Predictive Models for Making Patient Screening Decisions

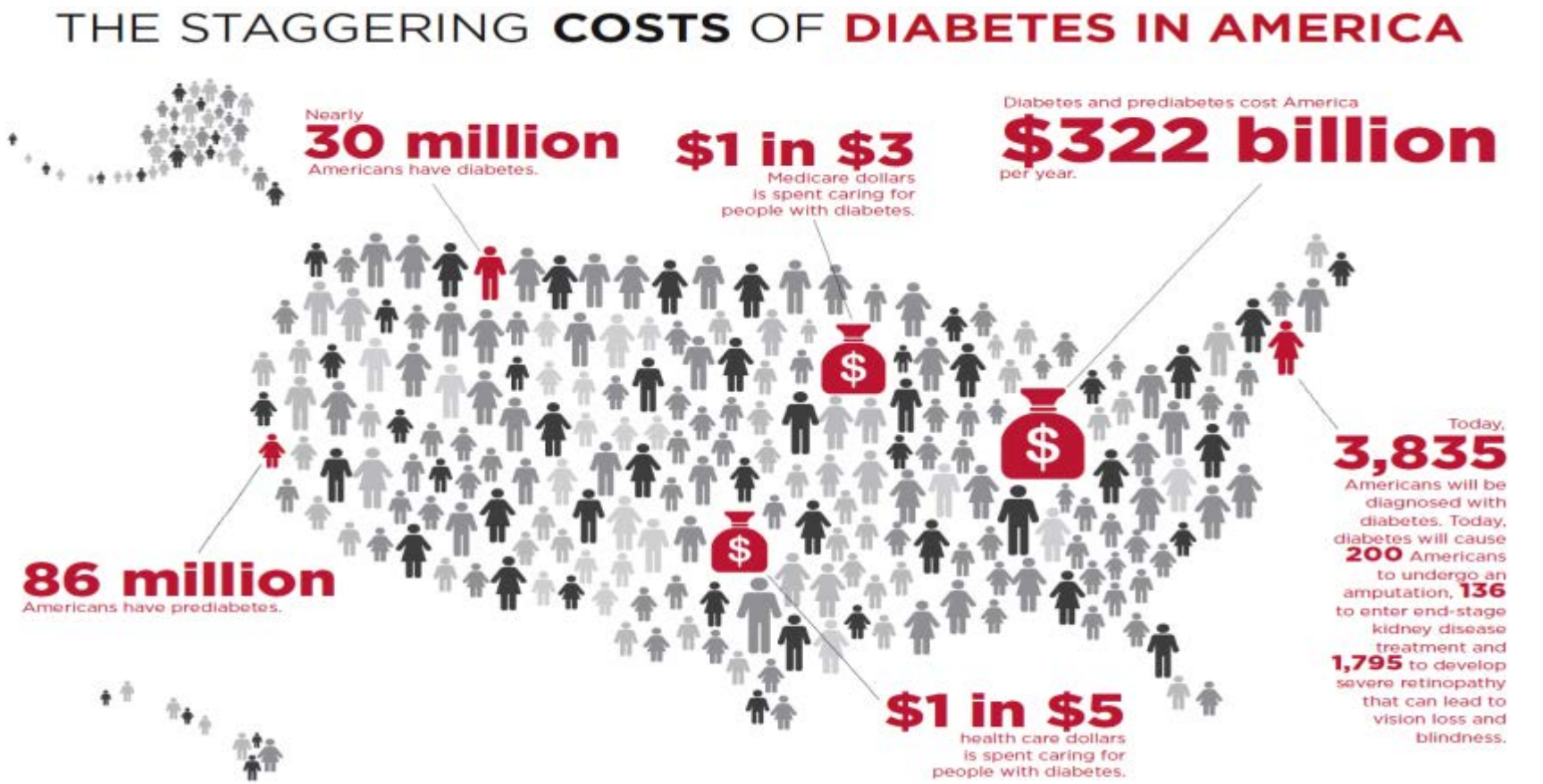
MICHAEL HAHSLER[1], VISHAL AHUJA[1], MICHAEL BOWEN[2], AND FARZAD KAMALZADEH[1]

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

World Changers Shaped Here

SMU

UT Southwestern Medical Center



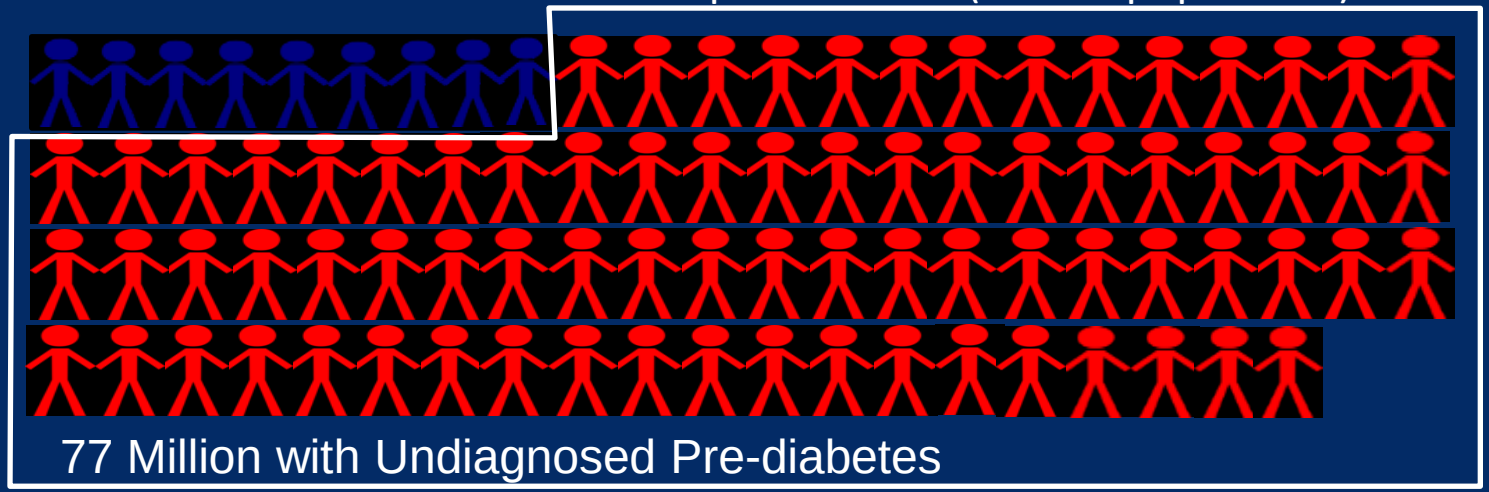
Source: http://www.diabetes.org/advocacy/news-events/cost-of-diabetes.html

American Diabetes Association.

