Preliminary Findings: Predictive Modeling Of SC PLWH's Retention-in-care

Bankole Olatosi Ph.D, Sharon Weissman, MD,

Xiaowen Sun, Jiajia Zhang, Ph.D & Xiaoming Li Ph.D

02/10/2020

Funding Support

This study is supported by the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under Award Number 1R01AI127203-01A1

Outline

1. Background
   - HIV in South Carolina
2. Aim
3. Data Sources
   - Multiple data sources in SC
4. Data analysis
   - Machine learning methods
5. Conclusions & Future Direction

Exciting Times….

- Focus now moving towards aging with HIV
- Improvements in treatment for HIV (ART)
- Access to rapid and effective treatment
- Immediate initiation of ART
- Evidence of sustained viral suppression
- 90-90-90 (90 QoL) objective
- Our best chance to end the epidemic…. but

Not So Exciting Times….

- PLWH undiagnosed
- Linkage to care issues
- Retention in care issues
- Vulnerable population issues
- 90-90-90 by 2020! Nope, we didn’t make it

Throwing Down The Gauntlet…. What If

"Percentage of newly diagnosed persons achieving viral suppression within 3 months of diagnosis"

HIV Care Continuum in SC

HIV care continuum is vital in containing HIV epidemic. Based on CDC, 2019, it includes

- ✓ timely diagnosis
- ✓ linkage to care
- ✓ retention-in-care
- ✓ ART adherence
- ✓ viral suppression

HIV care continuum in SC

- ✓ 93% are linked to care within 3months, 95% are linked within 6 months and 96% are linked within 1year.
- ✓ 68% of PLWH in SC receive HIV medical care, only **53%** received **continuous** HIV medical care
- ✓ **57%** of PLWH were virally suppressed at their most recent test

Data Sources

Department of Health and Environmental Control (DHEC)

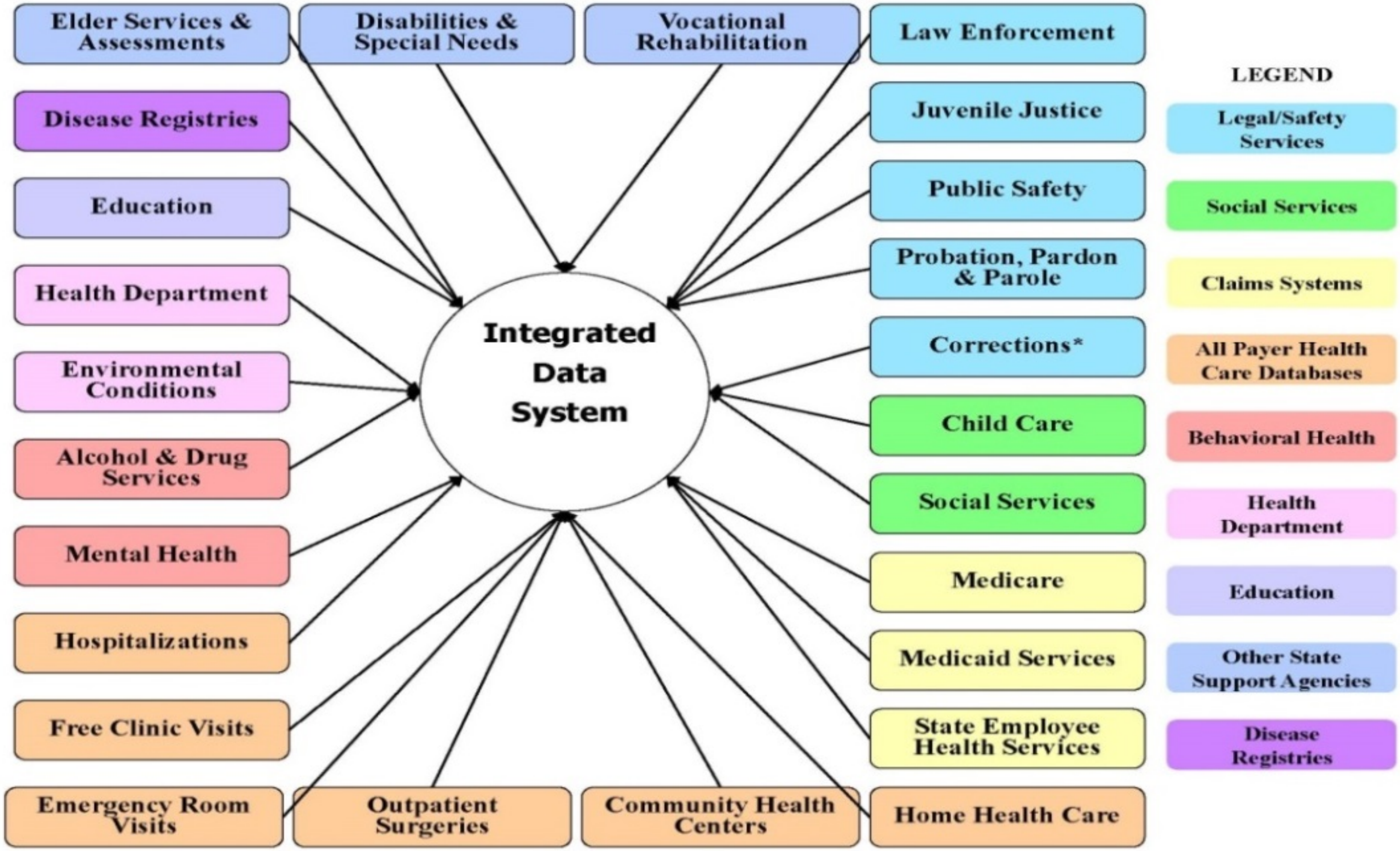
- SC statewide reporting of HIV/AIDS diagnosis began in February1986
- Reporting is done through SC DHEC HIV/AIDS electronic reporting system (e-HARS)
- e-HARS contains CD4 and viral load tests since January 1, 2004
- The Ryan White HIV/AIDS Program Data Report (RDR) provides annual report to DHEC capturing services provided

Health Sciences South Carolina (HSSC) consist of six of the state’s largest health systems

The SC Revenue and Fiscal Affairs Office (SC RFA) data oversight council collate and analyze data for different clients

