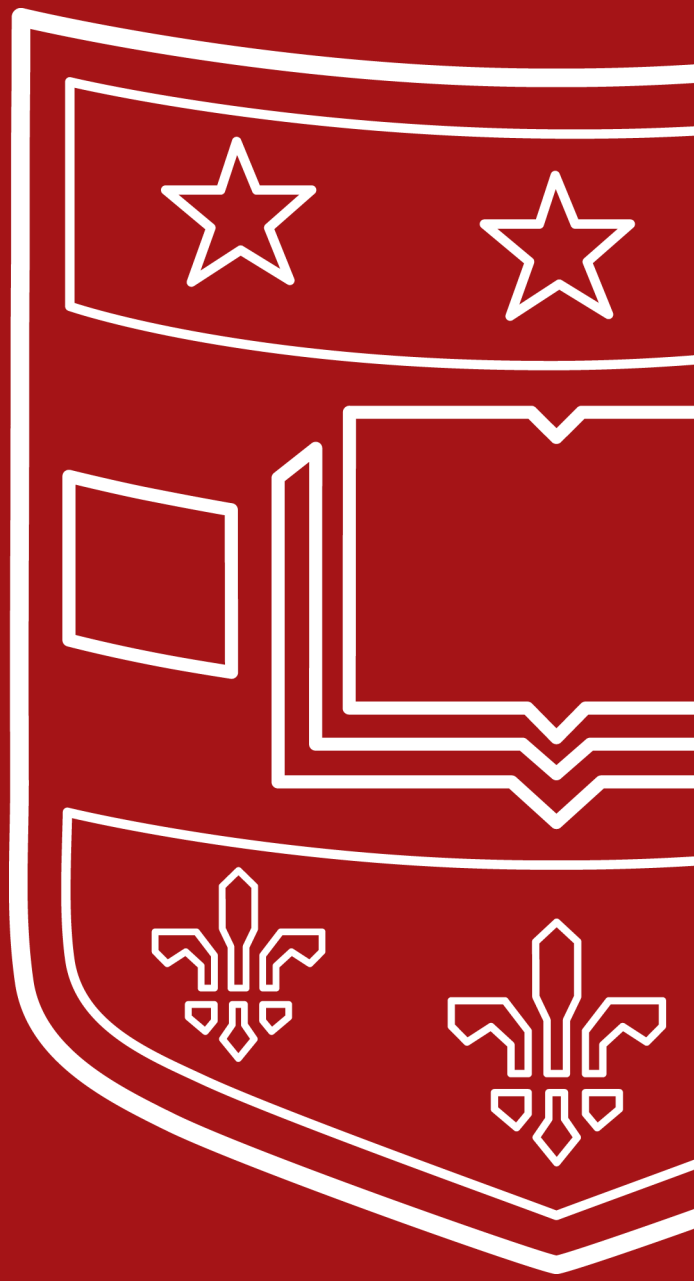
# Applying machine learning techniques in STATA to predict poor health outcomes using HIV-related data

Rachel Brathwaite, PhD

CAPS Methods Core Town Hall, UCSF

December 14th, 2021



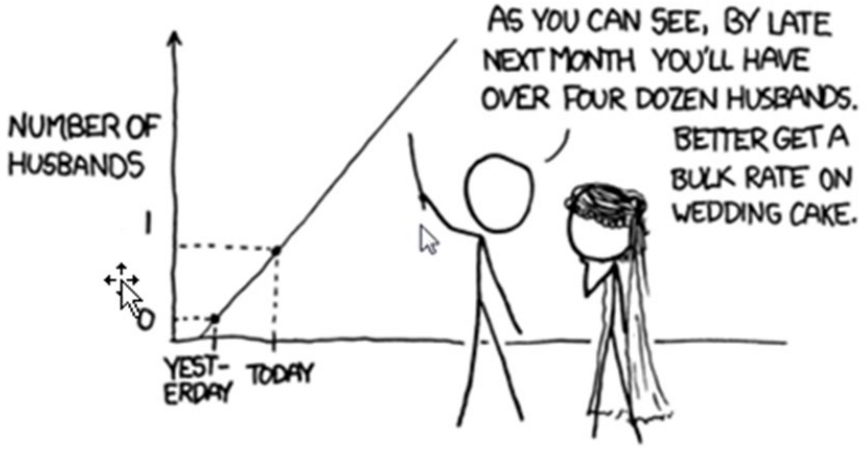
Washington University in St. Louis

# Outline

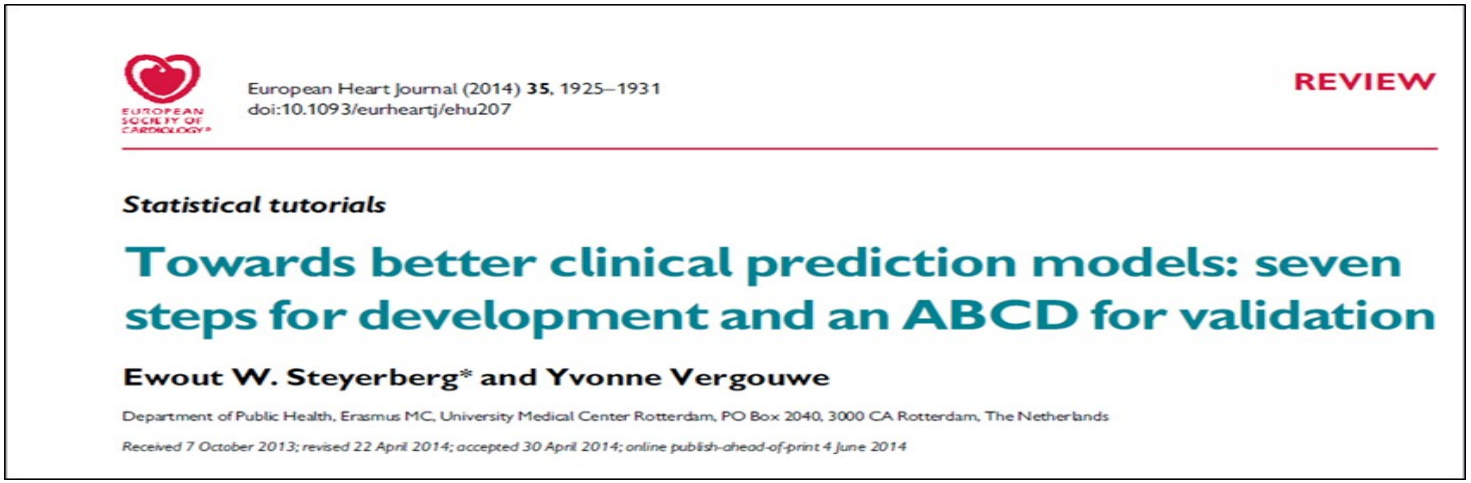
- Brief introduction to prediction modeling
- Brief background and introduction to hypothetical research question
- Lasso regression in STATA 17 (commands and interpretation of results)

# Prediction modeling definition and applications

- Prediction models are tools that predict an individual's risk of developing a health outcome.
- A prediction model/risk calculator is one tool that can provide estimated absolute risks of a future event for individuals based on their unique combination of characteristics (Moons et al., 2009).

