RISK STRATIFICATION ACROSS HEALTHCARE SYSTEMS

Amanda Zheutlin, PhD

Center for Genomic Medicine, Massachusetts General Hospital

PREDICTING BIPOLAR DISORDER

- Misdiagnosis rate up to 60%
- Longer duration of untreated illness predicts worse clinical outcomes
- Early and accurate diagnosis would improve the burden of disease

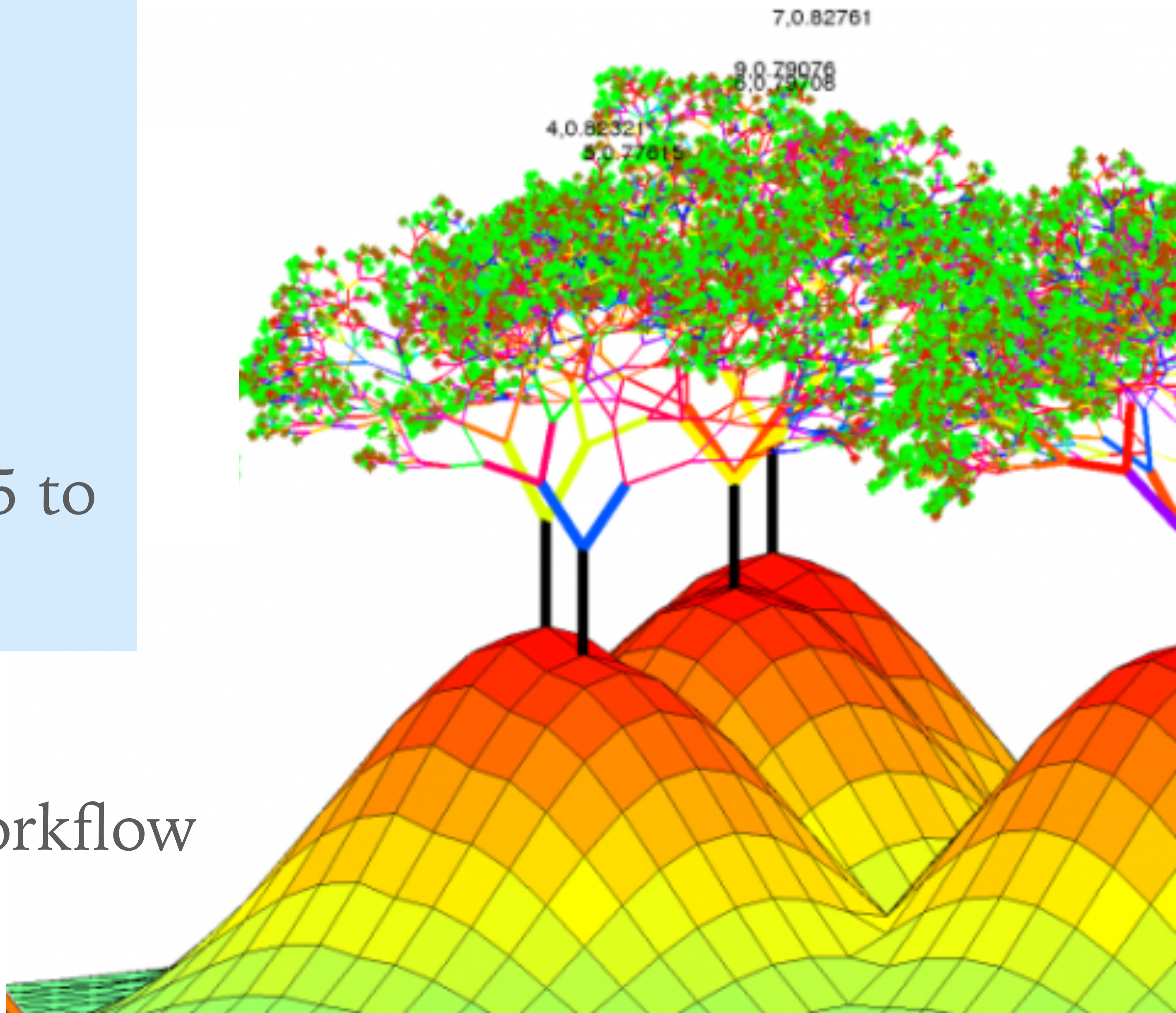
Geisinger
Health Plan

PARTNERS
HEALTHCARE

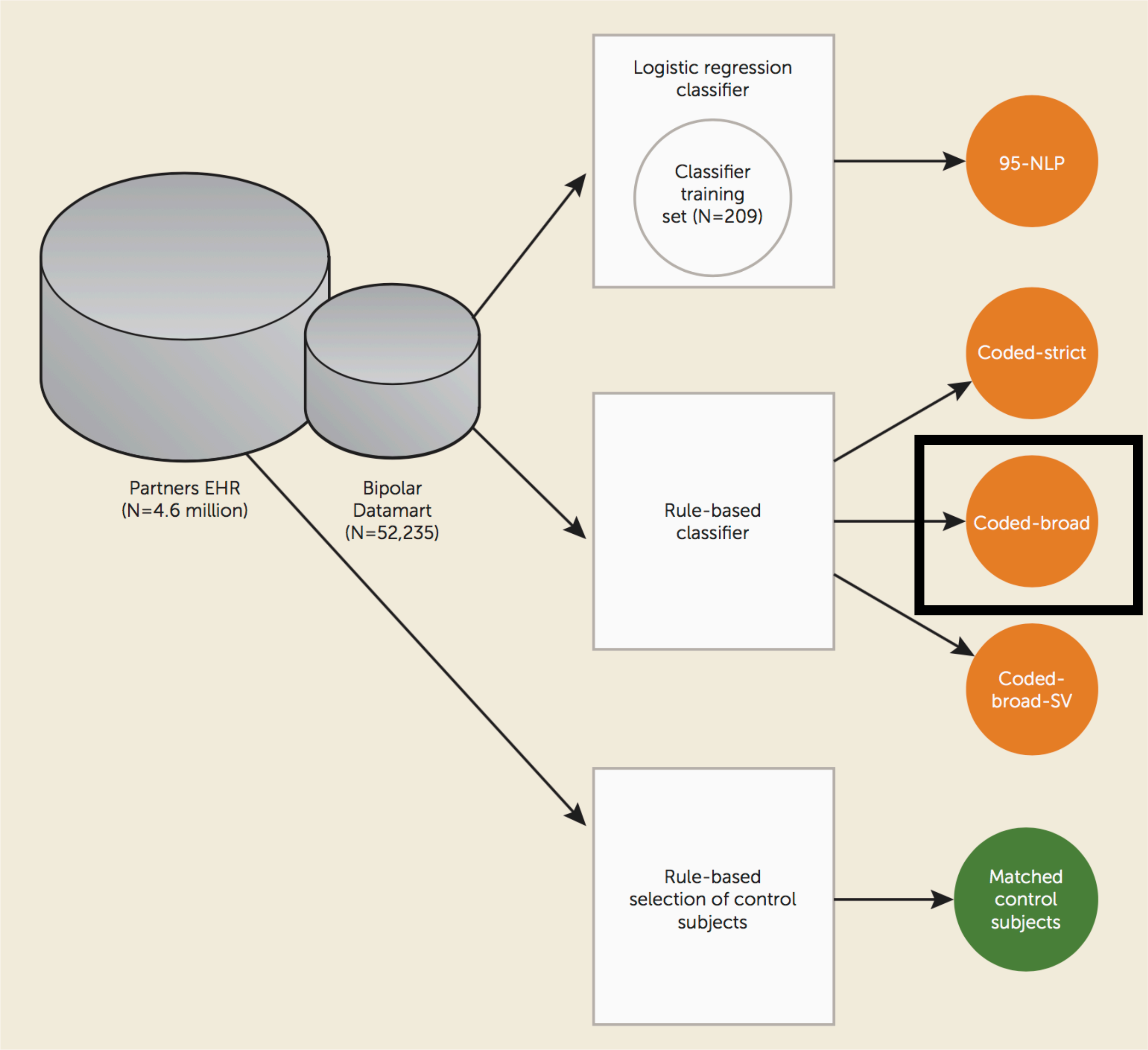
VANDERBILT

RISK STRATIFICATION PIPELINE

1. Define outcome
2. Extract data
3. Feature engineering
4. Machine learning
5. Test at external sites
6. Iterate through #4 and #5 to improve performance
7. Pilot for clinical use
8. Integrate with clinical workflow



IDENTIFYING BIPOLAR DISORDER CASES



Partners EHR (N=4.6 million)

Bipolar Datamart (N=52,235)

Logistic regression classifier

Classifier training set (N=209)

