Electronic Health Record Analytics: The Case of Optimal Diabetes Screening

Michael Hahsler1, Farzad Kamalzadeh[1]
Vishal Ahuja[1], and Michael Bowen[2]

[1] Southern Methodist University

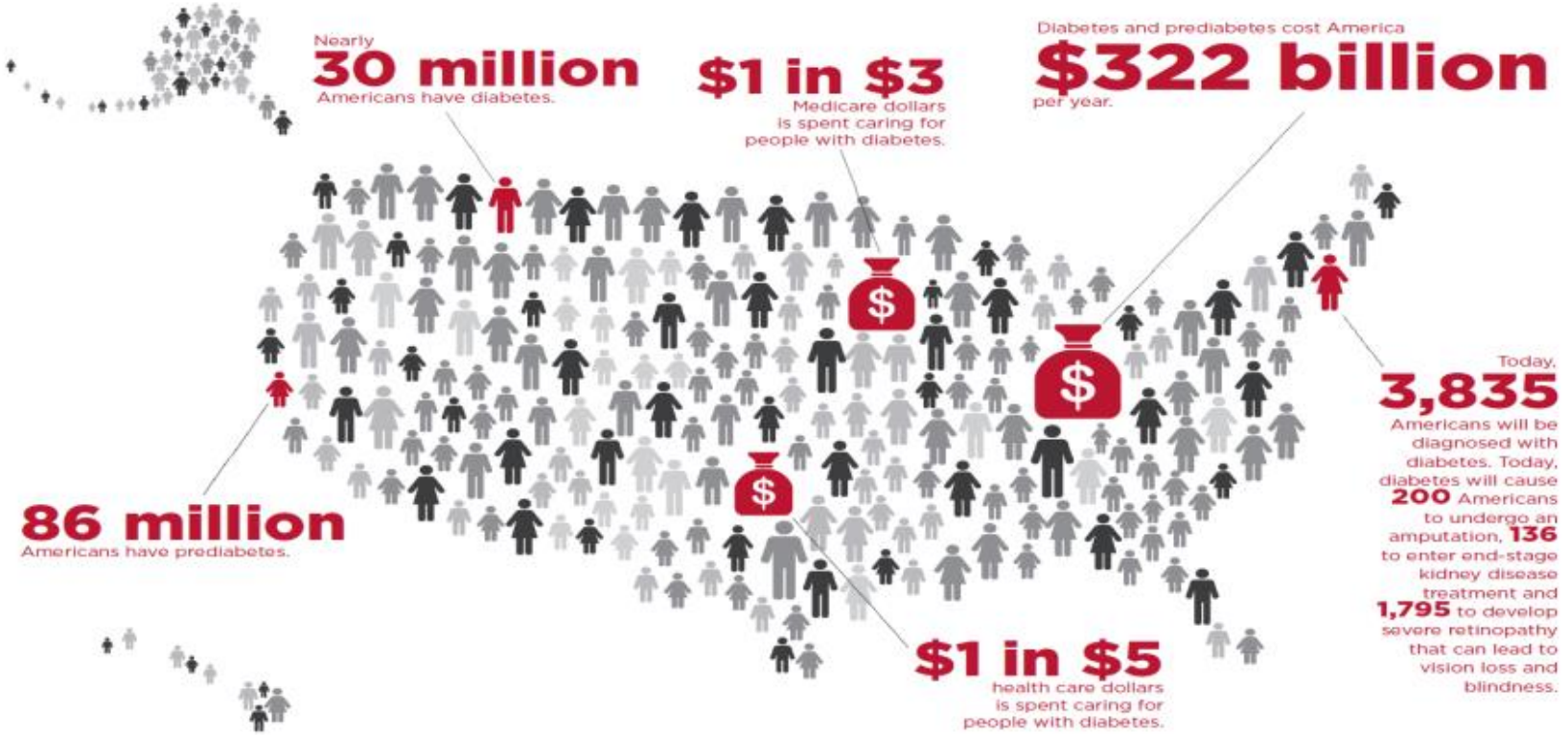
[2] UT Southwestern Medical Center and Parkland Health and Hospital System

April 13, 2018
Artificial Intelligence in Medicine Seminar Series
Division of Medical Physics and Engineering
UT Southwestern

World Changers Shaped Here

SMU

THE STAGGERING COSTS OF DIABETES IN AMERICA



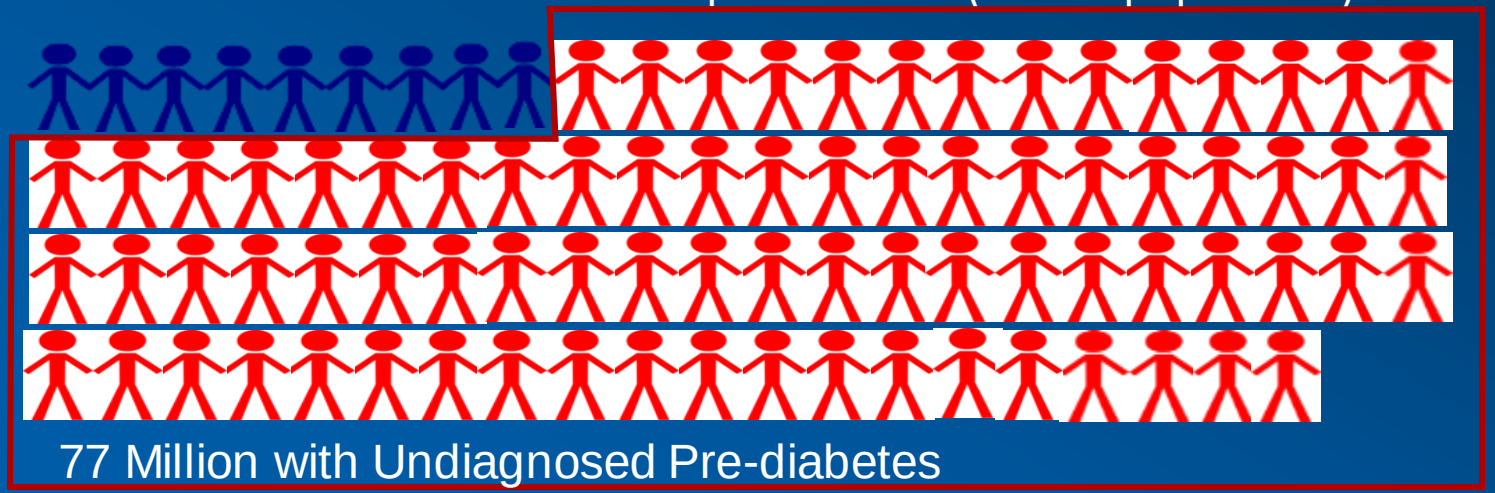
http://main.diabetes.org/dorg/images/infographics/adv-cost-of-diabetes.pdf;
Diabetes Care 2013; 36:1033-1046.

American Diabetes Association.

Prevalence of Diagnosed and Undiagnosed Type 2 Diabetes and Prediabetes

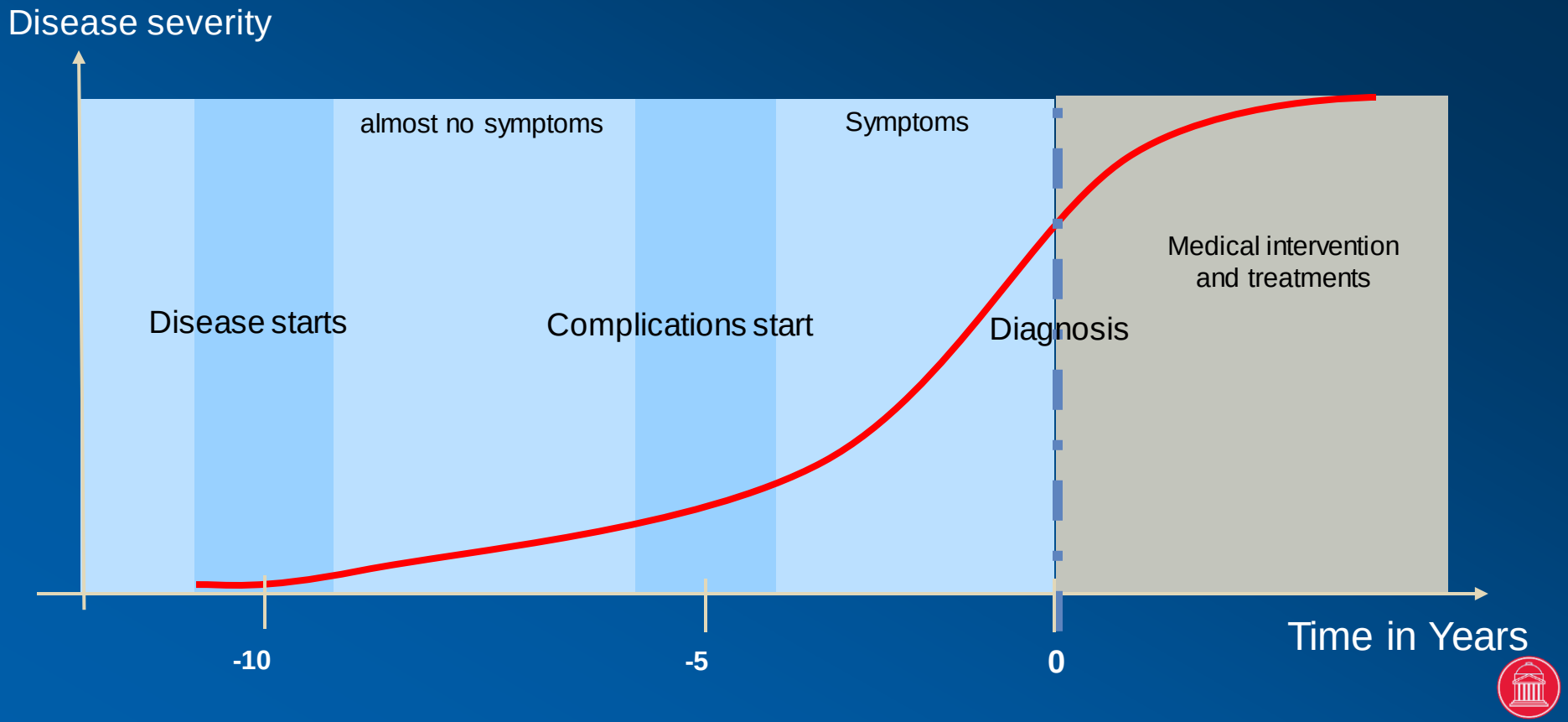
29.1 million people in the US have T2DM (9.3% of population)

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



CDC National Diabetes Statistics Report, 2014.