- Prediction modeling involves a series of steps designed to create the best model to predict an outcome.



European Heart Journal (2014) 35, 1925–1931
doi:10.1093/eurheartj/ehu207

REVIEW

*Statistical tutorials*

## Towards better clinical prediction models: seven steps for development and an ABCD for validation

**Ewout W. Steyerberg* and Yvonne Vergouwe**

Department of Public Health, Erasmus MC, University Medical Center Rotterdam, PO Box 2040, 3000 CA Rotterdam, The Netherlands

*Received 7 October 2013; revised 22 April 2014; accepted 30 April 2014; online publish-ahead-of-print 4 June 2014*

# 7 Steps in development of prediction model *(Steyerberg & Vergouwe, 2014)*

| Purpose | Description |
|---|---|
| 1. Problem definition and data inspection | Understand research question, outcome, define predictors, understand data available |
| 2. Coding of predictors | How to code predictors in the dataset, categorical or continuous |
| 3. Model specification | Which methods to use for selecting predictors (Lasso) |
| 4. Model estimation | Run the model to obtain regression coefficients for predictors |
| 5. Model performance | Evaluate how well the model works, discrimination and calibration |
| 6. Model validation | Conducting internal validation to reduce likelihood of overfitting |
| 7. Model presentation | How will model be presented for use- formula, computerized program, online version etc. |

# Applications of prediction models

- Can inform clinicians about:
  - Risk of developing a disease
  - Risk of presence of a disease
  - Risk of future course of an illness
- Applications:
  - Screening- identify signs of disease early
  - Personalized medicine- supports individualized clinical decision making. Who is most likely to benefit from treatment, Who is most at risk of adverse outcomes?
  - Minimize costs to healthcare system (efficient use of resources)

# Questions to ask prior to developing the model

1. Who is the target population we want to make predictions for?
   (Adolescents living with HIV in Uganda)
2. What is the outcome we wish to predict?
   (depression)
3. What statistical technique we wish to use?
   (Lasso regression)

*Prediction- Interested in correlation not causation

# Background to research question

- Adolescence is a critical developmental period
  - Increased vulnerability to neurological, mental and substance-use disorders
- Young people affected by HIV/AIDS in Uganda and other SSA countries are expected to face additional challenges as they transition through adolescence and young adulthood while living in resource-poor settings
- Living with HIV elevates one's risk of poor mental health outcomes

# Background

- Maintaining good mental health in childhood and adolescence is essential for avoiding poor mental health functioning, even suicide.
- For ALWHIV, Depression is associated with:
  - Poor linkage to care
  - Failing to initiate and adhere to ART medications
  - Increased viral load, susceptibility to opportunistic illnesses
  - Increased engagement in risky sexual behaviors—transmission to others

# Background

- Given that in SSA, there is only one child psychiatrist for every 4 million people.
- Alternative non-clinical tools that can help identify/predict which ALWHIV are at higher risk of being depressed are needed.
- Individualized risk assessment approaches have the potential to advance our understanding, specifically in identifying adolescents at risk for poor health outcomes even before these outcomes occur.

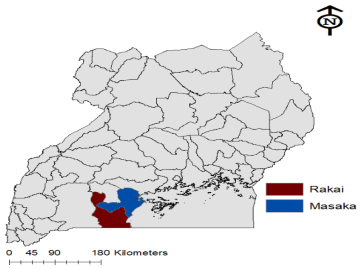
# Background

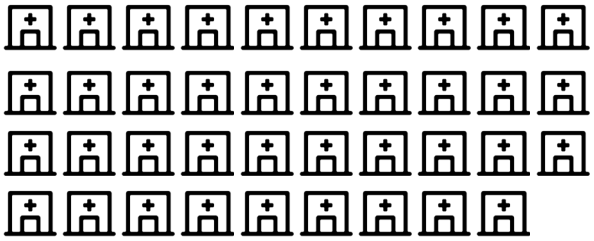
- This can help in identifying at-risk adolescents, and also to develop and/or implement interventions to mitigate this risk.
- A prediction model/risk calculator is one tool that can provide the estimated risk of depression for ALWHIV based on their unique combination of characteristics

Screening tool?

# Dataset: Suubi+Adherence study

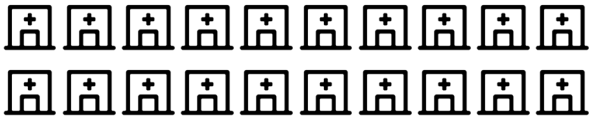
Recruitment areas in Uganda heavily affected by HIV/AIDS



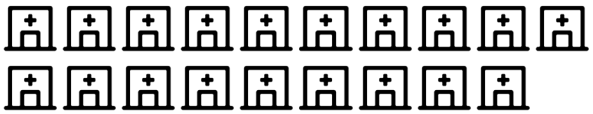



39 clinics

randomized



20 clinics to Control condition



19 clinics to Intervention condition

'Suubi+Adherence' dataset
702 adolescents living with HIV aged 10-16 at baseline

| Baseline 2012 | 12 months 2013 | 24 months 2014 | 36 months 2015 | 48 months 2016 |
|---|---|---|---|---|

# Inclusion criteria for Suubi+Adherence

1. Medically diagnosed with HIV and aware of their HIV status;
2. Living within a family (could be biological family or caregiver, but not an institution);
3. Aged 10 to 16 at baseline;
4. Prescribed ART medication and;
5. Receiving HIV care and treatment at one of 39 clinics enrolled.

