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

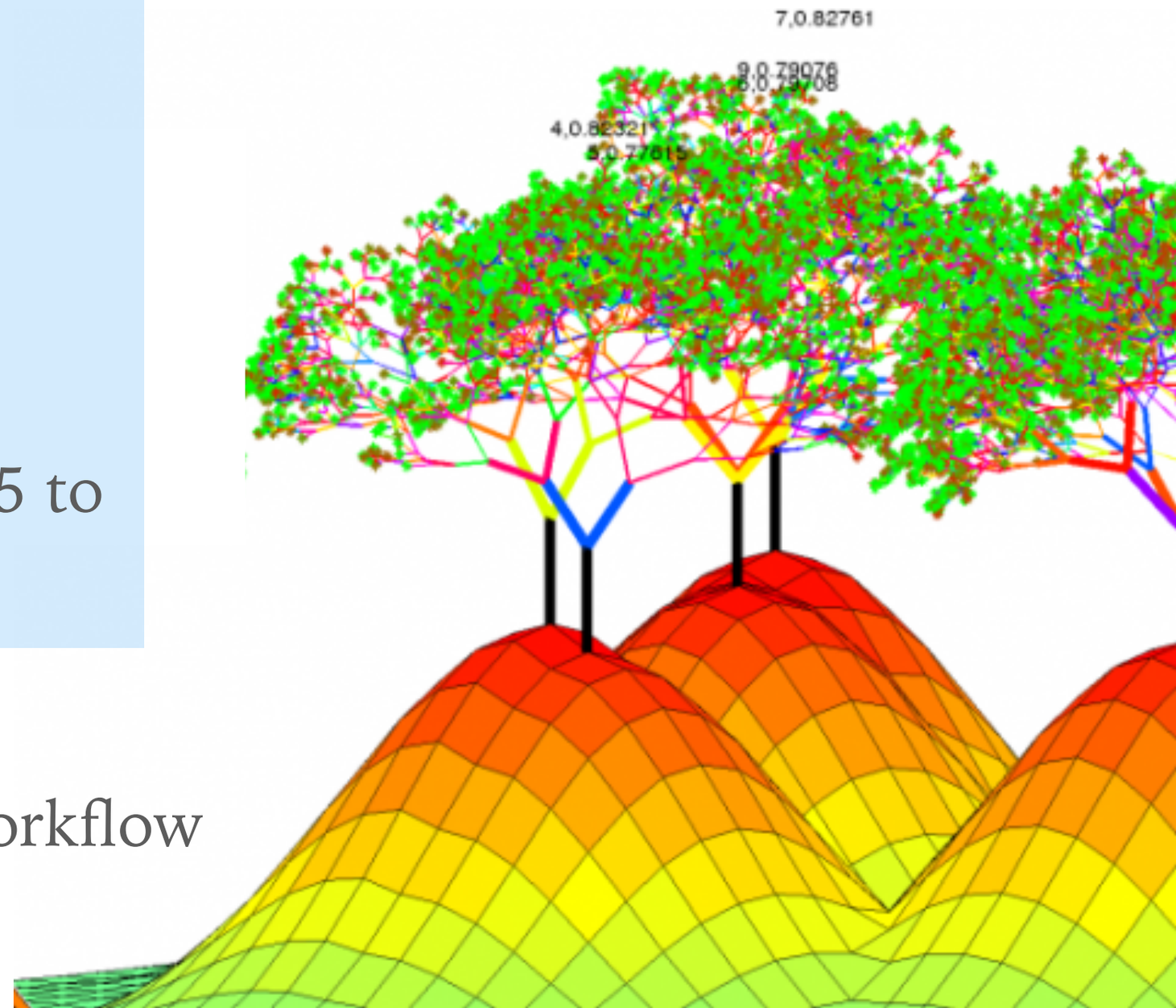
Geisinger Health Plan

PARTNERS HEALTHCARE

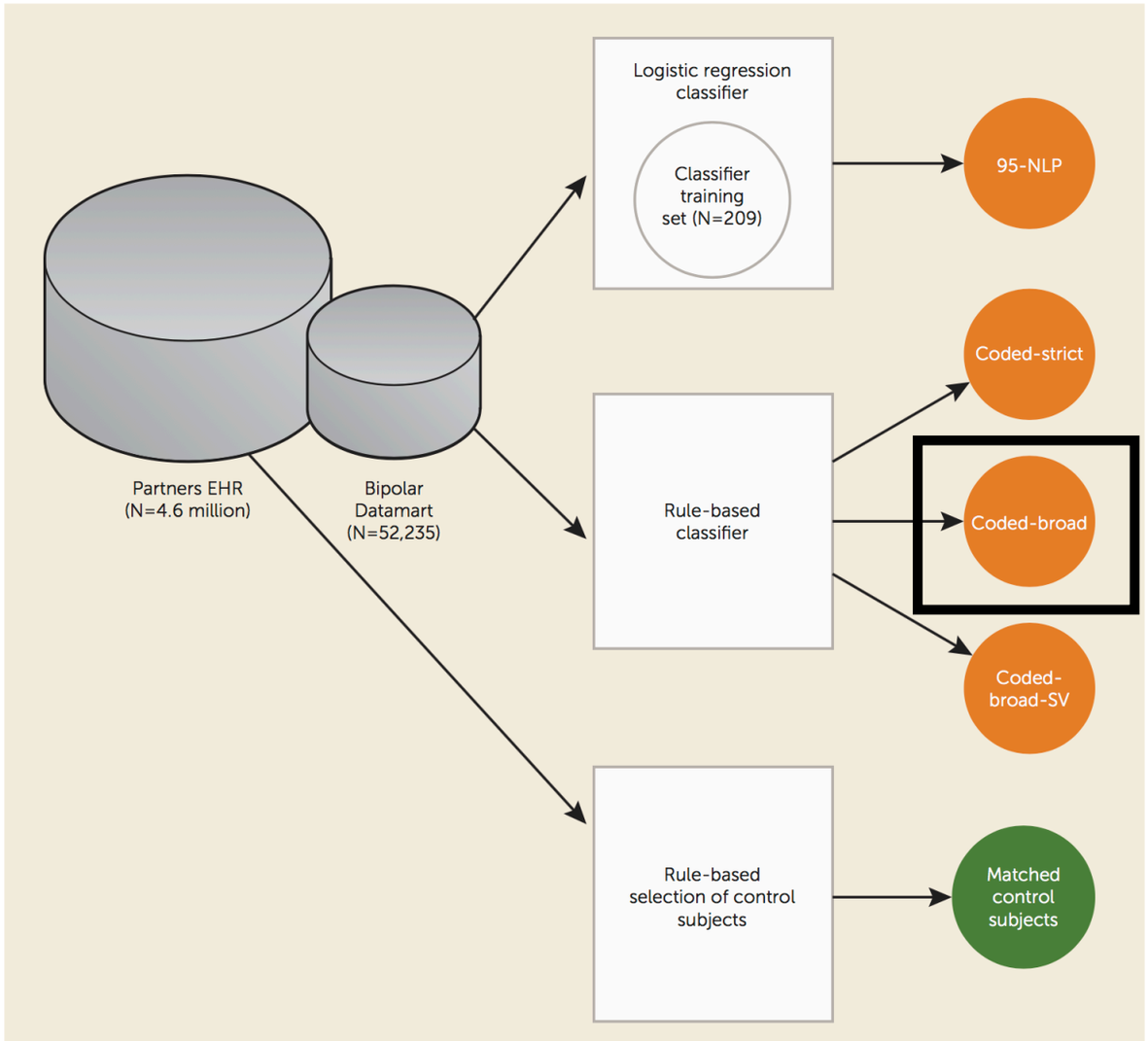
VANDERBILT

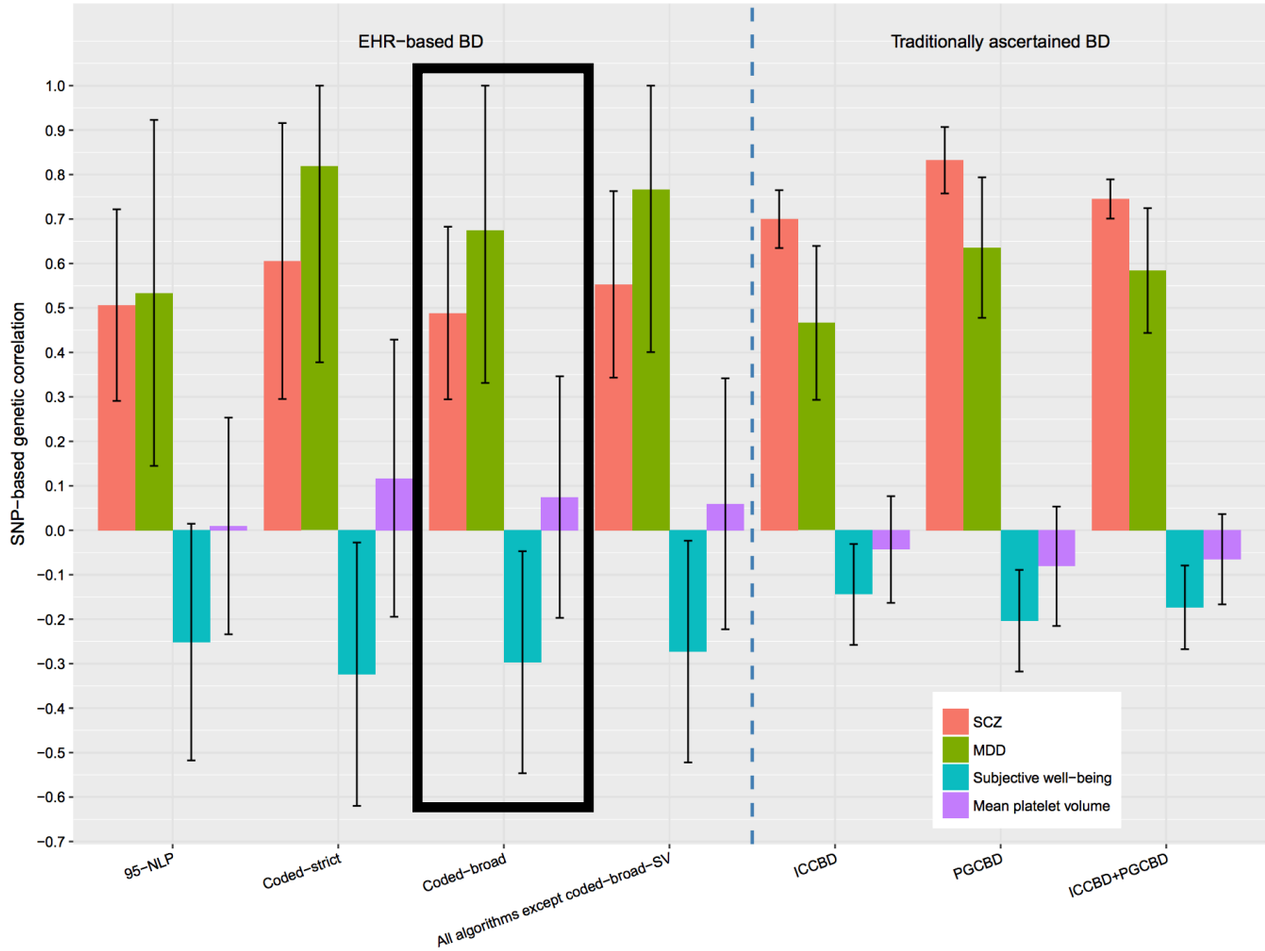
RISK STRATIFICATION PIPELINE

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



