# Selecting Clinical Populations

*Cristina Goldfain*

What criteria can be used to identify patients with hypertension? I will test the ability of different algorithms to pull out patients with hypertension from a large database of records. To do this, I am comparing which patients were selected by the algorithm with the patients manually diagnosed by experts examining the electronic health records. An ideal algorithm selects all the patients that the manual reviewers identified as having hypertension.

Hypertension is clinically defined as having Systolic blood pressure greater than 140 mmHg on more than two occasions, or Diastolic blood pressure higher than 90 mmHg on more than two occasions. It is reasonable to think that using these two rules will allow us to pull out patients with hypertension, but as seen below, that is not the case. One explanation is that patients can have elevated blood pressure in many conditions, not only because they are diagnosed with hypertension.

How good is an algorithm that uses lab measurements with the clinical threshold Systolic BP >= 140 mmHg or Diastolic BP>= 90 mmHg as criteria to classify hypertension patients?

```
## Confusion Matrix and Statistics
##
##                  HYPERTENSION
## systolic_2events  0  1
##                0 18 18
##                1 18 45
##
##                Accuracy : 0.6364
##                  95% CI : (0.5336, 0.7307)
##     No Information Rate : 0.6364
##     P-Value [Acc > NIR] : 0.5453
##
##                   Kappa : 0.2143
##  Mcnemar's Test P-Value : 1.0000
##
##             Sensitivity : 0.7143
##             Specificity : 0.5000
##          Pos Pred Value : 0.7143
##          Neg Pred Value : 0.5000
##              Prevalence : 0.6364
##          Detection Rate : 0.4545
##    Detection Prevalence : 0.6364
##       Balanced Accuracy : 0.6071
##
##        'Positive' Class : 1
##
```

```
## Confusion Matrix and Statistics
##
##                   HYPERTENSION
## diastolic_2events  0  1
##                 0 29 57
##                 1  7  6
##
##                Accuracy : 0.3535
##                  95% CI : (0.2601, 0.456)
```

```
##     No Information Rate : 0.6364
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : -0.0765
##  Mcnemar's Test P-Value : 9.068e-10
##
##             Sensitivity : 0.09524
##             Specificity : 0.80556
##          Pos Pred Value : 0.46154
##          Neg Pred Value : 0.33721
##              Prevalence : 0.63636
##          Detection Rate : 0.06061
##    Detection Prevalence : 0.13131
##       Balanced Accuracy : 0.45040
##
##        'Positive' Class : 1
##
```

How good is an algorithm that uses the ICD9 billing codes for essential hypertension as a criterion to classify patients with hypertension?

```
##                            HYPERTENSION
## has_hypertension_ICD9codes  0  1
##                           0 33 28
##                           1  3 35
```

```
## Confusion Matrix and Statistics
##
##                            HYPERTENSION
## has_hypertension_ICD9codes  0  1
##                           0 33 28
##                           1  3 35
##
##                Accuracy : 0.6869
##                  95% CI : (0.5859, 0.7764)
##     No Information Rate : 0.6364
##     P-Value [Acc > NIR] : 0.1739
##
##                   Kappa : 0.4111
##  Mcnemar's Test P-Value : 1.629e-05
##
##             Sensitivity : 0.5556
##             Specificity : 0.9167
##          Pos Pred Value : 0.9211
##          Neg Pred Value : 0.5410
##              Prevalence : 0.6364
##          Detection Rate : 0.3535
##    Detection Prevalence : 0.3838
##       Balanced Accuracy : 0.7361
##
##        'Positive' Class : 1
##
```

How good is an algorithm that uses the ICD9 4019 billing code for essential hypertension? ICD9 code 4019 is the most often used code in this database.

```
## Confusion Matrix and Statistics
##
##                               HYPERTENSION
## has_hypertension_ICD9code401.9  0  1
##                               0 33 30
##                               1  3 33
##
##                Accuracy : 0.6667
##                  95% CI : (0.5648, 0.7582)
##     No Information Rate : 0.6364
##     P-Value [Acc > NIR] : 0.3033
##
##                   Kappa : 0.3795
##  Mcnemar's Test P-Value : 6.011e-06
##
##             Sensitivity : 0.5238
##             Specificity : 0.9167
##          Pos Pred Value : 0.9167
##          Neg Pred Value : 0.5238
##              Prevalence : 0.6364
##          Detection Rate : 0.3333
##    Detection Prevalence : 0.3636
##       Balanced Accuracy : 0.7202
##
##        'Positive' Class : 1
##
```