Prevalence of Diagnosed and Undiagnosed Type 2 Diabetes and Prediabetes

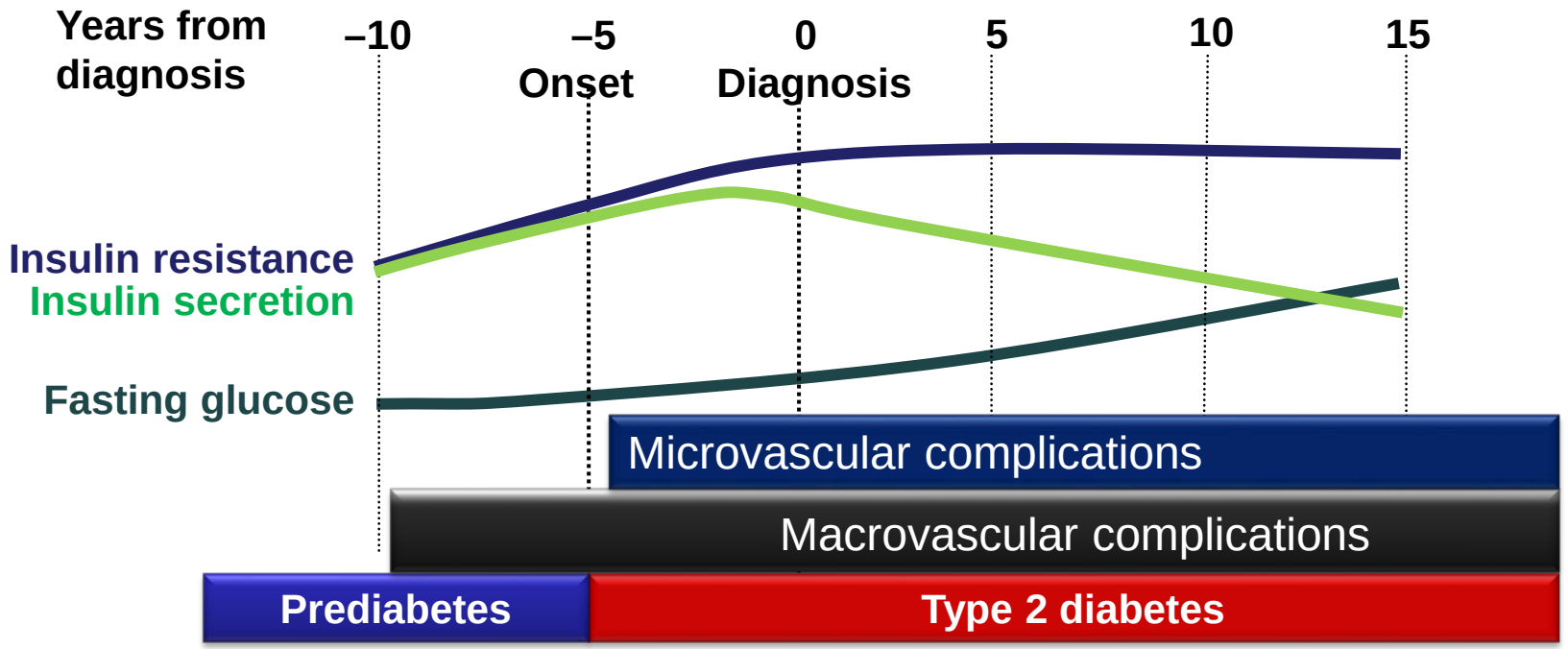
29.1 million people in the US have T2DM (9.3% of population)