95-NLP

Rule-based classifier

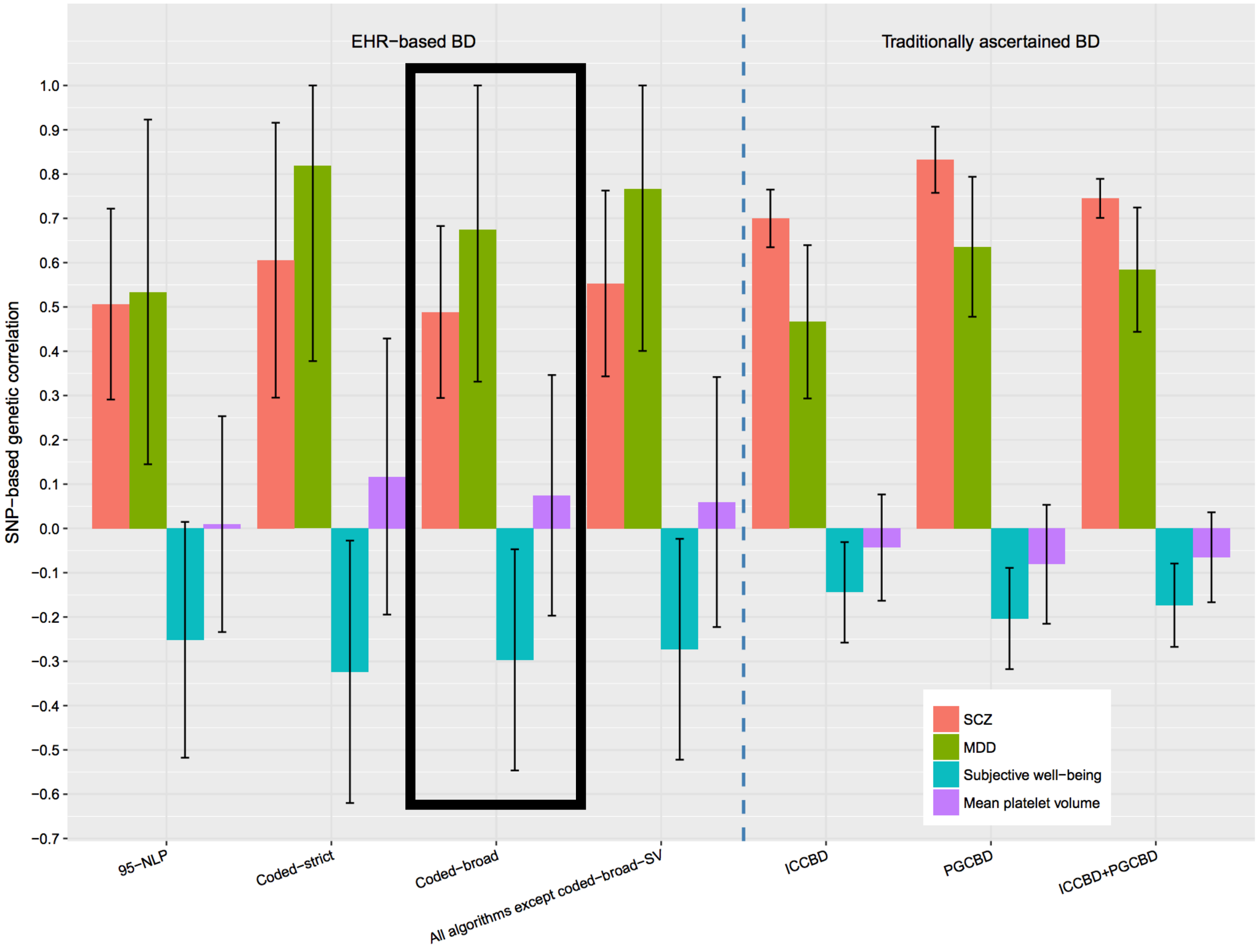
Coded-strict

Coded-broad

Coded-broad-SV

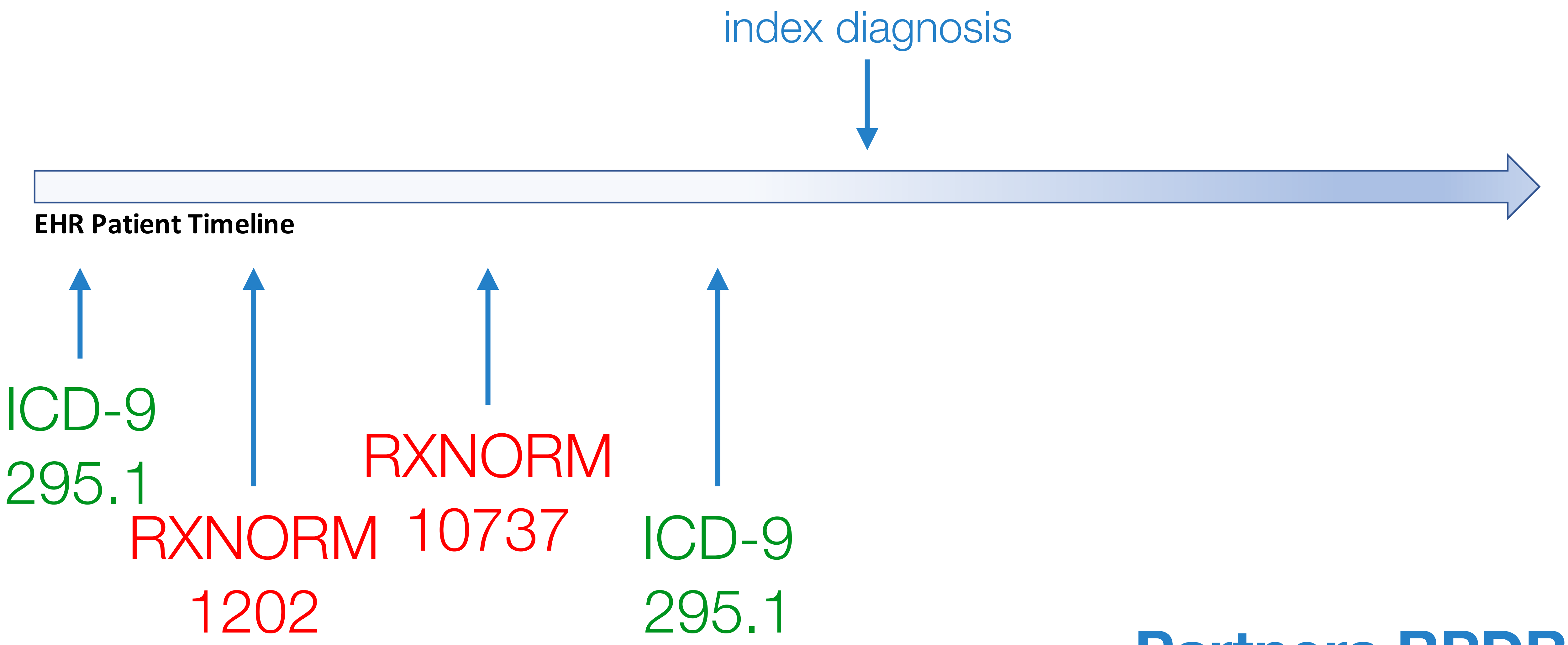
Rule-based selection of control subjects

