Peer-graded Assignment

Developing Computational Phenotyping Algorithm

This is a mandatory assignment to complete the Module 5,
Course 3 'Identifying Patient Populations',

Taught by

Laura K. Wiley, PhD, Associate Professor
Department of Biomedical Informatics
University of Colorado System

Palchamy Elango

Data Types – Identifying patients with hypertension

1. **Gold standard hypertension dataset**, (99 records)

   cases 63, controls 36

2. **ICD-9 Diagnosis Codes**, excluded ICD-9 Medication. Queried the below IDs against DIAGNOSES_ICD dataset

   401.0 (malignant),
   401.1 (benign), or
   401.9 (unspecified).

   Found 38 subjects diagnosed with hypertension,

3. **Searched the laboratory Data** (LABEVENTS) for the following itemids. Observed missing data for LDL.

row_id	ITEMID	LABEL	FLUID	CATEGORY	
105	50904	Cholesterol, HDL	Blood	Chemistry	2085-9
107	50906	Cholesterol, LDL, Measured	Blood	Chemistry	18262-6
108	50907	Cholesterol, Total	Blood	Chemistry	2093-3

Data Types – Manipulations of individual data types

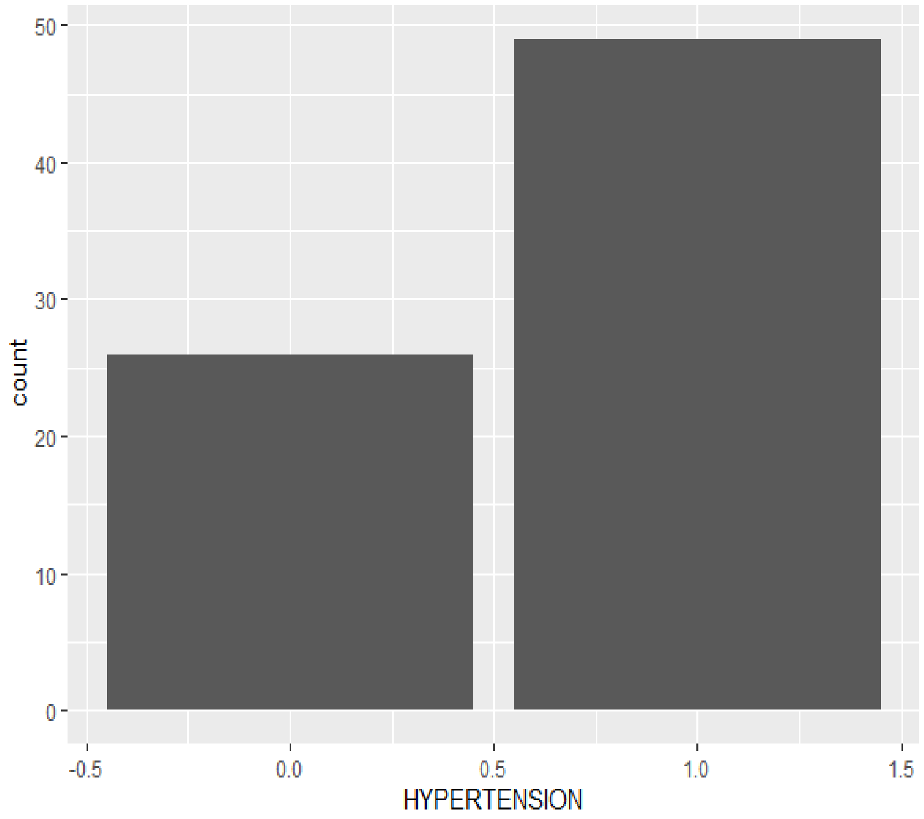
Found 38 subjects diagnosed with hypertension in the DIAGNOSIS_ICD dataset.

Appended these records with the origin gold-standard hypertension data.