# 7 Steps in prediction modelling

*(Steyerberg & Vergouwe, 2014)*

1. Define the prediction problem: define predictors and outcome of interest
2. Code predictors
3. Specify a model
4. Estimate model parameters
5. Model evaluation
6. Model validation
7. Presentation of the model

# Step 1: Defining problem, predictors and outcome of interest

- Research Question: Which combination of multi-level factors can predict depression in one year's time among ALWHIV in Uganda?
- Outcome: Depression (as measured by the 14-item version of the Children's Depression Inventory)
- In this hypothetical example depression is defined as those scoring above the mean depression score in the dataset.

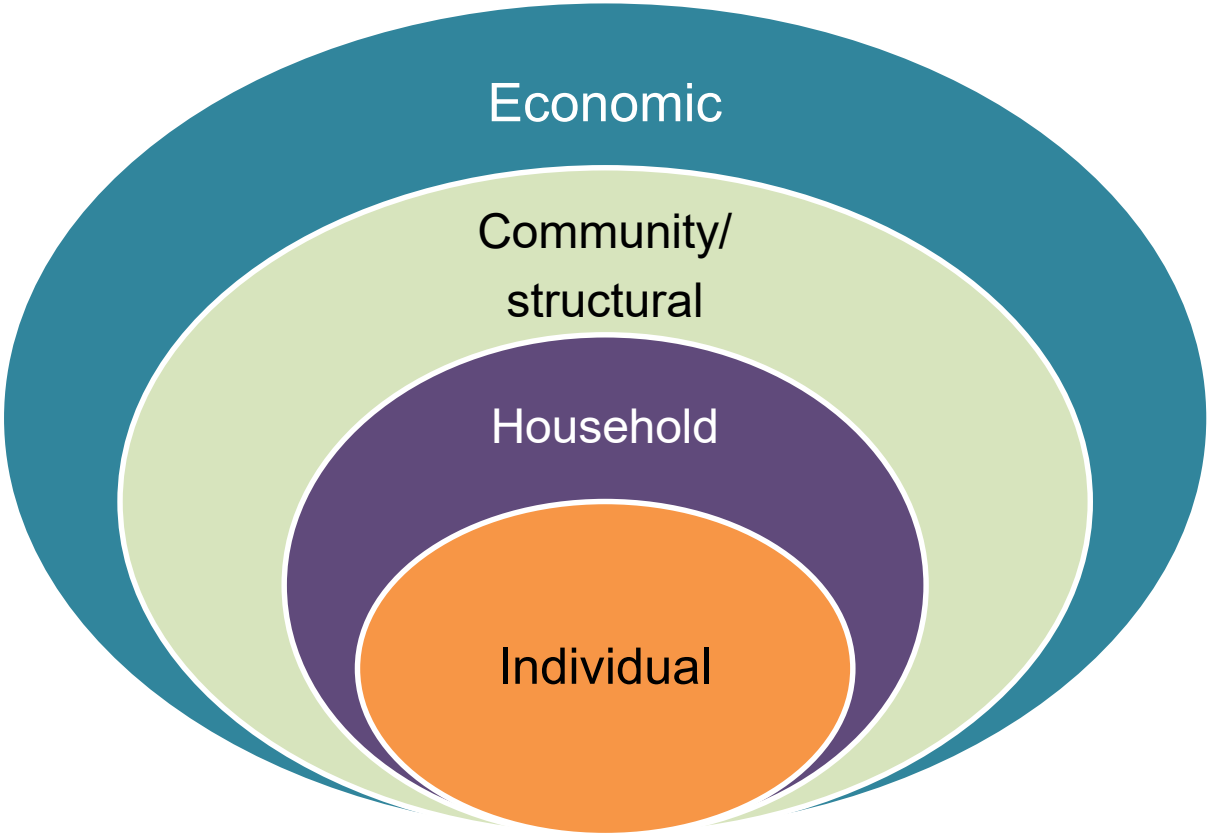
# Description of the dataset

```
. tab depression if depression !=. & disclosure_status !=. & HIV_stigma !=. & adherence_s
> elf_efficacy !=. & adherence_history !=. & hopelessness !=. &  caregiver !=. & social_s
> upport !=. & family_cohesion !=. & agegroup !=. & sex !=. & distance !=. & substance_us
> e !=. & ART_buddy !=. & intervention_group !=. & child_poverty !=. & assets !=.
```

| depression | Freq. | Percent | Cum. |
|---|---|---|---|
| notdepressed | 346 | 61.46 | 61.46 |
| depressed | 217 | 38.54 | 100.00 |
| Total | 563 | 100.00 | |

- We will use the dataset comprising complete cases in this example (N=563)
- 217 ALWHIV with depression outcome (hypothetical example)

# Defining predictors




Using Modified Social Ecological Model as a framework- multi-level factors influence outcome
Question: What is already known about the predictors from the literature?

# Defining predictors

| INDIVIDUAL LEVEL | |
|---|---|
| Demographics | |
| Age group | 13-17=0/≥18=1 |
| Biological sex | Male=0/female=1 |
| Behavioral | |
| Substance use | Never used=0/used=1 |
| History of ART Adherence | Good adherence=0/poor adherence=1 |
| Psychosocial | |
| Hopelessness | 20-item Beck Hopelessness Scale total score |
| Adherence self-efficacy | 12-item HIV treatment adherence Self-Efficacy Scale total score |
| HIV disclosure | None=0/few or some or all=1 |
| Health-related | |
| Viral load | <40 copies per ml=0/≥40=1 |

# Defining predictors

| HOUSEHOLD LEVEL | |
|---|---|
| ART treatment supporter | Yes=1/No=0 |
| Family cohesion | 6-item family environment scale total score |
| Caregiver type | Biological=0/non-biological=1 |

# Defining predictors

| COMMUNITY/STRUCTURAL LEVEL | |
|---|---|
| HIV-related stigma | 9-items from Berger Stigma Scale total score |
| Social Support Network | 12-items from adapted Social Support Behaviors Scale total score |
| Distance to health facility | Distance from home to hospital/health clinic: very near to near=0/ very far to far=1 |

# Defining predictors

| ECONOMIC LEVEL | |
|---|---|
| Asset ownership | 20-item asset index: high possession=0/low possession=1 |
| Child poverty | 6-item |
| Economic group assignment | Control=0/Intervention=1 |