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

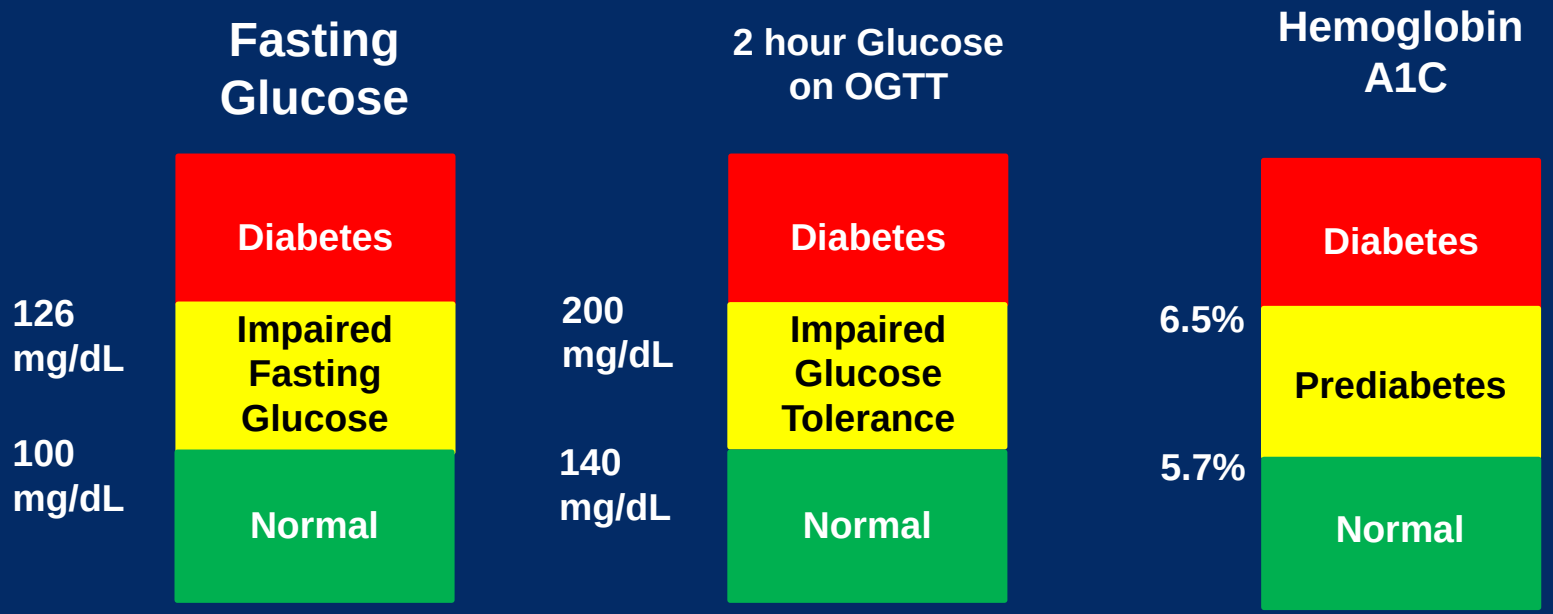


CDC National Diabetes Statistics Report, 2014.

Diabetes: Progression of Disease



Diabetes Screening and Diagnostic Tests



Diabetes Care. 2015;38 Suppl 1: S8-16.

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:
  - What predictive analytics models work best to produce a risk score?
  - What are the strongest predictors?
  - If information is missing, what information should be obtained next?
- Design a data-driven screening strategy to optimize operational consequences.

Related Literature

- Rich literature related to analytics and disease screening in healthcare – we cite a few examples:
- **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 Problem: Initial Screening Decision

12 month of observation
- **Office visits (vitals, ICD-9)**
- **Labs**
- **Medication**

Follow-up period >12 month

2012

2015

1st Encounter

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

Data Set

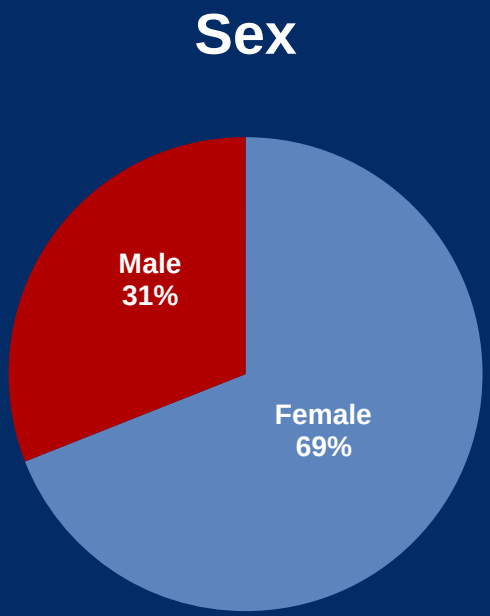
- **Retrospective cohort** (N = 34,297 patients)
- **Cohort Dates:** 2012-2015
- **Setting:** Parkland Health and Hospital System, a large integrated, safety-net healthcare system in North Texas.
- **Data Source:** Epic Electronic Medical Record (EMR)
- **Eligibility:**
  - Ages 18-65
  - ≥1 primary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

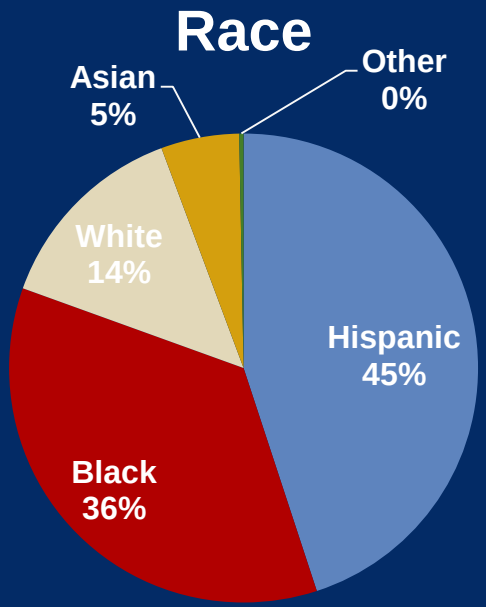
Available Data

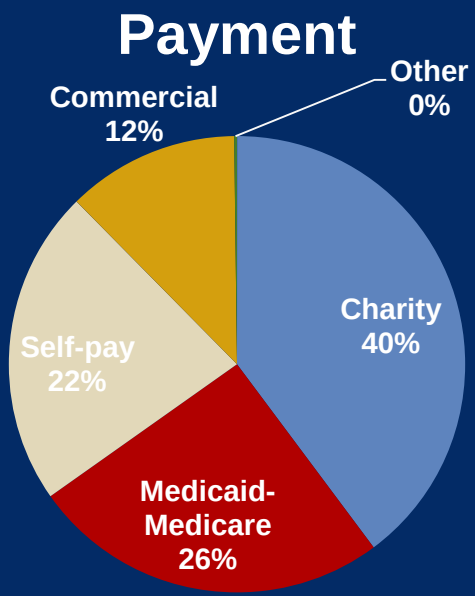
- **Demographic information:** Age, Gender, Race, etc.
- **BMI, vitals:** Blood pressure, etc.
- **Risk factors** (co-morbidities): Hypertension, family history, etc.
- **Lab values:** Cholesterol, random blood glucose, etc.
- **Medications** (prescribed): Blood pressure, cholesterol, etc.
- **Health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

Note: Only demographic information, BMI and vitals are widely available. Other information is often missing.