Matched control subjects



Castro et al., 2014; Chen et al., 2018

LEVERAGING ELECTRONIC HEALTH RECORDS



index diagnosis

EHR Patient Timeline

ICD-9 295.1

RXNORM 1202

RXNORM 10737

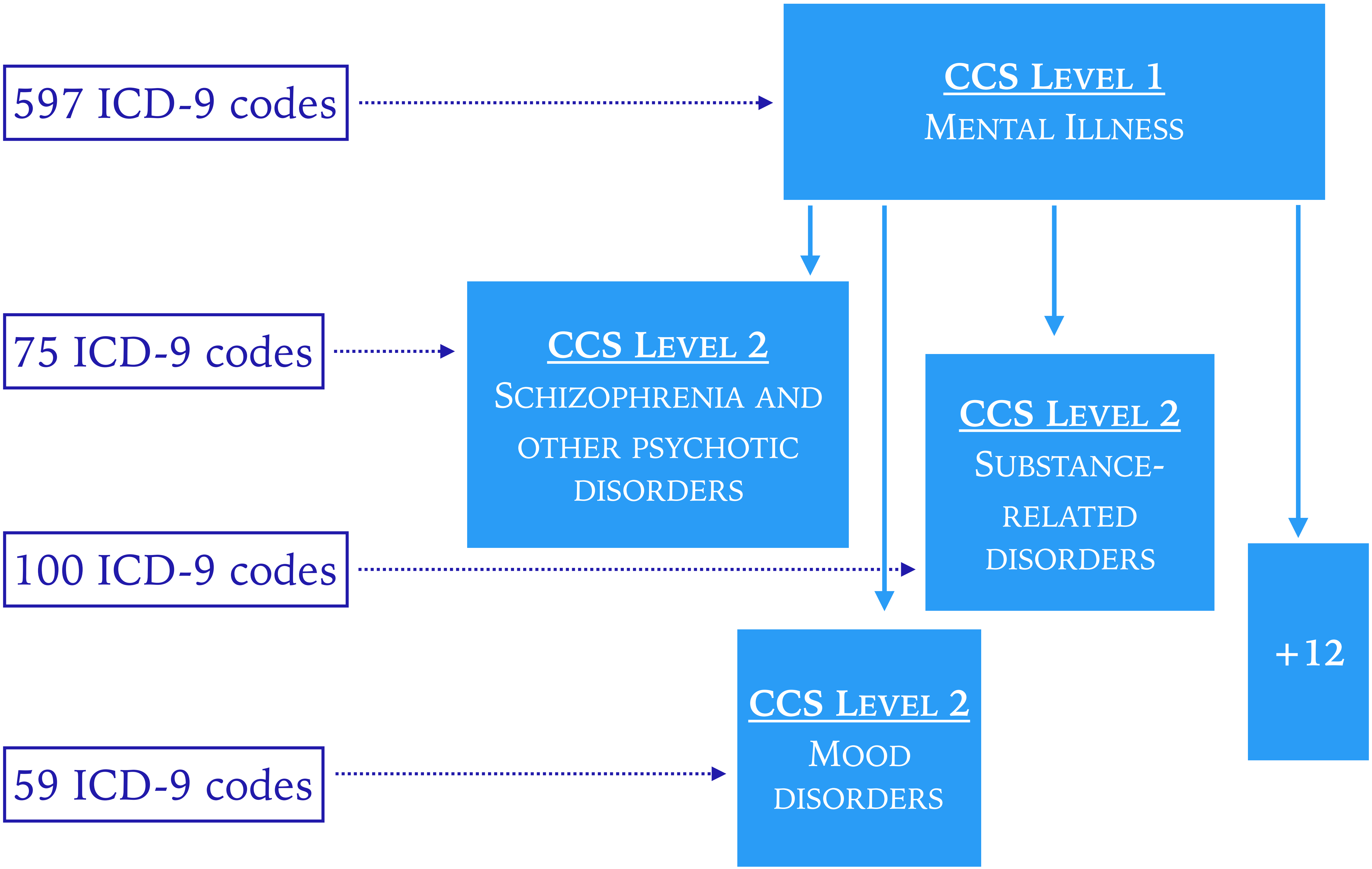
ICD-9 295.1

Partners RPDR

8,602 cases

1.8M controls

BILLING CODES TO DIAGNOSTIC CATEGORIES



597 ICD-9 codes

CCS LEVEL 1
MENTAL ILLNESS

75 ICD-9 codes

CCS LEVEL 2
SCHIZOPHRENIA AND OTHER PSYCHOTIC DISORDERS

CCS LEVEL 2
SUBSTANCE-RELATED DISORDERS

100 ICD-9 codes

+12

59 ICD-9 codes

CCS LEVEL 2
MOOD DISORDERS

BUILDING A FEATURE MATRIX IN R

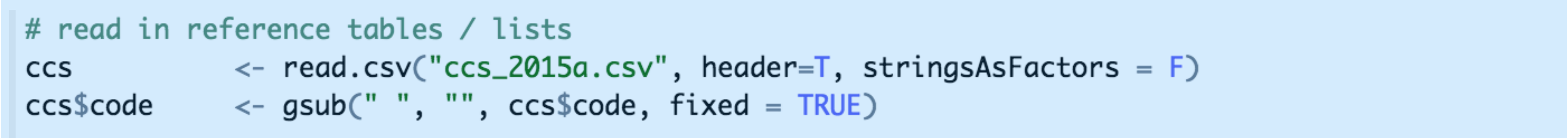
Patient #	Concept	Concept Date	Case Status
1	ICD9:295.1	7/8/13	FALSE
1	ICD9:296.2	7/13/13	FALSE
1	ICD9:296.2	8/16/13	FALSE
1	RXNORM: 10737	12/27/01	FALSE
1	RXNORM: 1202	2/19/01	FALSE
2	ICD9:295.1	12/8/14	FALSE
2	ICD9:296.2	1/13/15	FALSE
3	ICD9:333	8/10/14	TRUE
3	ICD9:395	8/20/14	TRUE
4	RXNORM: 101	3/3/03	FALSE
4	ICD9:103	4/15/06	FALSE
5	RXNORM: 10737	3/14/01	FALSE

457M rows x 4 columns

Patient#	Case Status	ICD9:295.1	ICD9:296.2	RXNORM: 10737
1	FALSE	1	1	1
2	FALSE	1	1	0
3	TRUE	0	0	0
4	FALSE	0	0	0
5	FALSE	0	0	1