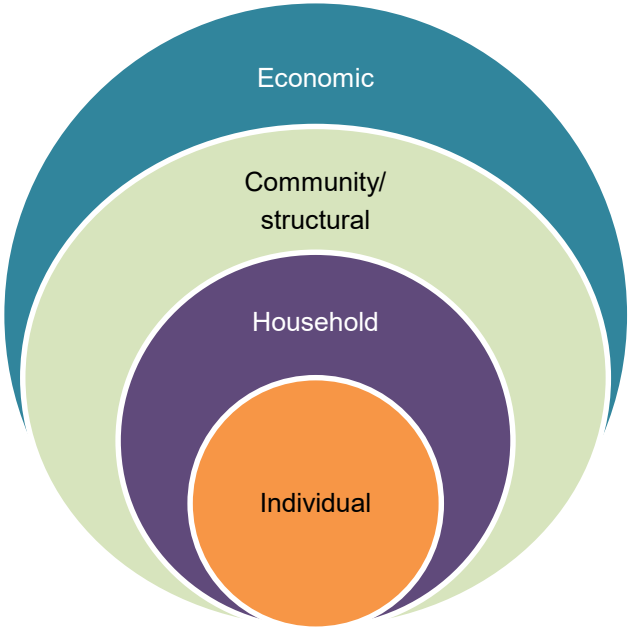
# 7 Steps in prediction modelling

*(Steyerberg & Vergouwe, 2014)*

1. Define the prediction problem: define predictors and outcome of interest
2. Code predictors
3. **Specify a model**
4. Estimate model parameters
5. Model evaluation
6. Model validation
7. Presentation of the model

# Step 3. Specify a model

- As described above we are referring to previous studies and clinical knowledge (guided by the social ecological model) to decide which predictors to include in the full model
- Predictors selected using lasso regression



# 7 Steps in prediction modelling

*(Steyerberg & Vergouwe, 2014)*

1. Define the prediction problem: define predictors and outcome of interest
2. Code predictors
3. Specify a model
4. **Estimate model parameters**
5. Model evaluation
6. Model validation
7. Presentation of the model

# Step 4: Model estimation

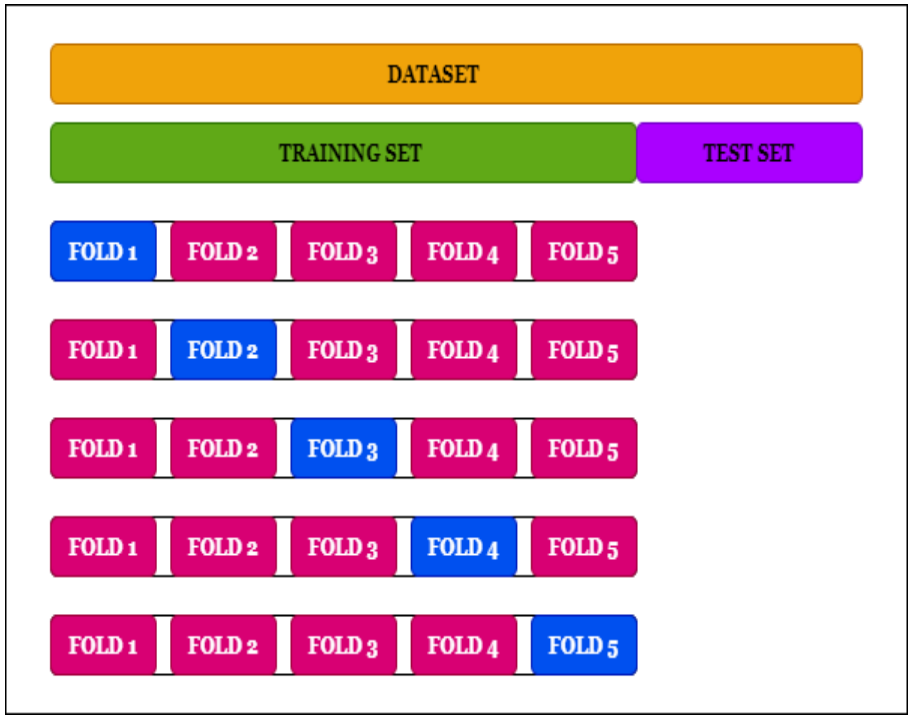
- All predictors may not be contributing to the outcome.
- We are interested in selecting the best subset of predictors of depression from the list above, hence we use the least absolute shrinkage selection operator (LASSO)
- Since we have a binary outcome (depressed vs not depressed) we specify a lasso logistic regression model

# Lasso vs Ordinary least square regression

- LASSO is a supervised machine learning method for prediction.
- In traditional Ordinary Least Square regression (coefficients estimated by minimizing least square, all predictors remain in the model, add variance to prediction of outcome)
- LASSO determines which predictors are relevant for predicting the outcome by applying a penalty term (lambda) to least square. This causes some regression coefficients to shrink to zero, excluding them from the model, resulting in a simpler model. As lambda increases, more variables excluded from model

# Cross-validation (resampling technique)

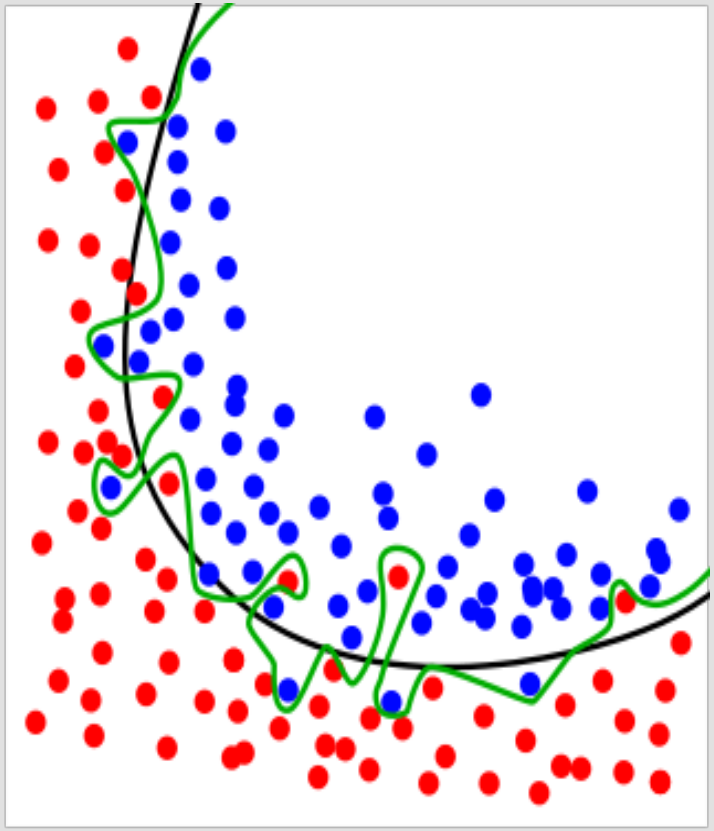
- Prediction modelling is concerned with how well model will perform in new cases (generalizability)
- k-fold (10-fold) cross-validation helps generate a more realistic estimate of predictive performance in new cases (allows model to learn underlying distribution better)
- CV prevents overfitting, improves ability to generalize to new data