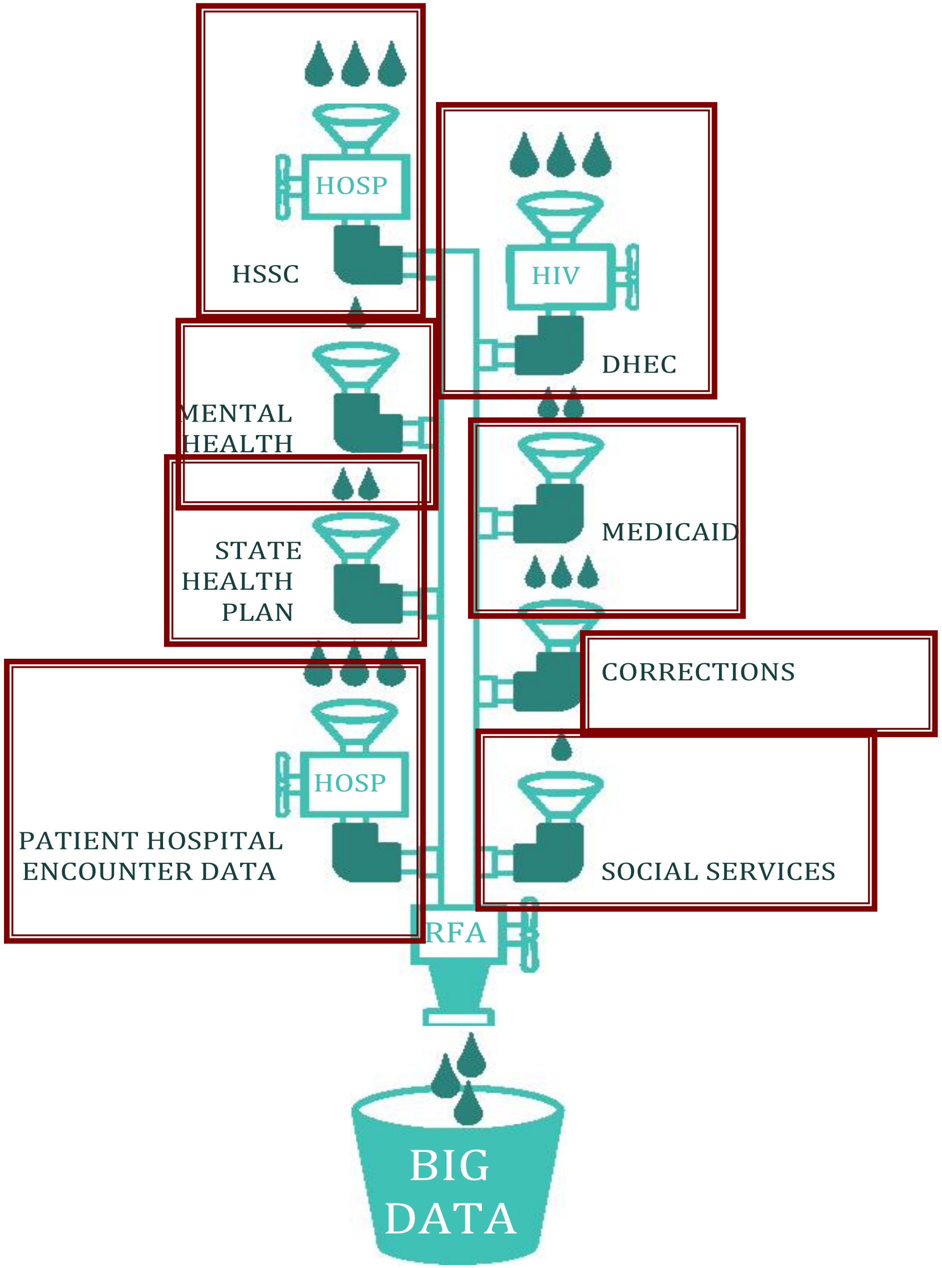
- Data linkage

Data Sources



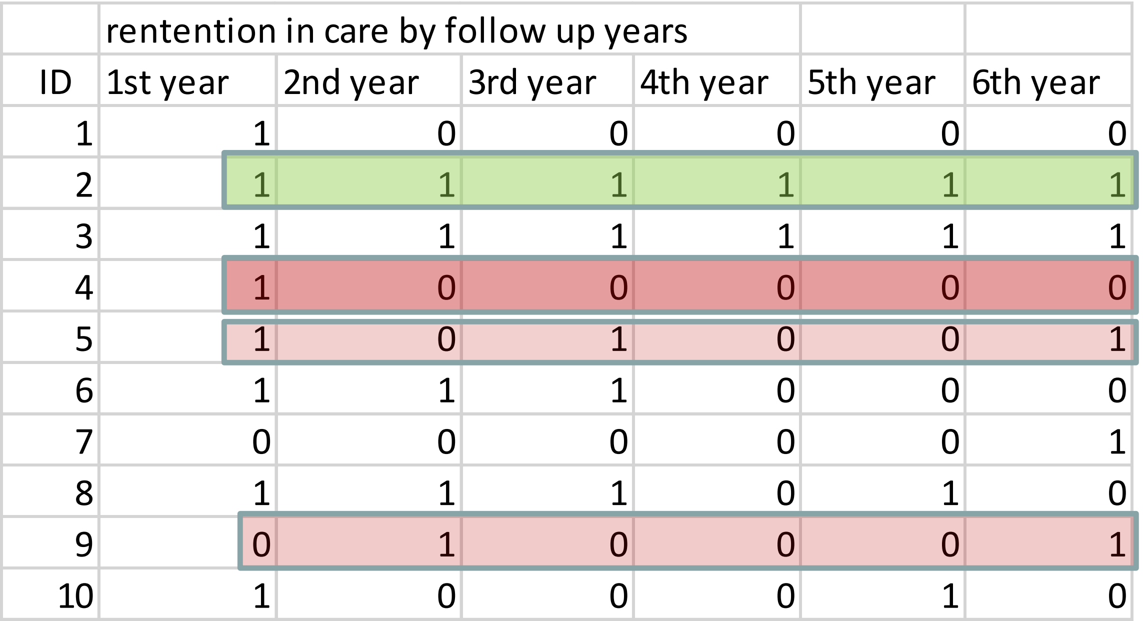
South Carolina Office of Revenue and Fiscal Affairs Integrated Data System

Data Linkage under the Big Data Project



Retention in HIV Care Status

- CDC defines retention in HIV medical care as documentation of at least 2 CD4 cell counts or viral load tests performed at least 3 months apart during the year of evaluation
- Retention in care status changes by follow up year



	rentention in care by follow up years					
ID	1st year	2nd year	3rd year	4th year	5th year	6th year
1	1	0	0	0	0	0
2	1	1	1	1	1	1
3	1	1	1	1	1	1
4	1	0	0	0	0	0
5	1	0	1	0	0	1
6	1	1	1	0	0	0
7	0	0	0	0	0	1
8	1	1	1	0	1	0
9	0	1	0	0	0	1
10	1	0	0	0	1	0

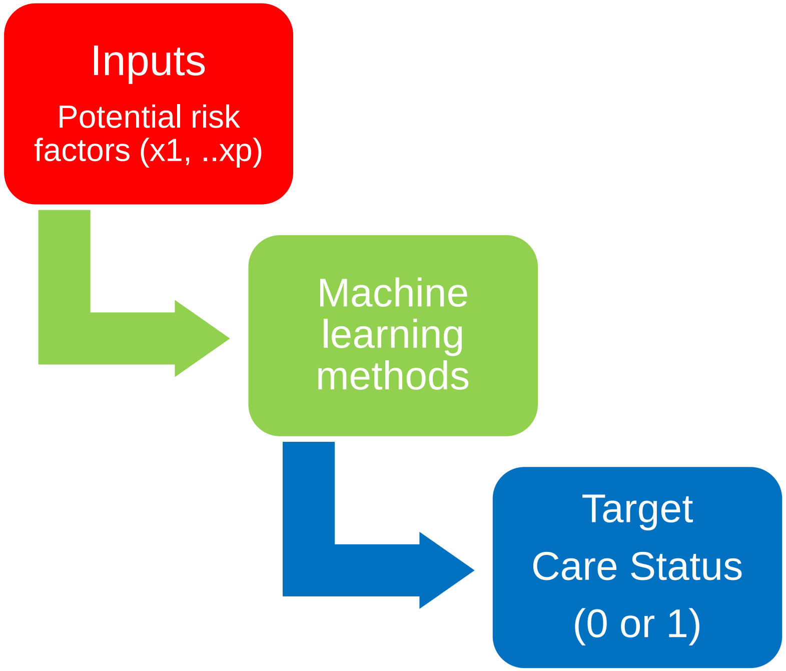
Question & Aim

Can we predict retention in care status after linkage to care?