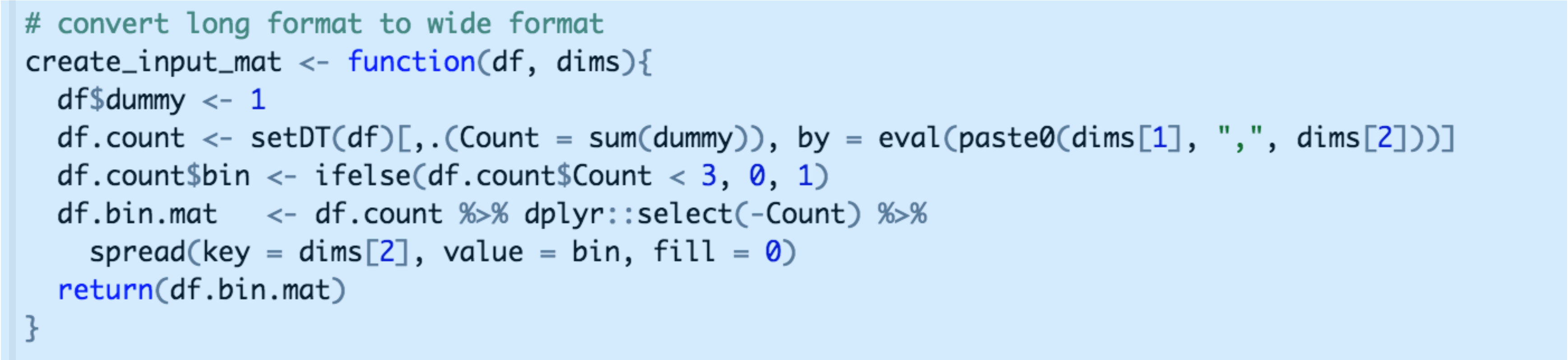
1.8M rows x 2150 columns

BUILDING A FEATURE MATRIX IN R



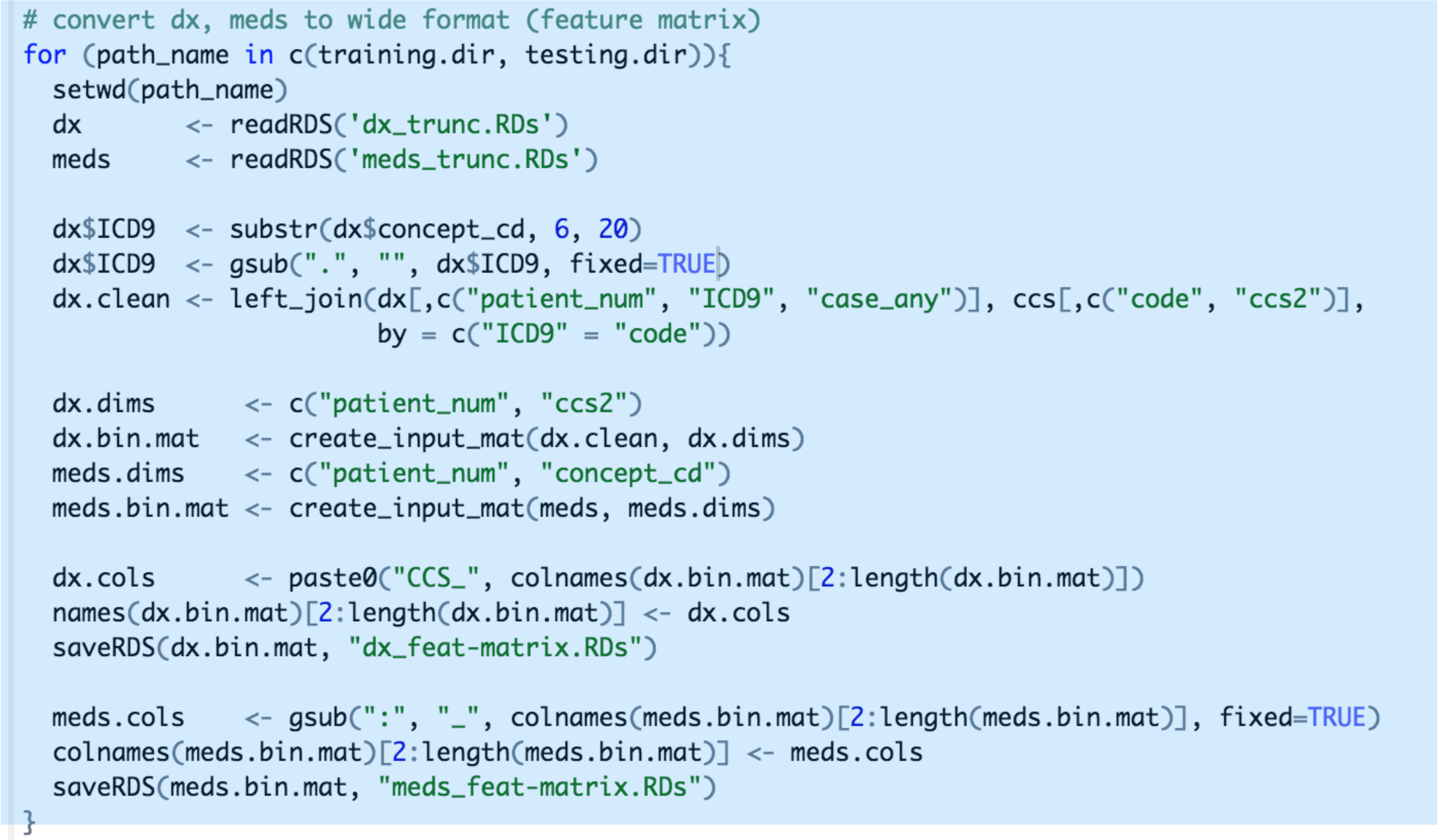
```r
# read in reference tables / lists
ccs          <- read.csv("ccs_2015a.csv", header=T, stringsAsFactors = F)
ccs$code     <- gsub(" ", "", ccs$code, fixed = TRUE)
```

Read in CCS map



```r
# convert long format to wide format
create_input_mat <- function(df, dims){
  df$dummy <- 1
  df.count <- setDT(df)[,.(Count = sum(dummy)), by = eval(paste0(dims[1], ",", dims[2]))]
  df.count$bin <- ifelse(df.count$Count < 3, 0, 1)
  df.bin.mat   <- df.count %>% dplyr::select(-Count) %>%
    spread(key = dims[2], value = bin, fill = 0)
  return(df.bin.mat)
}
```