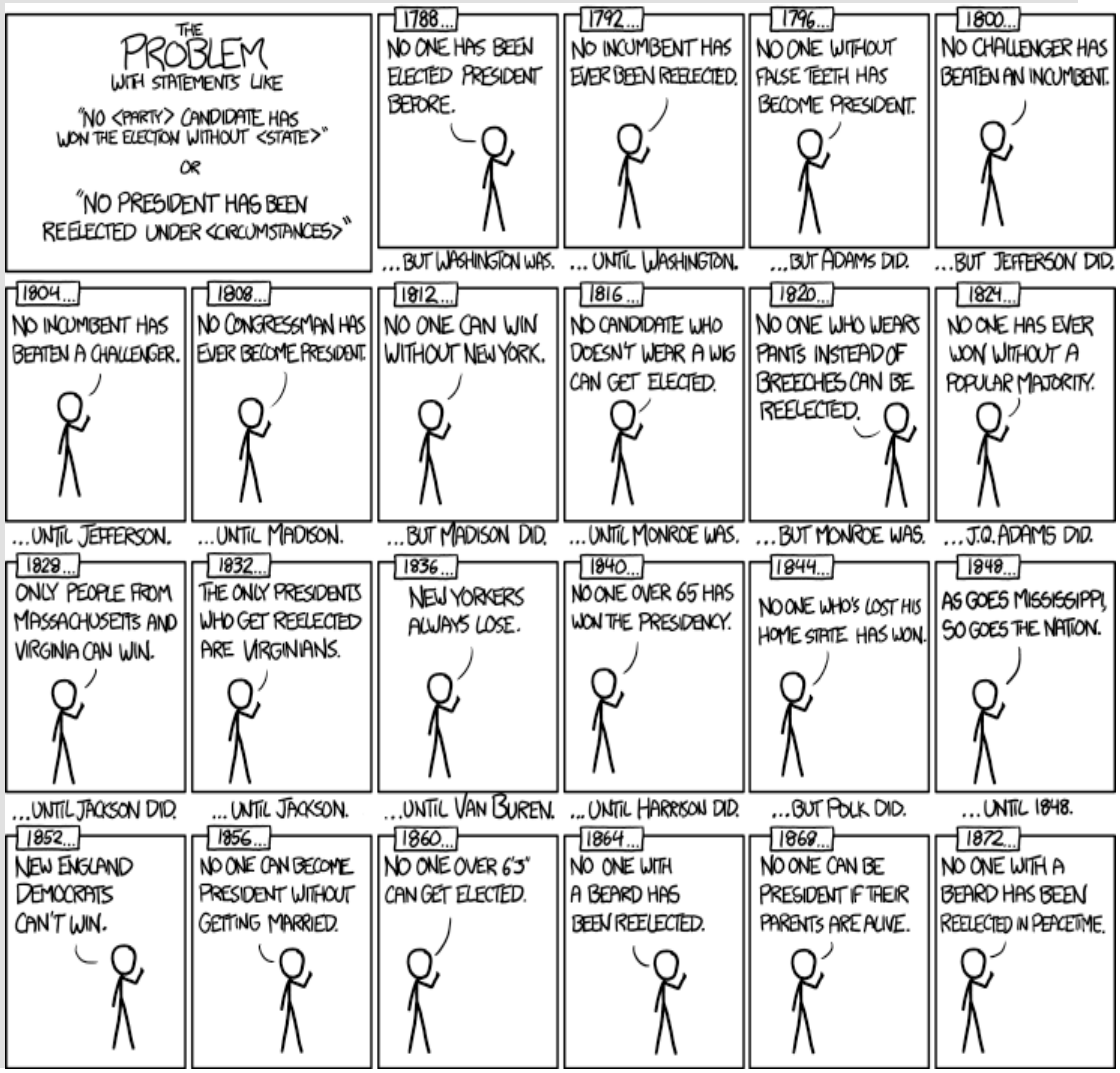

When we have limited data, dividing the dataset into Train and Validation sets may cause some data points with useful information to be excluded from the training procedure, and the model fails to learn the data distribution properly.

We will use 10-fold CV for selecting lambda during lasso regression and for determining the AUC

# Examples of overfitting



Black line - fits data well
Green line - is over fitted to the data



# Example of lasso logistic regression in Stata 17

```
. lasso logit depression i.disclosure_status i.caregiver i.agegroup i.sex i.distance i.substance_use
> i.ART_buddy i.intervention_group i.asset i.viral_load_t4 i.adherence_history child_poverty HIV_stig
> ma adherence_self_efficacy hopelessness social_support family_cohesion, selection(cv) rseed(1234) f
> olds(10) cluster(clinic_id)
```

cluster: accounts for potential correlation between observations in the same clinic

selection(cv) & folds(10): select lambda by 10-fold cross-validation

rseed(1234): set random-number seed to ensure reproducibility

```
Lasso logit model                          No. of obs         =        563
                                           No. of covariates  =         29
Cluster  : clinic_id                       No. of clusters    =         39
Selection: Cross-validation                No. of CV folds    =         10
```

| ID | Description | lambda | No. of nonzero coef. | Out-of-sample dev. ratio | CV mean deviance |
|---|---|---|---|---|---|
| 1 | first lambda | .1675651 | 0 | -0.0046 | 1.35559 |
| 15 | lambda before | .045554 | 5 | 0.1051 | 1.207525 |
| * 16 | selected lambda | .0415071 | 6 | 0.1058 | 1.206537 |
| 17 | lambda after | .0378198 | 8 | 0.1057 | 1.206694 |
| 21 | last lambda | .0260677 | 12 | 0.1010 | 1.213044 |

* lambda selected by cross-validation.

