Practical Application Project: Develop a Computational Phenotyping Algorithm to Identify Patients with Hypertension

Setting up the environment

```
library(tidyverse)
library(magrittr)
library(bigrquery)
library(caret)
con <- DBI::dbConnect(drv = bigquery(), project =
"learnclinicaldatascience")
hypertension <- tbl(con, "course3_data.hypertension_goldstandard")
training <- hypertension %>% collect() %>% sample_n(80)
testing <- hypertension %>% filter(!SUBJECT_ID %in%
training_population$SUBJECT_ID)
getStats <- function(df, ...){df %>% select_(.dots = lazyeval::lazy_dots(...)) %>%
mutate_all(funs(factor(., levels = c(1,0)))) %>% table() %>% confusionMatrix()}
```

Test individual data types

- ICD9 CODES for hypertension: "4019", "4011", "36504", "5723", "3482", "64610", "40591", "40501", "4010", "4160", "45939", "64612", "64611", "64203", "64204", "64292", "64291", "40509", "64622", "64621", "64620", "40599", "64272", "64270", "64233", "64234", "40511", "40519", "45933", "45931", "45932", "45930", "64613", "64614", "7962", "64200", "64202", "64201", "64290", "64293", "64294", "64223", "64221", "64224", "64220", "64222", "64623", "64624", "99791", "64273", "64274", "64271", "64230", "64232", "64231"
- ITEMIDs for Systolic blood pressure: 3317, 6, 6701, 3323, 3321, 455, 3325, 3319, 442, 666, 3313, 492, 3315, 51, 7643, 482, 484, 480, 228152, 220059, 226852, 226850, 220050, 227243, 220179, 225309, 224167
- Medication: Lisinopril

ICD9 CODES

```
diagnoses_icd <- tbl(con, "mimic3_demo.DIAGNOSES_ICD")


icd <- diagnoses_icd %>% filter(ICD9_CODE %in% c("4019", "4011", "36504",
"5723", "3482", "64610", "40591", "40501", "4010", "4160", "45939", "64612",
"64611", "64203", "64204", "64292", "64291", "40509", "64622", "64621",
"64620", "40599", "64272", "64270", "64233", "64234", "40511", "40519",
"45933", "45931", "45932", "45930", "64613", "64614", "7962", "64200", "64202",
"64201", "64290", "64293", "64294", "64223", "64221", "64224", "64220",
"64222", "64623", "64624", "99791", "64273", "64274", "64271", "64230",
"64232", "64231")) %>% distinct(SUBJECT_ID) %>% mutate(icd = 1) %>% collect()


training %<>% left_join(icd, copy = TRUE) %>% mutate(icd = coalesce(icd, 0)) %>%
collect() %>% getStats(icd, HYPERTENSION)
```

ICD9 CODES

		HYPERTENSION	
		1	**0**
ICD	**1**	27	3
	0	25	25

Sensitivity : 0.5192

Specificity : 0.8929

Pos Pred Value : 0.9000

Neg Pred Value : 0.5000

ITEM IDs

```
chartevents <- tbl(con, "mimic3_demo.CHARTEVENTS")


systolic <- chartevents %>% filter(ITEMID %in% c(3317, 6, 6701, 3323,
3321, 455, 3325, 3319, 442, 666, 3313, 492, 3315, 51, 7643, 482, 484,
480, 228152, 220059, 226852, 226850, 220050, 227243, 220179,
225309, 224167)) %>% distinct(SUBJECT_ID) %>% mutate(systolic = 1)


training %<>% left_join(systolic, copy = TRUE) %>% mutate(systolic =
coalesce(systolic, 0)) %>% collect() %>% getStats(systolic,
HYPERTENSION)
```

ITEM IDs

	HYPERTENSION		
SYSTOLIC		**1**	**0**
	1	53	25
	0	0	2

Sensitivity : 1.00000

Specificity : 0.07407

Pos Pred Value : 0.67949

Neg Pred Value : 1.00000

Lisinopril

```
prescriptions <- tbl(con, "mimic3_demo.PRESCRIPTIONS")

lisinopril <- prescriptions %>% filter(tolower(DRUG) %like% "%lisinopril%") %>% distinct(SUBJECT_ID) %>% mutate(lisinopril = 1)

training %<>% left_join(lisinopril, copy= TRUE) %>% mutate(lisinopril = coalesce(lisinopril, 0)) %>% collect() %>% getStats(lisinopril, HYPERTENSION)
```

Lisinopril

		HYPERTENSION	
LISINOPRIL		**1**	**0**
	1	13	3
	0	40	24

Sensitivity : 0.2453

Specificity : 0.8889

Pos Pred Value : 0.8125

Neg Pred Value : 0.3750

Data manipulations

- Temporal manipulation- First Value: First instance of systolic blood pressure above 140mm Hg

- Frequency and Value manipulations- Threshholding: 2+ counts of systolic blood pressure above 140mm Hg