Nature of Chronic Diseases



Existing Guidelines and Risk Scores

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

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 (EHR)
- **Eligibility:**
  - Ages 18-65
  - Established patients (≥1 primary care visit every 18 month)
  - Only unscreened patients with no known diabetes during first 12 month

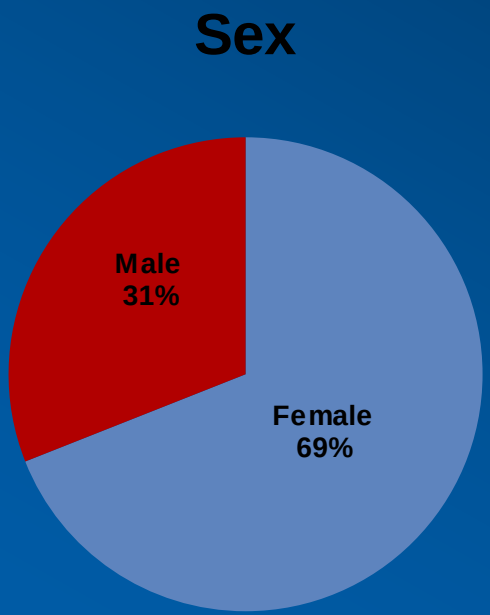
Available Data

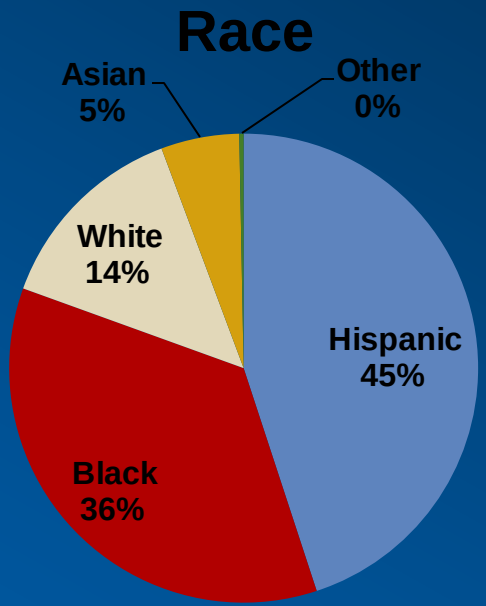
105 Features extracted including

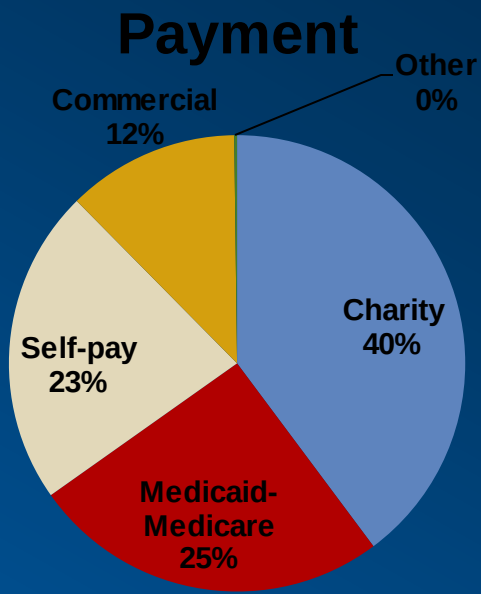
- **Demographic information:** Age, Gender, Race, etc.
- **BMI, vitals:** Blood pressure, etc.
- **Risk factors** (co-morbidities): Hypertension, family history, etc.
- **Lab values:** Cholesterol, random blood glucose, etc.
- **Medications** (prescribed): Blood pressure, cholesterol, etc.
- **Health care utilization:** Office encounters, ER visits, etc.
- **Screening results:** Hemoglobin A1C

Only demographic information, BMI and vitals are widely available.
>20% of the data values are missing overall.
>50% of lab values missing.

Cohort Specifics

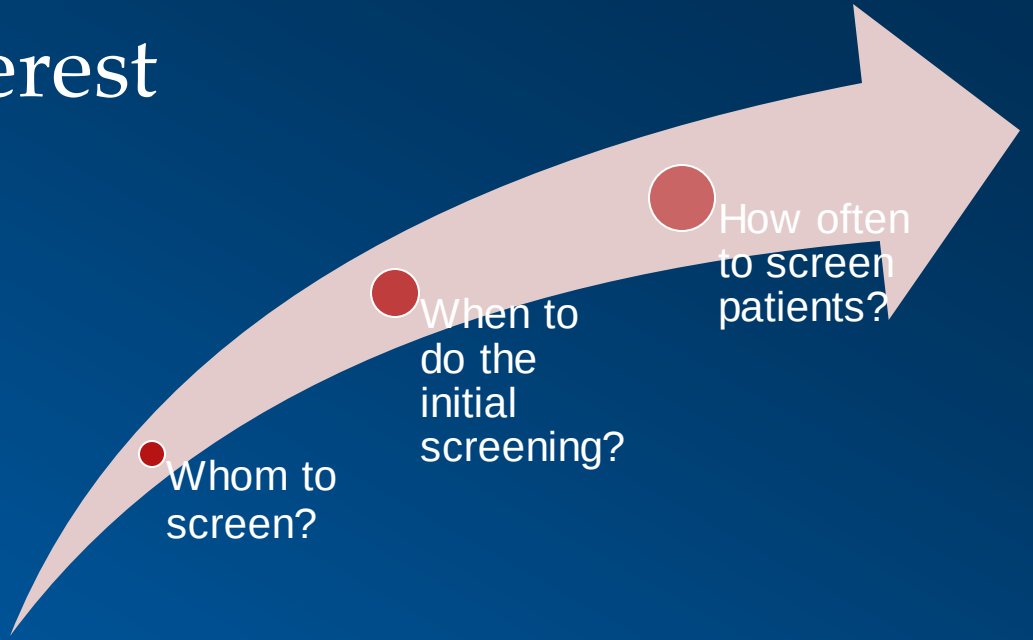






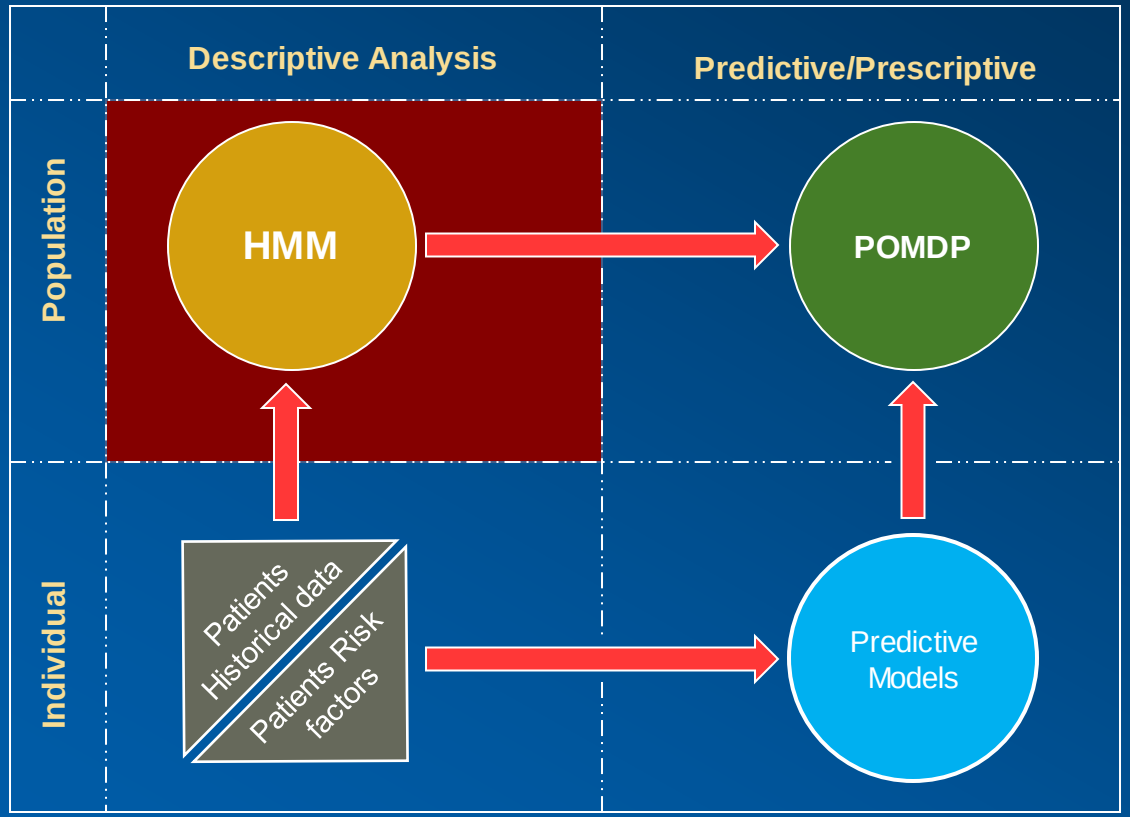
Median age: 46.9 years

Questions of interest

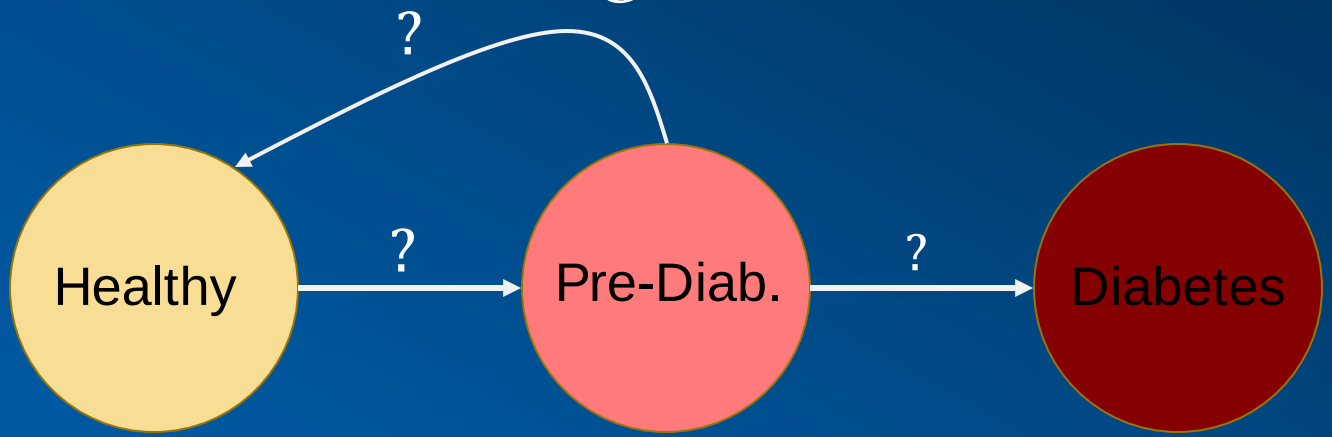


- Optimal screening decision under constraints
  - Constraints on resources and patient availability. Screening almost everyone (e.g., follow ADA Guidelines) is not feasible.
  - Individualize the decision for each patient
  - Focus on catching the disease at earlier stages (such as pre-diabetes)

Framework


Descriptive Analysis
Predictive/Prescriptive
Population
Individual
HMM
POMDP
Patients Historical data
Patients Risk factors
Predictive Models