# Post-estimation commands: *estimates store*

```
. estimates store depression_cluster
```

To store the results in memory

To display penalized coefficients of selected unstandardized variables, sorted by penalized coefficients of unstandardized variables after fitting lasso model

```
. lassocoef depression_cluster, display(coef, penalized) sort(coef, penalized)
```

| | depression_cluster |
|---|---|
| _cons | .1771601 |
| hopelessness | .1124357 |
| HIV_stigma | .0418476 |
| social_support | -.0385868 |
| child_poverty | -.0383331 |
| family_cohesion | -.0052862 |
| adherence_self_efficacy | -.0000434 |

Legend:
  b - base level
  e - empty cell
  o - omitted

# Post-estimation commands: *predict*

```
. predict double depression_clusterPR3, pr
(option penalized assumed; Pr(depression) with penalized coefficients)
```

- **predict:** creates a new variable containing probabilities (logit model)
- **double:** type of variable
- **depression_clusterPR3:** name of new variable (predicted probability of the outcome)
- **pr:** probability of a positive outcome

# 7 Steps in prediction modelling

*(Steyerberg & Vergouwe, 2014)*

1. Define the prediction problem: define predictors and outcome of interest
2. Code predictors
3. Specify a model
4. Estimate model parameters
5. **Model evaluation**
6. **Model validation**
7. Presentation of the model

# Model evaluation: calibration

Binary outcome

Predicted probability of outcome

```
. calibrationbelt depression depression_clusterPR3, devel("internal") cLevel1(0.95) cLevel2(0.99) maxDeg(4) thres(
> 0.95)
-------------------------------------------------------------
                GiViTI Calibration Belt

Calibration belt and test for internal validation:
the calibration is evaluated on the training sample.

Sample size: 563
Polynomial degree: 2
Test statistic:  2.66
p-value: 0.1028
-------------------------------------------------------------
```

Type of calibration: internal/ external

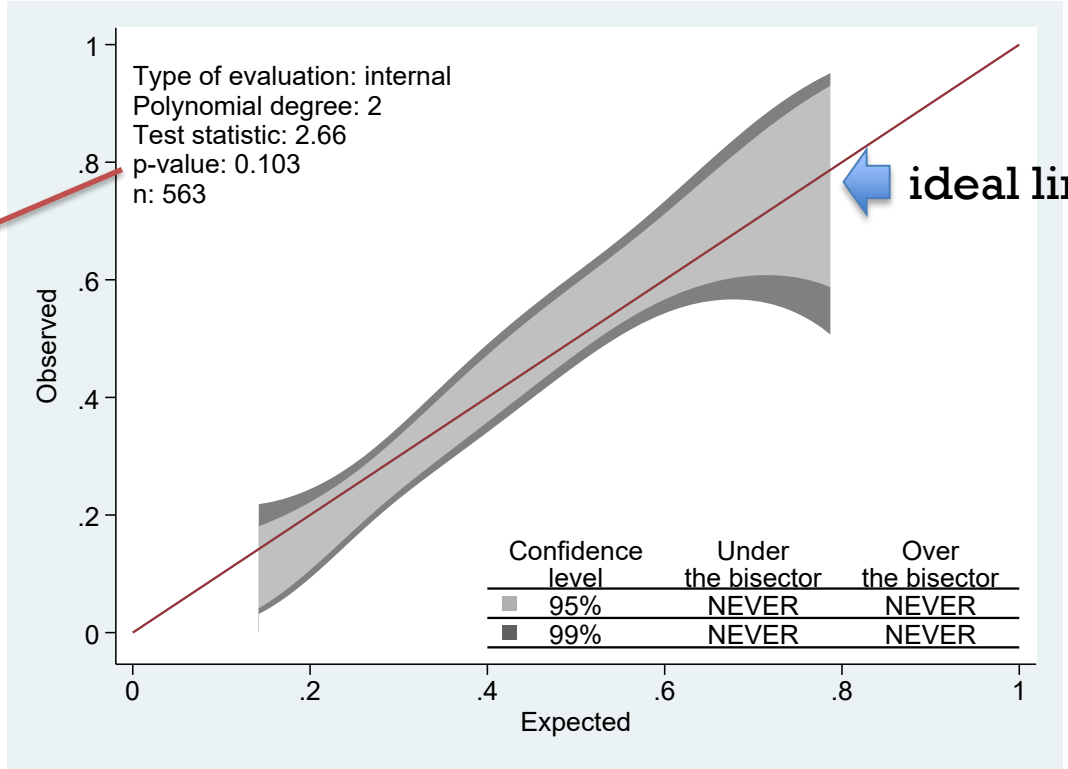
95% Confidence level

99% Confidence level

- **Calibration:** agreement between observed rates of depression and predicted probabilities

# Model evaluation: calibration

Large p-value indicates no statistical difference between model predictions and the 45° line. We want this!




- **Calibration belt:** generates the calibration plot along with associated statistical test

# Model evaluation: discrimination

- Discrimination: how well the model can differentiate ALWHIV with depression from those without
- Measured using Area under the receiver operator characteristic curve (AUC)/ C-statistic
- Prediction modelling: how well model will perform in new cases (generalizability)
- 10-fold cross-validation helps generate a more realistic estimate of predictive performance in new cases