Examine **promising** machine learning methodologies to predict retention in care at the individual level

Machine Learning

- ✓ A practical and effective approach that allows computers to learn from the past patterns/behaviors to perform a specific task
- ✓ A statistical model is built based on sample data (history data) in order to make predictions or decisions



Inputs

Potential risk factors (x1, ..xp)

Machine learning methods

Target

Care Status

(0 or 1)

Components Needed

In order to build a good prediction model, we need

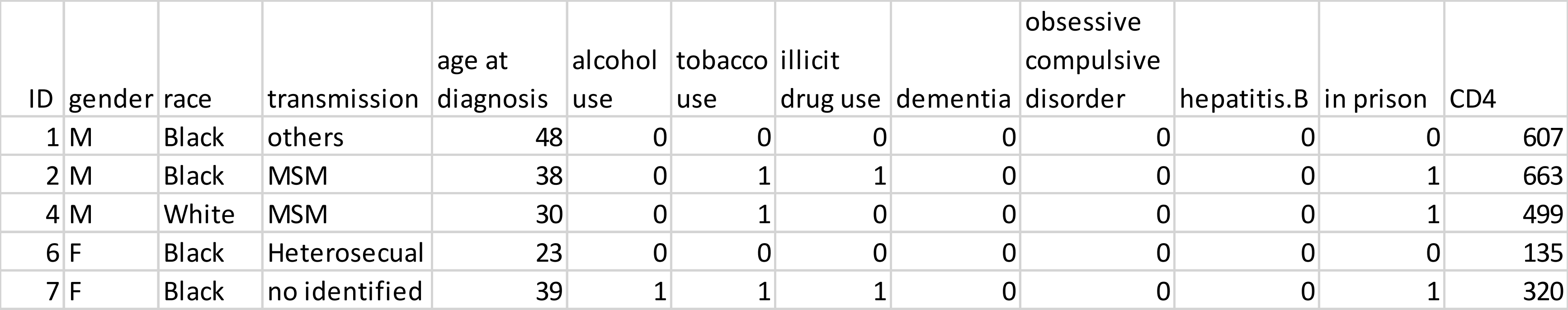
- Comprehensive ***database*** of PLWH with most potential risk factors

- Advanced modelling approach
  - Different candidate methods (5 selected)
  - For each method
    - Training, validation datasets
    - Validation procedure
  - Select the model with the best performance

- Prediction (keep tailoring model)

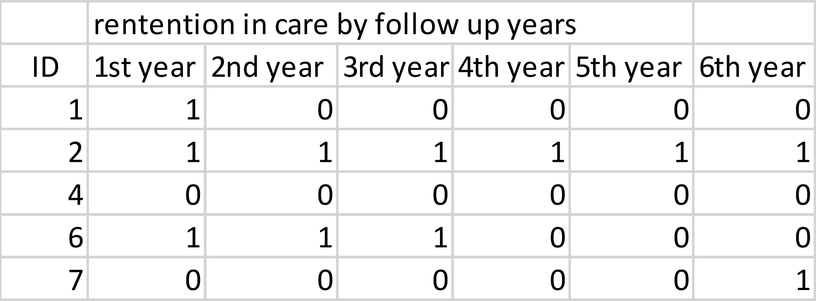
Hypothesis (Aim)

Snapshot 5 patients (observed data)

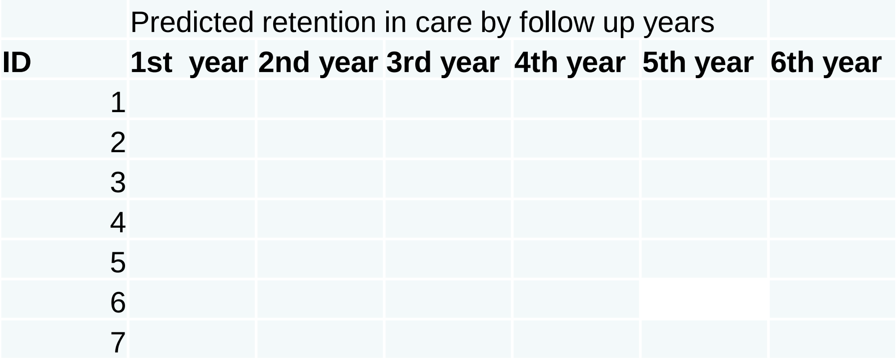


ID	gender	race	transmission	age at diagnosis	alcohol use	tobacco use	illicit drug use	dementia	obsessive compulsive disorder	hepatitis.B	in prison	CD4
1	M	Black	others	48	0	0	0	0	0	0	0	607
2	M	Black	MSM	38	0	1	1	0	0	0	1	663
4	M	White	MSM	30	0	1	0	0	0	0	1	499
6	F	Black	Heterosecual	23	0	0	0	0	0	0	0	135
7	F	Black	no identified	39	1	1	1	0	0	0	1	320

What will happen next year?
What can predict what happen next year?



	rentention in care by follow up years					
ID	1st year	2nd year	3rd year	4th year	5th year	6th year
1	1	0	0	0	0	0
2	1	1	1	1	1	1
4	0	0	0	0	0	0
6	1	1	1	0	0	0
7	0	0	0	0	0	1



	Predicted retention in care by follow up years					
ID	**1st year**	**2nd year**	**3rd year**	**4th year**	**5th year**	**6th year**
1						
2						
3						
4						
5						
6						
7						

Rationale

Current Gaps

- ✓ Lack of data
- ✓ Short time span: PLWH’s retention in care as in care status in few years (mostly 1 & 3 years)
- ✓ Mixed study populations: Including patients who were not linked to care as PLWH not in retention
- ✓ Lack of prediction models using advanced technique

Added Value of current study

- ✓ PLWH’s time span (i.e. from the patients’ diagnosis year till the most recent data)
- ✓ Splice for different groups -only include the patients linked to care
- ✓ Use machine leaning to predict PLWH retention in care overtime

Study Population

A total sample size of 8263 PLWH in SC for final analysis

- ✓ All PLWH in SC diagnosed in 2005-2016 (10025)
- ✓ Population Inclusion criteria:
  - ➢ Age ≥13 at HIV/AIDS diagnosis year (10025-44=9981)
  - ➢ With ≥1 CD4 or viral load test after the laboratory test date in the HIV diagnosis month (9981-1431=8550)
- ✓ Excluding the participants with missing information of covariate (287)

Covariates

Demographics

- ✓ Gender
- ✓ Race
- ✓ Age at diagnosis
- ✓ Driving time from home to facility
- ✓ Marital status
- ✓ Education
- ✓ CD4 cell count

HIV risk factors

- ✓ HIV transmission risk
- ✓ Alcohol use
- ✓ Tobacco use
- ✓ Illicit drug use
- ✓ HIV Opportunistic infections (Hepatitis B & C)

Covariates

Mental health condition (ICD 9)

- ✓ With one of following condition (anxiety, depression, bipolar disorder, persistent-mood affective disorder),
- ✓ Personality disorder
- ✓ Obsessive compulsive disorder
- ✓ With one of following condition (schizoaffective disorders, schizophrenia),
- ✓ Dementia

Care status in previous years

- ✓ Longitudinal care status indicator for all previous years

Data Analysis

- ✓ **LASSO:** (least absolute shrinkage and selection operator): based on linear regression, and restrict some coefficients being exactly 0
- ✓ **CART:** (classification and regression tree): classify each observation to the region of most commonly occurring class
- ✓ **Random Forest**: use trees as building blocks to construct more powerful prediction models
- ✓ **SVM:** (support vector machine): construct a hyper plane or set of hyper planes in a high or infinite dimensional space
- ✓ **KNN** (k-nearest neighbors): classified by assigning the label which is most frequent among the k training samples nearest to that query point

Data Analysis

Examine the relationship between retention in care and risk factors using 5 machine learning methodologies. For each method

- ✓ Split the data into training data set (80%) and test data set (20%)

Cross validation is used to choose the best tuning parameter

- ✓ Cross validation is the process of training: Using one set of data for training learner and testing it using a different set.
- ✓ Parameter tuning is the process of model selection in cross validation: selecting the values for a model's parameters that maximize the accuracy of the model.

The best prediction model ID chosen using the AUC criteria.

- ✓ AUC: area under the receiver operating curve (ROC) and the larger the better.