Cohort Specifics





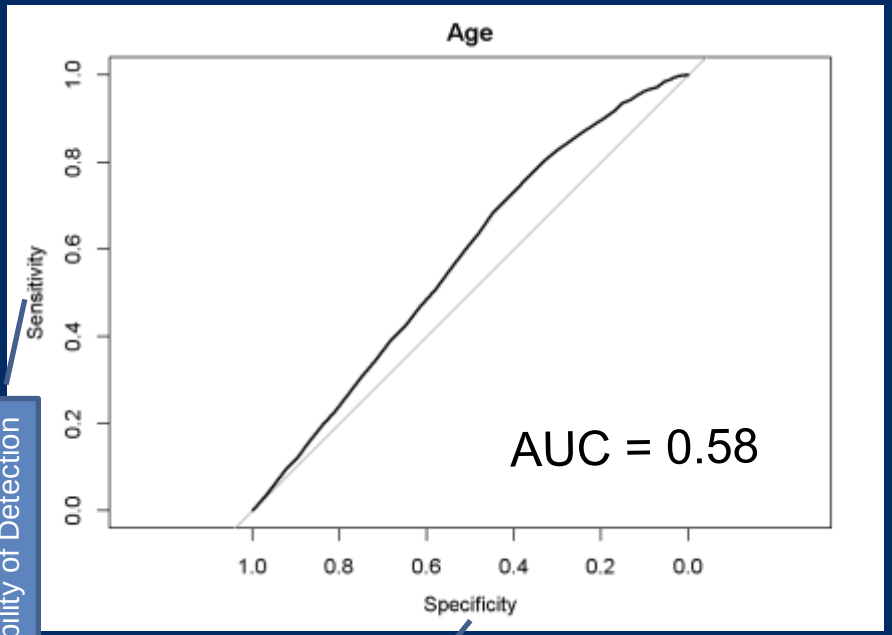


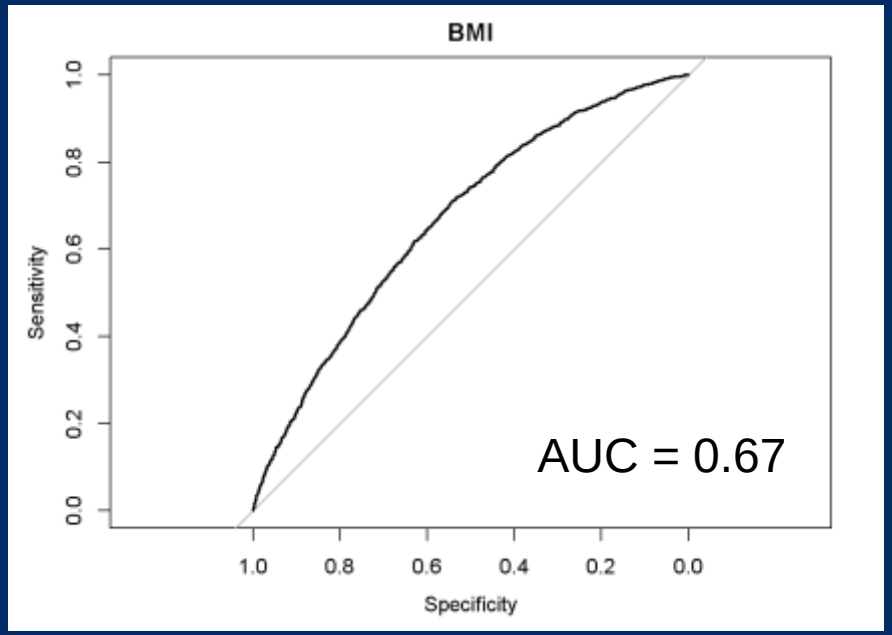
Median age: 46.9 years

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.
- We use a 20% holdout sample.

