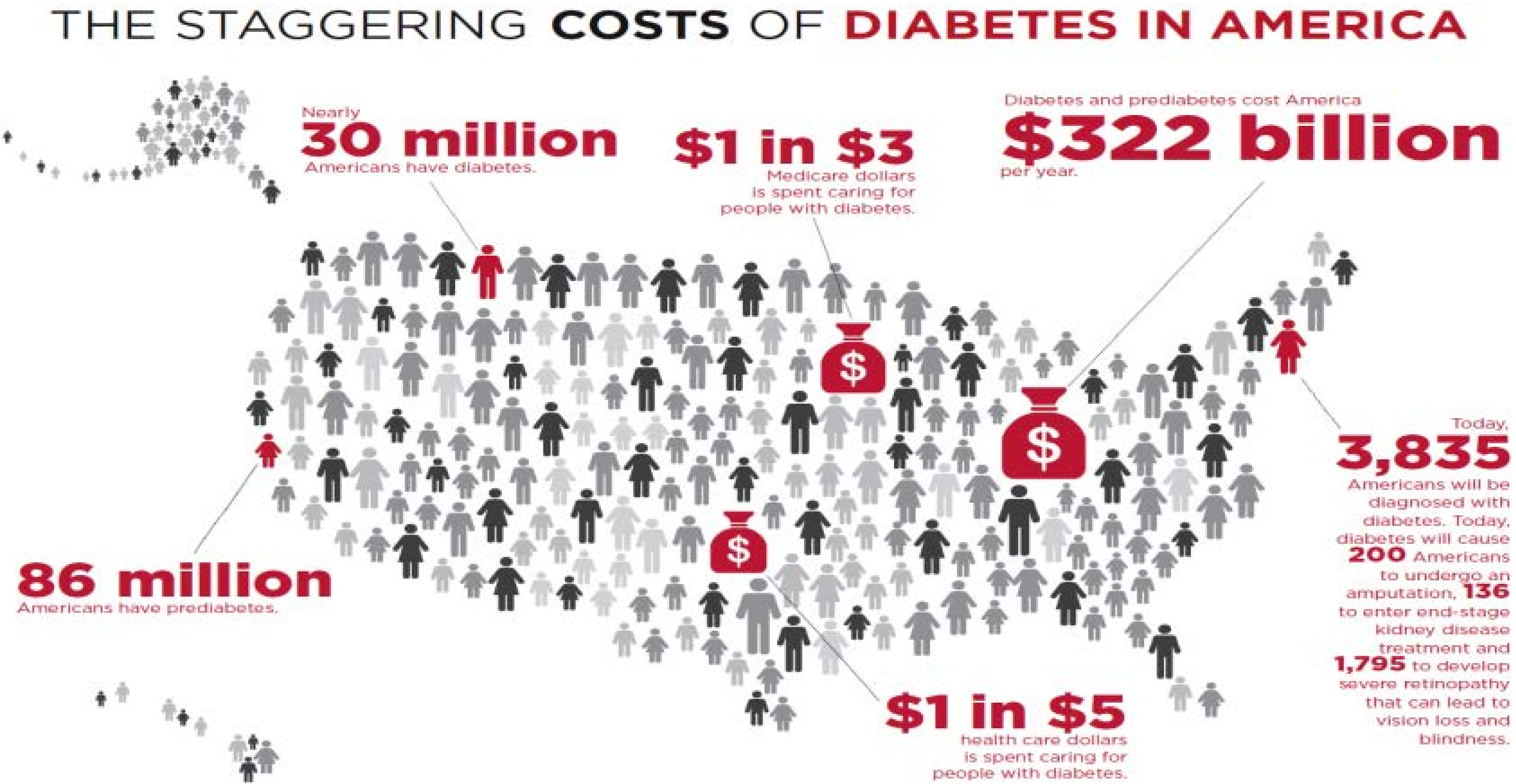
Predictive Models for Making Patient Screening Decisions

MICHAEL HAHSLER[1], VISHAL AHUJA[1], MICHAEL BOWEN[2], AND FARZAD KAMALZADEH[1]

[1] Southern Methodist University,
[2] UT Southwestern Medical Center and Parkland Health and Hospital System

World Changers Shaped Here SMU

UT Southwestern Medical Center

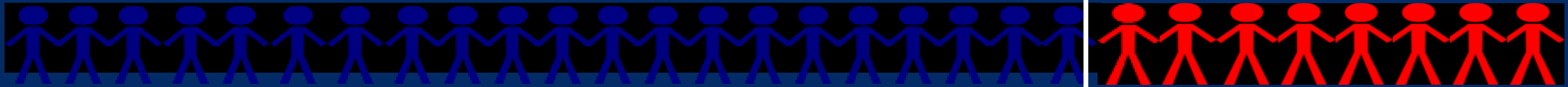


Source: http://www.diabetes.org/advocacy/news-events/cost-of-diabetes.html

American Diabetes Association®

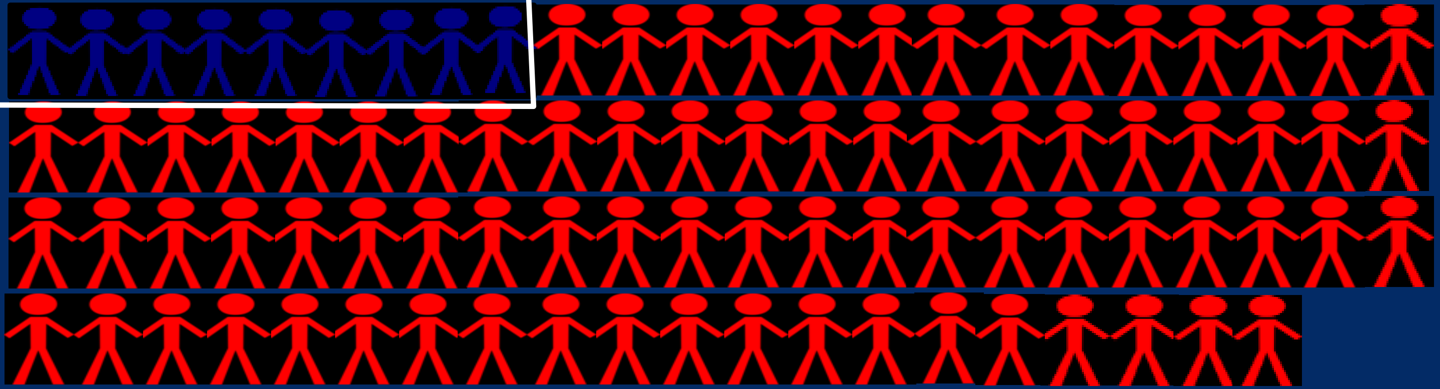
Prevalence of Diagnosed and Undiagnosed Type 2 Diabetes and Prediabetes

29.1 million people in the US have T2DM (9.3% of population)



8.1 Million Undiagnosed

Over 86 million adults in the US with pre-diabetes (37% of population)



77 Million with Undiagnosed Pre-diabetes

CDC National Diabetes Statistics Report, 2014.