# AUC generated using 10-fold cross-validation (internal validation)

Binary outcome

Predicted probability of outcome

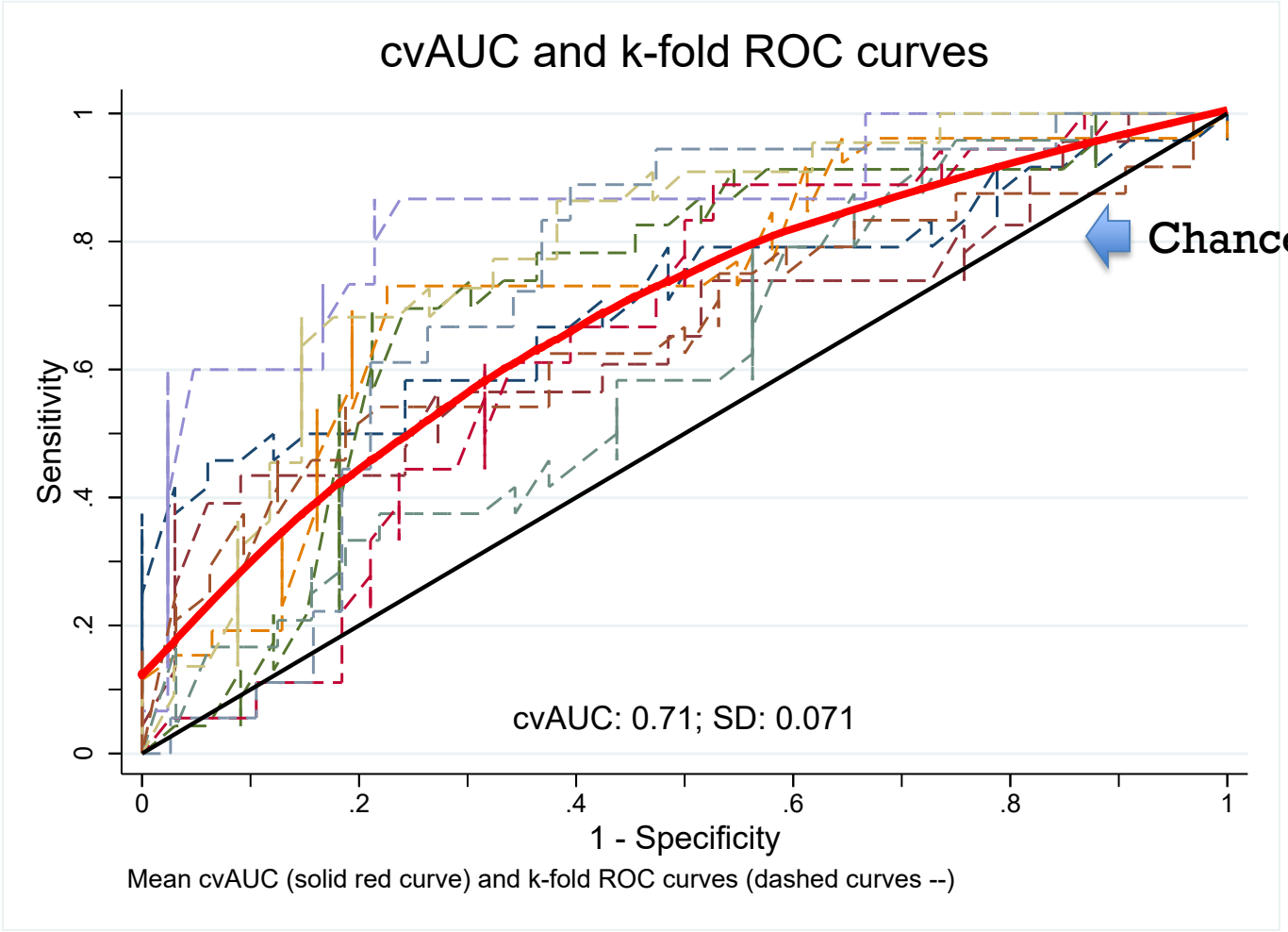
```
. cvauroc depression depression_clusterPR3, kfold(10) seed(1972) fit detail graphlowess
1-fold (N=57).........AUC =  0.708
2-fold (N=56).........AUC =  0.656
3-fold (N=56).........AUC =  0.719
4-fold (N=57).........AUC =  0.732
5-fold (N=56).........AUC =  0.608
6-fold (N=56).........AUC =  0.649
7-fold (N=57).........AUC =  0.848
8-fold (N=56).........AUC =  0.790
9-fold (N=56).........AUC =  0.663
10-fold (N=56).........AUC =  0.730

Model:logistic

Seed:1972

_____________________________________________________________________
Cross-validated (cv) mean AUC, SD and Bootstrap Bias Corrected 95%CI
_____________________________________________________________________
cvMean AUC:                          | 0.7103
Bootstrap bias corrected 95%CI:      | 0.6574, 0.7500
cvSD AUC:                            | 0.0711
_____________________________________________________________________


_____________________________________________________________________
Mean cross-validated Sen, Spe and false(+) at depression predicted values
_____________________________________________________________________


Prevalence of  depression: 38.54%
```

Split data into 10 random folds

Ensures reproducibility

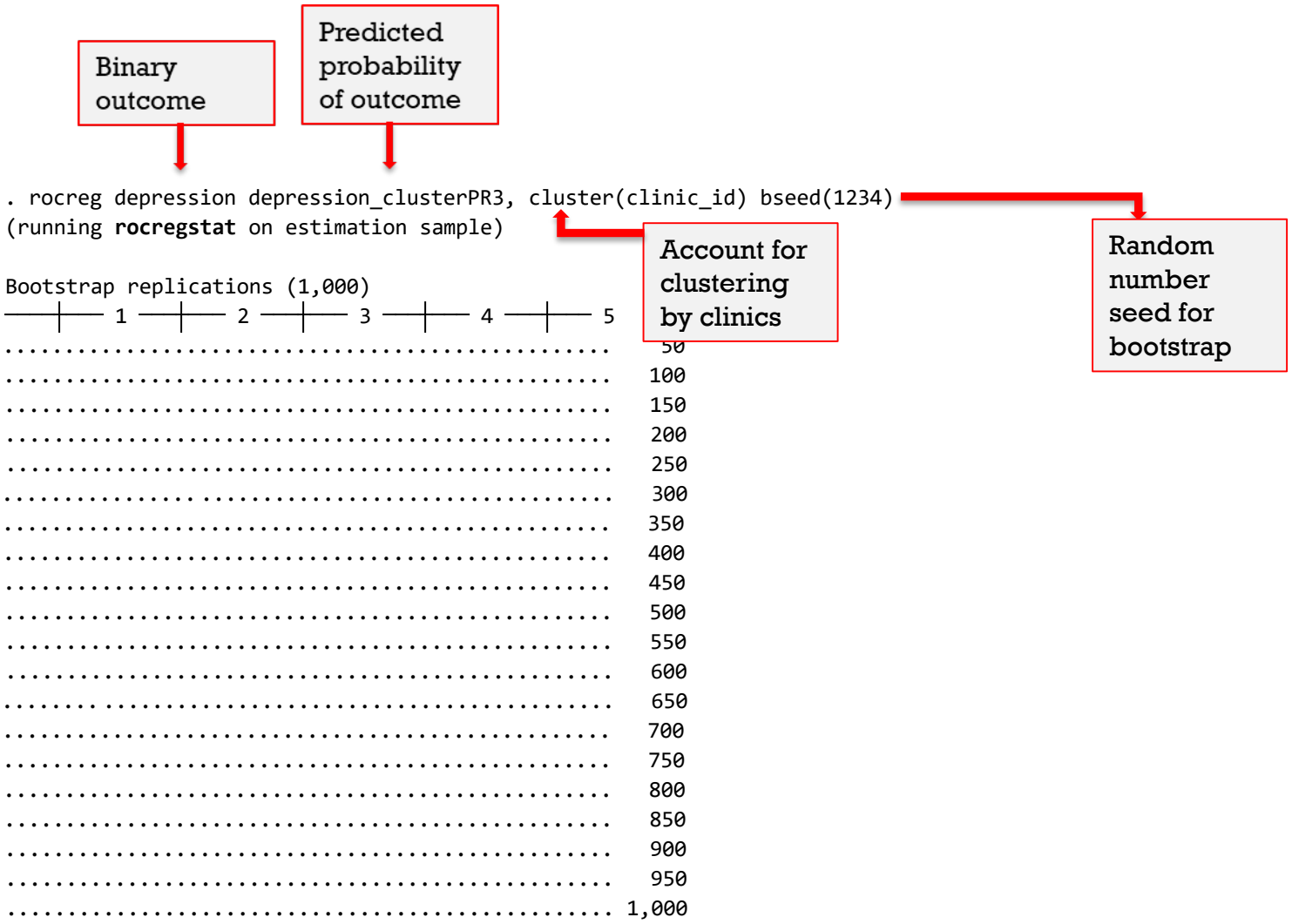
Shows the mean AUC curve and the 10 curves for each fold

```
. cvauroc depression depression_clusterPR3, kfold(10) seed(1972) fit detail graphlowess
```




