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)

CAD data features

```r
data(ehr_data)
data <- PhecapData(ehr_data, "healthcare_utilization",
                   "label", 75,
                   "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
```

- Label: whether the patient has CAD, **extracted from chart review by a clinician**

Read in the CAD data

- Features:
  - "main_ICD", "main_NLP" refers to total number of billing codes or NLP mentions of the disease
  - "healthcare_utilization" refers to total number of notes the patient has
  - "CODx" ($n = 10$) refers to the counts of a specific code
  - "NLPx" ($n = 574$) refers to the counts of a NLP term

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

```
## 'data.frame':    10000 obs. of  11 variables:
##  $ patient_id            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ label                 : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ main_ICD              : int  1 41 4 0 0 4 1 1 1 0 ...
##  $ main_NLP              : int  0 157 1 0 0 2 0 3 0 0 ...
##  $ healthcare_utilization: int  25 187 138 67 9 4 477 182 7 71 ...
##  $ NLP10                 : int  5 72 44 9 4 0 2172 211 0 87 ...
##  $ NLP11                 : int  0 0 1 0 0 0 7 0 0 0 ...
##  $ NLP12                 : int  0 2 0 0 0 0 2 1 0 0 ...
##  $ NLP13                 : int  0 22 8 1 0 0 20 10 1 2 ...
##  $ NLP14                 : int  0 0 13 0 0 0 30 12 3 0 ...
##  $ NLP15                 : int  0 47 12 2 0 0 178 4 1 0 ...
```

Basic descriptives

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

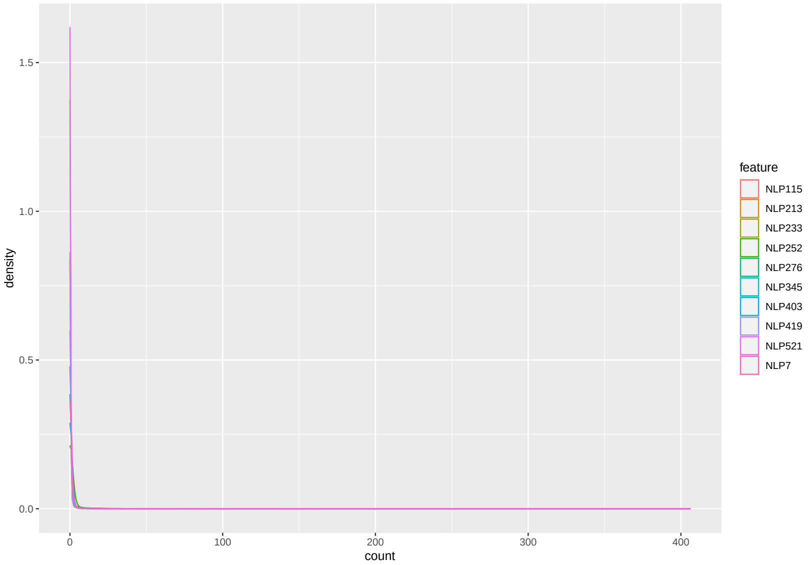
```
## [1] "label"
```

```r
# 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 transform all features with $x \rightarrow log(x + 1)$ 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