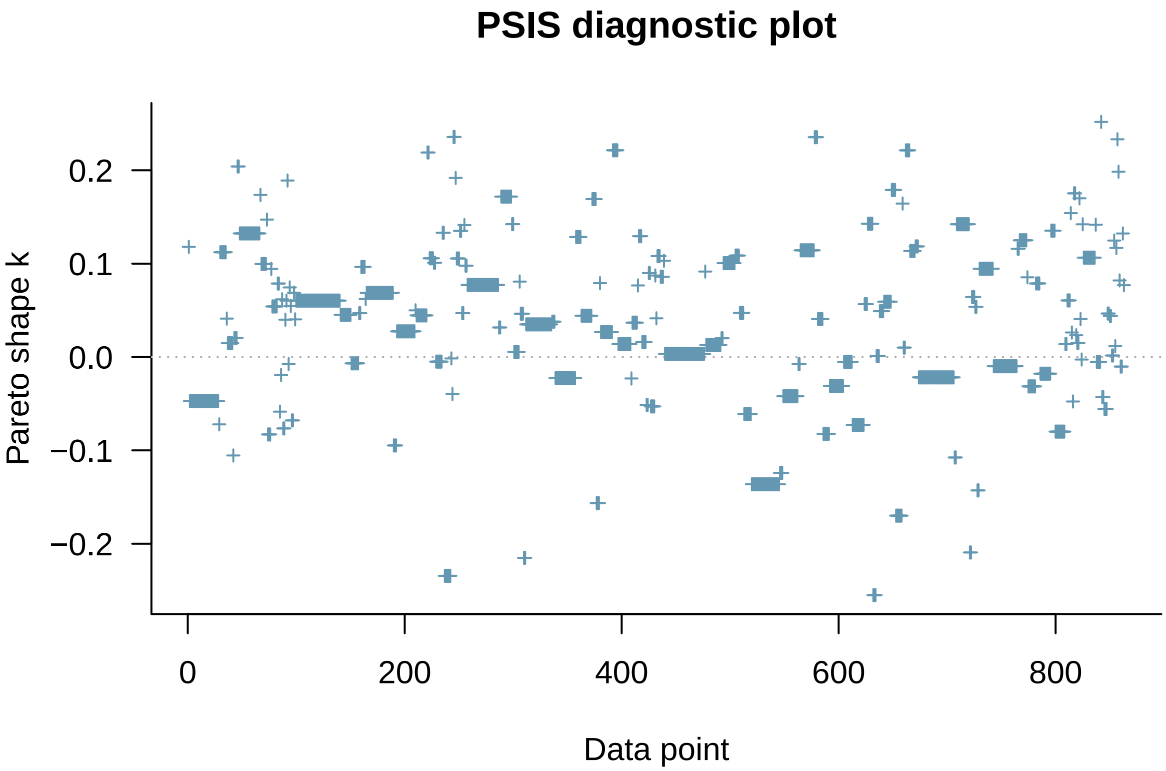
```
## Accuracy:    0.8
## Precision:   0.5
## Recall:      0.1
## F1 Score:    0.1
```

**Accuracy**: Accuracy measures how often a classifier makes the correct prediction. It is the ratio of the number of correct predictions to the total number of predictions. The model's accuracy of 0.8 is decent.