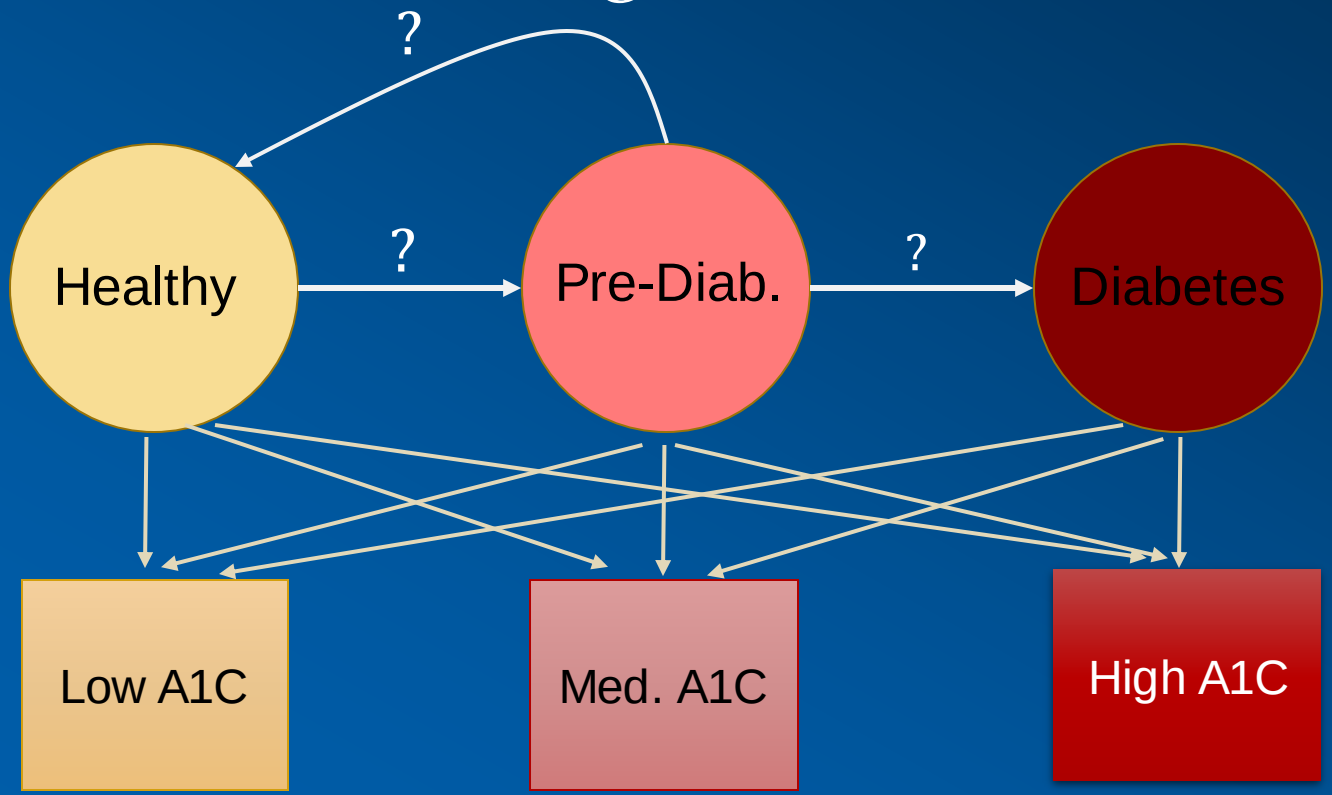
A simple Markov Model for Diabetes Progression



States
with transitions

A simple Hidden Markov Model (HMM) for Diabetes Progression

HMM

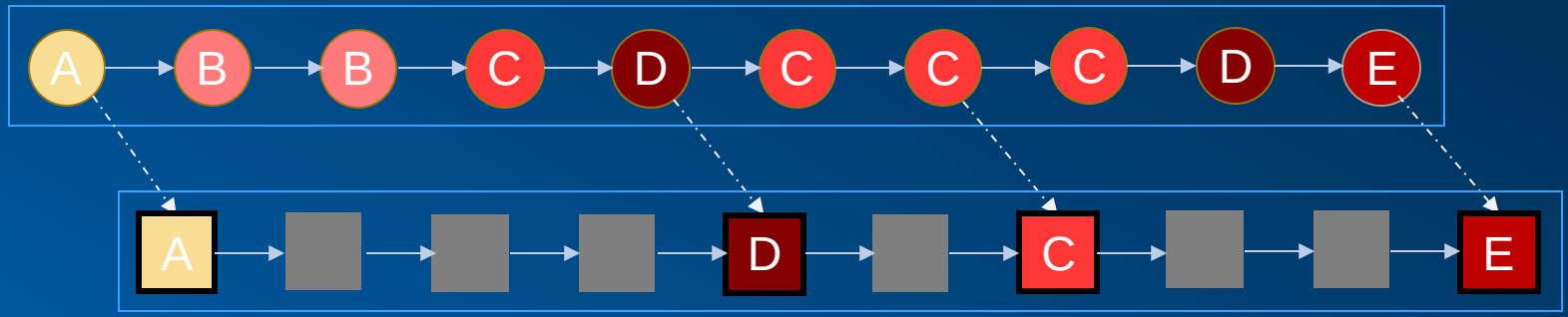


Hidden States
with transitions

Observations

Transition Parameter Estimation

HMM



	t=1	t=2	t=3	t=4	t=5	t=6	t=7	t=8	t=9	t=10
Patient 1	A				D		C			E
Patient 2	C		D	D		E				
Patient 3	A					B	A		A	
Patient 4		B		C						C
Patient 5	C		C					E		
Patient 6		D					D			E
Patient 7			B			C			B	

Baum–Welch algorithm

HMM

$\lambda = (A, B, \pi)$

for each sequence

while desired level of convergence not acquired

for t= 1 to T

for i in S

$$\alpha_i(t) = P(Y_1 = y_1, Y_2 = y_2, \dots, Y_t = y_t | X_t = i, \lambda)$$

the probability of seeing the $Y_1 = y_1, Y_2 = y_2, \dots, Y_t = y_t$ and being in state i at time t

$$\beta_i(t) = P(Y_{t+1} = y_{t+1}, Y_{t+2} = y_{t+2}, \dots, Y_T = y_T | X_t = i, \lambda)$$

the probability of the ending partial sequence $Y_{t+1} = y_{t+1}, Y_{t+2} = y_{t+2}, \dots, Y_T = y_T$ given starting state i at time t

$$\gamma_i(t) = P(X_t = i | Y, \lambda) = \frac{\alpha_i(t).\beta_i(t)}{\sum_{j=1}^{N} \alpha_j(t).\beta_j(t)}$$

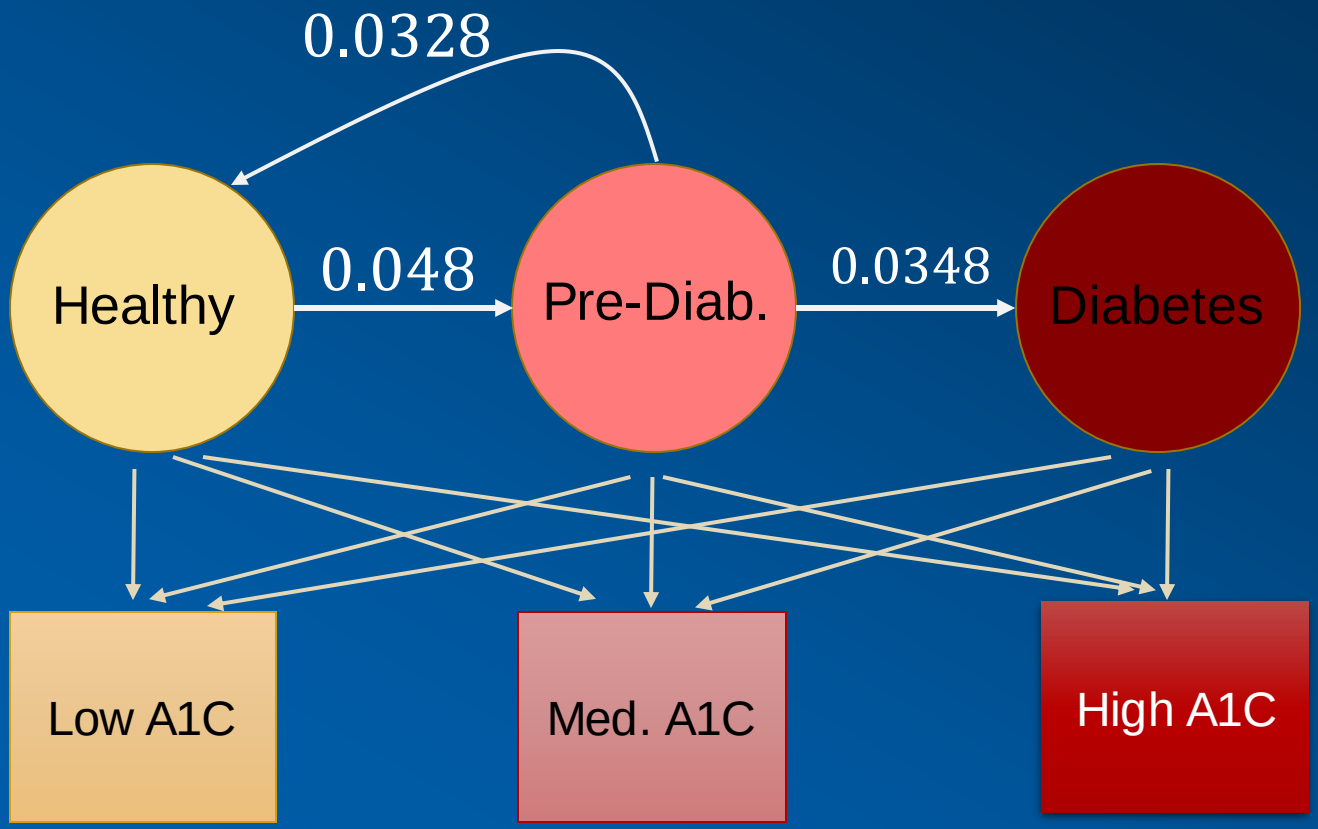
the probability of being in state i at time t given the observed sequence Y and the parameters λ

$$\delta_{ij}(t) = P(X_t = i, X_{t+1} = \mathrm{j} | Y, \lambda) = \frac{\alpha_i(t) a_{ij}.\beta_i(t+1) b_j(y_{t+1})}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i(t) a_{ij}.\beta_i(t+1) b_j(y_{t+1})}$$

