Module 1: Introduction

CAD data overview

This data mart is a random sample of patients in the Mass General Brigham (formerly Partner's Healthcare) EHR database who had at least one note of 500 characters and met an initial filter for **coronary artery disease** (CAD) defined as:

- $\geq$ 1 ICD9 code related to CAD (410.x, 411.x, 412.x, 413.x, 414.x).
- $\geq$ 1 mention for any CAD related concepts (eg. CAD, CAD procedures, CAD biomarkers, positive stress test).

Read in the CAD data

"label": whether the patient has CAD, **extracted from chart review by a clinician**.

```
data(ehr_data)
data <- PhecapData(ehr_data, "healthcare_utilization", "label", 75, patient_id = "patient_id", seed = 123)
data
```

```
## PheCAP Data
## Feature: 10000 observations of 587 variables
## Label: 119 yes, 62 no, 9819 missing
## Size of training samples: 106
## Size of validation samples: 75
```

Read in the CAD data

"main_ICD", "main_NLP": total number of billing codes or NLP mentions of the disease.

```
head(ehr_data[, c(1:5, 25:30)])
```

```
##   patient_id label main_ICD main_NLP healthcare_utilization NLP10 NLP11 NLP12
## 1          1    NA        1        0                      25     5     0     0
## 2          2    NA       41      157                     187    72     0     2
## 3          3    NA        4        1                     138    44     1     0
## 4          4    NA        0        0                      67     9     0     0
## 5          5    NA        0        0                       9     4     0     0
## 6          6    NA        4        2                       4     0     0     0
##   NLP13 NLP14 NLP15
## 1     0     0     0
## 2    22     0    47
## 3     8    13    12
## 4     1     0     2
## 5     0     0     0
## 6     0     0     0
```

Read in the CAD data

"healthcare_utilization": total number of notes the patient has.

```
head(ehr_data[, c(1:5, 25:30)])
```

```
##   patient_id label main_ICD main_NLP healthcare_utilization NLP10 NLP11 NLP12
## 1          1    NA        1        0                      25     5     0     0
## 2          2    NA       41      157                     187    72     0     2
## 3          3    NA        4        1                     138    44     1     0
## 4          4    NA        0        0                      67     9     0     0
## 5          5    NA        0        0                       9     4     0     0
## 6          6    NA        4        2                       4     0     0     0
##   NLP13 NLP14 NLP15
## 1     0     0     0
## 2    22     0    47
## 3     8    13    12
## 4     1     0     2
## 5     0     0     0
## 6     0     0     0
```

Read in the CAD data

"CODx" (n = 10): the counts of a specific code. "NLPx" (n = 574): the counts of a NLP term.

```
head(ehr_data[, c(1:5, 25:30)])
```

```
##   patient_id label main_ICD main_NLP healthcare_utilization NLP10 NLP11 NLP12
## 1          1    NA        1        0                      25     5     0     0
## 2          2    NA       41      157                     187    72     0     2
## 3          3    NA        4        1                     138    44     1     0
## 4          4    NA        0        0                      67     9     0     0
## 5          5    NA        0        0                       9     4     0     0
## 6          6    NA        4        2                       4     0     0     0
##   NLP13 NLP14 NLP15
## 1     0     0     0
## 2    22     0    47
## 3     8    13    12
## 4     1     0     2
## 5     0     0     0
## 6     0     0     0
```

Basic descriptives

```
# Check for missing data.
colnames(ehr_data)[which(colMeans(is.na(ehr_data)) > 0)]
```

```
## [1] "label"
```

Basic descriptives

```
# Check for missing data.
colnames(ehr_data)[which(colMeans(is.na(ehr_data)) > 0)]
```