World Changers Shaped Here SMU.

Who Should We Screen?

1. Screening Guidelines
   - U.S. Preventive Services Task Force (USPSTF) 2015
     - Adults 40-70 AND BMI≥25
   - American Diabetes Association (ADA)
     - All adults over age 45 OR any age if BMI ≥ 25 (or ≥ 23 in Asians) AND an additional risk factor
2. Diabetes Risk Score (not widely used in the US)
   - Incident Risk Scores: predict development of diabetes in the future
   - Prevalent Risk Scores: assess the current probability of having undiagnosed diabetes

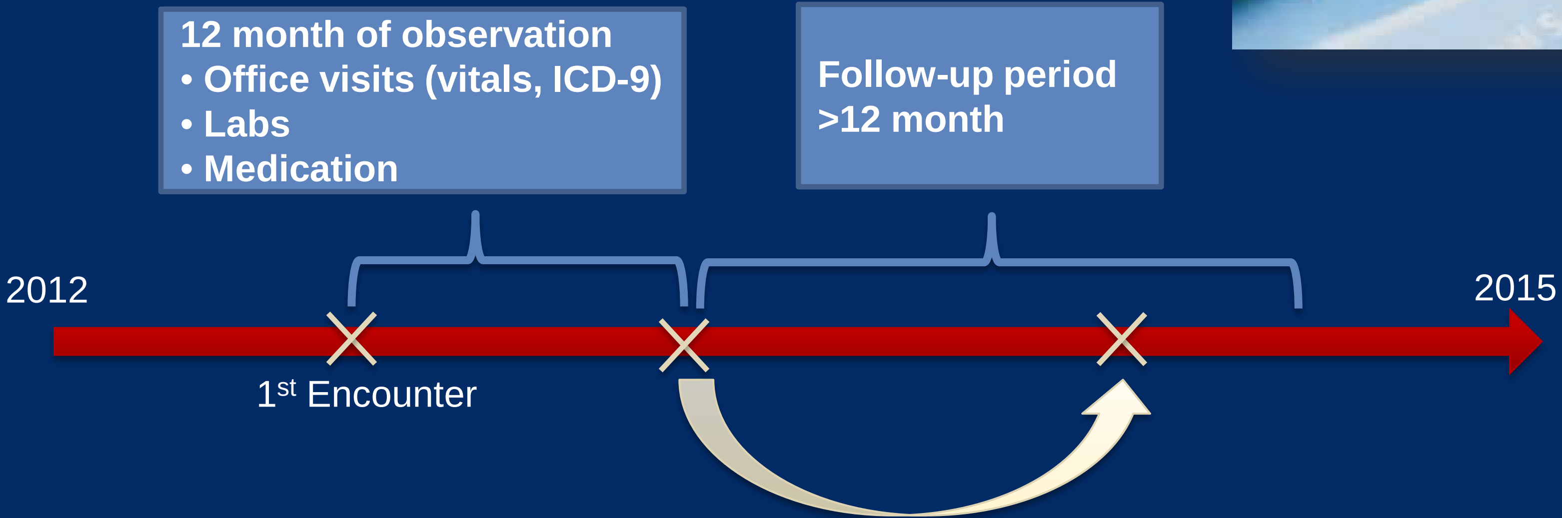
Aim

- Assist in clinical decision-making in terms of screening patients at “highest” risk of developing diabetes.
- Our key questions are:
  - How to produce simple predictive models for risk scores?
  - How do we deal with a large quantity of missing data?
- Desired properties:
  - Applicable to all patients, no matter how much information we have.
  - Tells us what information about the patient to elicit next.

Related Literature

- **Analytics in Healthcare**
  - Bertsimas (2016) – Personalized Diabetes Management
  - Bertsimas (2012) – Cancer Therapy
  - Shams et al. (2015) – 30-day readmissions
- **Healthcare screening decisions**
  - Lee et al. (working) - Screening for Hepatocellular Carcinoma
  - Maillart et al. (2008) – breast cancer screening
  - Deo et al. (2015) – HIV screening
- **Predictive models for type 2 diabetes**
  - Baan et al. (1999) - Performance of a predictive model to identify undiagnosed diabetes in a health care setting
  - Collins et al. (2011) - Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting (survey)
  - Jahani & Mahdavi (2016) - Comparison of Predictive Models for the Early Diagnosis of Diabetes (neural networks)

Predictive Problem: Initial Screening Decision



Predict if the patient
has or will develop diabetes
and should be screened

Data Set

- **Retrospective cohort** (N = 34,297 patients)
- **Cohort Dates:** 2012-2015
- **Setting:** Parkland Health and Hospital System, a large integrated, safety-net healthcare system in North Texas.
- **Data Source:** Epic Electronic Medical Record (EMR)
- **Eligibility:**
  - Ages 18-65
  - ≥1 primary care visit every 18 month
  - Only unscreened patients with no known diabetes during first 12 month

Available Data

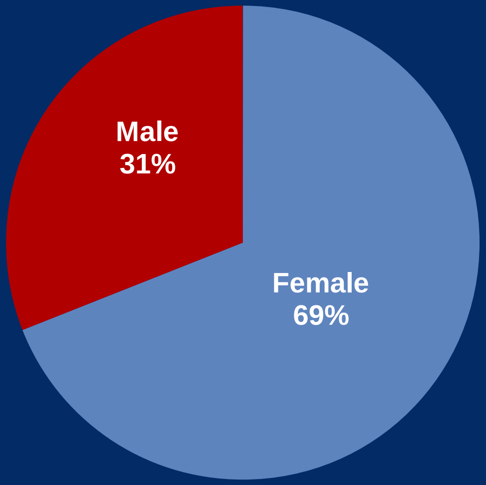
105 Features including

- **Demographic information:** Age, Gender, Race, etc.
- **BMI, vitals:** Blood pressure, etc.
- **Risk factors** (co-morbidities): Hypertension, family history, etc.
- **Lab values:** Cholesterol, random blood glucose, etc.
- **Medications** (prescribed): Blood pressure, cholesterol, etc.
- **Health care utilization:** Office encounters, ER visits, etc.