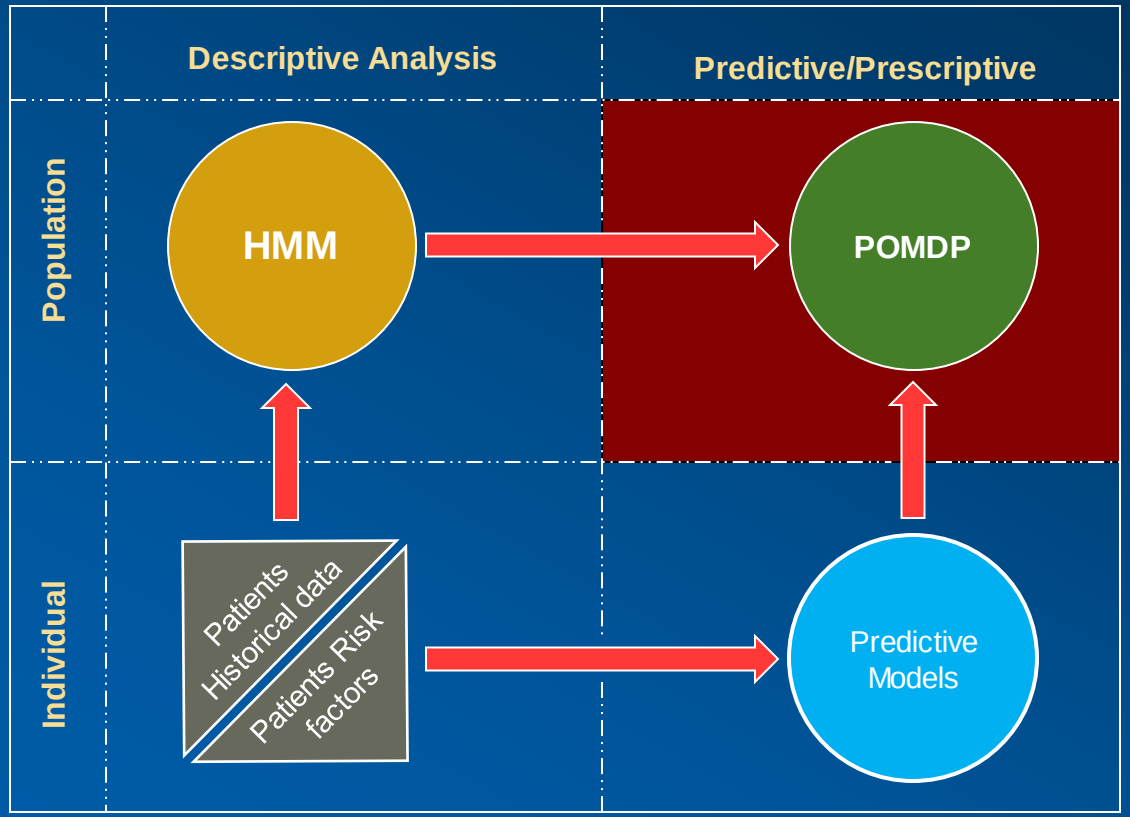
the probability of being in state i and j at times t and t+1 respectively given the observed sequence Y and parameters λ

$$update: \qquad \pi_i = \gamma_i(1) \qquad a_{ij} = \frac{\sum_{t=1}^{T-1} \delta_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)} \qquad b_i(v_k) = \frac{\sum_{t=1}^{T} 1_{y_t = v_k} \gamma_i(t)}{\sum_{t=1}^{T} \gamma_i(t)}$$

Result of Baum–Welch algorithm



Framework


Descriptive Analysis
Predictive/Prescriptive
Population
Individual
HMM
POMDP
Patients Historical data
Patients Risk factors
Predictive Models

POMDP for Diabetes

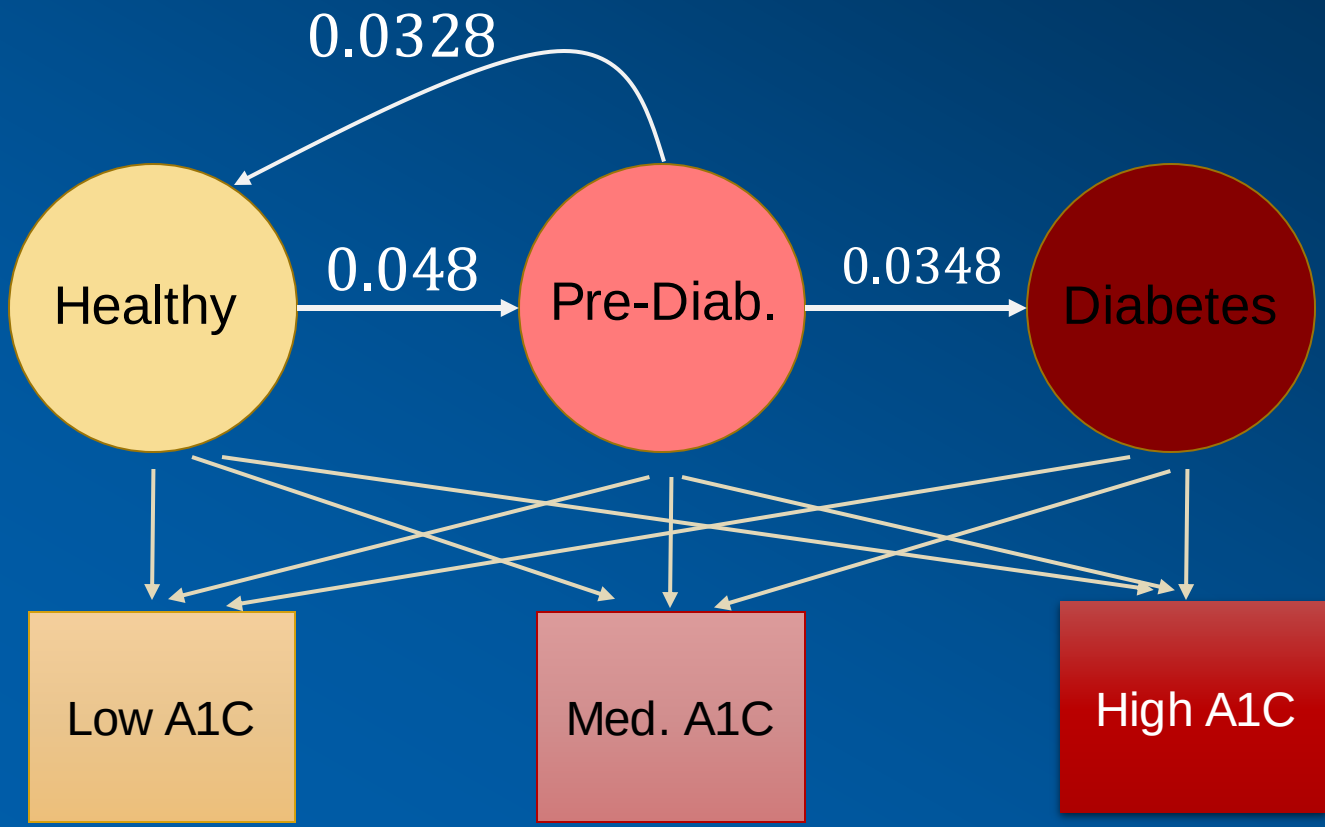
POMDP

- A Markov decision process (MDP) adds the following elements to a Markov model:
  1. **Actions** which affect transition between states.
  2. **Rewards** for actions in different states.

- The goal is to find an **optimal policy.** I.e., what action to take in each state to maximize the expected reward.

- Partially observable MDP (POMDP):
  States are not directly observable like in HMMs. POMDP keeps track of belief states.

POMDP

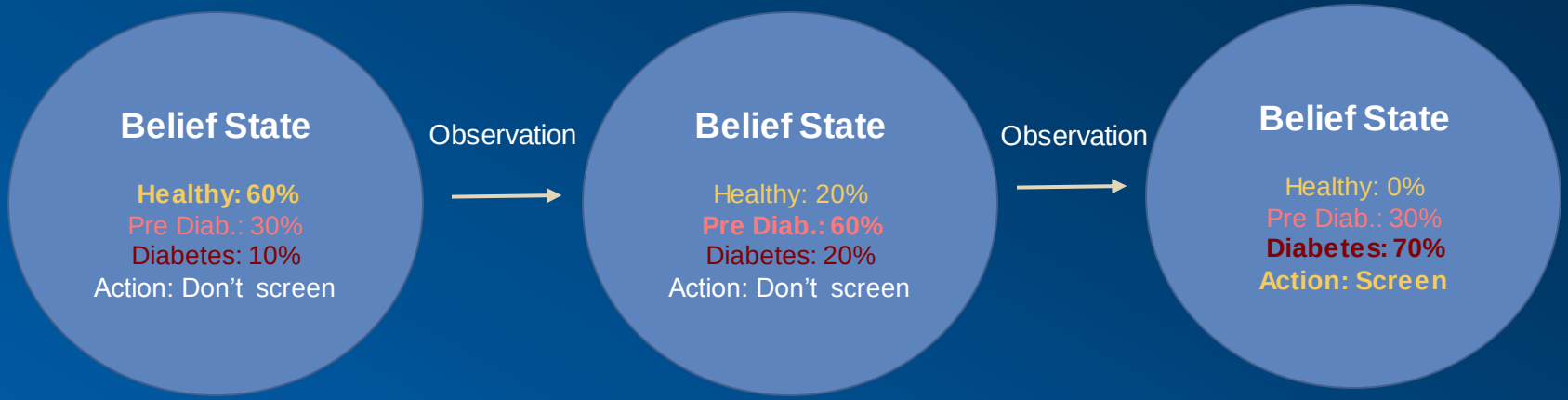


+ Actions

+ Rewards

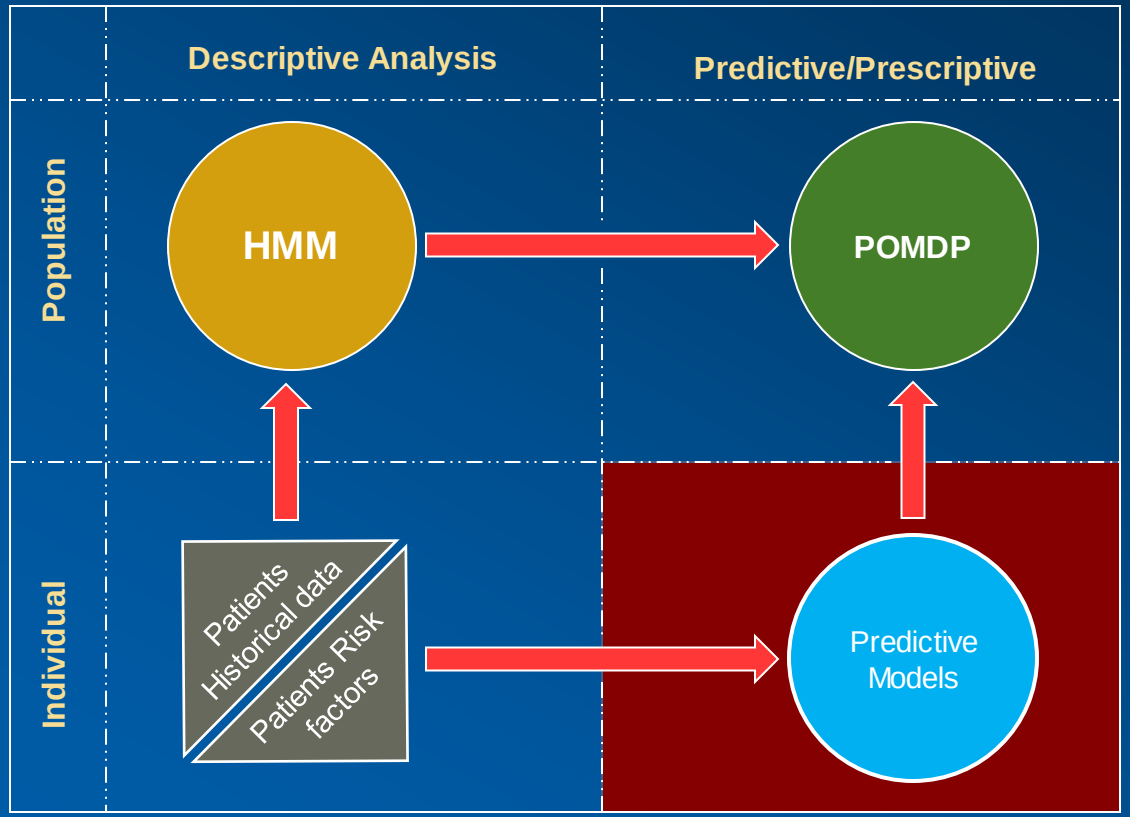
Belief States and Policy

POMDP



- Belief states represent our "belief" about in what state the patient currently is.
- Observations change the belief state.
- Belief states have associated actions that maximize the expected reward.