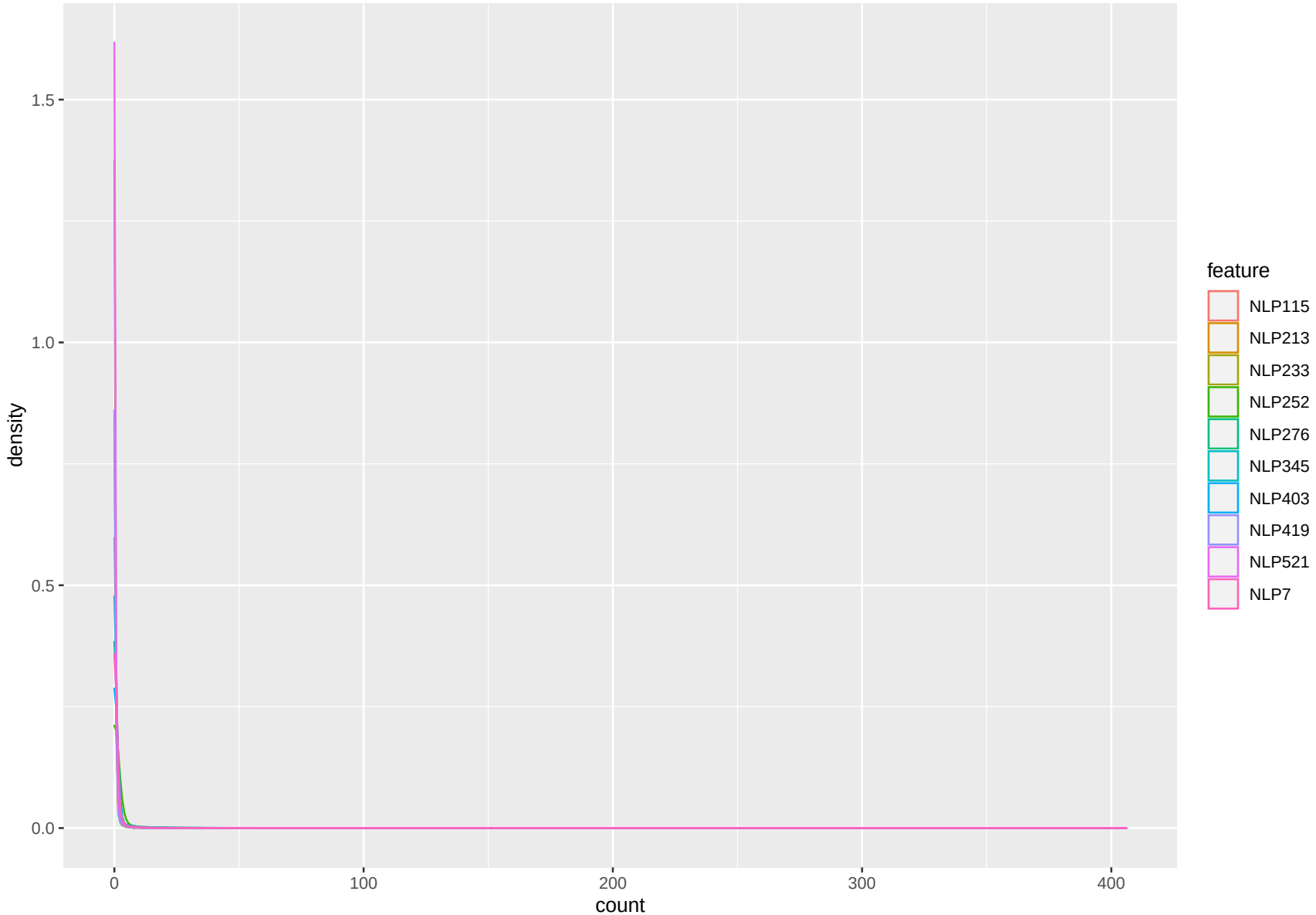
```
## [1] "label"
```

```
# Prevalence of the label.
mean(ehr_data$label, na.rm = TRUE)
```

```
## [1] 0.6574586
```

Feature distributions

The features are highly skewed.



Prepare the data for model fitting

- We log transform all the features as they are highly skewed
- We orthogonalize all features against health care utilization before fitting as patients with higher healthcare utilization have higher feature counts

(Please find more details in the `.Rmd` file.)