Temporal manipulation- First value: First instance of systolic blood pressure above 140mm Hg

```
chartevents <- tbl(con, "mimic3_demo.CHARTEVENTS")
d_items <- tbl(con, "mimic3_demo.D_ITEMS")


systolic_over140_first <- chartevents %>% inner_join(d_items, by = c("ITEMID" = "ITEMID"), suffix = c("_c","_i")) %>% filter(ITEMID %in% c(3317, 6, 6701, 3323, 3321, 455, 3325, 3319, 442, 666, 3313, 492, 3315, 51, 7643, 482, 484, 480, 228152, 220059, 226852, 226850, 220050, 227243, 220179, 225309, 224167)) %>% group_by(SUBJECT_ID) %>% mutate(earliest_pressure = min(CHARTTIME, na.rm = TRUE)) %>% filter(CHARTTIME == earliest_pressure) %>% mutate(systolic_over140_first = case_when(VALUENUM >= 140 ~ 1, TRUE ~ 0)) %>% select(SUBJECT_ID, systolic_over140_first)


training %>% left_join(systolic_over140_first, copy = TRUE) %>% mutate(systolic_over140_first = coalesce(systolic_over140_first, 0)) %>% collect() %>% getStats(systolic_over140_first, HYPERTENSION)
```

Temporal manipulation- First value: First instance of systolic blood pressure above 140mm Hg

	HYPERTENSION		
SYSTOLIC_OVER140_FIRST		**1**	**0**
	1	20	8
	0	31	26

Sensitivity : 0.3922

Specificity : 0.7647

Pos Pred Value : 0.7143

Neg Pred Value : 0.4561

Frequency and Value manipulations- Threshholding: 2+ counts of systolic blood pressure above 140mm Hg

```
chartevents <- tbl(con, "mimic3_demo.CHARTEVENTS")
d_items <- tbl(con, "mimic3_demo.D_ITEMS")


systolic_over140_min2 <- chartevents %>% inner_join(d_items, by = c("ITEMID" = "ITEMID"), suffix = c("_c","_i")) %>% filter(ITEMID %in% c(3317, 6, 6701, 3323, 3321, 455, 3325, 3319, 442, 666, 3313, 492, 3315, 51, 7643, 482, 484, 480, 228152, 220059, 226852, 226850, 220050, 227243, 220179, 225309, 224167)) %>% group_by(SUBJECT_ID) %>% mutate(systolic_over140_counter = case_when(VALUENUM >= 140 ~ 1, TRUE ~0)) %>% summarise(systolic_over140_count = sum(systolic_over140_counter, na.rm = TRUE)) %>% mutate(systolic_over140_min2 = case_when(systolic_over140_count >= 2 ~ 1, TRUE ~ 0)) %>% select(SUBJECT_ID, systolic_over140_min2)


training %<>% left_join(systolic_over140_min2, copy = TRUE) %>% mutate(systolic_over140_min2 = coalesce(systolic_over140_min2, 0)) %>% collect()


training %>% collect() %>% getStats(systolic_over140_min2, HYPERTENSION)
```

Frequency and Value manipulations- Threshholding: 2+ counts of systolic blood pressure above 140mm Hg

	HYPERTENSION		
SYSTOLIC_OVER140_MIN2		**1**	**0**
	1	38	15
	0	12	15

Sensitivity : 0.7600

Specificity : 0.5000

Pos Pred Value : 0.7170

Neg Pred Value : 0.5556

Data combinations

- 2+ counts of systolic blood pressure above 140mm Hg AND any ICD9 CODE

- 2+ counts of systolic blood pressure above 140mm Hg OR any ICD9 CODE

2+ counts of systolic blood pressure above 140mm Hg AND any ICD9 CODE

```
diagnoses_icd <- tbl(con, "mimic3_demo.DIAGNOSES_ICD")

chartevents <- tbl(con, "mimic3_demo.CHARTEVENTS")

d_items <- tbl(con, "mimic3_demo.D_ITEMS")


icd <- diagnoses_icd %>% filter(ICD9_CODE %in% c("4019", "4011", "36504", "5723", "3482", "64610", "40591", "40501", "4010", "4160", "45939", "64612", "64611", "64203", "64204", "64292", "64291", "40509", "64622", "64621", "64620", "40599", "64272", "64270", "64233", "64234", "40511", "40519", "45933", "45931", "45932", "45930", "64613", "64614", "7962", "64200", "64202", "64201", "64290", "64293", "64294", "64223", "64221", "64224", "64220", "64222", "64623", "64624", "99791", "64273", "64274", "64271", "64230", "64232", "64231")) %>% distinct(SUBJECT_ID) %>% mutate(icd = 1) %>% collect()


systolic_over140_min2 <- chartevents %>% inner_join(d_items, by = c("ITEMID" = "ITEMID"), suffix = c("_c","_i")) %>% filter(ITEMID %in% c(3317, 6, 6701, 3323, 3321, 455, 3325, 3319, 442, 666, 3313, 492, 3315, 51, 7643, 482, 484, 480, 228152, 220059, 226852, 226850, 220050, 227243, 220179, 225309, 224167)) %>% group_by(SUBJECT_ID) %>% mutate(systolic_over140_counter = case_when(VALUENUM >= 140 ~ 1, TRUE ~0)) %>% summarise(systolic_over140_count = sum(systolic_over140_counter, na.rm = TRUE)) %>% mutate(systolic_over140_min2 = case_when(systolic_over140_count >= 2 ~ 1, TRUE ~ 0)) %>% select(SUBJECT_ID, systolic_over140_min2)


training %>% left_join(icd, copy = TRUE) %>% left_join(systolic_over140_min2, copy = TRUE) %>% mutate(icd = coalesce(icd, 0), systolic_over140_min2 = coalesce(systolic_over140_min2, 0)) %>% mutate(icd_and_systolic_over140_min2 = case_when(icd == 1 && systolic_over140_min2 == 1 ~ 1, TRUE ~ 0)) %>% collect() %>% getStats(icd_and_systolic_over140_min2, HYPERTENSION)
```