How good is an algorithm that classifies patients with hypertension because they were prescribed medications commonly used to treat hypertension? As you can see below, many of the patients with hypertension were prescribed these drugs, however the problem is that many other patients were prescribed anti hypertensives for other reasons.

```
## Confusion Matrix and Statistics
##
##                        HYPERTENSION
## has_hypertension_drugs  0  1
##                       0 12 11
##                       1 24 52
##
##                Accuracy : 0.6465
##                  95% CI : (0.544, 0.7399)
##     No Information Rate : 0.6364
##     P-Value [Acc > NIR] : 0.46219
##
##                   Kappa : 0.172
##  Mcnemar's Test P-Value : 0.04252
##
##             Sensitivity : 0.8254
##             Specificity : 0.3333
##          Pos Pred Value : 0.6842
##          Neg Pred Value : 0.5217
##              Prevalence : 0.6364
##          Detection Rate : 0.5253
##    Detection Prevalence : 0.7677
##       Balanced Accuracy : 0.5794
```

```
##
##        'Positive' Class : 1
##
```

How good is an algorithm that classifies patients with hypertension because they were prescribed medications commonly used to treat hypertension ten or more times? Here I'm trying to improve the Specificity of the algorithm by screening out patients that might have been prescribed these drugs for other conditions a few times.

```
## Confusion Matrix and Statistics
##
##                       HYPERTENSION
## has_hypertension_drugs  0  1
##                      0 30 45
##                      1  6 18
##
##                Accuracy : 0.4848
##                  95% CI : (0.3832, 0.5875)
##     No Information Rate : 0.6364
##     P-Value [Acc > NIR] : 0.9993
##
##                   Kappa : 0.0966
##  Mcnemar's Test P-Value : 1.032e-07
##
##             Sensitivity : 0.2857
##             Specificity : 0.8333
##          Pos Pred Value : 0.7500
##          Neg Pred Value : 0.4000
##              Prevalence : 0.6364
##          Detection Rate : 0.1818
##    Detection Prevalence : 0.2424
##       Balanced Accuracy : 0.5595
##
##        'Positive' Class : 1
##
```

How good is an algorithm that classifies patients with hypertension based on medication records showing they were prescribed an anti-hypertensive and they had an ICD9 billing code for hypertension?

```
## Confusion Matrix and Statistics
##
##                   HYPERTENSION
## has_drug_and_codes  0  1
##                  0 33 31
##                  1  3 32
##
##                Accuracy : 0.6566
##                  95% CI : (0.5544, 0.7491)
##     No Information Rate : 0.6364
##     P-Value [Acc > NIR] : 0.3804
##
##                   Kappa : 0.3639
##  Mcnemar's Test P-Value : 3.649e-06
##
##             Sensitivity : 0.5079
##             Specificity : 0.9167
```

```
##          Pos Pred Value : 0.9143
##          Neg Pred Value : 0.5156
##              Prevalence : 0.6364
##          Detection Rate : 0.3232
##    Detection Prevalence : 0.3535
##       Balanced Accuracy : 0.7123
##
##        'Positive' Class : 1
##
```