## Model comparisons using ELPD

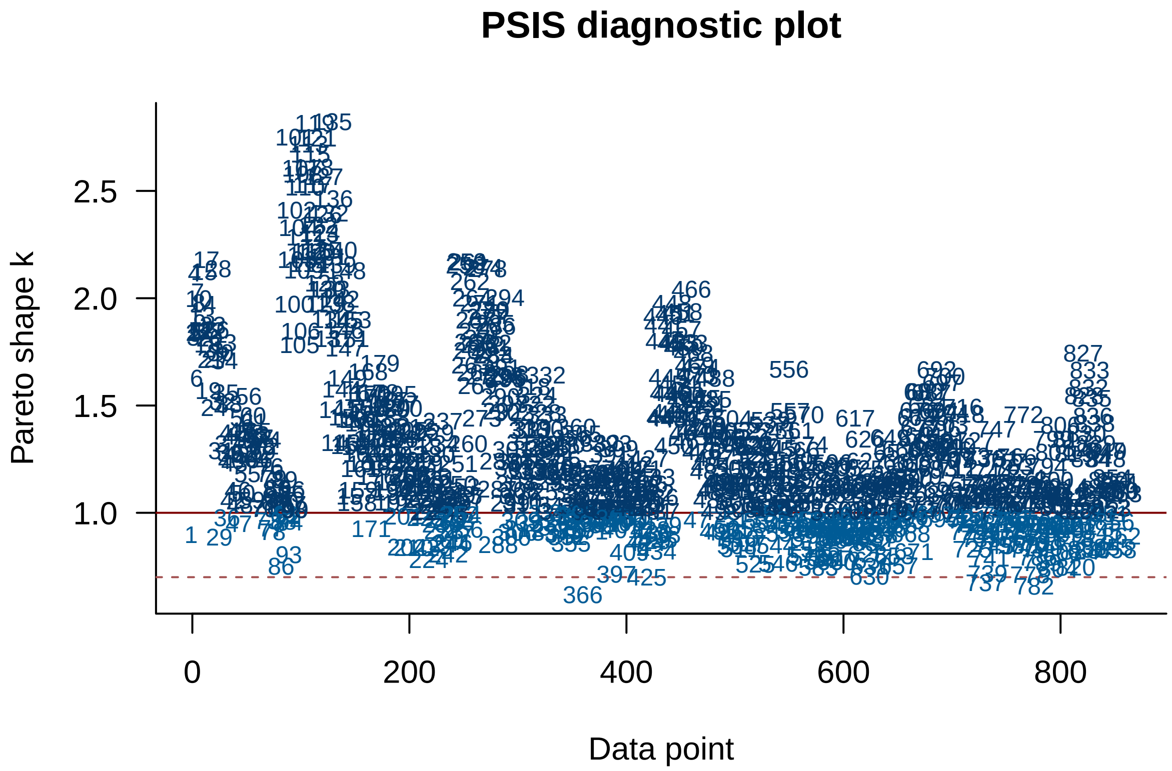
```
loo1 <- loo::loo(model)
loo2 <- loo::loo(model2)
```

```
plot(loo1, label_points=TRUE)
```



All Pareto $\hat{k}$ are smaller than 0.5. The distribution of raw importance rations has finite variance.

```
plot(loo2, label_points=TRUE)
```



We can see that almost all data points have Pareto $\hat{k}$ bigger than 0.7. It is due to our model having only one ID observation for each patient meaning that LOO cross validation does not provide a big insight.

```
loo::loo_compare(loo1, loo2)
```

```
##        elpd_diff se_diff
## model2    0.0       0.0
## model  -267.5      12.5
```

The loo comparison indicates that the hierarchical model is better than the pooled logistic regression. However, looking at posterior predictive checks, we can clearly see that is not the case. Leave-one-out is faulty in this case, as the hierarchical model has 1 observation per ID. To solve the problem with loo of the hierarchical model we would need to redo the ELPD 800-900 times or acquire more data about each patient.

Therefore, based on the posterior predictive checks and classification accuracy, pooled logistic regression is a better, more suitable model.

## Possible improvements

- Both of the models have quite few false negatives when we try to predict the cancer on test data. This is not desired in cancer diagnosis. Therefore, the model should be improved to try and limit the number of false negatives. One possible solution to that is gathering more data, as our data set is quite small and the rate of cancer also is not very high.

- Gather new data, our data set is quite limited right now and if we could gather more information about each patient it would help our analysis a lot. For example getting the tobacco, alcohol consumption and age of each patient would be helpful. This would allow as to introduce more complicated models that could fit and explain the data better.

## Conclusion

- Probability of getting esophageal cancer in each age group. How significant is the impact of the alcohol and tobacco as risk factors in being diagnosed with the cancer.
- The tobacco and alcohol are risk factors and not direct causes of the cancer. This with the low probability of getting the cancer, makes it difficult to accurately predict that somebody will develop the disease.

## Reflections

Throught the project we managed to deepen our understanding of Bayesian data analysis, as well as improve in many technical aspects of the analysis. Our skills in Stan, R or many different libraries that are associated with them excelled by a lot. However, everything did not go smoothly. We definitely learned how to work under pressure as we needed to change our models quite a bit in the final week of the project work due to some wrong assumptions. Moreover, our pick for data set was also quite flawed and did not allow us to express all of the ideas we had for the project.

## References

[1]. Vehtari, Aki, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. 2021. "Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of Mcmc." Bayesian Analysis. https://doi.org/10.1214/20-BA1221.

[2]. Esophagus cancer: Esophageal cancer (no date) Esophageal Cancer | American Cancer Society. Available at: https://www.cancer.org/cancer/types/esophagus-cancer.html (Accessed: 30 November 2023).

## Appendix