Framework


Descriptive Analysis
Predictive/Prescriptive
Population
Individual
HMM
POMDP
Patients Historical data
Patients Risk factors
Predictive Models

Observations via Predictive Modeling

PM

- POMDP needs observations, but health status cannot be directly observed unless we screen!
- **Idea:** Use other clinical observations recorded in EHRs as a proxy and learn the relationship to the A1C using predictive modeling.

- Our key questions are:
  - How to produce **simple predictive models** to guide screening using only already available data?
  - How do we deal with a large quantity of **missing data**?
- Desired properties:
  - Applicable to all patients, no matter how much information we have.
  - Can guide us to what missing patient information would be most valuable.

Related Literature

PM

Collins et al. (2011): Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting.

- Surveys **39 studies** with 43 risk prediction models
- Models use **4-64 predictors** (most common: age, family history, BMI, hypertension, fasting glucose)
- Most common modeling method: **Logistic regression**

- Missing data: Almost all (50%) **remove incomplete cases** or do not mention missing data. One study uses imputation.

Predictive Problem: Initial Screening Decision

PM

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

Follow-up period >12 month

2012

2015

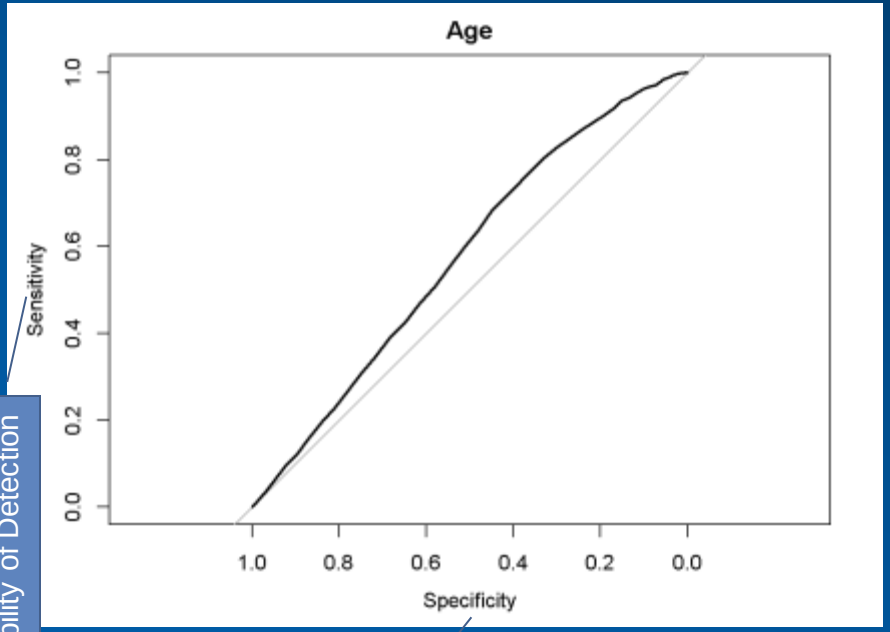
1st Encounter

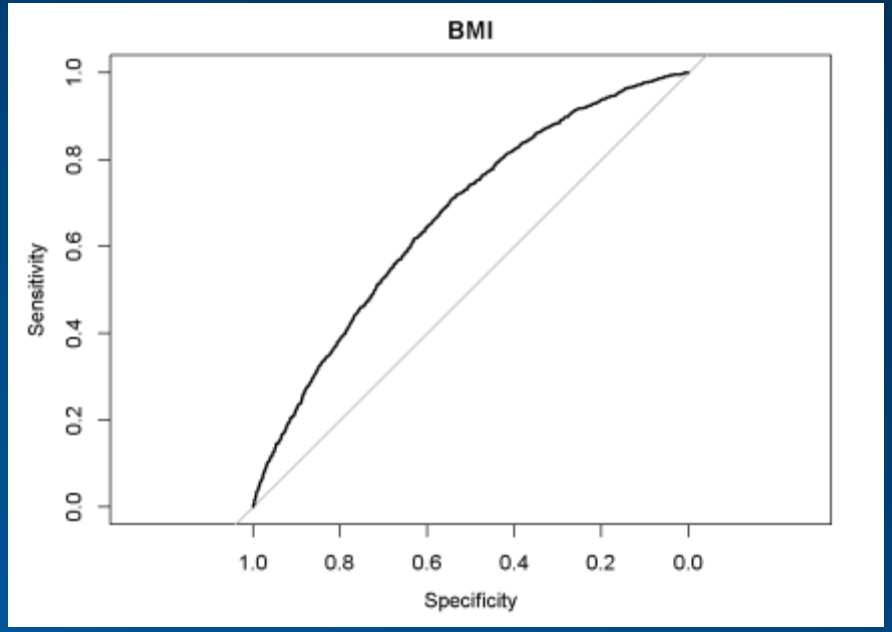
Predict if the patient has or will develop diabetes and should be screened

13.6% in the cohort are diagnosed with diabetes in the follow-up period.