Single Factor Threshold Models
Usual risk factors: Age and BMI



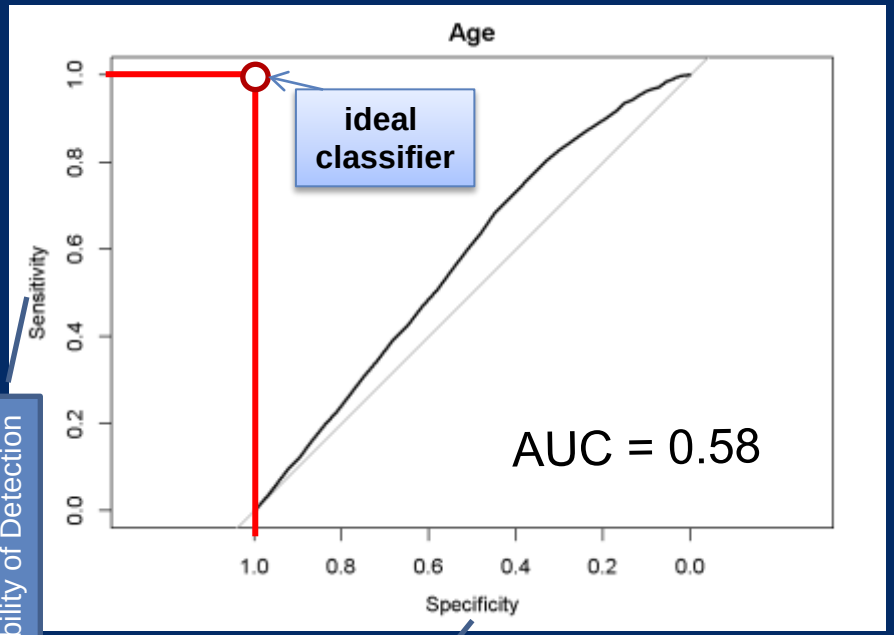


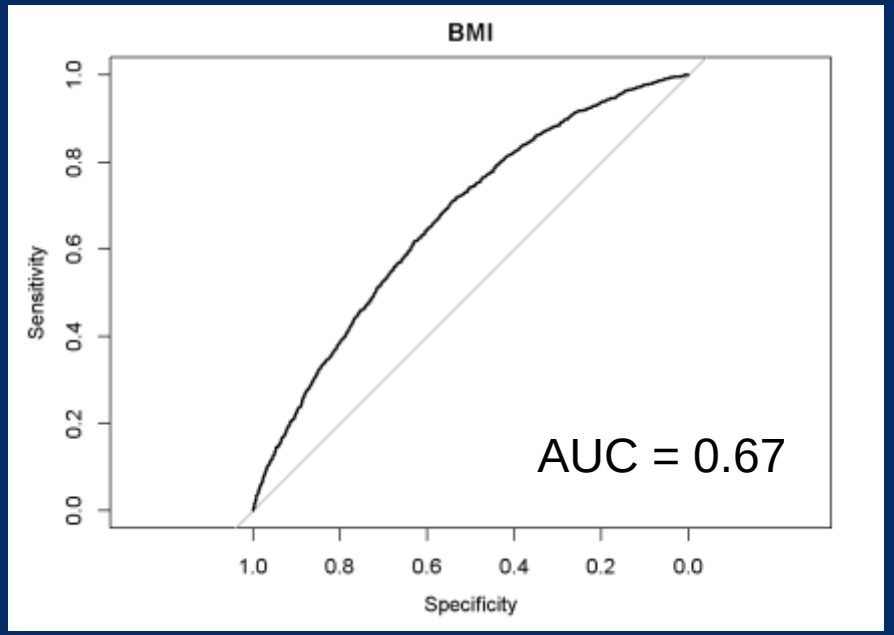
Probability of Detection

1- False Alarm Rate

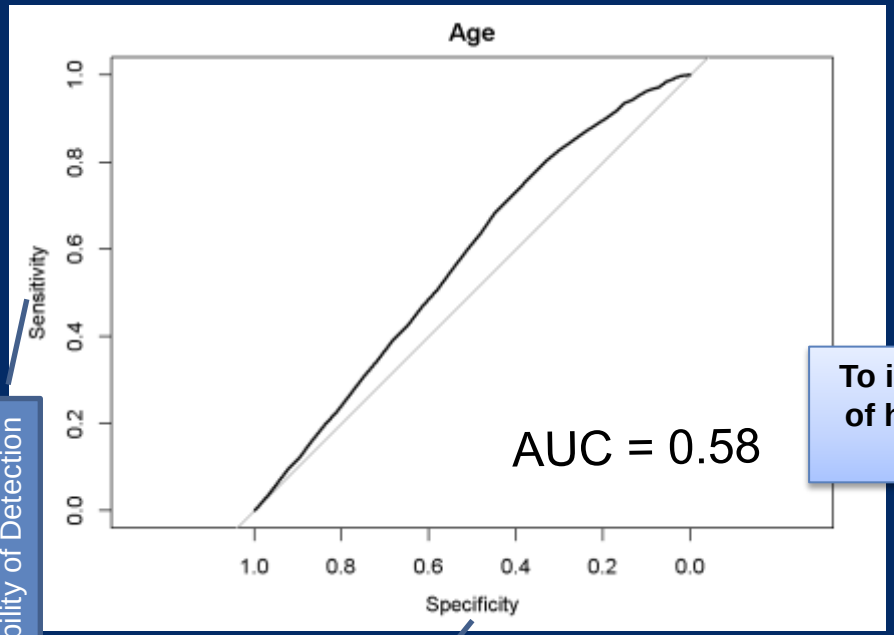
Available for 100% of patients

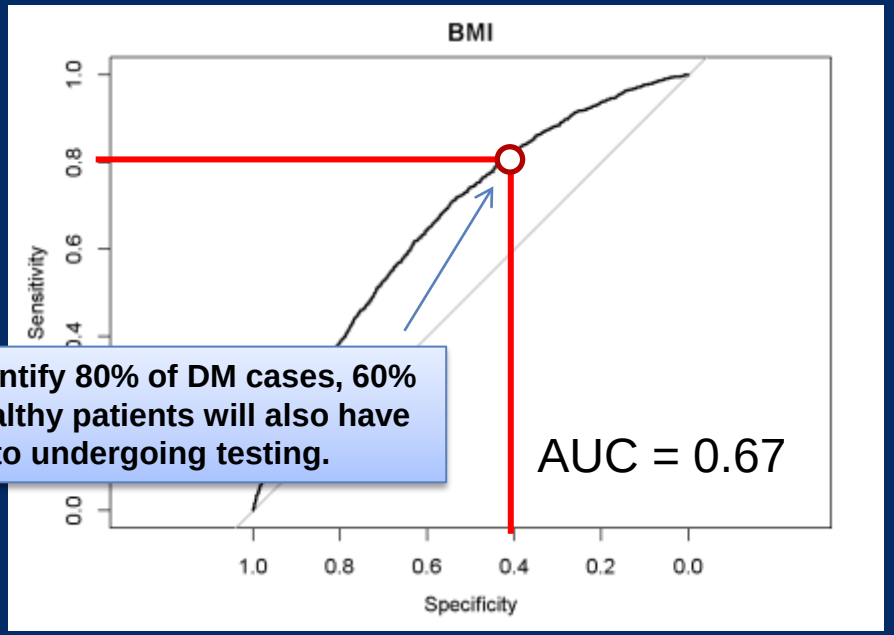
Single Factor Threshold Models
Usual risk factors: Age and BMI