Function to turn long format to wide format

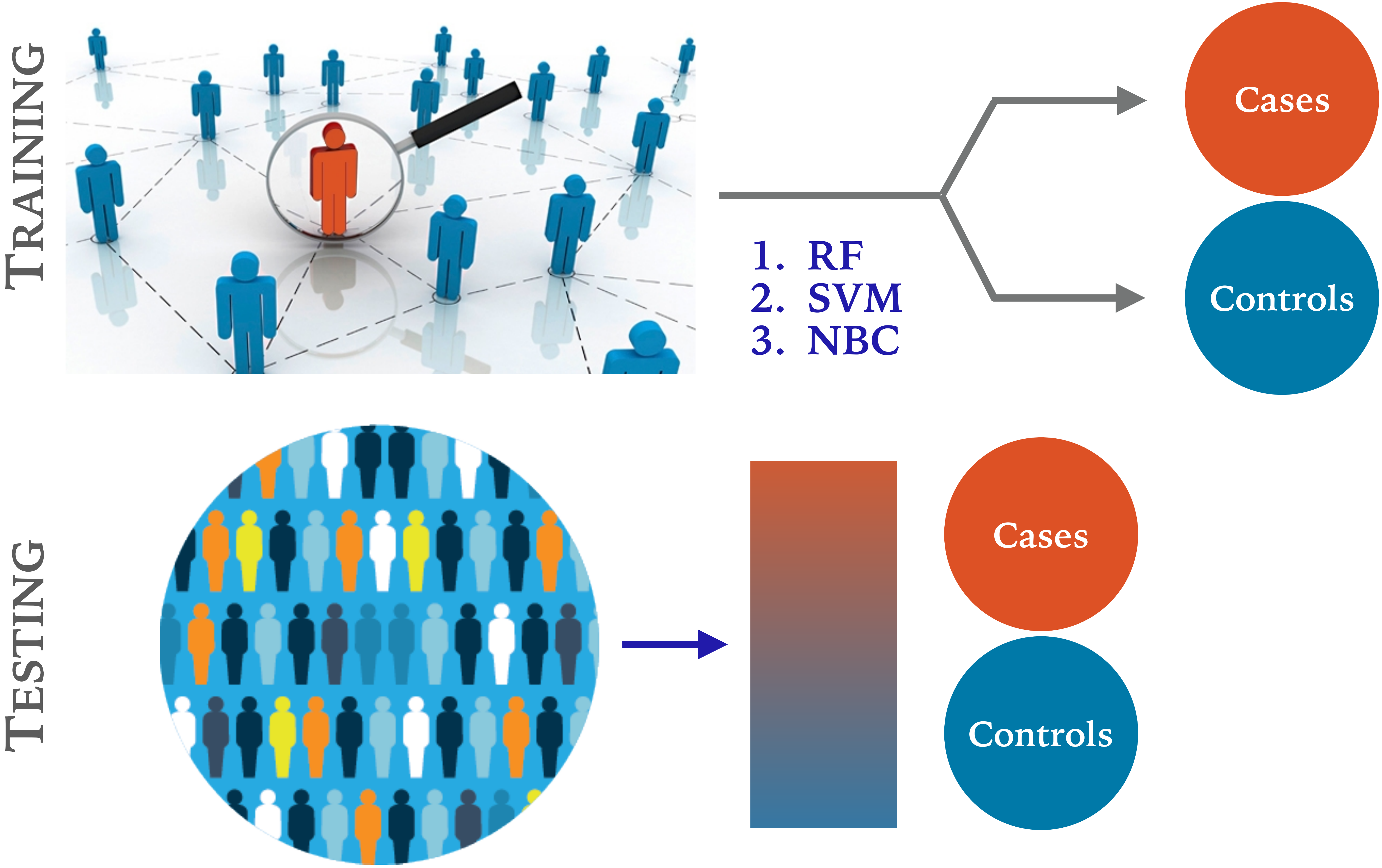


```r
# convert dx, meds to wide format (feature matrix)
for (path_name in c(training.dir, testing.dir)){
  setwd(path_name)
  dx       <- readRDS('dx_trunc.RDs')
  meds     <- readRDS('meds_trunc.RDs')

  dx$ICD9  <- substr(dx$concept_cd, 6, 20)
  dx$ICD9  <- gsub(".", "", dx$ICD9, fixed=TRUE)
  dx.clean <- left_join(dx[,c("patient_num", "ICD9", "case_any")], ccs[,c("code", "ccs2")],
                        by = c("ICD9" = "code"))

  dx.dims      <- c("patient_num", "ccs2")
  dx.bin.mat   <- create_input_mat(dx.clean, dx.dims)
  meds.dims    <- c("patient_num", "concept_cd")
  meds.bin.mat <- create_input_mat(meds, meds.dims)

  dx.cols      <- paste0("CCS_", colnames(dx.bin.mat)[2:length(dx.bin.mat)])
  names(dx.bin.mat)[2:length(dx.bin.mat)] <- dx.cols
  saveRDS(dx.bin.mat, "dx_feat-matrix.RDs")

  meds.cols    <- gsub(":", "_", colnames(meds.bin.mat)[2:length(meds.bin.mat)], fixed=TRUE)
  colnames(meds.bin.mat)[2:length(meds.bin.mat)] <- meds.cols
  saveRDS(meds.bin.mat, "meds_feat-matrix.RDs")
}
```