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

PM



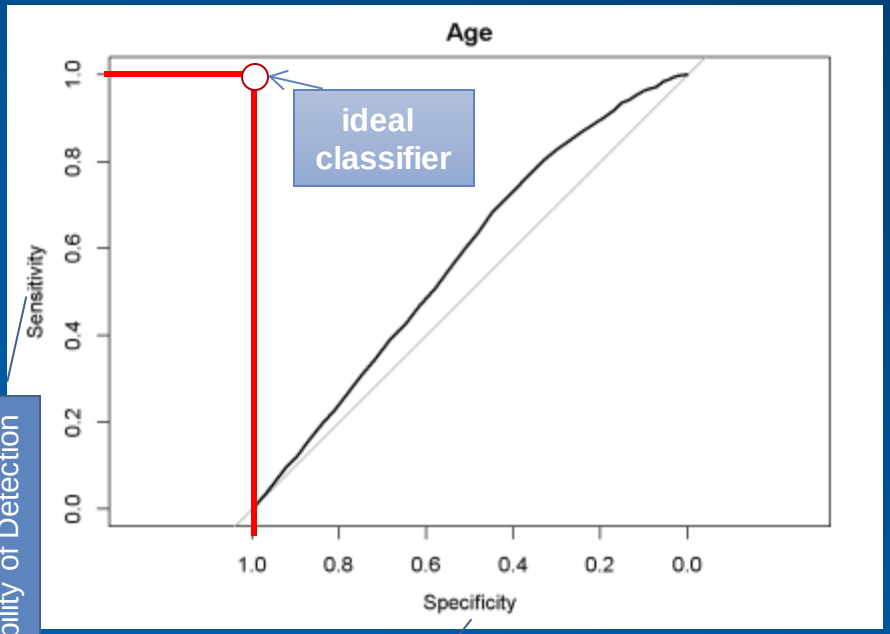


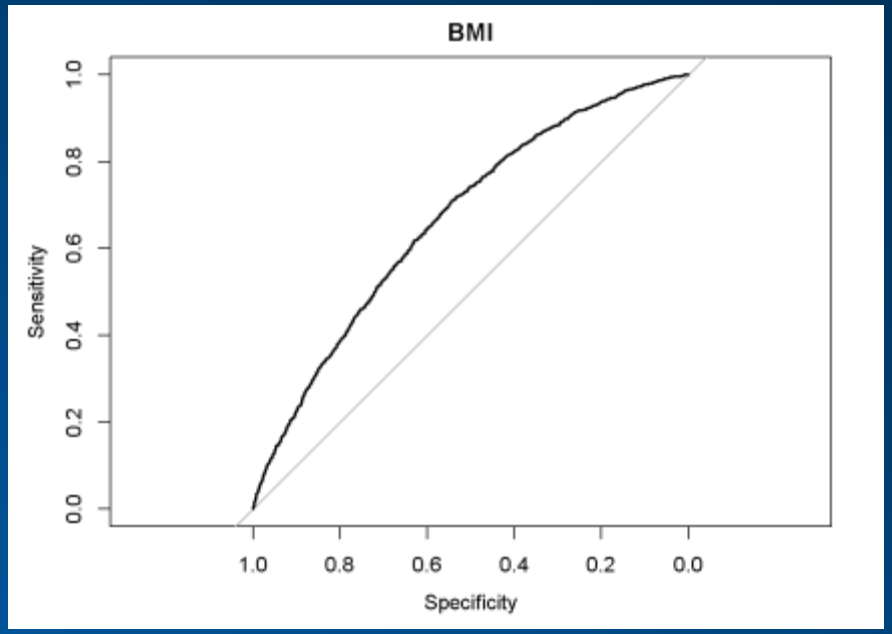
Probability of Detection

1- False Alarm Rate

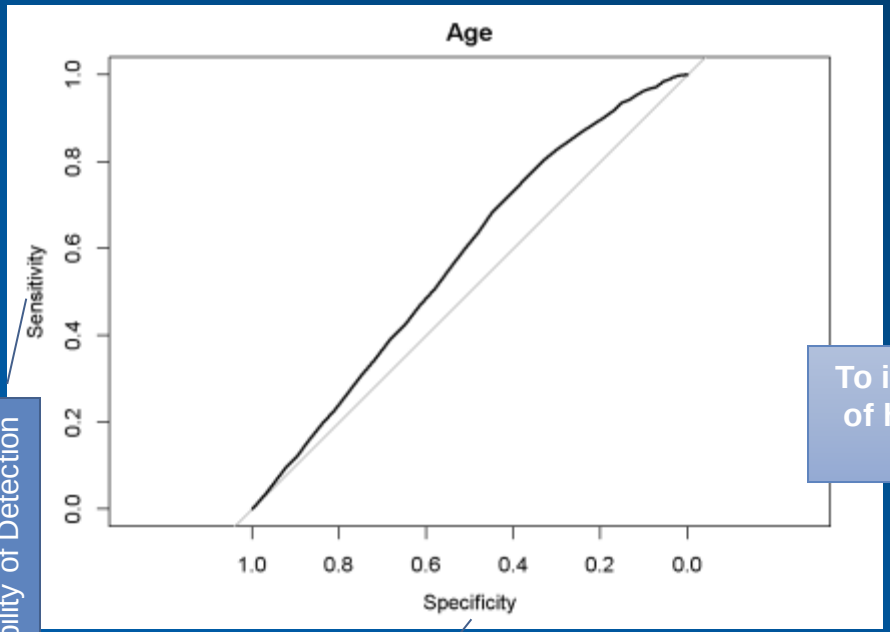
Available for 87-100% of patients

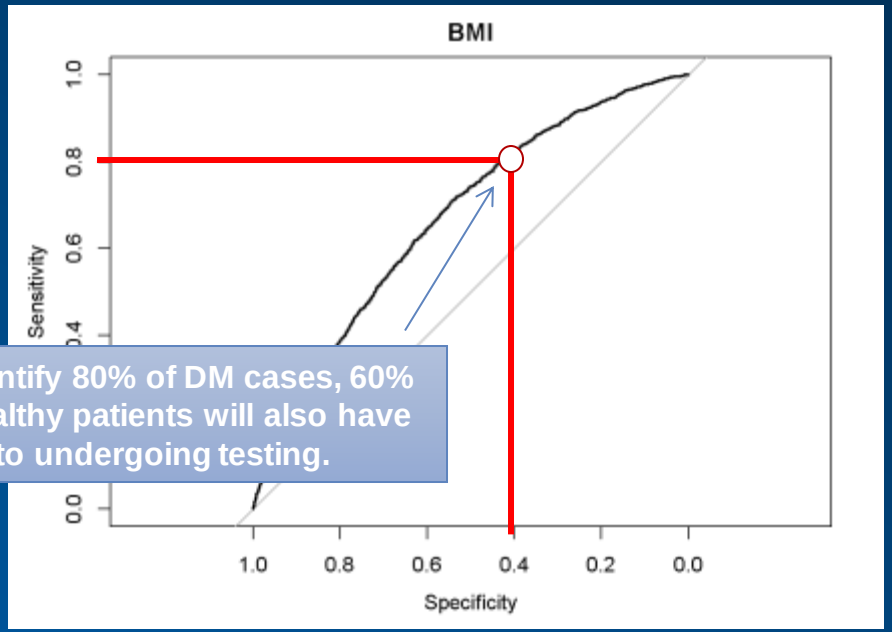
Single-Factor Threshold Models
Usual risk factors: Age and BMI





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



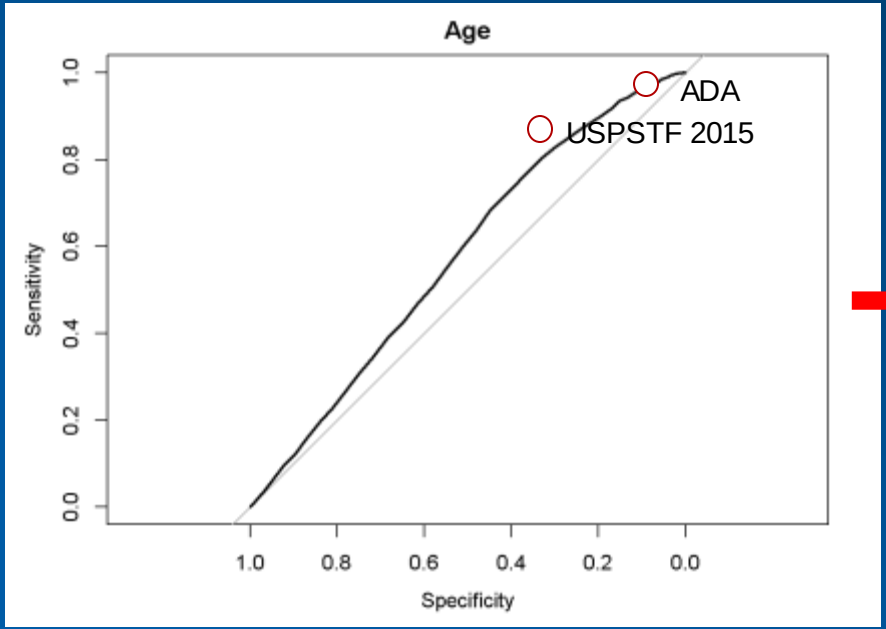


Probability of Detection

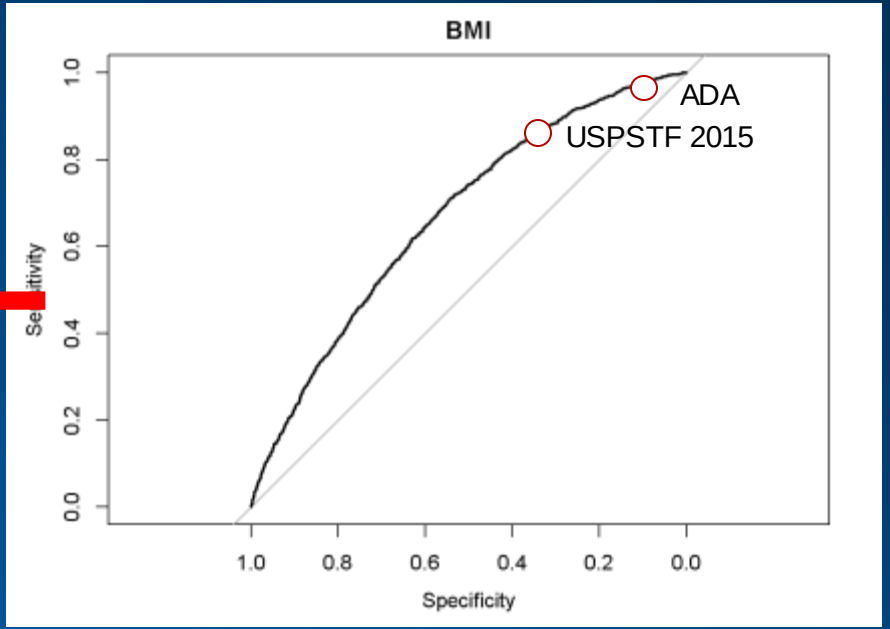
1- False Alarm Rate

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



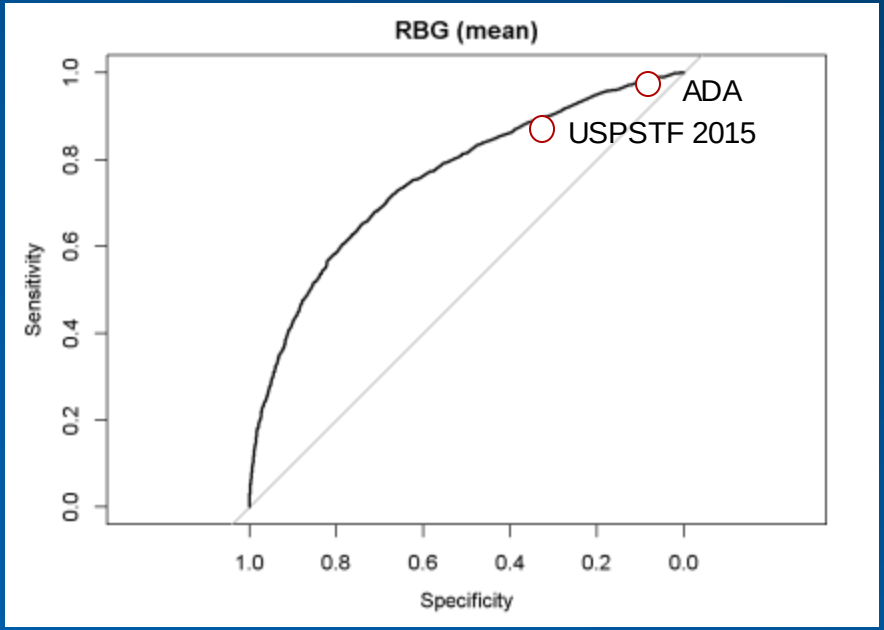
+

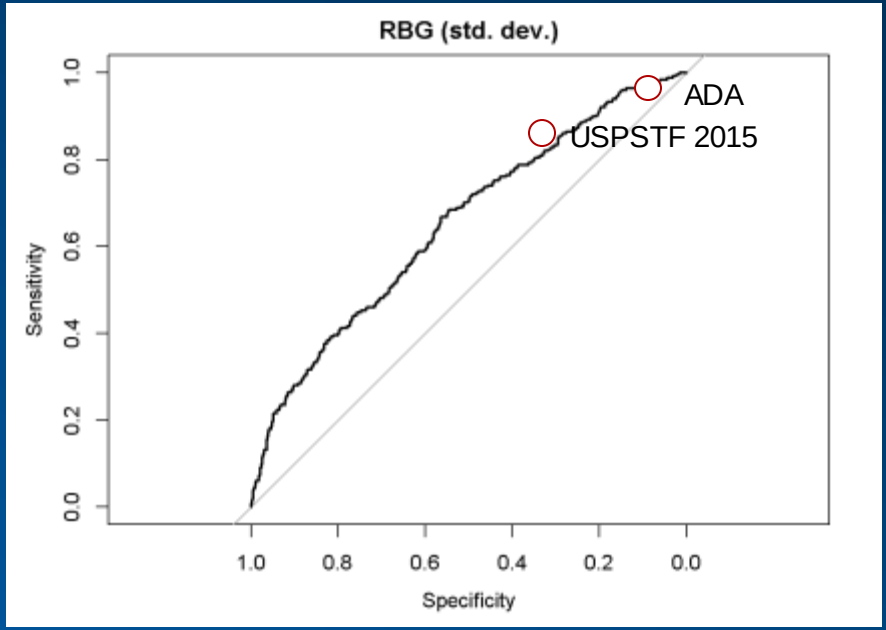


= USPSTF 2015 (Age>40, BMI>25)

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

Drawbacks for Single-Factor Models

PM

- Ignores important available information.
- What if exactly the needed factor is not available (e.g., no blood test)?

Multi-Factor Models

PM

- 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 feature selection
  - Logistic regression with LASSO regularization

- Both models apply feature selection, but dealing with missing data needs more consideration.
- We will use a 20% holdout sample for testing.