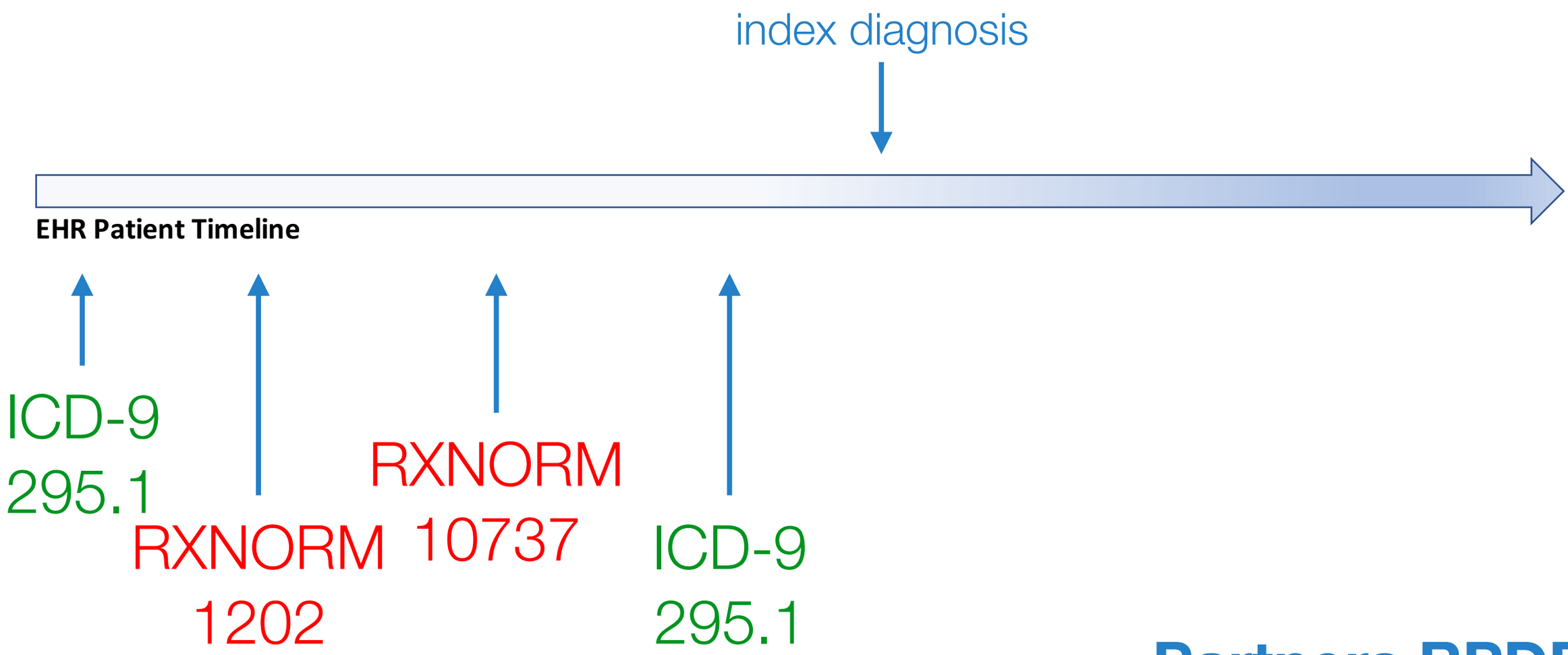
IDENTIFYING BIPOLAR DISORDER CASES





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

LEVERAGING ELECTRONIC HEALTH RECORDS

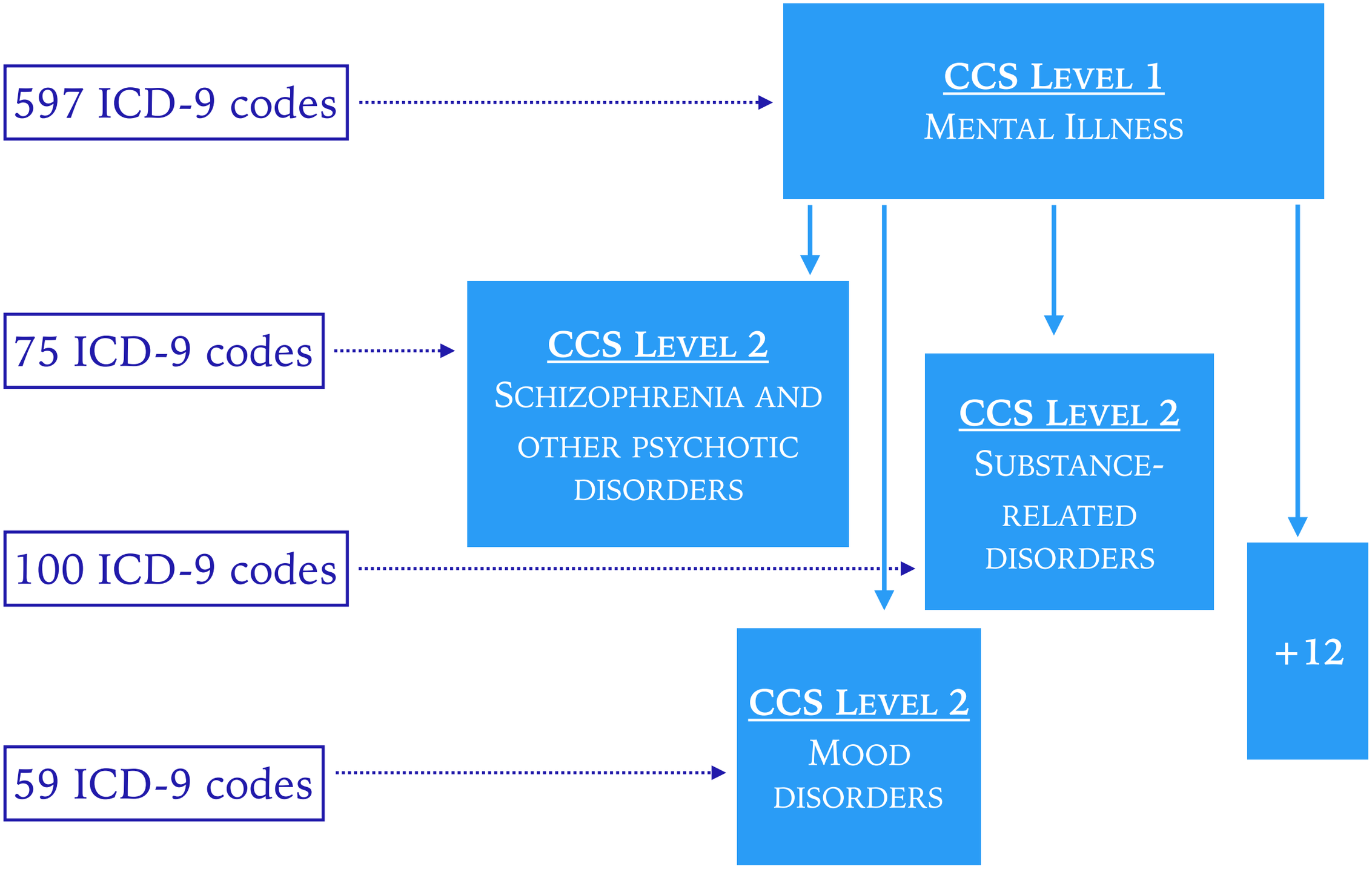


Partners RPDR

8,602 cases

1.8M controls