Preliminary Analysis

Preliminary Table

Covariates	Mean or Frequency(%)
Age at diagnosis:	35.5
Gender:	
Male	**6175(75%)**
Female	2088(25%)
Race:	
Black	**5963(72%)**
White	1844(22%)
Hispanic	429(5%)
Other	27(1%)
AIDS indicator:	
AIDS	4115(50%)
Not AIDS	4148(50%)
Transmission:	
MSM	**4143(50%)**
Heterosexual	1799(22%)
No Identifiable risk	1476(18%)
IDU	298(3%)
MSM/IDU	172(2%)
Others	375(5%)
Initial CD4 count	377.1

Preliminary Table

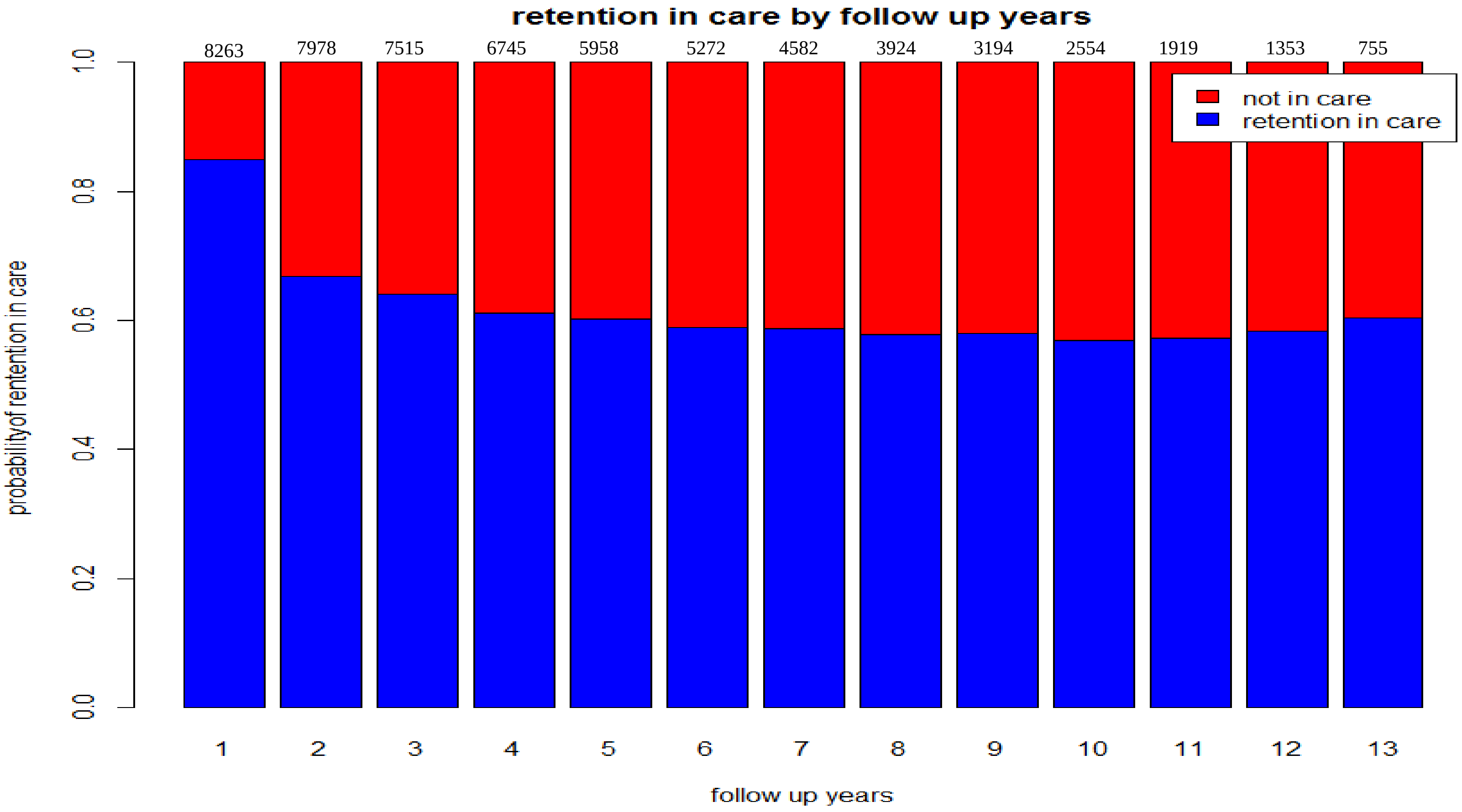
Covariates	Mean or Frequency(%)
Alcohol use:	
yes	1167(14%)
no	7096(86%)
Tobacco use:	
yes	3948(48%)
no	4315(52%)
Illicit drug use:	
yes	1433(17%)
no	6830(83%)
Hepatitis C:	
yes	394(5%)
no	7869(95%)
Hepatitis B:	
yes	185(2%)
no	8078(98%)
In prison:	
yes	616(7%)
no	7647(93%)

Preliminary Table

Covariates:	Frequency
With one of following condition (anxiety, depression, bipolar disorder, persistent-mood affective disorder	
Yes	**2230(27%)**
No	6033(73%)
With one of following condition (schizoaffective disorders, schizophrenia)	
Yes	257(3%)
No	8006(97%)
Personality disorder:	
Yes	190(2%)
No	8073(98%)
Obsessive compulsive disorder:	
Yes	23(1%)
No	8240(99%)
Dementia:	
Yes	118(2%)
No	8145(98%)