Function to

1. Map ICD-9 to CCS
2. Run above function
3. Save files

MACHINE LEARNING IN R



Training

1. RF
2. SVM
3. NBC

Cases

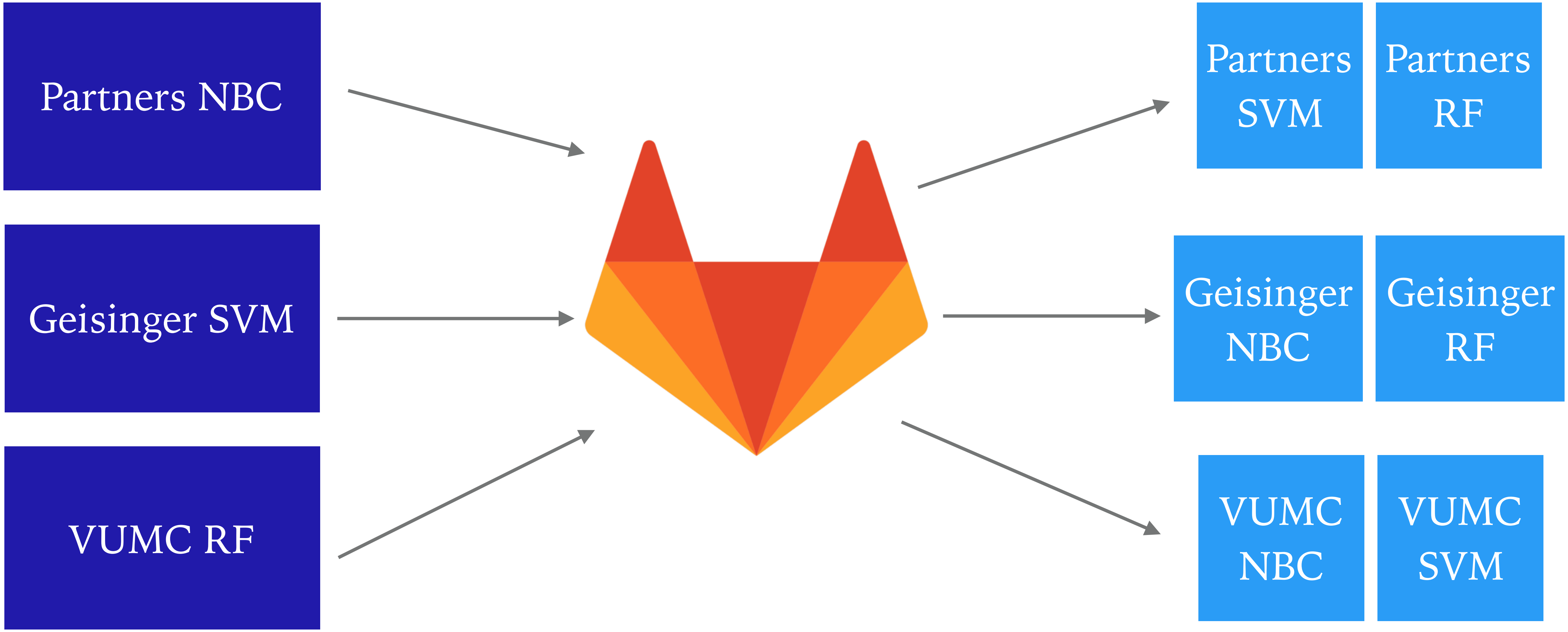
Controls

Testing

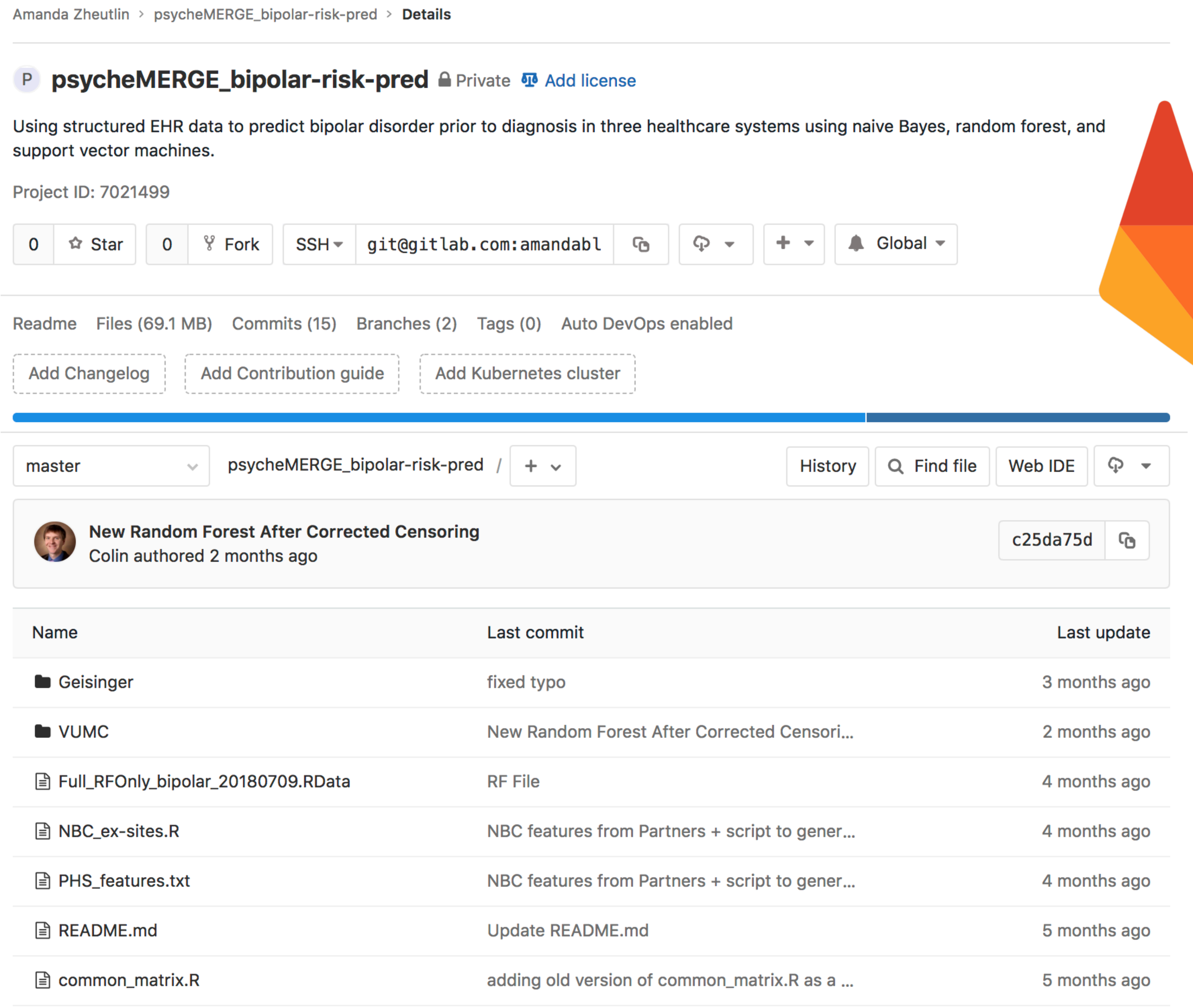
Cases

Controls

PORTABLE CODE AND HOW TO SHARE IT



PORTABLE CODE AND HOW TO SHARE IT



Amanda Zheutlin › psycheMERGE_bipolar-risk-pred › **Details**

psycheMERGE_bipolar-risk-pred

Using structured EHR data to predict bipolar disorder prior to diagnosis in three healthcare systems using naive Bayes, random forest, and support vector machines.

Project ID: 7021499

New Random Forest After Corrected Censoring
Colin authored 2 months ago

c25da75d

Name	Last commit	Last update
Geisinger	fixed typo	3 months ago
VUMC	New Random Forest After Corrected Censori...	2 months ago
Full_RFOnly_bipolar_20180709.RData	RF File	4 months ago
NBC_ex-sites.R	NBC features from Partners + script to gener...	4 months ago
PHS_features.txt	NBC features from Partners + script to gener...	4 months ago
README.md	Update README.md	5 months ago
common_matrix.R	adding old version of common_matrix.R as a ...	5 months ago