Single Factor Threshold Models
Usual risk factors: Age and BMI



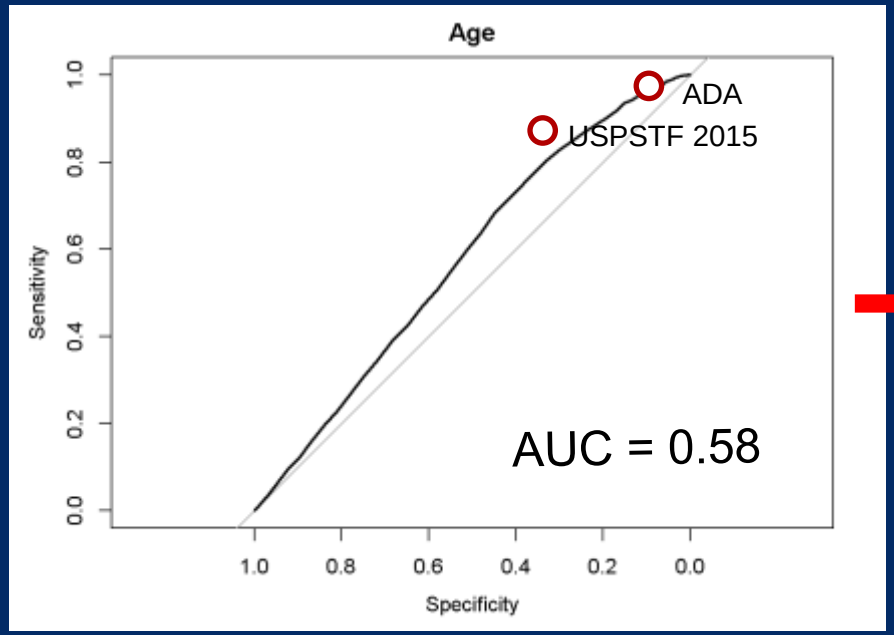


Probability of Detection

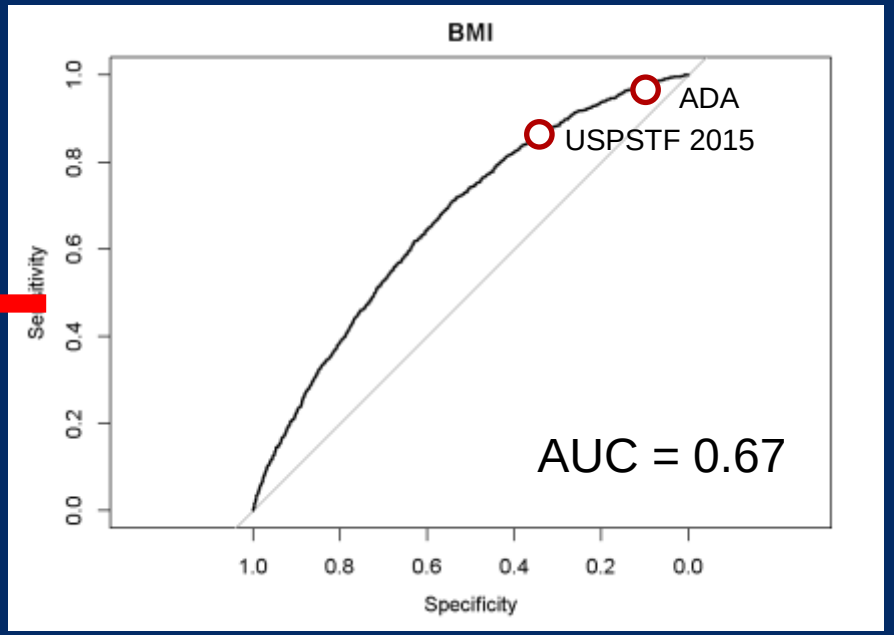
1- False Alarm Rate

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

Single Factor Threshold Models
Usual risk factors: Age and BMI



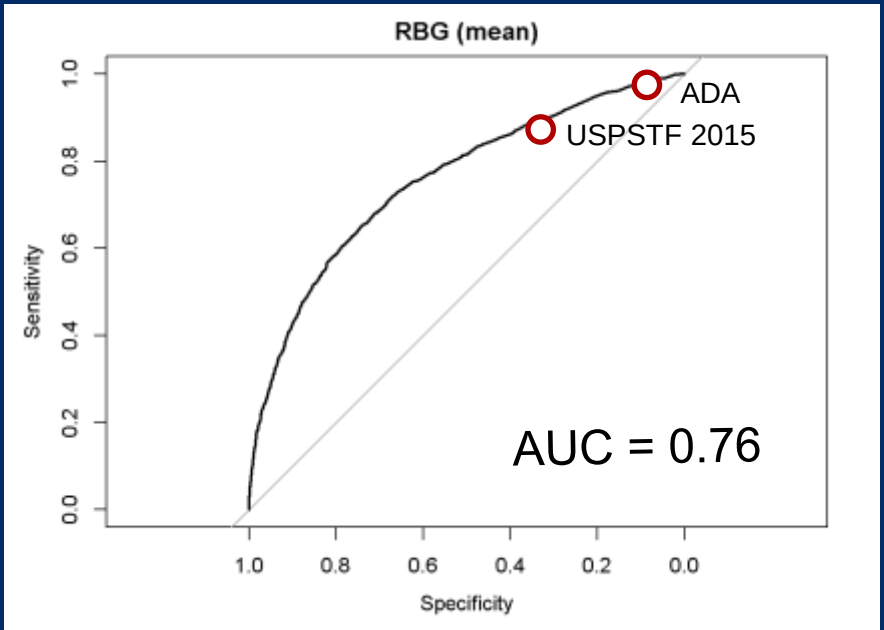
+

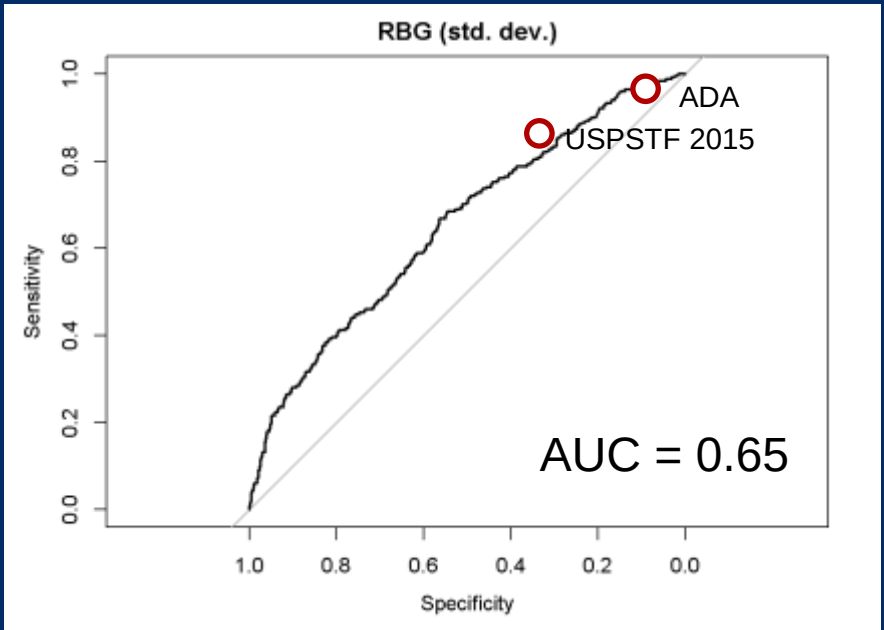


= USPSTF 2015 (Age>40, BMI>25)

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

Naïve Bayes Classifier

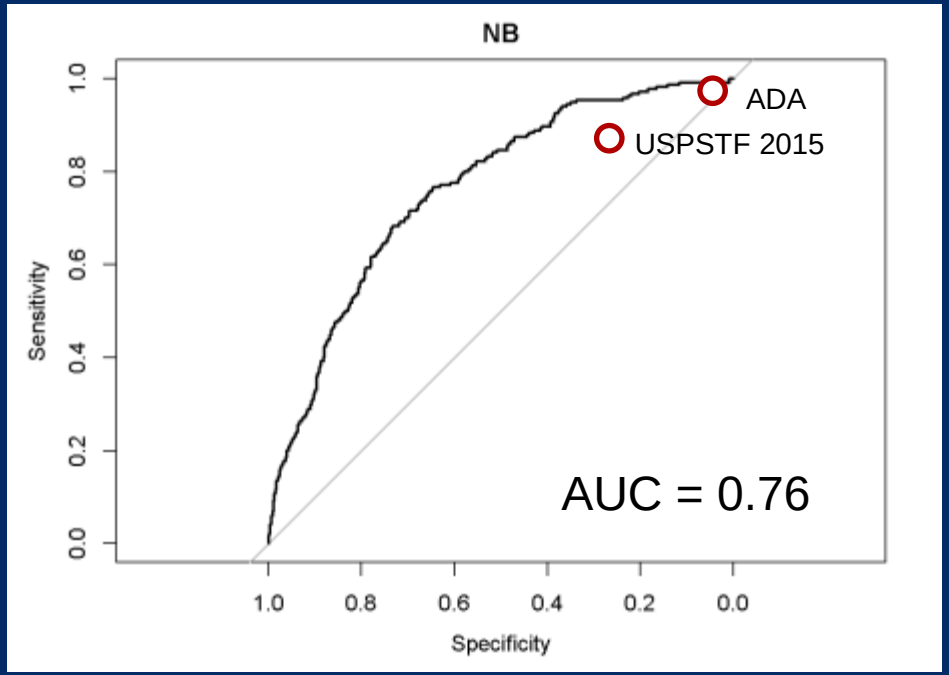
- Applies Bayes' theorem with strong (naive) independence assumptions between the features.

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \; p(C_k) \prod_{i=1}^{n} p(x_i | C_k).$$