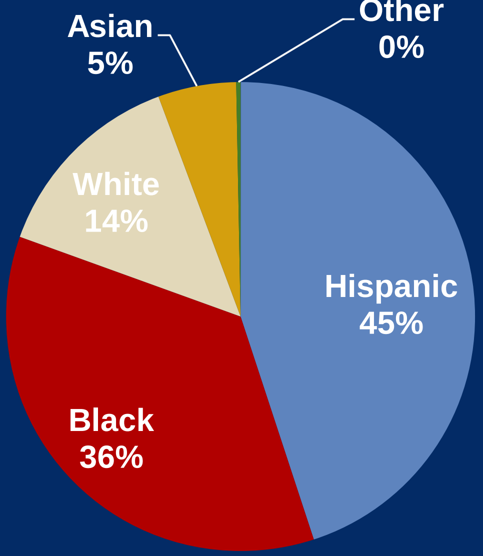
Note: Only demographic information, BMI and vitals are widely available. **19% of the data is missing.**

Cohort Specifics

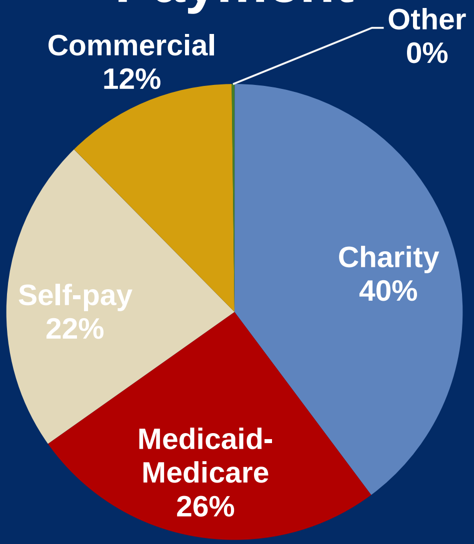
Sex



Race



Payment

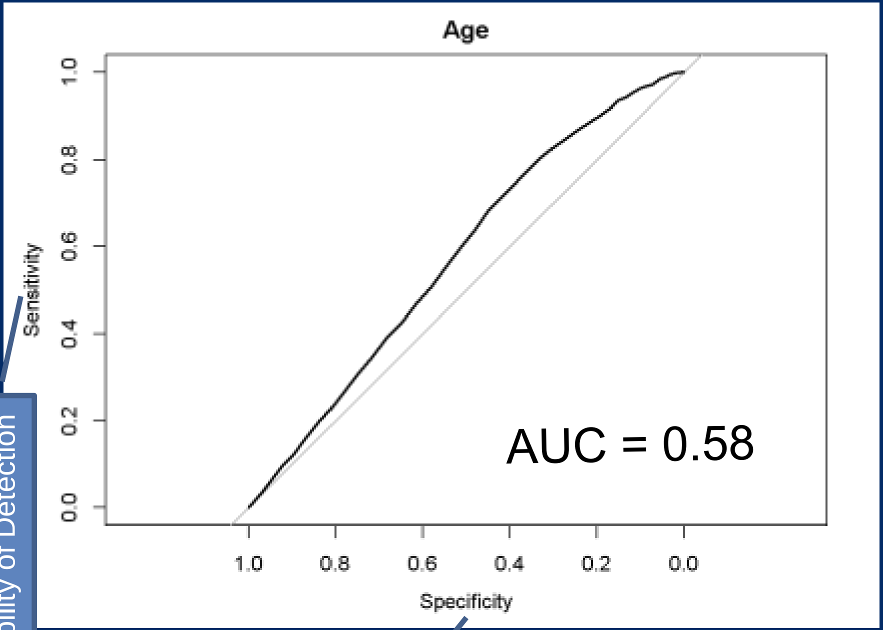


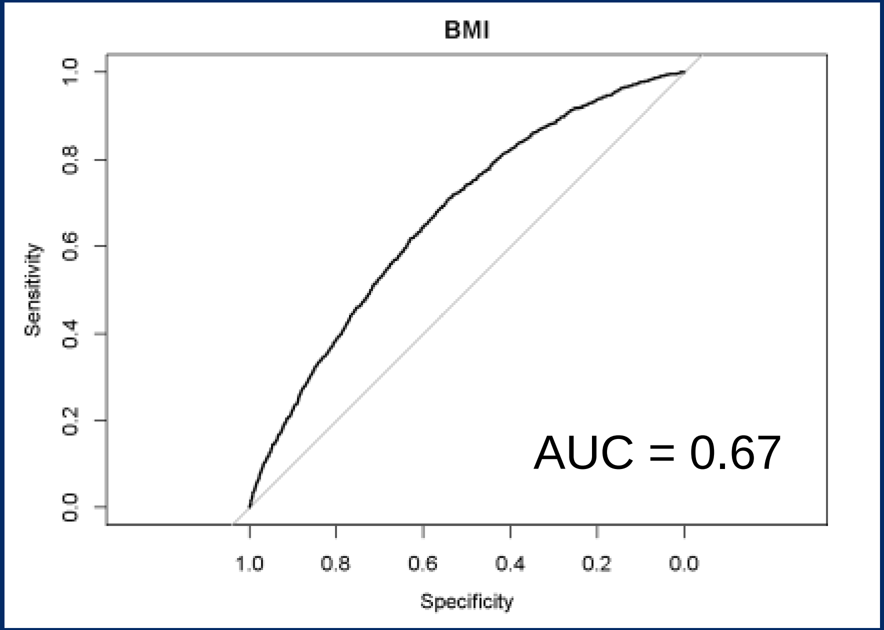
Median age: 46.9 years

Single-Factor Threshold Models

- Set a threshold on an individual predictor (risk factor) to distinguish between the classes:
  - Diabetes 13.6%
  - Not Diabetes 86.7%
- This represents the strategy used by current screening guidelines. Note that screening guidelines use several risk factors at a time.