BILLING CODES TO DIAGNOSTIC CATEGORIES



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)
```

```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)
}
```

Read in CCS map

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

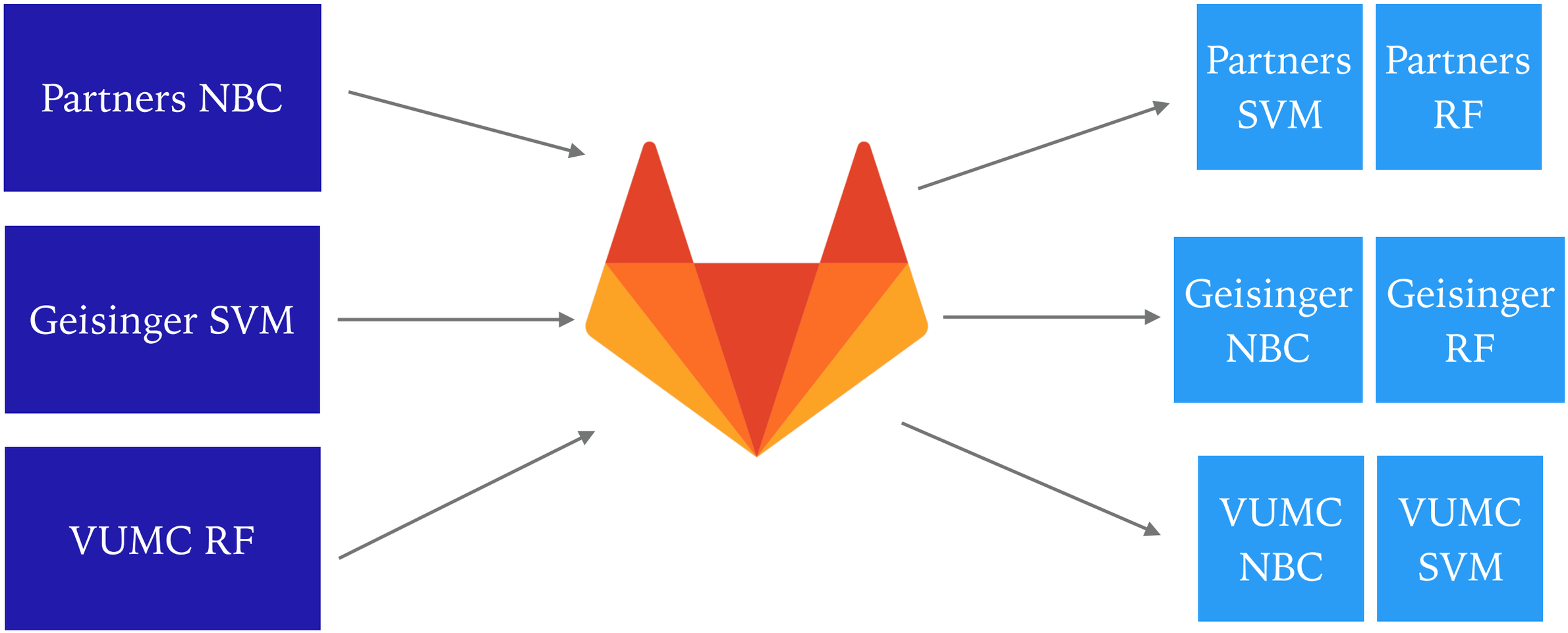
Controls