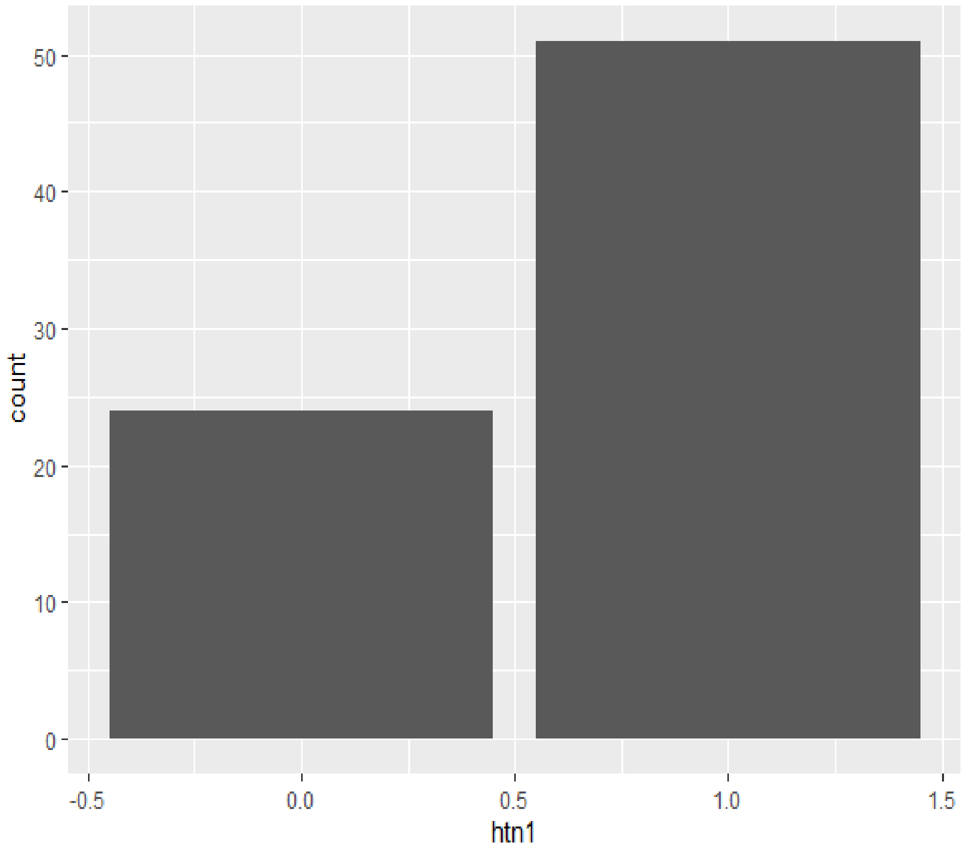
Gold-standard hypertension (manual review)	Hypertension	Freq
controls	0	36
cases	1	64

Diagnosis_ICD	Hypertension	Freq
controls	0	61
cases	1	38

.



Manual Review (original) Hypertension



Original + ICD code + Lipid measures

Manual review data combined with ICD code

Confusion Matrix and Statistics

		Manual Review of Hypertension	
		+	-
ICD Codes 401.0 401.1 401.9	+	35	3
	-	28	33

```
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
```

Data Types – Adding additional data types

Previous data (manual review + icd code) has been updated with Laboratory data.

Searched laboratory Data (LABEVENTS) for the following ITEMIDs

row_id	**ITEMID**	LABEL	FLUID	CATEGORY	loinc_code
105	50904	Cholesterol, HDL	Blood	Chemistry	2085-9
107	50906	Cholesterol, LDL, Measured	Blood	Chemistry	18262-6
108	50907	Cholesterol, Total	Blood	Chemistry	2093-3

Filtered the data based on the following conditions:

Total cholesterol >=240 mg/dL
HDL <35

LDL tests were missing in the laboratory data

Created ‘cholesterol’ datatype, and appended with the previous data,
New diagnosis variable ‘htn’ was derived based on other diagnosis columns, now the **hypertension cases increased to 51**

Manual review data combined with ICD code and Lab tests

```
Confusion Matrix and Statistics

Accuracy : 0.9733
      95% CI : (0.907, 0.9968)
 No Information Rate : 0.6533
 P-Value [Acc > NIR] : 1.122e-11
      Kappa : 0.94
Mcnemar's Test P-Value : 0.4795

    Sensitivity : 1.0000
    Specificity : 0.9231
   Pos Pred Value : 0.9608
   Neg Pred Value : 1.0000
    Prevalence : 0.6533
   Detection Rate : 0.6533
 Detection Prevalence : 0.6800
  Balanced Accuracy : 0.9615

  'Positive' Class : 1
```

		Manual Review of Hypertension	
		+	-
ICD Codes + Laboratory Lipid data	+	49	2
	-	0	24

Conclusion

Adding additional data types increased both sensitivity and specificity.

Treatment (medication) datatype was not used, hence the specificity was marginally low.

Algorithm performance was very high, less complexity in implementations and the portability of the algorithm is moderate.