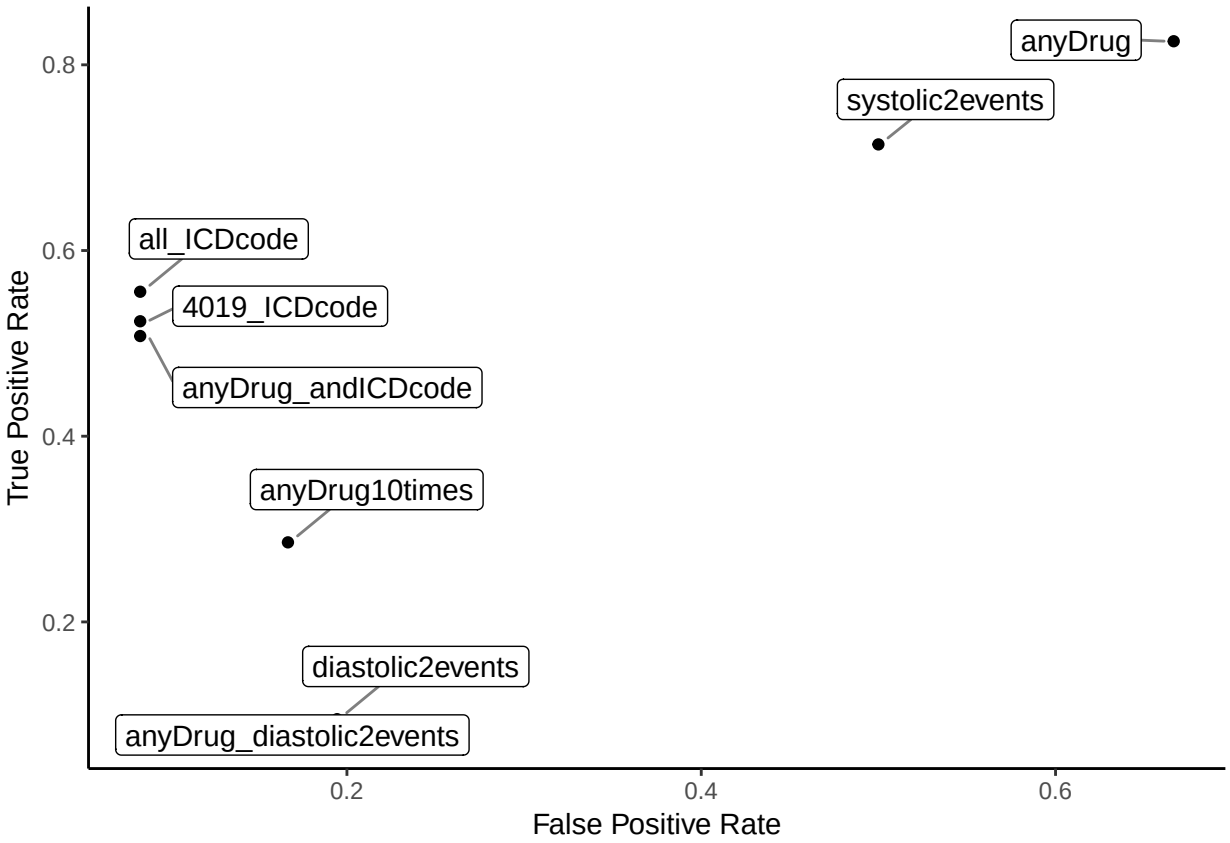
How good is an algorithm that classifies patients with hypertension because they were prescribed an anti-hypertensive and had two diastolic measurements above 90mmHg ?

```
## Confusion Matrix and Statistics
##
##                        HYPERTENSION
## has_drug_and_diastolic  0  1
##                      0 31 58
##                      1  5  5
##
##                Accuracy : 0.3636
##                  95% CI : (0.2693, 0.4664)
##     No Information Rate : 0.6364
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : -0.0452
##  Mcnemar's Test P-Value : 5.701e-11
##
##             Sensitivity : 0.07937
##             Specificity : 0.86111
##          Pos Pred Value : 0.50000
##          Neg Pred Value : 0.34831
##              Prevalence : 0.63636
##          Detection Rate : 0.05051
##    Detection Prevalence : 0.10101
##       Balanced Accuracy : 0.47024
##
##        'Positive' Class : 1
##
```

Summary

As expected and seen in the graph below, there is a trade of between the ability of the algorithms to identify all patients with hypertension (True Positive Rate) and to classify other patients mistakenly as having hypertension. Using all ICD9 codes associated with hypertension as a selection criteria seems to be a good start, though much improvement is needed.

Future steps are testing these criteria on a new test population to make sure we did not over fit to our current data and improving our models. Random forest approaches are particularly good at classification, however, they do require more labeled data (manually sorted records) to be trained.


anyDrug
systolic2events
all_ICDcode
4019_ICDcode
anyDrug_andICDcode
anyDrug10times
diastolic2events
anyDrug_diastolic2events
0.8
0.6
0.4
0.2
True Positive Rate
0.2
0.4
0.6
False Positive Rate