Dealing With Missing Values

PM

- 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., patient does not undergo a test because of financial considerations.
  - **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:
  a. Ignore feature with missing values
  b. Ignore observations with missing values

  } Not practical for the data set. No data left.

  c. Pairwise deletion (ignore just the missing values) – needs to be supported by the method
  d. Imputation (e.g., mean imputation)
  e. Imputation + indicator for missingness

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

Naïve Bayes Classifier

PM

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

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

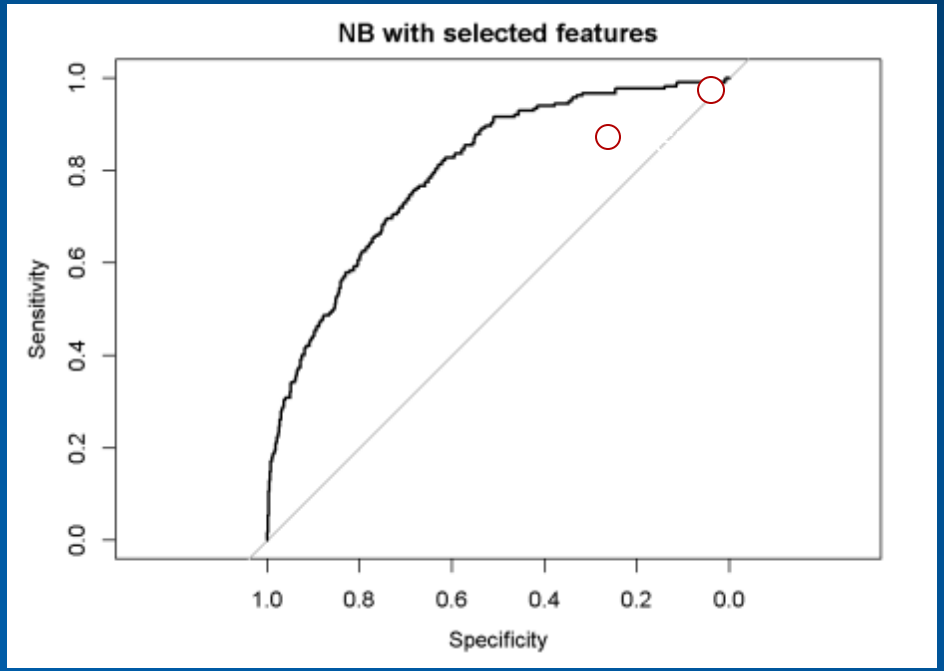
- C_k is the class, $\boldsymbol{x}$ is a feature vector. We use a threshold on $p(C_{diabetes}|\boldsymbol{x})$ to produce a biased classifier.
- Metric predictors: we assume Gaussian distributions (given the target class).
- **Missing values:**
  - Method supports pairwise deletion: leave out missing values for the computation of the probability factors and omit components for prediction.
  - Implies MCAR!
  - Missing indicator can potentially preserve information for MNAR.

Russell, Intelligence: A Modern Approach

Multi Factor Model NB – Forward Feature Selection

2 of top 10 predictors are not in current guidelines

PM



Available for 100% of patients

Forward Feature Selection

	Feature	AUC
1	BMI	64.74%
2	LAB_RANDOM_GLUCOSE_MEAN	69.72%
3	BP_SYSTOLIC	71.27%
4	LAB_HIGH_DENSITY_CHOL	72.19%
5	AGE	72.75%
6	LAB_ALANINE_AMINOTRANSFERASE	73.23%
7	MED_CHOL	73.56%
8	MED_DM	73.81%
9	PULSE	74.08%
10	PATIENT_RACE_White	74.26%

- Mean imputation hurts the results.
- Missing indicators improves the results from 0.758 to 0.762.

Generalized Linear Model with LASSO

PM

- GLM for binomial response with L1 regularization.

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{N} \sum_{i=1}^{N} Cost\left(h_{\boldsymbol{\beta}}(\boldsymbol{x_i}), y_i\right) \right\} \qquad s.t. \|\boldsymbol{\beta}\|_1 \leq t$$

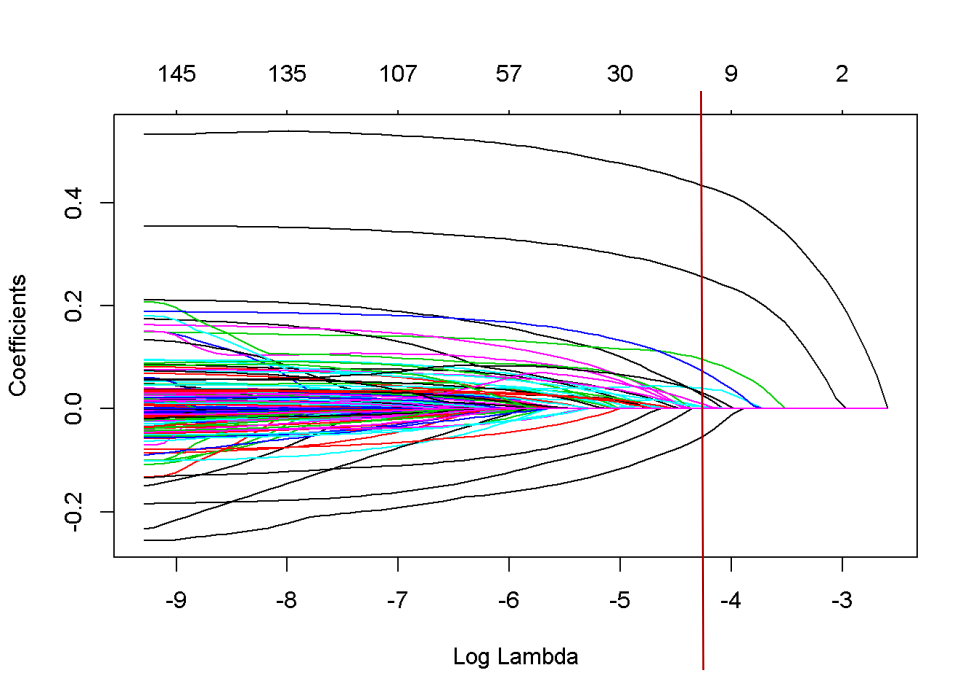
- All variables are scaled to Z-scores.
- **Missing values:**
  - Method needs imputation.
  - Numeric values: Mean imputation and add a dummy indicator variable.
  - Nominal variables: add an additional value for missing data.

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

Logistic Regression with LASSO

Most important of top 10 predictors is not in current guidelines

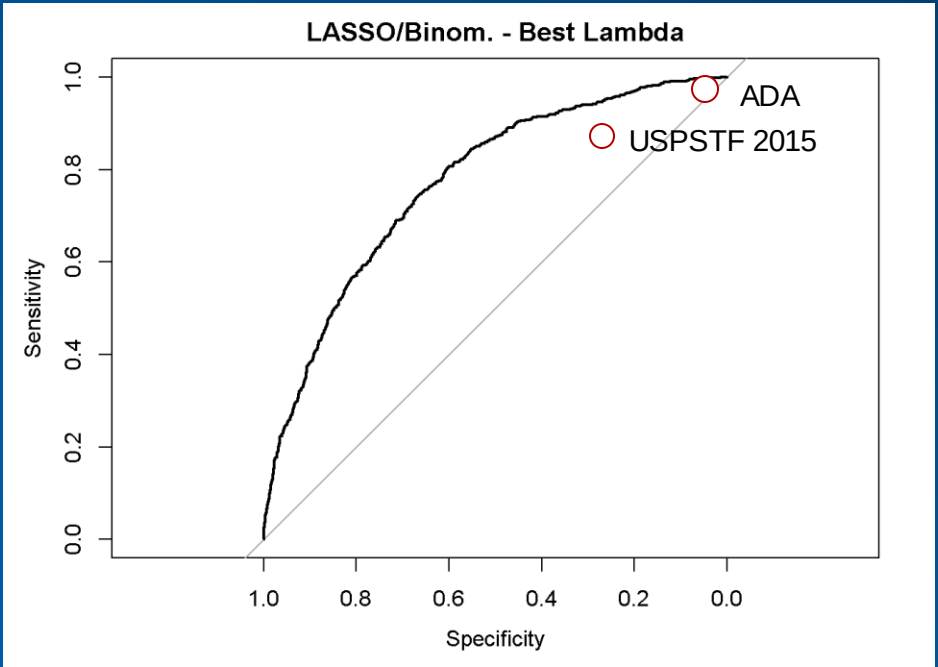
PM



First 10 features

	Feature	OR	AUC
1	LAB_RANDOM_GLUCOSE_MEAN	1.67	65.53%
2	BMI	1.40	68.50%
3	BP_SYSTOLIC	1.14	71.17%
4	COMORB_HYPERTENSION	1.04	72.10%
5	COMORB_FAMILY_HIST	1.19	72.10%
6	LAB_HIGH_DENSITY_CHOL.	0.85	72.60%
7	AGE	1.19	72.87%
8	MED_BP	1.06	72.87%
9	MED_CHOL.	1.09	73.15%
10	LAB_CHOLESTEROL_HDL_RATIO	1.02	73.42%

Logistic Regression - LASSO



Available for 100% of patients

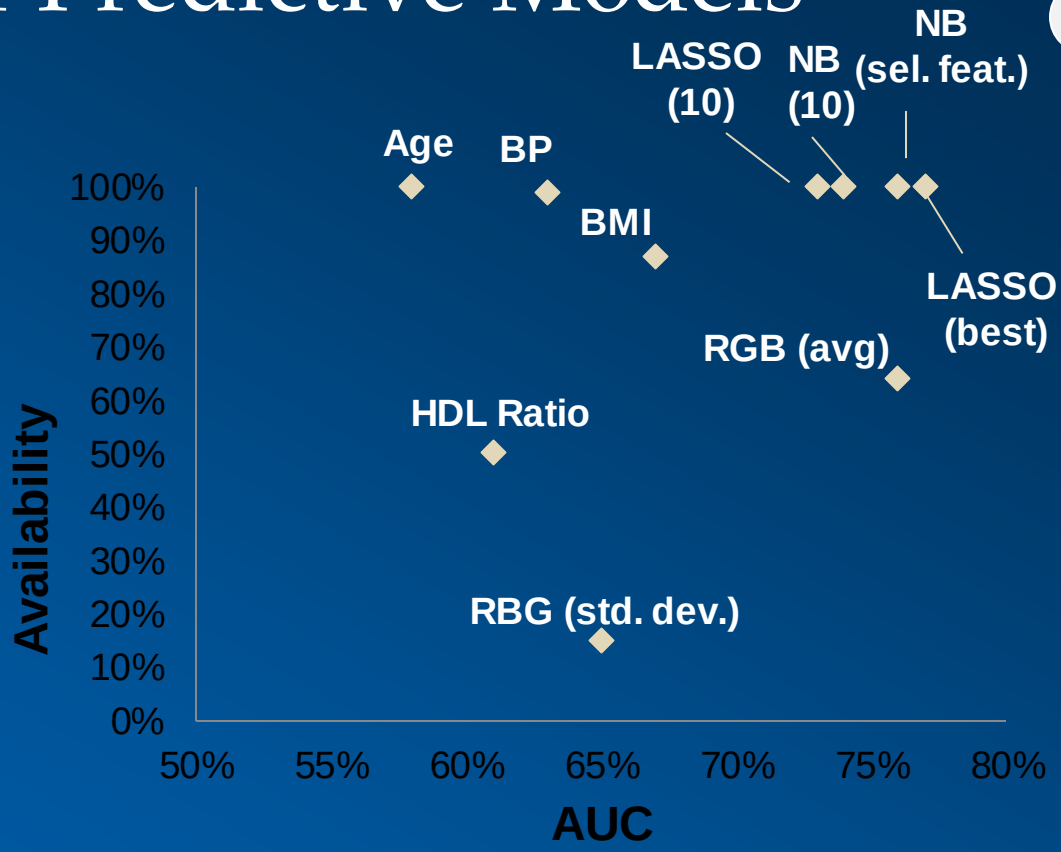
Cross Validated lambda selection chooses 41 features.