- The classifier usually predicts the class C_k (in our case diabetes or not) using the MAP (maximum a posteriori) decision rule. That is, the class with the highest posterior probability given the evidence as features x_i.
- We use a threshold on $p(diabetes)$ to produce a biased classifier.
- Advantages:
  - Missing features in training and test data are allowed.
  - The classifier is known to produce good results, even if the independence assumption is violated.

Russell, Stuart; Norvig, Peter (2003) [1995]. Artificial Intelligence: A Modern Approach

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

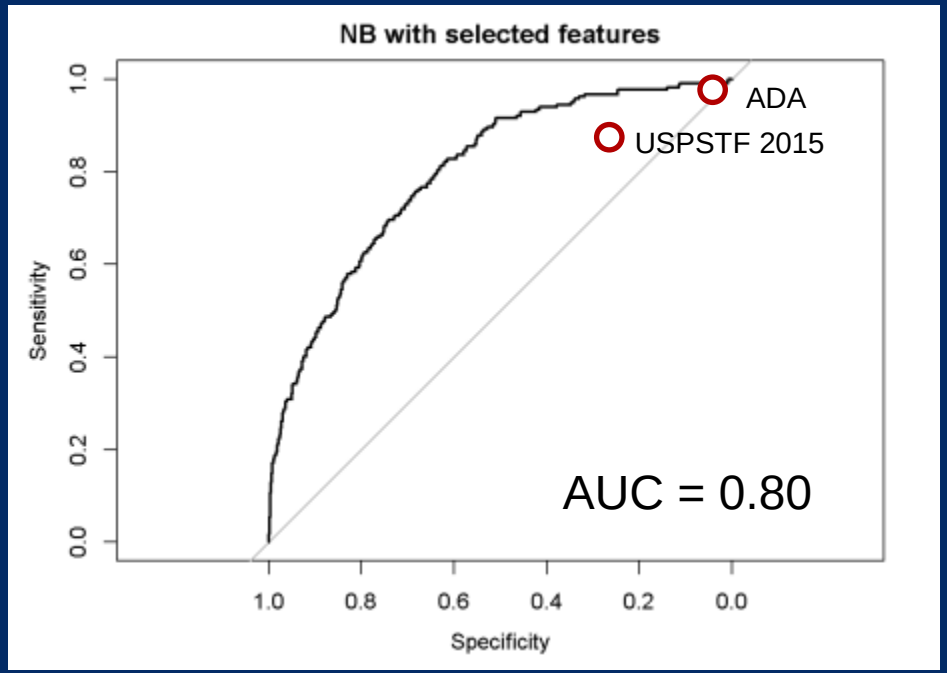


Makes predictions for **all patients**, even if information is missing (e.g., no blood test)

Results are as good as with blood test.