Single-Factor Threshold Models
Usual risk factors: Age and BMI



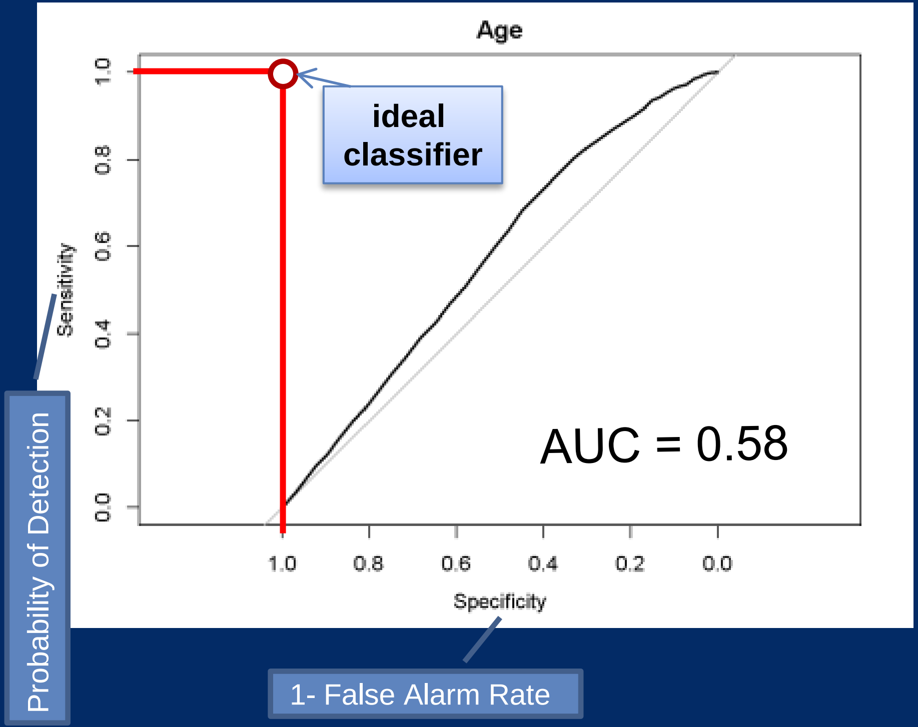


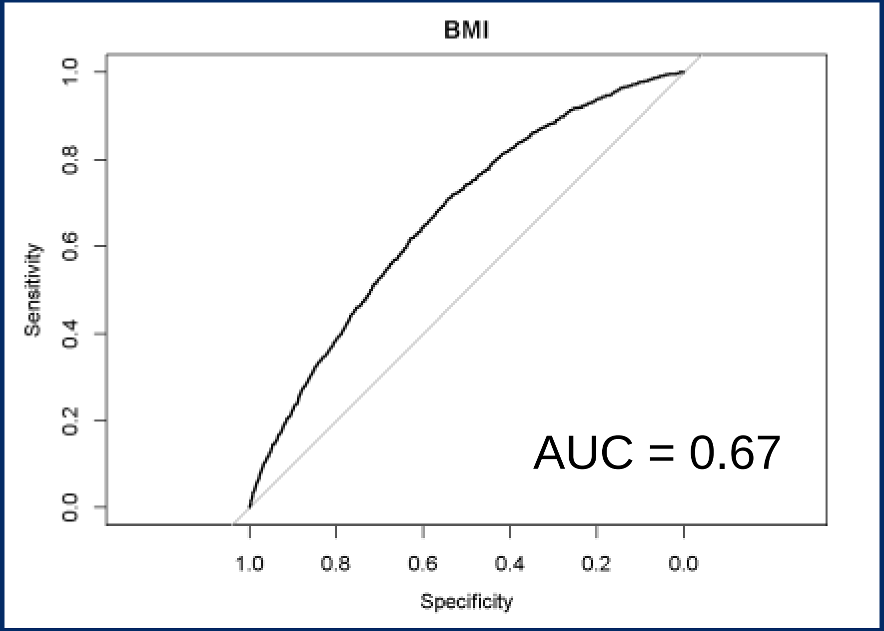
Probability of Detection

1- False Alarm Rate

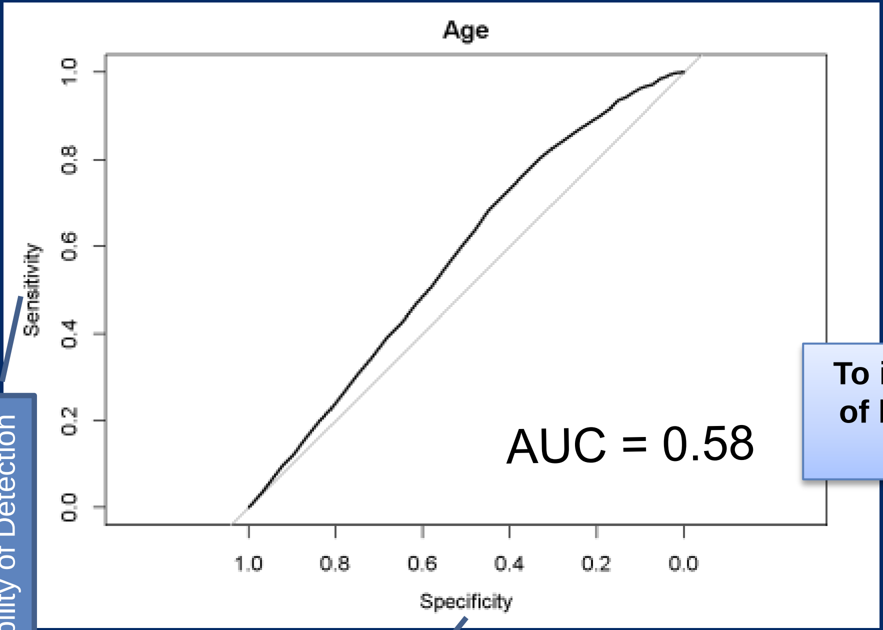
Available for 100% of patients

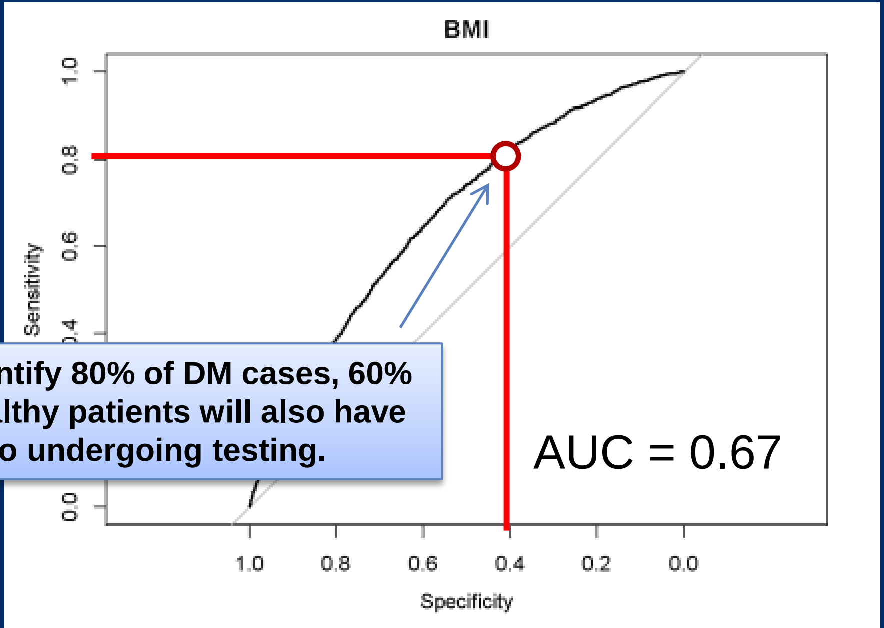
Single-Factor Threshold Models
Usual risk factors: Age and BMI





Single-Factor Threshold Models
Usual risk factors: Age and BMI



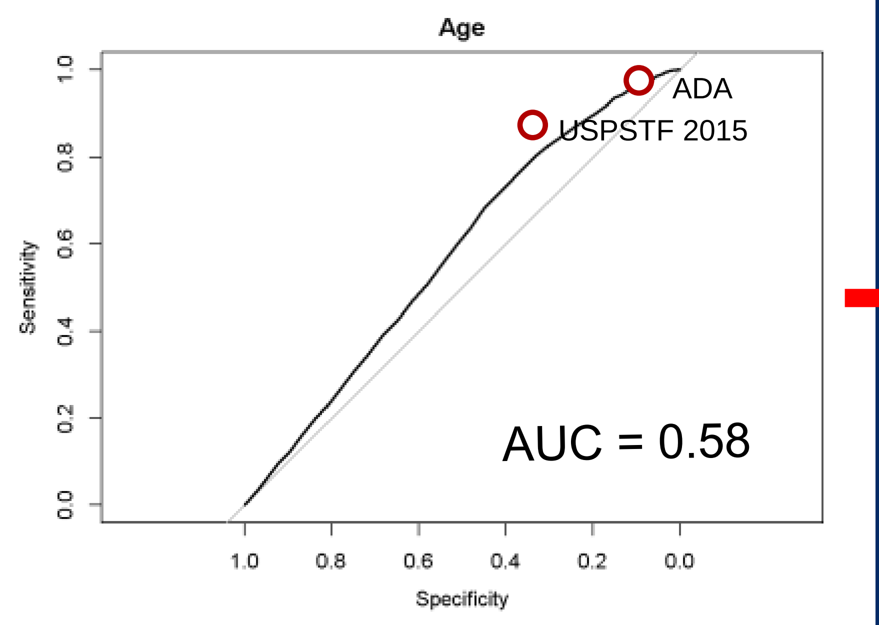


To identify 80% of DM cases, 60% of healthy patients will also have to undergoing testing.