Missing data

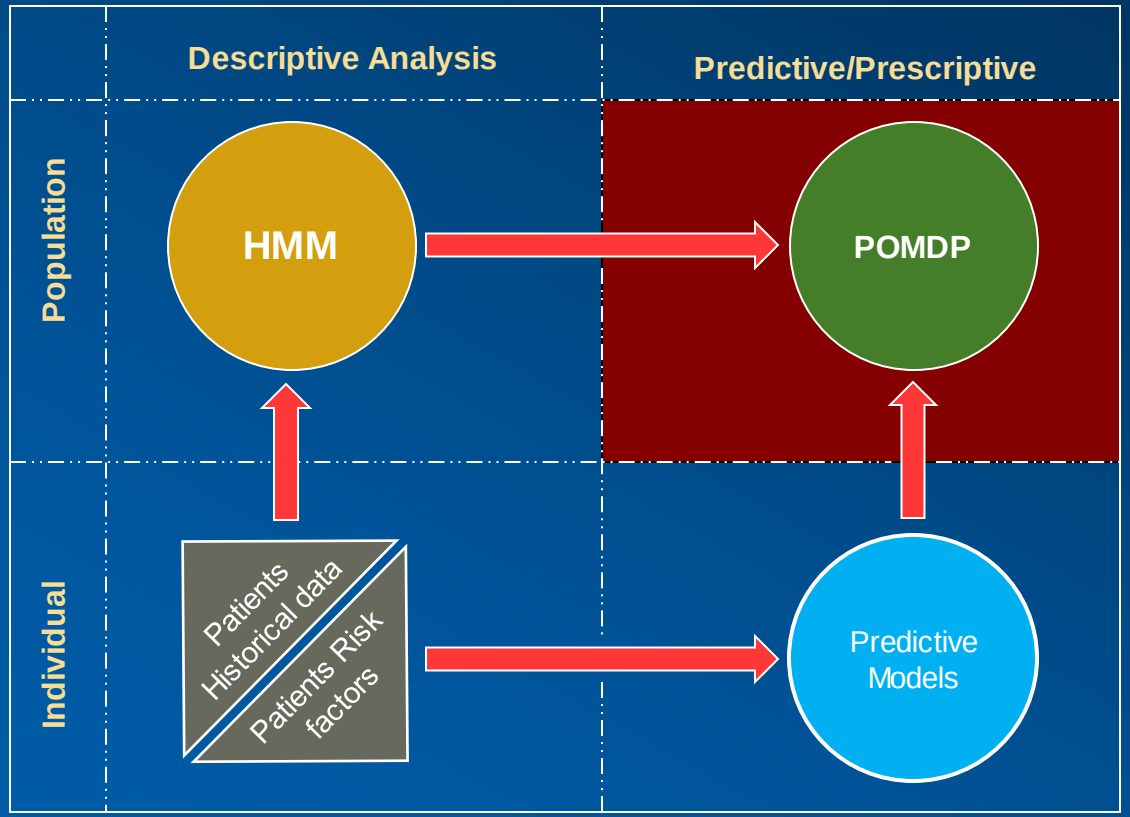
- Imputation is necessary
- Missing indicator improves the results from 0.765 to .772
- Important missing indicators have to do with missing lab values. E.g.,
  - missing platelet count
  - missing HDL values

Comparison of Predictive Models

	AUC	Availability
LASSO (best)	77%	100%
NB (select feat.)	76%	100%
NB (10)	74%	100%
LASSO (10)	73%	100%
RGB (avg)	76%	64%
BMI	67%	87%
RGB (std. dev.)	65%	15%
BP	63%	99%
HDL Ratio	61%	50%
Age	58%	100%

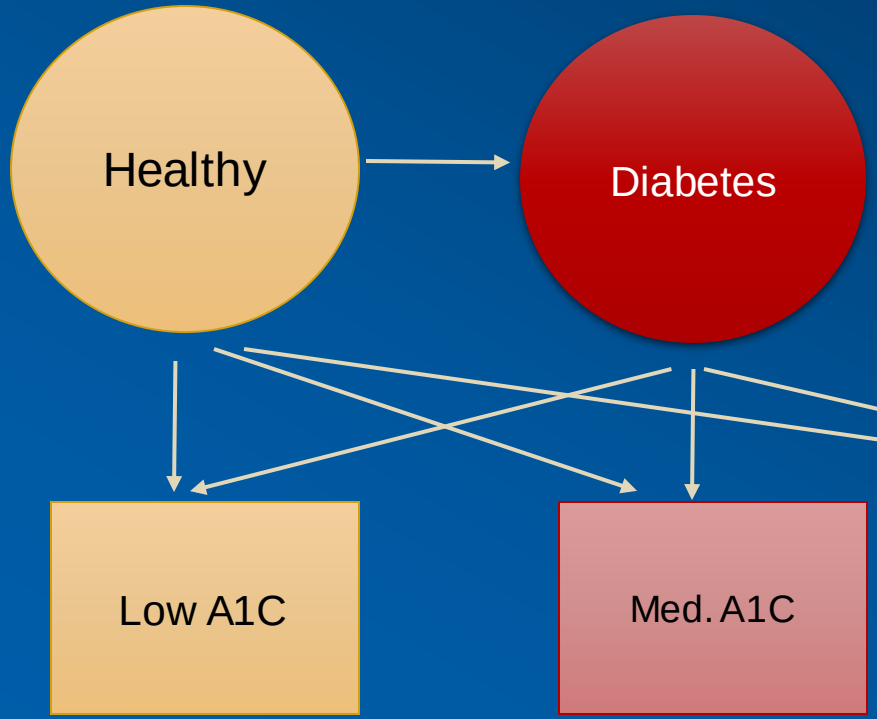


Framework


Descriptive Analysis
Predictive/Prescriptive
Population
Individual
HMM
POMDP
Patients Historical data
Patients Risk factors
Predictive Models

Simple Markov Model for Diabetes Progression

POMDP



Hidden States

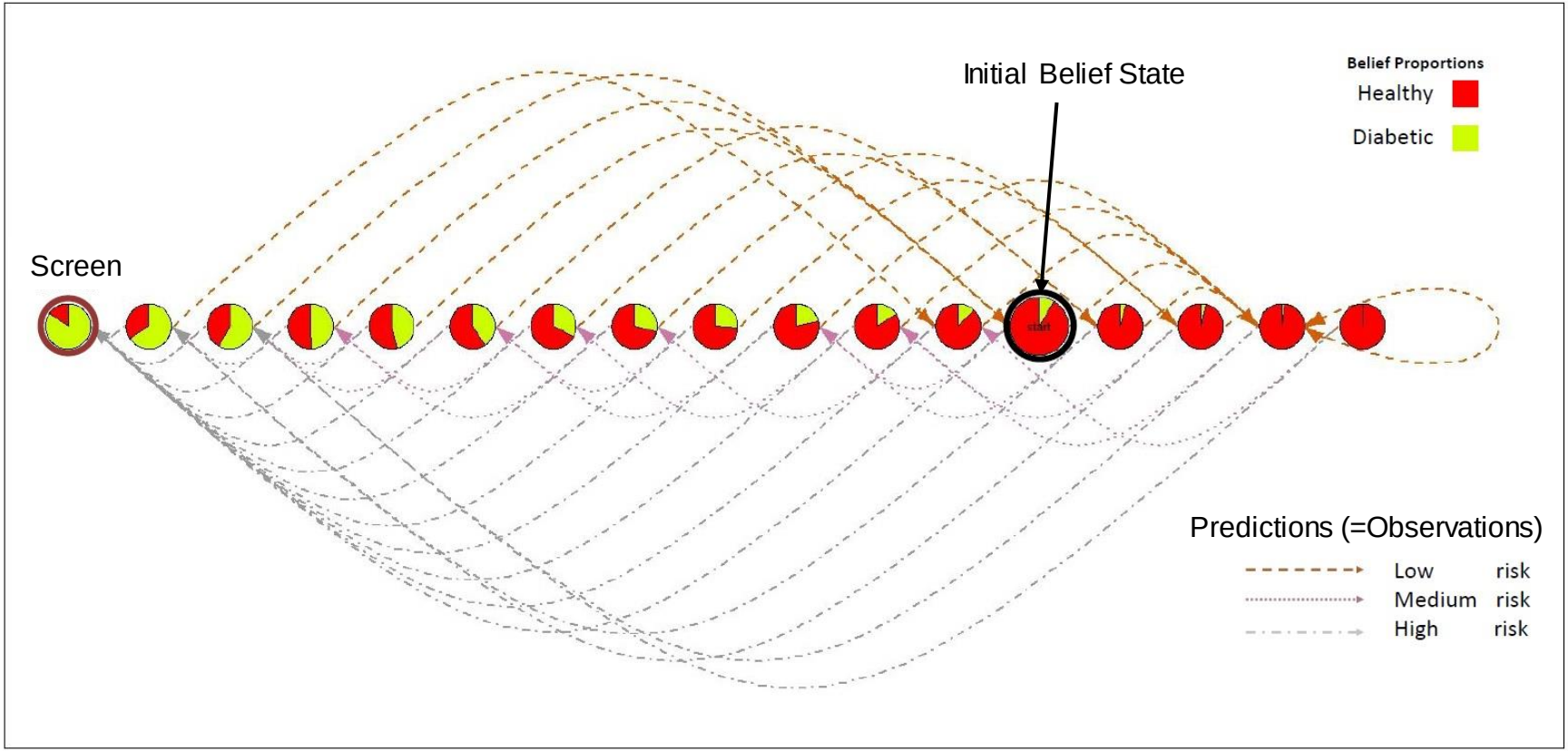
Patients likely to be in the "Diabetes" state should be screened.

Observations

obtained from predictive Model

Solution of the POMDP: Optimal Screening Strategy

POMDP



Limitations and Future Steps

- **HMM:** Estimation of transition probabilities may be biased because it is based on actually screened patients.
- **Predictive Model:** Missing data!
- **POMDP**
  - Cost/reward structure in POMDP (e.g., cost does not increase linearly)
  - Other dimensions for the state space? Makes the model harder to solve due to an explosion of belief states.
  - Set of possible/available actions (e.g., other interventions including diet and exercise changes).
  - Rescreening: Reset the belief state after negative screening.