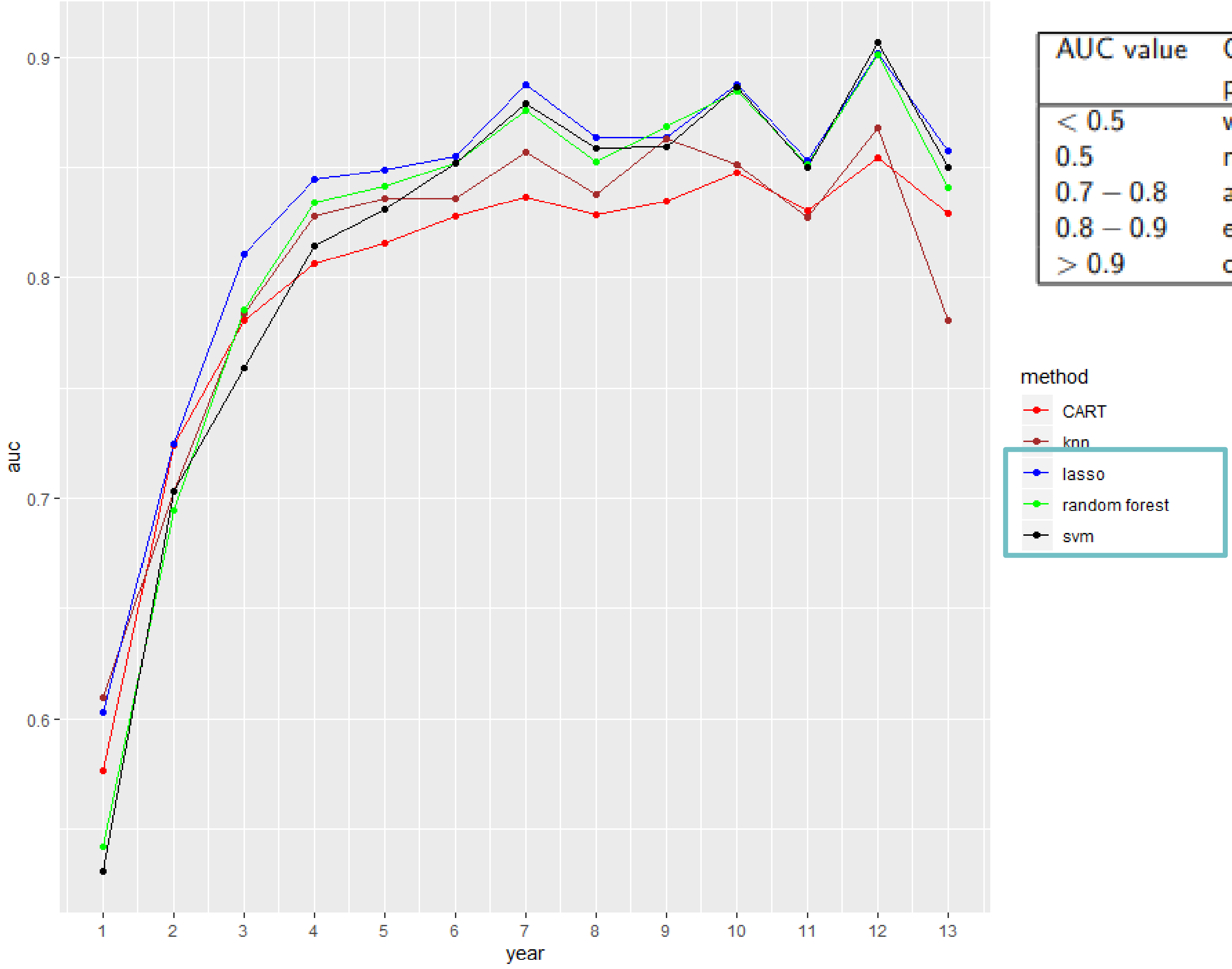
Retention in care by follow up years

Outcome Variables: Retention in care is defined as having ≥2 CD4 or VL results at least 3 months apart after having been diagnosed with HIV.



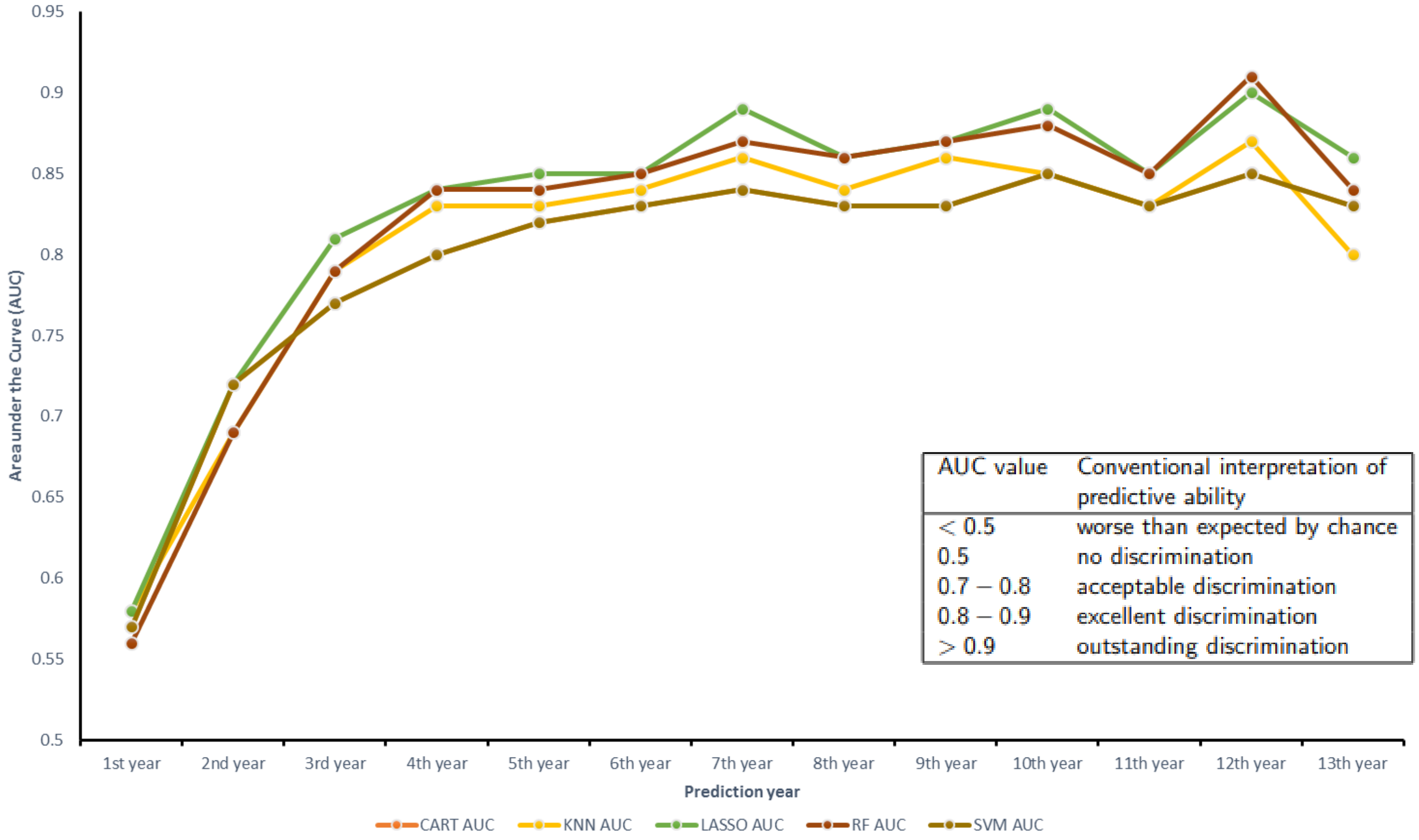
Data analysis for overall sample

Validation based on AUC



AUC value	Conventional interpretation of predictive ability
< 0.5	worse than expected by chance
0.5	no discrimination
0.7 – 0.8	acceptable discrimination
0.8 – 0.9	excellent discrimination
> 0.9	outstanding discrimination

Fig 1: AUC Model Performance Comparison for Machine Learning Models Use for Individual Prediction of Care Status, SC PLWH 2005-2016



AUC value	Conventional interpretation of predictive ability
< 0.5	worse than expected by chance
0.5	no discrimination
0.7 – 0.8	acceptable discrimination
0.8 – 0.9	excellent discrimination
> 0.9	outstanding discrimination

- Predictive performance improved over time (AUC > 0.80) by year 4 for all algorithms.
- By year 12, RF, LASSO, and CART were the top model performers based on AUC (Fig 1).

Prediction based on Lasso

The most important variables were obtained by ranking the absolute value of the coefficients.

1st year

1	obsessive compulsive disorder
2	transmission
3	race
4	with AIDS
5	mental health group 1

2nd year

1	whether in care 1st year
2	dementia
3	personality disorder
4	mental health group 1
5	race

3rd year

1	whether in care 2nd year
2	whether in care 1st year
3	transmission
4	hepatitis.B
5	mental health group 1

6th year

1	whether in care 5th year
2	obseesive compulsive disorder
3	whether in care 4th year
4	whether in care 3th year
5	race