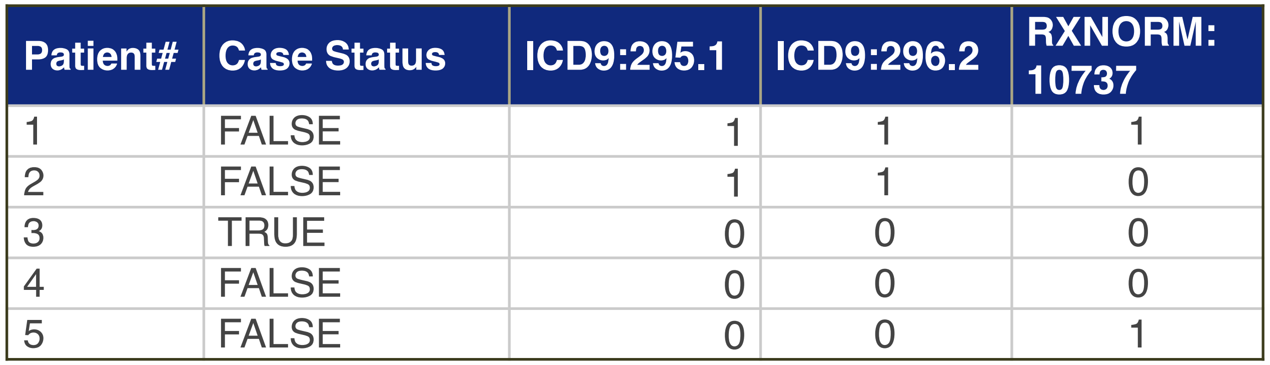
NEXT STEPS: BOOSTING PERFORMANCE

- ➤ Sampling
- ➤ Feature engineering
- ➤ Additional features

Partners RPDR

8,602 cases

1.8M controls

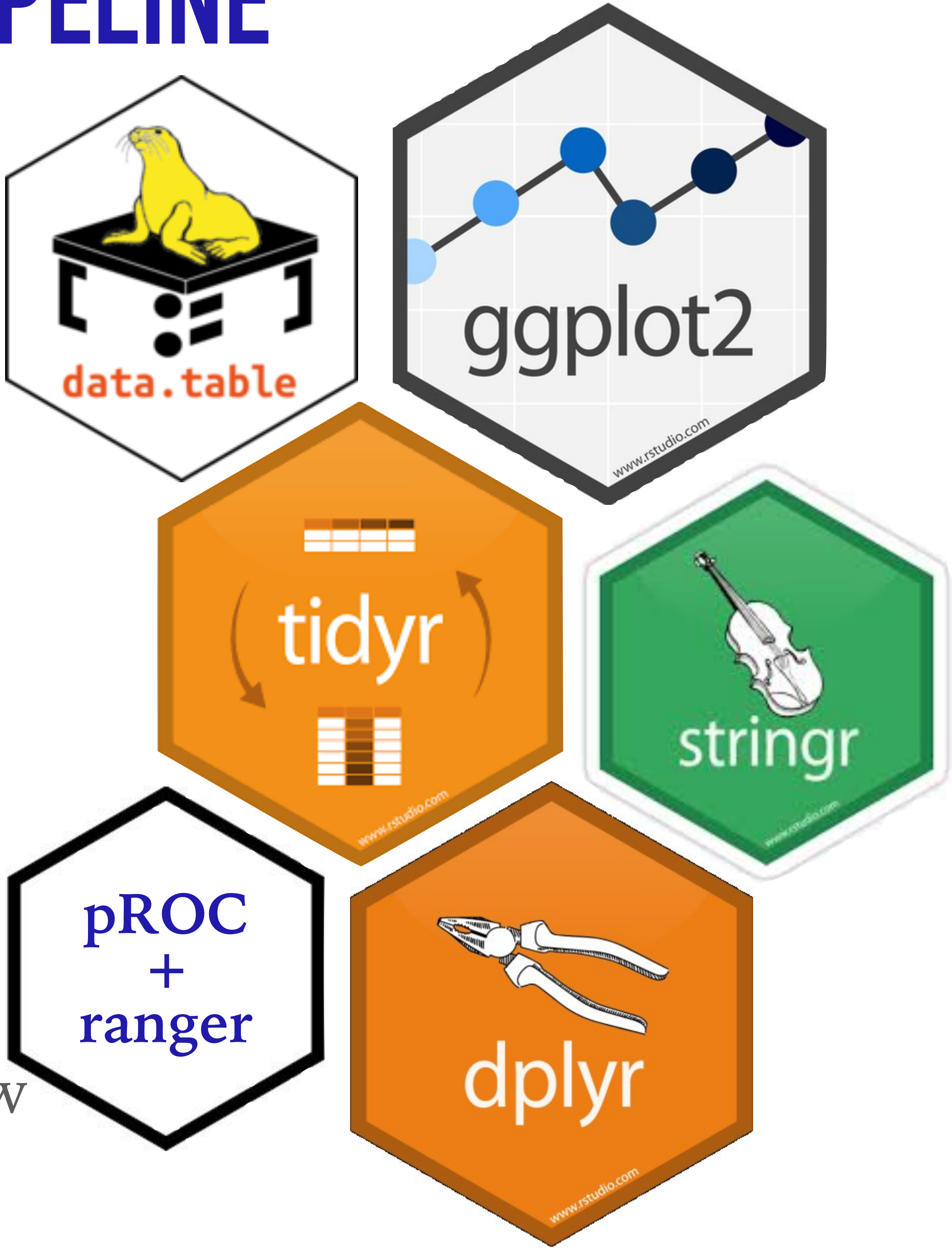


Patient#	Case Status	ICD9:295.1	ICD9:296.2	RXNORM: 10737
1	FALSE	1	1	1
2	FALSE	1	1	0
3	TRUE	0	0	0
4	FALSE	0	0	0
5	FALSE	0	0	1

SUMMARY

RISK STRATIFICATION PIPELINE

1. Define outcome
2. Extract data
3. Feature engineering
4. Machine learning
5. Test at external sites
6. Iterate through #4 and #5 to improve performance
7. Pilot for clinical use
8. Integrate with clinical workflow



ACKNOWLEDGEMENTS

THANK YOU!

PsycheMERGE

- Jordan Smoller
- Lea Davis
- Chris Chabris

Partners

- Victor Castro

VUMC

- Colin Walsh
- Doug Ruderfer

Geisinger

- Mariusz Butkiewicz
- Iris Hu
- Les Kirchner

Get in touch! azheutlin@mgh.harvard.edu & @amandabluezzz