**Rule of thumb-**AUC of 0.5= same as chance; >0.7= good model; >0.8= strong model; 1= perfect model

# Alternative way to estimate AUC




Rocreg estimates AUC using bootstrap resampling

# rocreg

```
Bootstrap results                                       Number of obs =    563
                                                        Replications  = 1,000

Nonparametric ROC estimation

Control standardization: empirical
ROC method             : empirical

Area under the ROC curve

   Status    : depression
   Classifier: depression_clusterPR3
                              (Replications based on 39 clusters in clinic_id)
```

| AUC | Observed coefficient | Bias | Bootstrap std. err. | [95% conf. | interval] | |
|---|---|---|---|---|---|---|
| | .7140593 | .0012518 | .0239153 | .6671862 | .7609324 | (N) |
| | | | | .666503 | .7616271 | (P) |
| | | | | .6643847 | .7601535 | (BC) |

Binary outcome

Predicted probability of outcome

normal-approximation CI

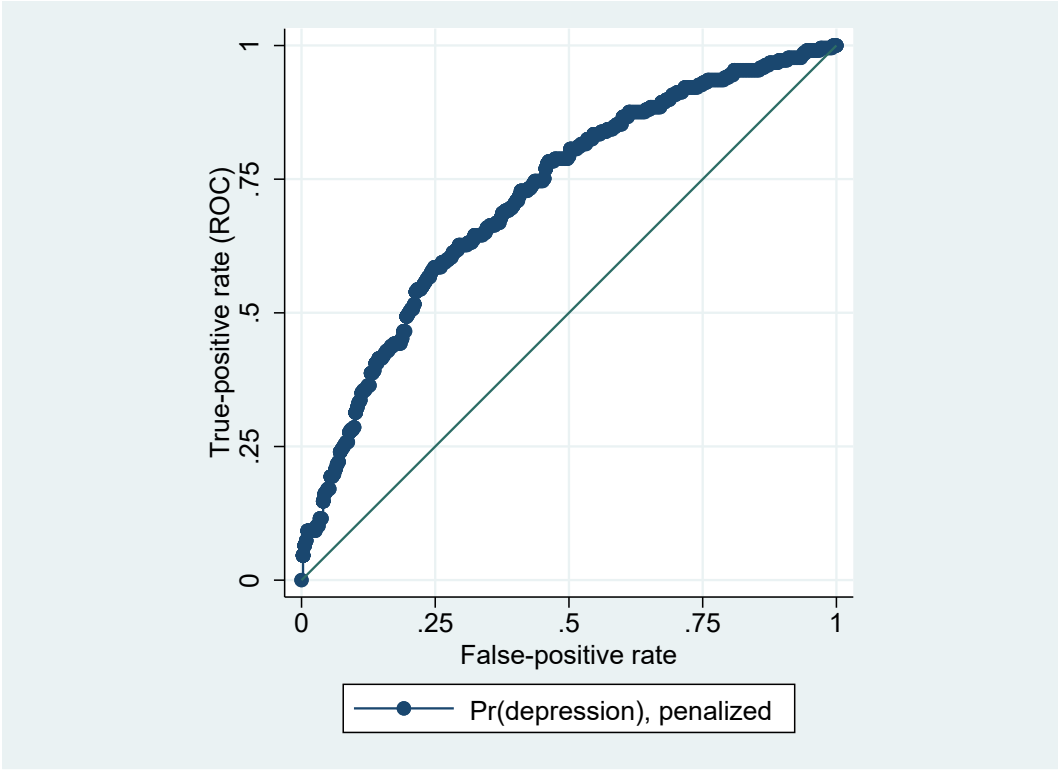
percentile CI

bias-corrected CI

Rocreg estimates AUC using bootstrap resampling

# AUC curve estimated using rocreg

rocregplot



Brathwaite R. et al. Journal of the International AIDS Society 2021, 24:e25756
http://onlinelibrary.wiley.com/doi/10.1002/jia2.25756/full | https://doi.org/10.1002/jia2.25756

JIAS
JOURNAL OF THE INTERNATIONAL AIDS SOCIETY

RESEARCH ARTICLE

# Predicting the individualized risk of poor adherence to ART medication among adolescents living with HIV in Uganda: the Suubi+Adherence study

Rachel Brathwaite[1,5], Fred M Ssewamala[1], Torsten B Neilands[2], Moses Okumu[3], Massy Mutumba[4], Christopher Damulira[1,5], Proscovia Nabunya[1], Samuel Kizito[6], Ozge Sensoy Bahar[1], Claude A Mellins[7] and Mary M McKay[8]

# Thank you and questions


- Suubi+Adherence study was funded by the Eunice Kennedy Shriver National Institutes of Child Health and Human Development (NICHD) (Grant #1R01HD074949-01, PI: Fred M. Ssewamala).
- Dr. Fred Ssewamala (mentor)
- Dr. Tor Neilands (statistical guidance)
- ICHAD Uganda for data collection