9th year

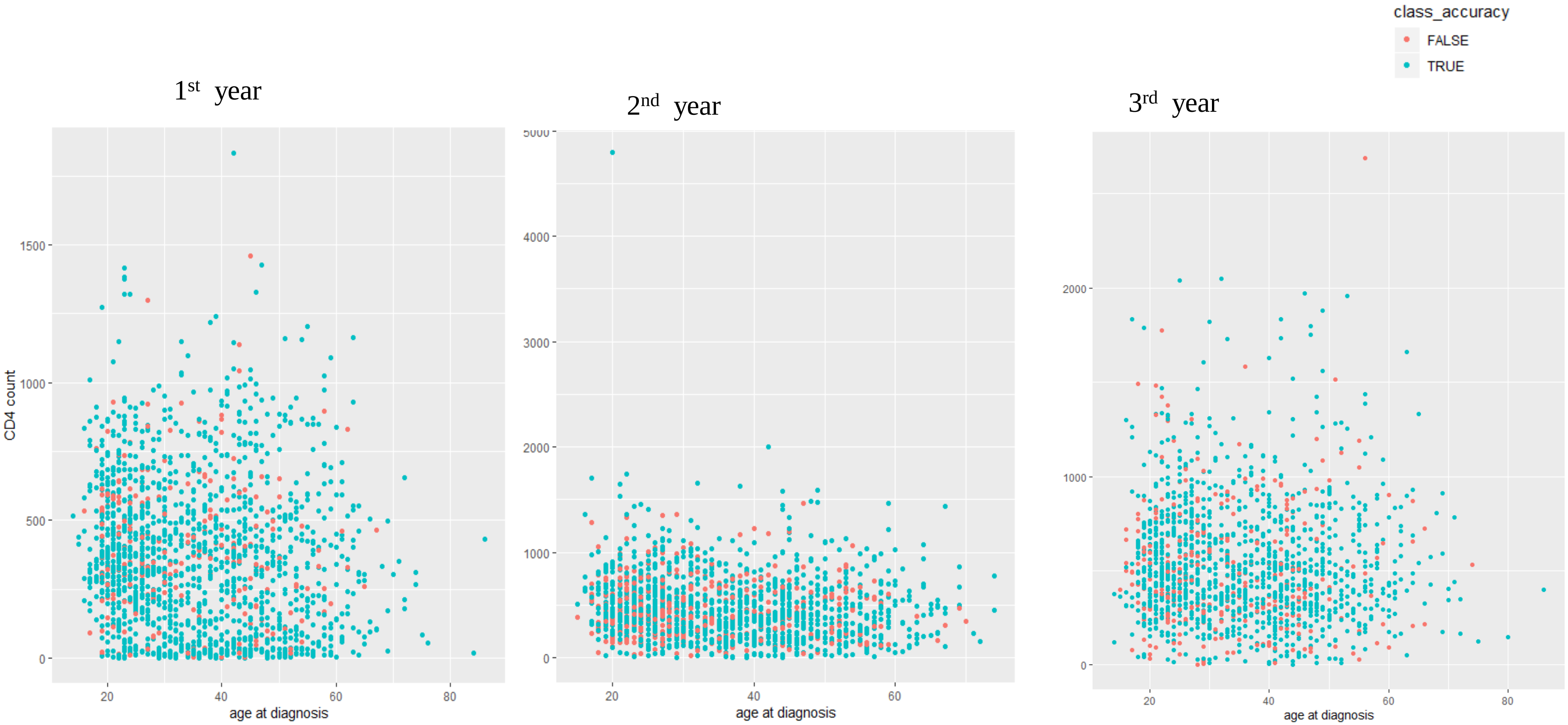
1	whether in care 8th year
2	whether in care 7th year
3	whether in care 6th year
4	hepatitis.B
5	whether in care 3th year

12th year

1	obsessive compulsive disorder
2	whether in care in 11th year
3	whether in care in 10th year
4	whether in care in 3rd year
5	transmission

Mental health group 1: with one of following condition (anxiety, depression, bipolar disorder, persistent-mood affective disorder)

Prediction based on Lasso



1st year

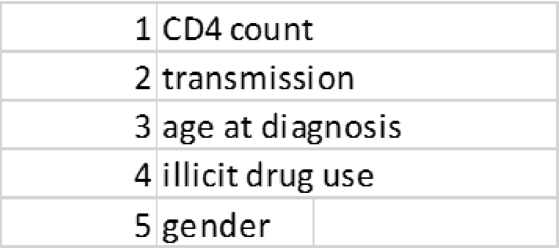
2nd year

3rd year

Prediction based on Random Forest

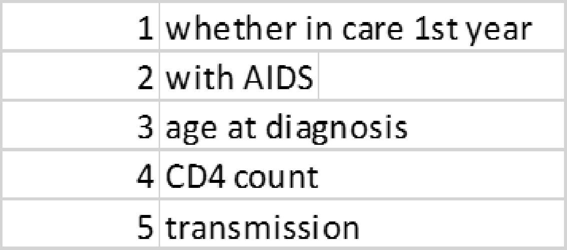
The most important variables is based on how much the accuracy decreases when the variable is excluded.

1st year



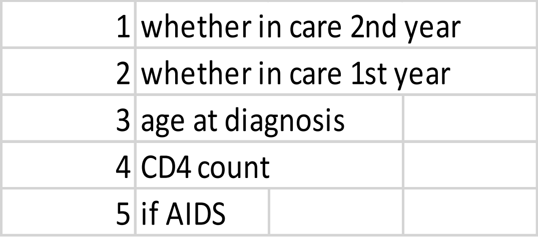
1	CD4 count
2	transmission
3	age at diagnosis
4	illicit drug use
5	gender

2nd year



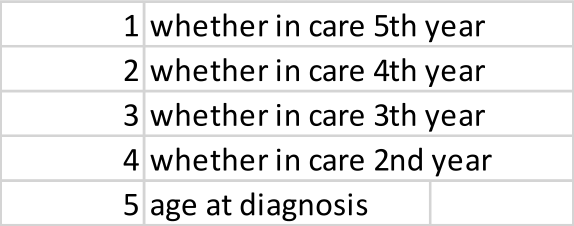
1	whether in care 1st year
2	with AIDS
3	age at diagnosis
4	CD4 count
5	transmission

3rd year



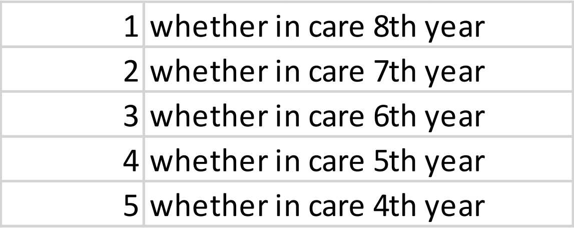
1	whether in care 2nd year
2	whether in care 1st year
3	age at diagnosis
4	CD4 count
5	if AIDS

6th year



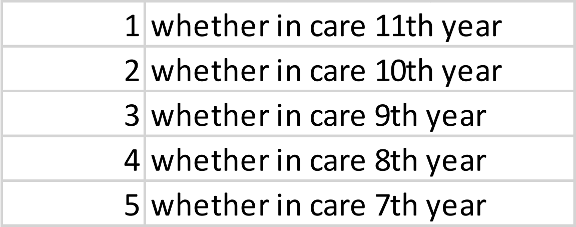
1	whether in care 5th year
2	whether in care 4th year
3	whether in care 3th year
4	whether in care 2nd year
5	age at diagnosis