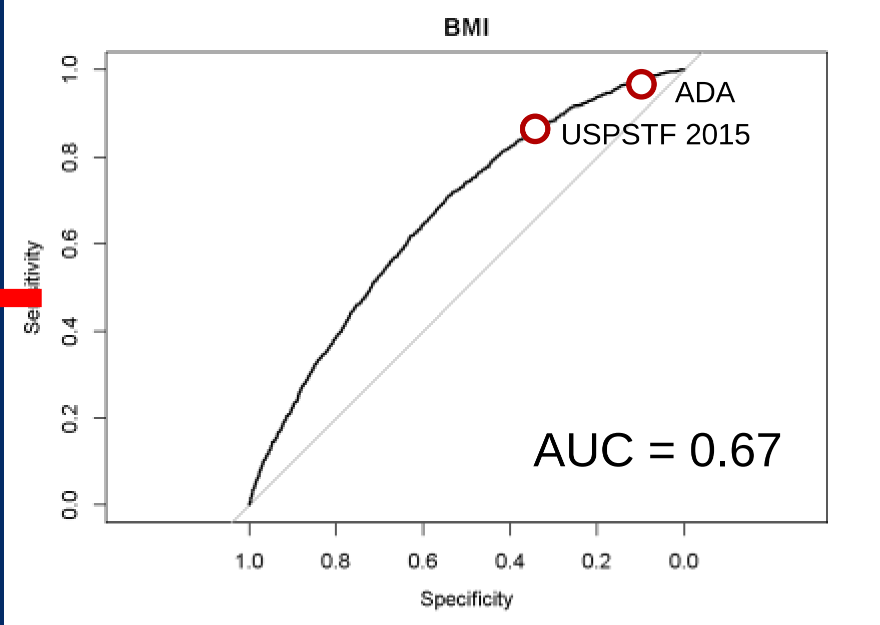
Probability of Detection

1- False Alarm Rate

Single-Factor Threshold Models
Usual risk factors: Age and BMI



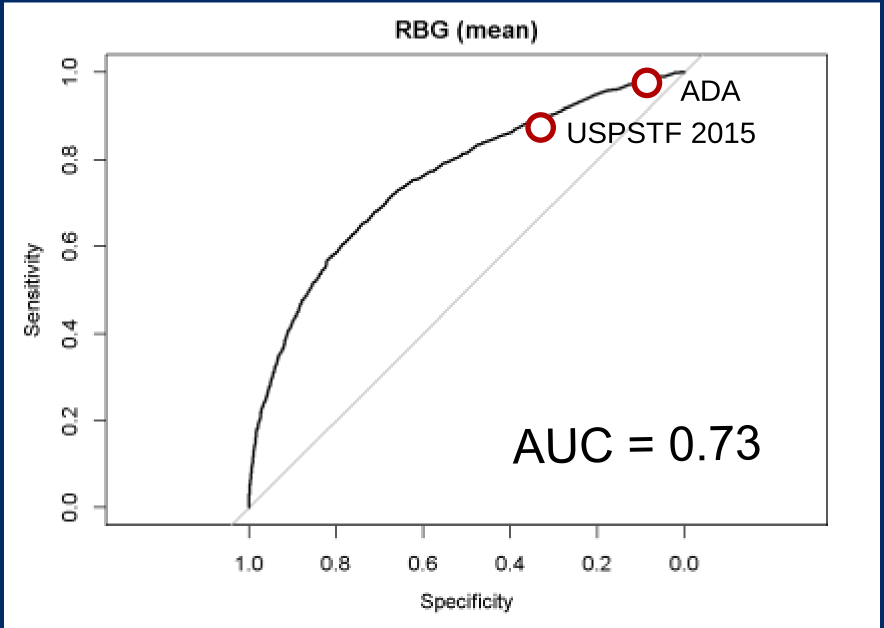
+

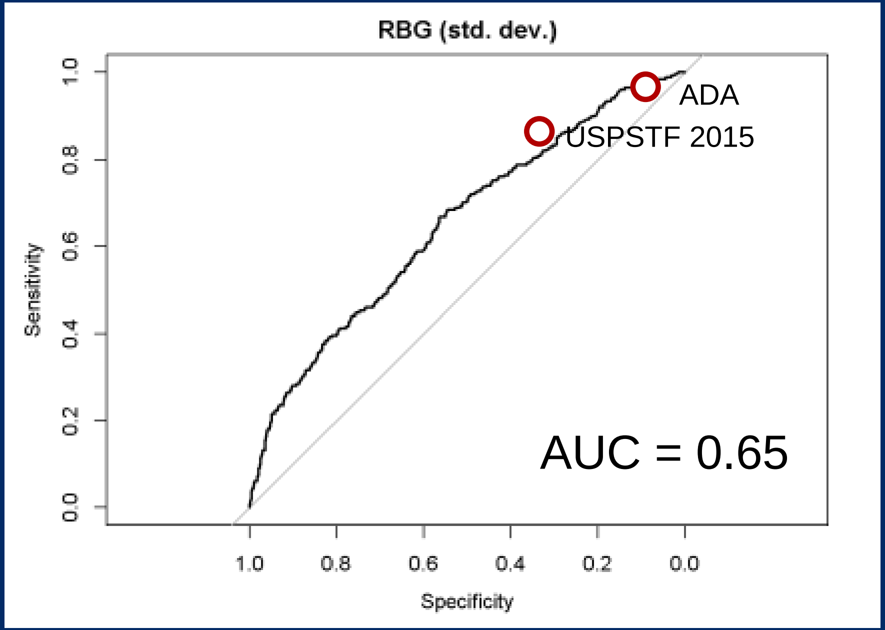


= USPSTF 2015 **(Age>40, BMI>25)**

Sensitivity : 0.817 Specificity : 0.377

Single-Factor Threshold Models
Uncommon risk factor: Random Blood Glucose





Available for 64% of patients

Available for 15% of patients

RBG was suggested in Bowen et al. J Clin Endocrinol Metab 2015;100(4):1503-1510

Multi-Factor Models

- For multi-factor models we have to deal with
  - Large number of features, but for practical decisions a small number of predictors is preferred.
  - Large part of the data is missing.
- We consider here two models
  - Naïve Bayes Classifier with backward feature selection
  - Logistic regression with LASSO regularization
- Both models apply feature selection, but dealing with missing data needs more consideration.