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 Private Add license

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

0 Star | 0 Fork | SSH | git@gitlab.com:amandabl | Global

Readme Files (69.1 MB) Commits (15) Branches (2) Tags (0) Auto DevOps enabled

Add Changelog | Add Contribution guide | Add Kubernetes cluster

master | psycheMERGE_bipolar-risk-pred / | History | Find file | Web IDE

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

EARLY RESULTS

Algorithm	Site	AUC	Sens	Spec	PPV	Cases	Controls
RF	VUMC	0.97	0.97	0.89	0.04	6,949	1.1M
SVM	GHS	0.89	0.24	0.97	0.05	1,735	59K
NBC	Partners	0.72	0.08	0.99	0.04	8,602	1.8M
RF	GHS	0.88	0.85	0.76	0.09	1,735	59K
RF	Partners	0.54	0.12	0.90	0.01	8,602	1.8M
SVM	VUMC	0.59	0.35	0.82	0.01	6,949	1.1M
SVM	Partners	0.58	0.30	0.86	0.01	8,602	1.8M
NBC	VUMC	0.49	0.90	0	0.01	6,949	1.1M
NBC	GHS	0.52	0.41	0.65	0.03	1,735	59K

NEXT STEPS: BOOSTING PERFORMANCE

- Sampling
- Feature engineering
- Additional features

Partners RPDR

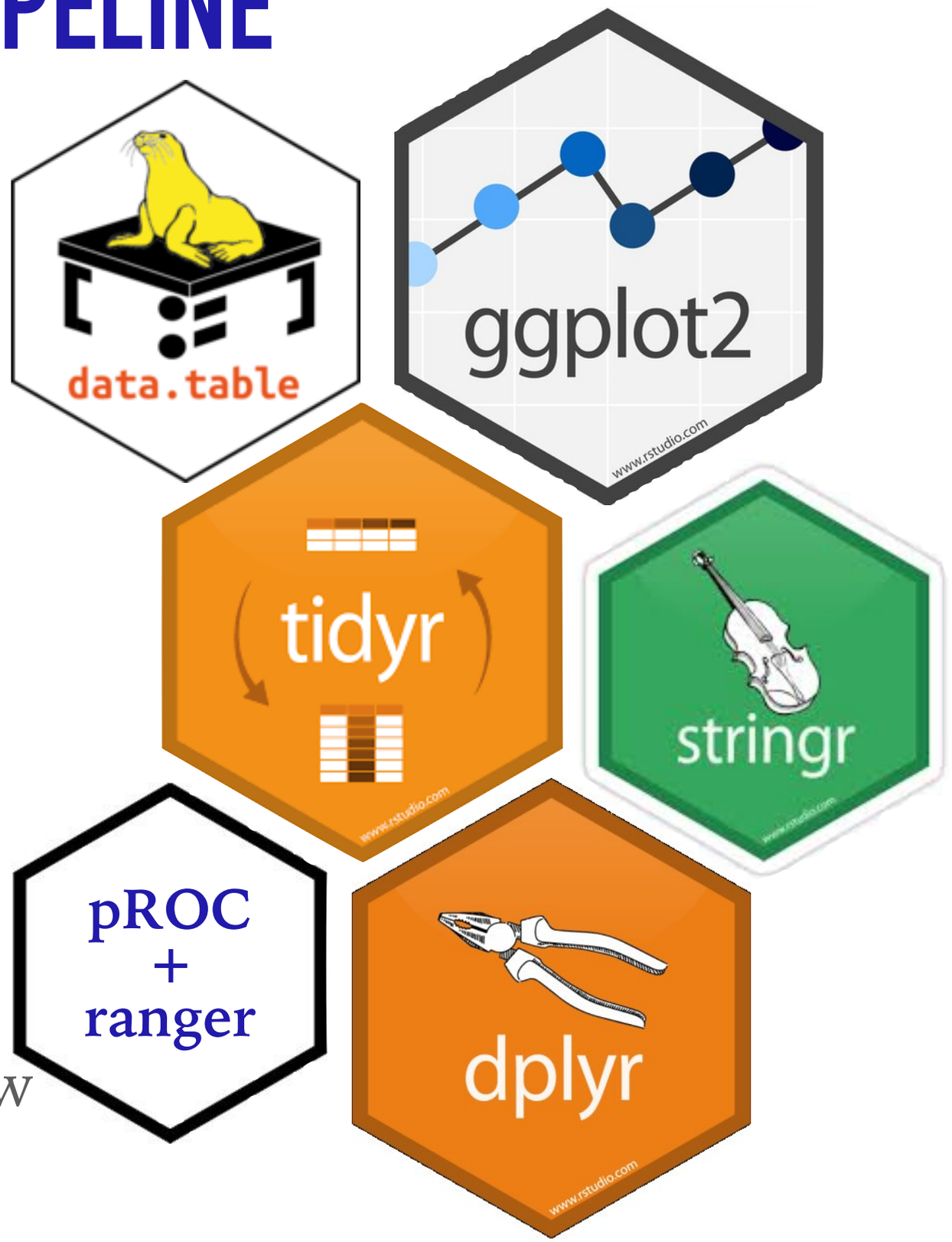
8,602 cases

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

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



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