9th year



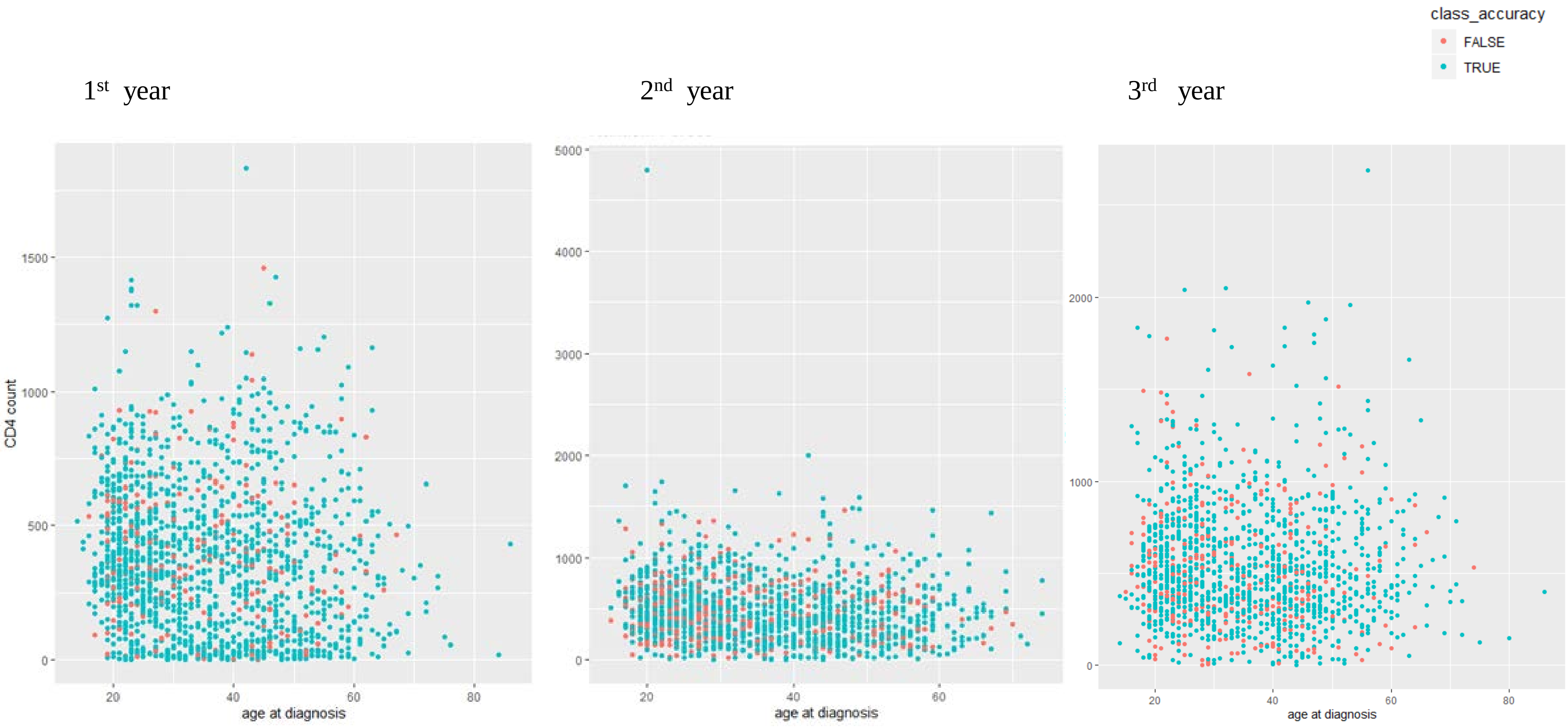
1	whether in care 8th year
2	whether in care 7th year
3	whether in care 6th year
4	whether in care 5th year
5	whether in care 4th year

12th year



1	whether in care 11th year
2	whether in care 10th year
3	whether in care 9th year
4	whether in care 8th year
5	whether in care 7th year

Prediction based on Random Forest



1st year

2nd year

3rd year

Data analysis for male

Prediction based on Lasso for Male

Most important variables varied by time

1st year

1	obsessive compulsive disorder
2	transmission
3	Race
4	with AIDS
5	mental health group 1

2nd year

1	whether in care 1st year
2	obsessive compulsive disorder
3	dementia
4	race
5	transmission

3rd year

1	whether in care 2nd year
2	whether in care 1st year
3	transmission
4	mental health 1
5	if AIDS

6th year

1	whether in care 5th year
2	whether in care 4th year
3	obsessive compulsive disorder
4	whether in care 3 year
5	race

9th year

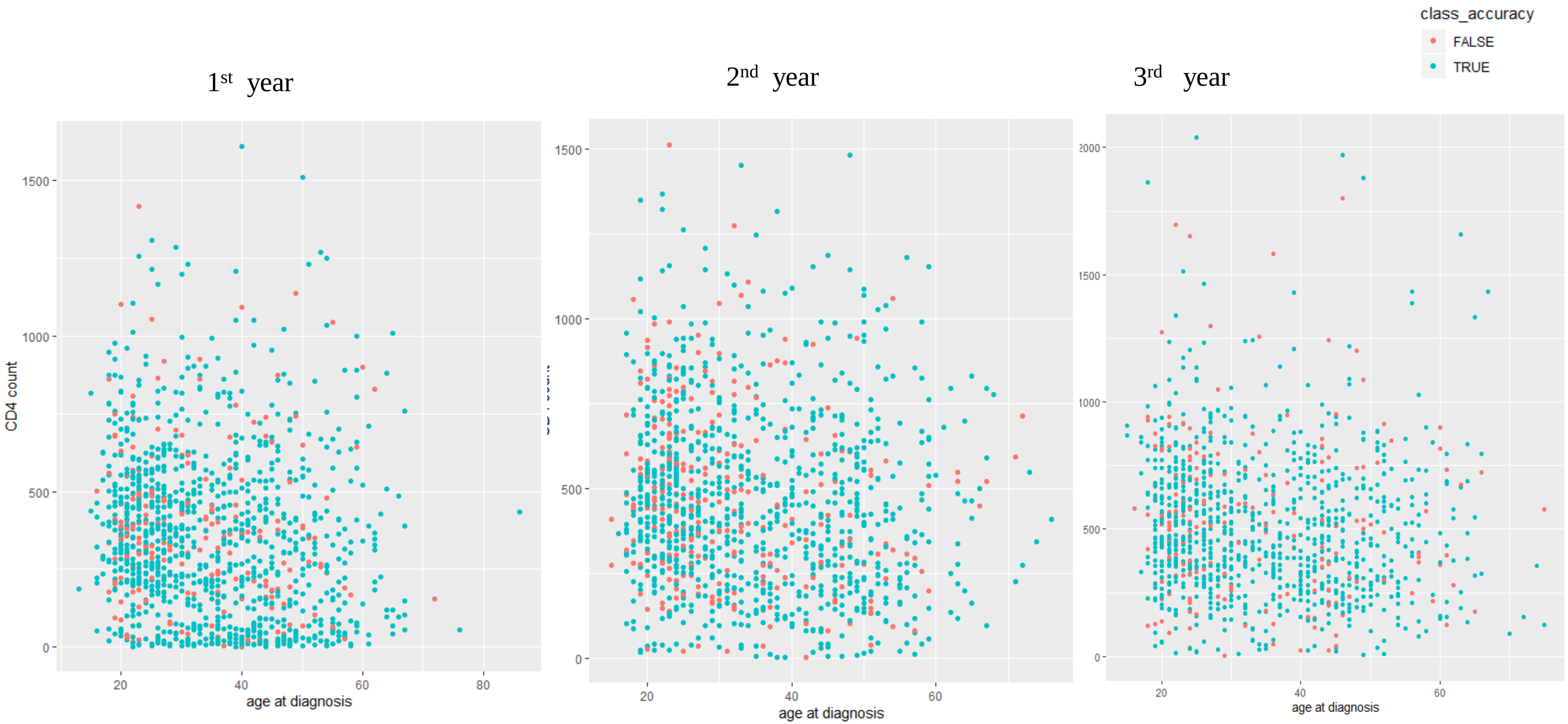
1	obsessive compulsive disorder
2	whether in care 8th year
3	whether in care 7th year
4	hepatitis.B
5	illicit drug use

12th year

1	whether in care 11th year
2	obsessive compulsive disorder
3	whether in care 10th year
4	whether in care 3rd year
5	whether in care 8th year

Mental health group 1: with one of following condition (anxiety, depression, bipolar disorder, persistent-mood affective disorder)

Prediction based on Lasso for Male



1st year

2nd year

3rd year

Prediction based on Random Forest for Male

1st year

1	age at diagnosis
2	cd4 count
3	illicit drug use
4	with AIDS
5	transmission

2nd year

1	whether in care 1st year
2	cd4 count
3	if aAIDS
4	age at diagnosis
5	transmission

3th year

1	whether in care 2nd year
2	whether in care 1st year
3	age at diagnosis
4	cd4 count
5	transmission

6th year

1	whether in care 5th year
2	whether in care 4th year
3	whether in care 3th year
4	whether in care 2th year
5	whether in care 1th year