Dealing with missing values

- Different types of missingness:
  - **Missing completely at random (MCAR):** missingness is unrelated to any study variable.
  - **Missing at random (MAR):** non-randomness of missingness can be explained by other variables, but is not related to the response variable. E.g., PCP does not order a specific test for a person with a low BMI.
  - **Missing not at random (MNAR):** missingness is related to the response variable value. E.g., overweighed patient does not perform test for fear of a bad test result.

- Need methods robust to missingness (do not introduce bias). Options:
  - Ignore feature with missing values
  - Ignore observations with missing values

    Not practical with many missing values
    And introduce bias for all but MCAR
  - Just ignore the missing value (pairwise deletion) – needs to be supported by the method
  - Imputation

Enders, Craig K. (2010). Applied Missing Data Analysis (1st ed.)

Naïve Bayes Classifier

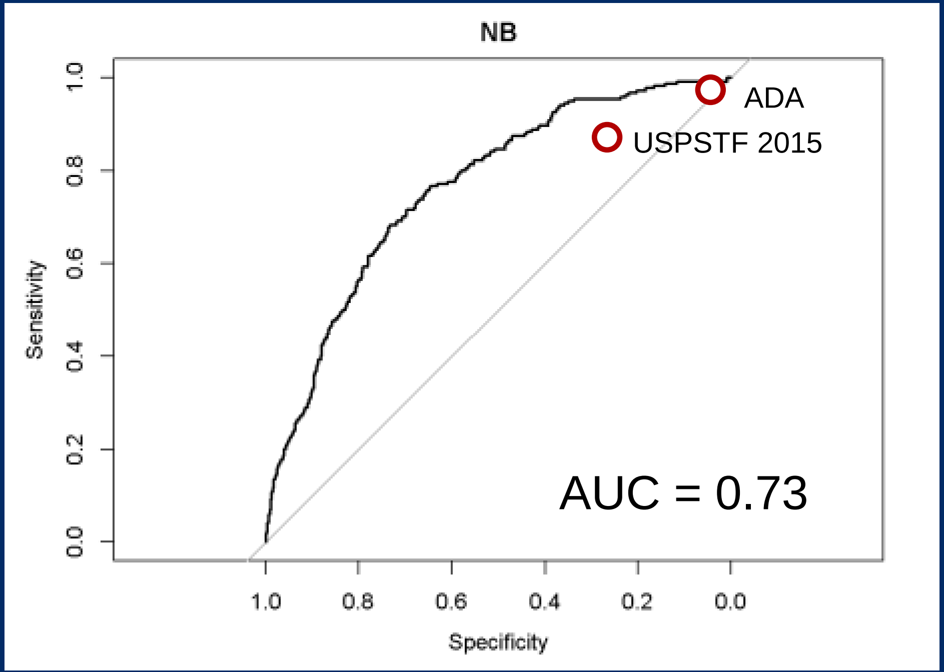
- Applies Bayes' theorem with a (naive) assumption of independence between features.

$$p(C_k | \boldsymbol{x}) = \frac{p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)}{p(\boldsymbol{x})}$$

- C_k is the class, $\boldsymbol{x}$ is a feature vector. We use a threshold on $p(C_{diabetes}|\boldsymbol{x})$ to produce a biased classifier.
- Metric predictors: we assume Gaussian distributions (given the target class).
- **Missing values:** Ignore missing values (pairwise deletion). Implies MCAR!
  - Learning: leave out missing values for the computation of the probability factors.
  - Prediction: omit corresponding table entries are omitted for prediction.

Russell, Intelligence: A Modern Approach

Multi Factor Model
Naive Bayes Model (NB) - All 105 Predictors



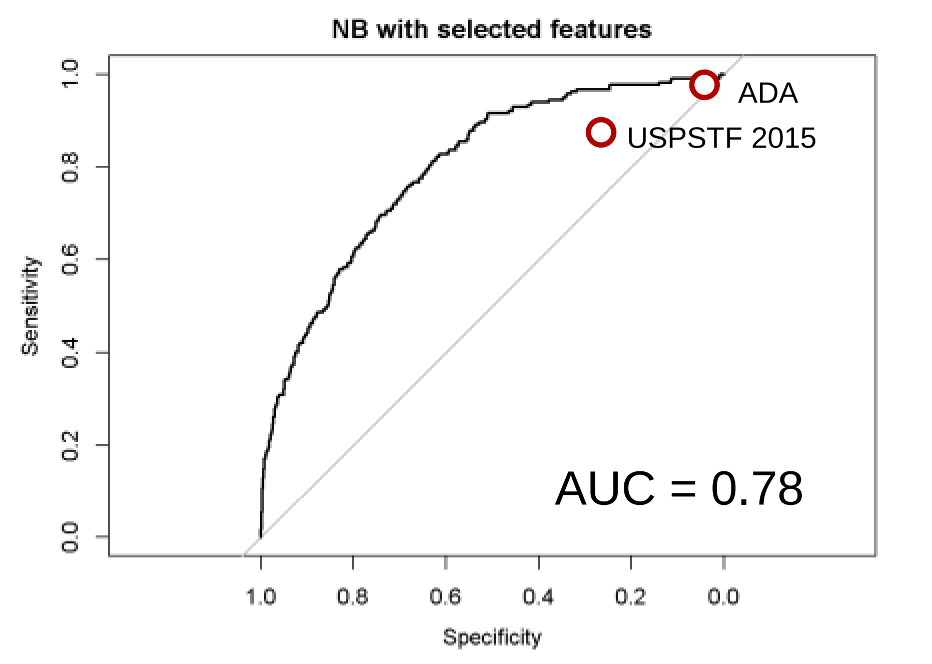
Available for 100% of patients

Makes predictions for **all patients**, even if information is missing (e.g., no blood test)

Results are as good as with blood test.