2+ counts of systolic blood pressure above 140mm Hg AND any ICD9 CODE

	HYPERTENSION		
ICD_AND_SYSTOLIC_OVER140_MIN2		**1**	**0**
	1	0	0
	0	51	29

Sensitivity : 0.0000

Specificity : 1.0000

Pos Pred Value : NaN

Neg Pred Value : 0.3625

2+ counts of systolic blood pressure above 140mm Hg OR any ICD9 CODE

```
diagnoses_icd <- tbl(con, "mimic3_demo.DIAGNOSES_ICD")

chartevents <- tbl(con, "mimic3_demo.CHARTEVENTS")

d_items <- tbl(con, "mimic3_demo.D_ITEMS")


icd <- diagnoses_icd %>% filter(ICD9_CODE %in% c("4019", "4011", "36504", "5723", "3482", "64610", "40591", "40501", "4010", "4160", "45939", "64612", "64611", "64203", "64204", "64292", "64291", "40509", "64622", "64621", "64620", "40599", "64272", "64270", "64233", "64234", "40511", "40519", "45933", "45931", "45932", "45930", "64613", "64614", "7962", "64200", "64202", "64201", "64290", "64293", "64294", "64223", "64221", "64224", "64220", "64222", "64623", "64624", "99791", "64273", "64274", "64271", "64230", "64232", "64231")) %>% distinct(SUBJECT_ID) %>% mutate(icd = 1) %>% collect()


systolic_over140_min2 <- chartevents %>% inner_join(d_items, by = c("ITEMID" = "ITEMID"), suffix = c("_c","_i")) %>% filter(ITEMID %in% c(3317, 6, 6701, 3323, 3321, 455, 3325, 3319, 442, 666, 3313, 492, 3315, 51, 7643, 482, 484, 480, 228152, 220059, 226852, 226850, 220050, 227243, 220179, 225309, 224167)) %>% group_by(SUBJECT_ID) %>% mutate(systolic_over140_counter = case_when(VALUENUM >= 140 ~ 1, TRUE ~0)) %>% summarise(systolic_over140_count = sum(systolic_over140_counter, na.rm = TRUE)) %>% mutate(systolic_over140_min2 = case_when(systolic_over140_count >= 2 ~ 1, TRUE ~ 0)) %>% select(SUBJECT_ID, systolic_over140_min2)


training %>% left_join(icd, copy = TRUE) %>% left_join(systolic_over140_min2, copy = TRUE) %>% mutate(icd = coalesce(icd, 0), systolic_over140_min2 = coalesce(systolic_over140_min2, 0)) %>% mutate(icd_or_systolic_over140_min2 = case_when(icd == 1 | systolic_over140_min2 == 1 ~ 1, TRUE ~ 0)) %>% collect() %>% getStats(icd_or_systolic_over140_min2, HYPERTENSION)
```

2+ counts of systolic blood pressure above 140mm Hg OR any ICD9 CODE

	HYPERTENSION		
ICD_OR_SYSTOLIC_OVER140_MIN2		**1**	**0**
	1	40	17
	0	7	16

Sensitivity : 0.8511

Specificity : 0.4848

Pos Pred Value : 0.7018

Neg Pred Value : 0.6957

Conclusion

I would like to compare 4 algorithms. They are- ‘ICD9 CODE only’, ‘2+ instances of blood pressure above 140mm Hg’, ‘2+ counts of systolic blood pressure above 140mm Hg AND any ICD9 CODE’ and ‘2+ counts of systolic blood pressure above 140mm Hg OR any ICD9 CODE’

- If we were looking for high specificity, I would choose ‘2+ counts of systolic blood pressure above 140mm Hg AND any ICD9 CODE’ but the sensitivity is 0.0000
- ‘2+ counts of systolic blood pressure above 140mm Hg OR any ICD9 CODE’ has slightly more balanced sensitivity and specificity
- Since a simpler algorithm is a better algorithm, I would choose ‘ICD9 CODE only’, if we were looking for higher specificity and ‘2+ instances of blood pressure above 140mm Hg’, if we were looking for higher sensitivity.