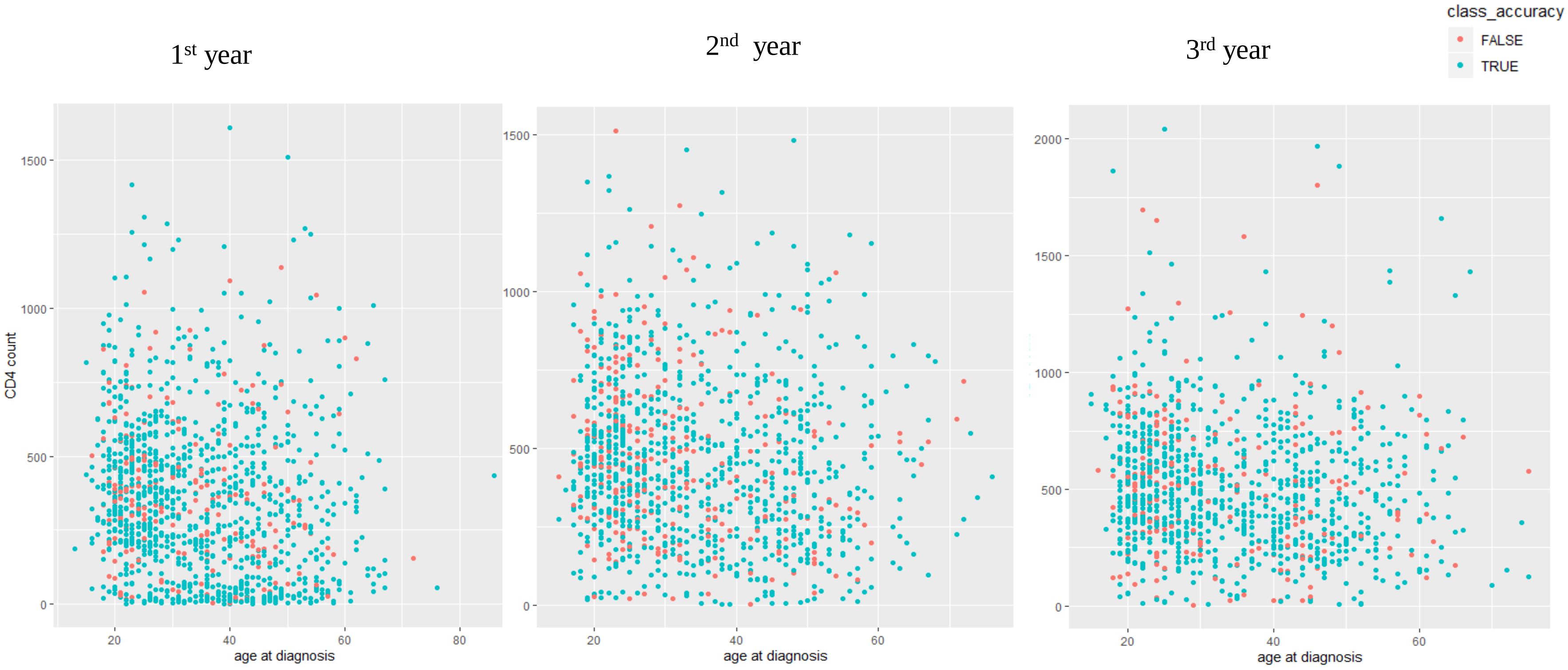
9th year

1	whether in care 8th year
2	whether in care 7th year
3	whether in care 6th year
4	whether in care 5th year
5	whether in care 4th year

12th year

1	whether in care 11th year
2	whether in care 10th year
3	whether in care 9th year
4	whether in care 8th year
5	whether in care 7th year

Prediction based on Random Forest for Male

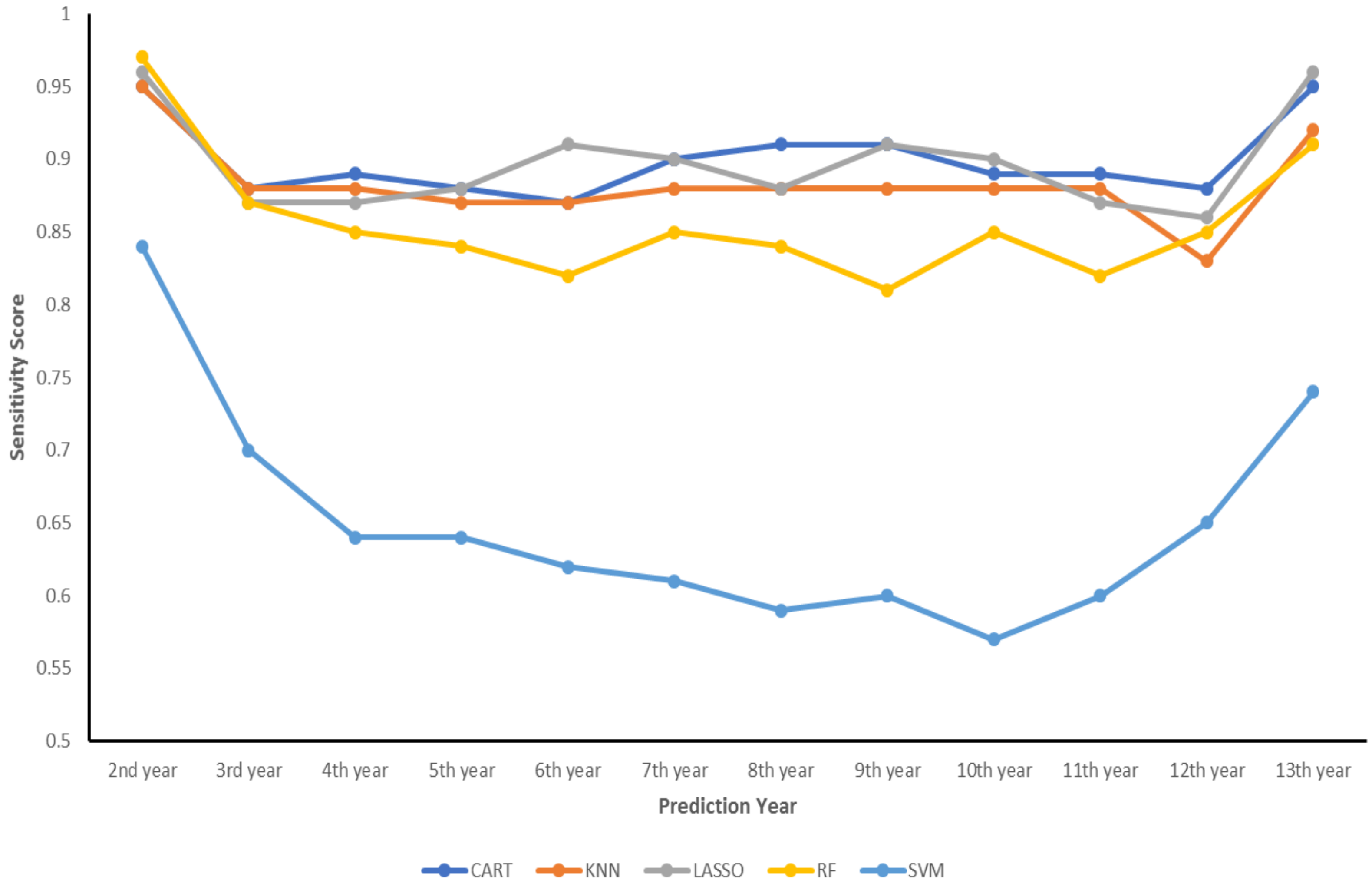


1st year

2nd year

3rd year

Fig 2: Sensitivity Comparisons for Machine Learning Models Used for Individual Prediction of Care Status, SC PLWH 2005-2016



Discussion

Longitudinal prediction of the HIV care status

- AUC curves summarize the prediction accuracy for each method by year. The prediction accuracy improves by year. After 3rd year, the prediction accuracy is large enough for practical use.

- The most important factors to predict the retention in care changes by time also. After the third year, the retention in care history is a good indicator for the next year retention status.

- More potential factors will be needed to improve the prediction accuracy for the first three-year HIV care status prediction.

Ongoing directions

Longitudinal prediction of the HIV care status

- Pre-diagnosis –Missed opportunities
- Post-diagnosis –Missed opportunities
- Predictive algorithms (clusters and individuals)

Thank You