Multi Factor Model NB – Backward Feature Selection

5 of 10 predictors are not in current guidelines



Available for 100% of patients

Backward Feature Selection

1. **LAB_RANDOM_GLUCOSE_MEAN**
2. **LAB_RANDOM_GLUCOSE_SD**
3. BMI
4. BP_SYSTOLIC
5. **LAB_ALANINE_AMINOTRANSFERASE***
6. LAB_CHOLESTEROL_HDL_RATIO
7. AGE
8. **LAB_ASPARTATE_AMINOTRANSFERASE***
9. **LAB_RED_BLOOD_COUNT****
10. COMORB_FAMILY_HIST

* Liver enzyme

** Relationship needs to be studied

Generalized Linear Model with LASSO

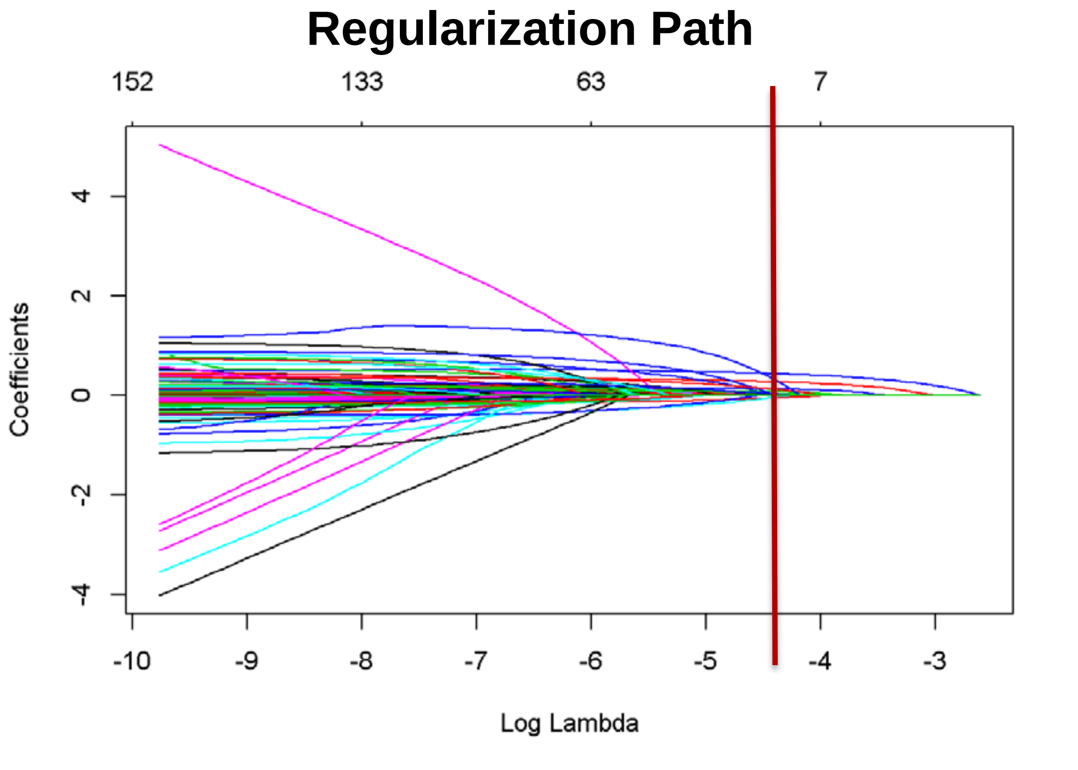
- GLM with binomial response and L1 regularization.

$$\min_{\boldsymbol{\beta}}\left\{\frac{1}{N}\left\|\boldsymbol{y}-\frac{1}{1-\exp(-X\boldsymbol{\beta})}\right\|_2^2\right\} \qquad s.t.\ \|\boldsymbol{\beta}\|_1 \leq t$$

- **Missing values:**
  - Numeric values: Mean imputation and add a dummy indicator variable.
  - Nominal variables: add an additional value for missing data.
- All variables are scaled to Z-scores.

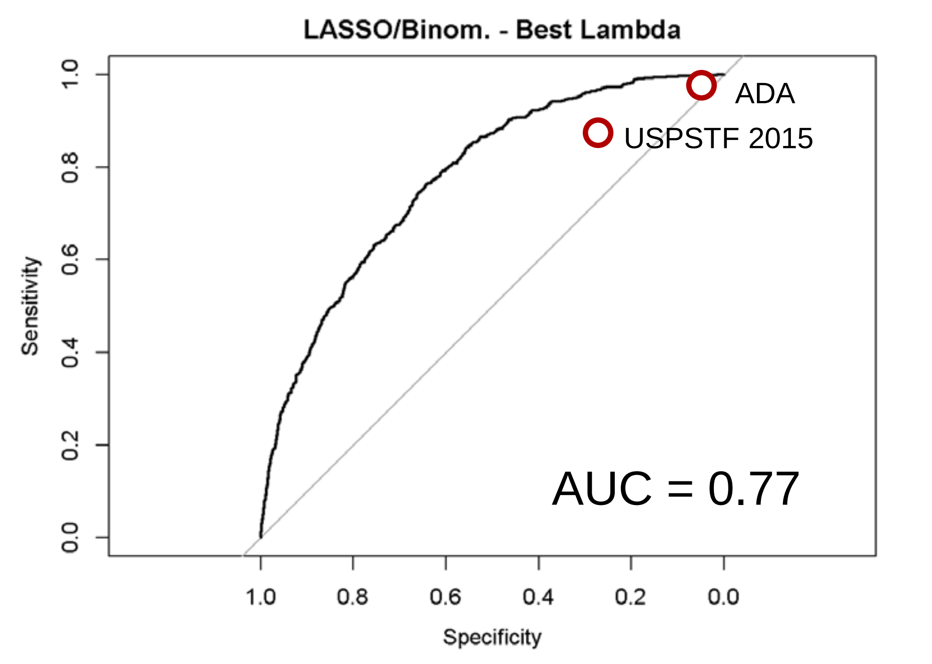
Tibshirani, Robert. 1996. “Regression Shrinkage and Selection via the lasso”

GLM - LASSO



Feature	Odds Ratio
LAB_RANDOM GLUCOSE_MEAN	1.49
BMI	1.26
BP_SYSTOLIC	1.10
COMORB HYPERTENSION	1.03

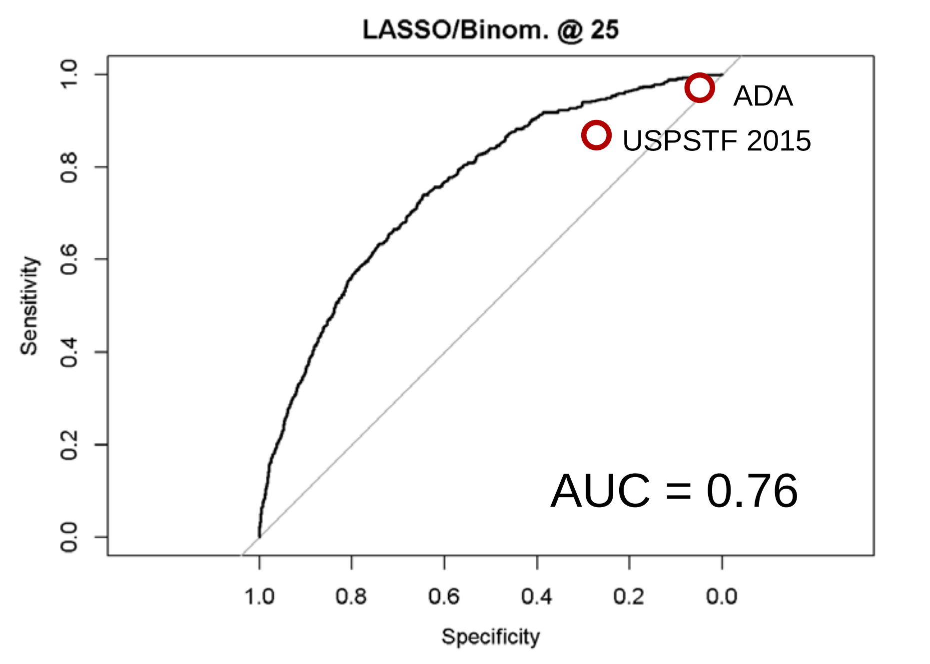
GLM – LASSO
Cross Validated selection of lambda (43 features)