Available for 100% of patients

Multi Factor Model
Naive Bayes Model (NB) - Feature Selection



Available for 100% of patients

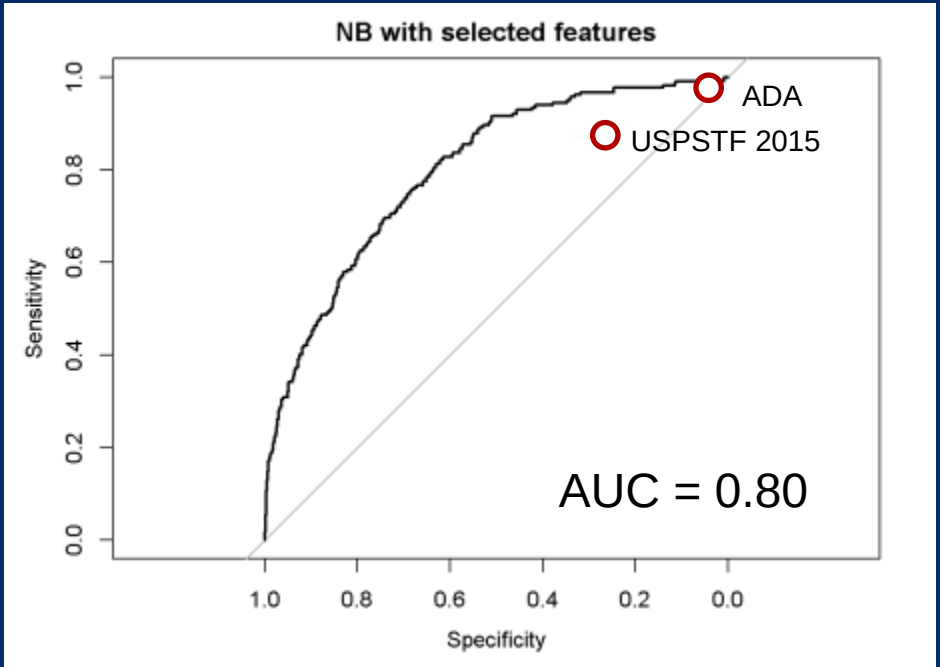
Backward Feature Selection

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

Multi Factor Model
NB - Feature Selection

5 of 10 predictors are not in current guidelines



Available for 100% of patients

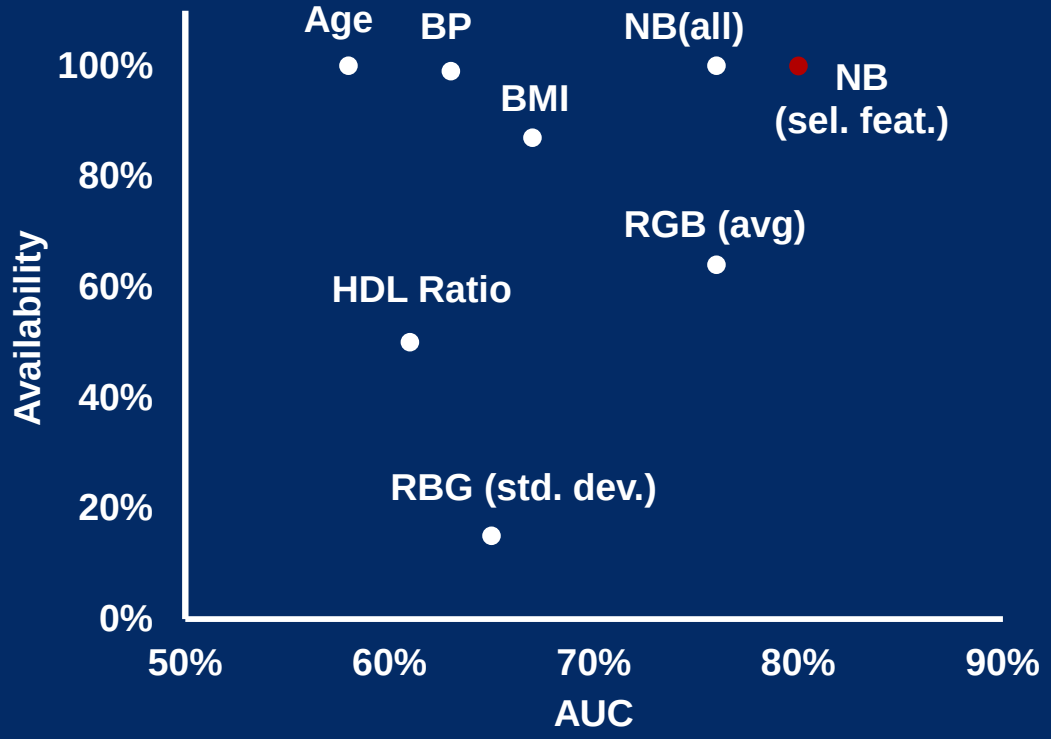
Backward Feature Selection

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

Comparison of Predictive Models

	AUC	Availability
NB (select feat.)	80%	100%
NB (all features)	76%	100%
RGB (avg)	76%	64%
BMI	67%	87%
RGB (std. dev.)	65%	15%
BP	63%	99%
HDL Ratio	61%	50%
Age	58%	100%



Conclusion

- Our naïve Bayes model
  - Reaches on our data an AUC of 80%.
  - Makes predictions for all patients.
  - Uses 10 factors typically already available in electronic health records.
  - 5 of the 10 factors (including the top 2) are currently NOT considered in guidelines.
  - The method easily generalizes to other patient population.

Future Research

- Consider other classification methods.
- 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.