Feature	Odds Ratio
LAB_NON_HDL_CHOLESTEROL_NA	1.68686548773596
LAB_RANDOM_GLUCOSE_MEAN	1.66887952164283
BMI	1.39449863570232
LAB_PLATELET_NA	1.21866934500033
COMORB_FAMILY_HIST	1.18536725238753
AGE	1.17605474716379
BP_SYSTOLIC	1.14286625129544
LAB_CHOLESTEROL_NA	1.11313974786007
LAB_RED_BLOOD_COUNT	1.10245828551814
MED_DM_biguanide	1.08600204482025
.	
.	
.	

Available for 100% of patients

GLM – LASSO
22 Features



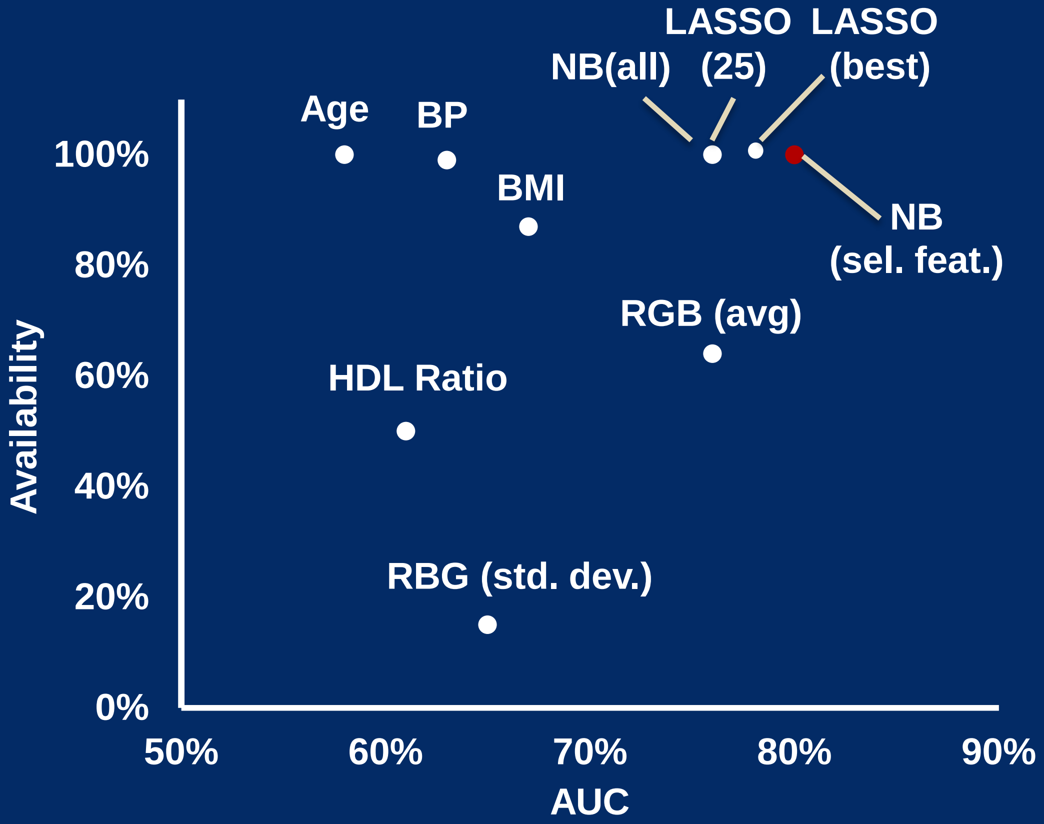
Available for 100% of patients

Features	Odds Ratio
LAB_RANDOM_GLUCOSE_MEAN	1.59082760110094
BMI	1.33973853414858
LAB_NON_HDL_CHOLESTEROL_NA	1.23002244516896
COMORB_FAMILY_HIST	1.13082194704404
LAB_PLATELET_NA	1.12206673237976
BP_SYSTOLIC	1.12117287327861
AGE	1.08818742855664
MED_DM_biguanide	1.06234757949966
COMORB_HYPERTENSION	1.04415582927392
MED_CHOL	1.04296022231027
LAB_RED_BLOOD_COUNT	1.04078442129921
COMORB_ABNORMAL_GLUCOSE	1.02970198207208
LAB_ALANINE_AMINOTRANSFERASE	1.02542623470265
MED_BPTRUE	1.02200770344732
LAB_CHOLESTEROL_HDL_RATIO	1.01698704953933
LAB_ALANINE_AMINOTRANSF_NA	1.01503915167147
LAB_HEMATOCRIT_NA	1.01441072109075
LAB_TRIGLYCERIDES	1.01243800612454
MED_DM_sulfonylurea	1.0114135696557
PULSE	1.01123780611578
LAB_RED_CELL_DIAMETER_WIDTH	1.00685382575705
COMORB_HYPERLIPIDEMIA	1.00582177220577

World Changers Shaped Here SMU

Comparison of Predictive Models

	AUC	Availability
NB (select feat.)	78%	100%
LASSO (best)	77%	100%
NB (all features)	76%	100%
LASSO (25)	76%	100%
RGB (avg)	76%	64%
BMI	67%	87%
RGB (std. dev.)	65%	15%
BP	63%	99%
HDL Ratio	61%	50%
Age	58%	100%



World Changers Shaped Here SMU

Conclusion

- Missing values are an issue for medical data.
- Naïve Bayes model with backwards feature selection and pairwise deletion
  - Reaches on our data an AUC of 78%.
  - Uses 10 factors typically already available in electronic health records (5 of the 10 factors are currently NOT considered in guidelines).
- Logistic Regression with LASSO, mean imputation and missing indicators
  - Reaches a AUC of 77%.
  - Uses 43 factors.

Future Research

- Consider other classification methods and evaluate model simplicity and influence of missing data.
- Automatic assistance in clinical decision making. Individualized optimal order of most critical risk factors.
- Investigate the operational effects of using the method to shift patients to outpatient care instead of ER visits.

Acknowledgements

- Dr. Bowen is supported by K23: NIDDK DK104065 and the Dedman Family Scholars in Clinical Care.

- Farzad Kamalzadeh is supported by a fellowship from the Niemi Center, Cox School of Business, SMU.

- We would also like to acknowledge Joanne Sanders for help on accessing and preparing the data.