Early Diabetic Risk Prediction with NHANES Data

Agenda

Background

Research Questions

Research Hypotheses

Methods
- Data Processing
- Targets/Independent Variables
- Feature Selection
- Model Exploration

Results

Conclusions/Limitations

Recommendations

Background: What is Diabetes?

- Blood glucose (sugar) levels higher than normal
- Not enough insulin to allow glucose to enter cells to be used for energy
- Diabetes types: Type 1, Type 2, Gestational diabetes
- Diagnosed by blood test
  - Fasting glucose >= 126 mg/dl
  - OGTT >= 200 mg/dl
  - HbA1c >= 6.4%

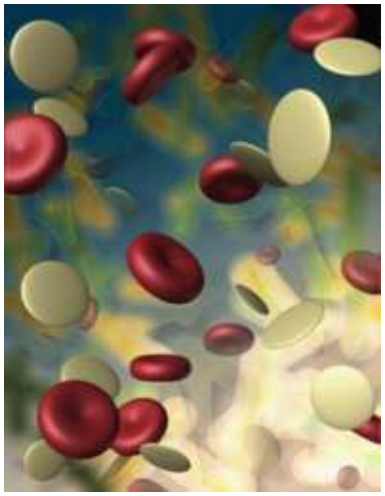
Background: Diabetes type 2 is a growing concern

- An estimated 7.0% of the American population have diabetes
- 1/3 of diabetics remain undiagnosed
- It is the 6th leading cause of death in America
- 54 million Americans have prediabetes
- 1 in 3 children born today will develop diabetes

The number of people with diabetes is expected to double by the year 2025!!!

Background: Prediabetes represents the gray area between normoglycemia and diabetes

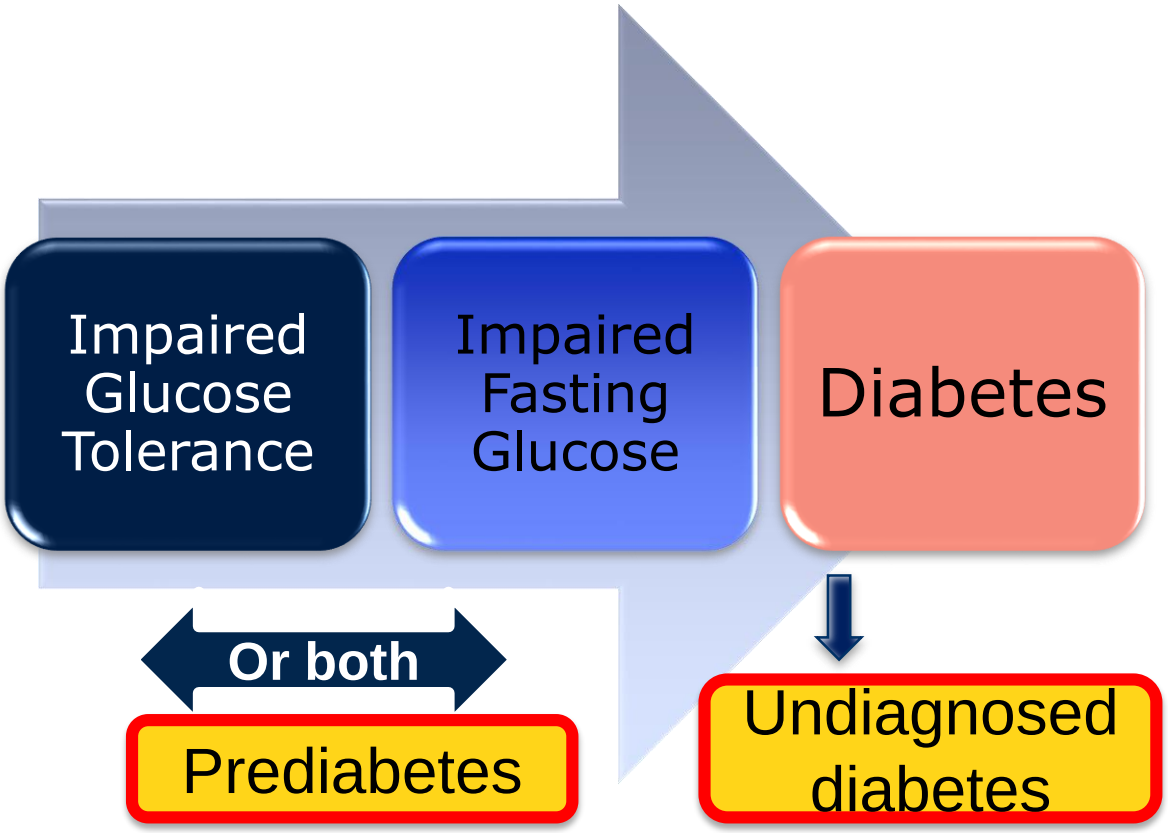
- Comes before type 2 diabetes
- Blood glucose higher than normal, but not yet diabetes
- You can have pre-diabetes and not know it
- Screening Test Results
  - OGTT 140-199 mg/dl
  - FPG 100-125 mg/dl
  - HbA1c 5.7-6.4%



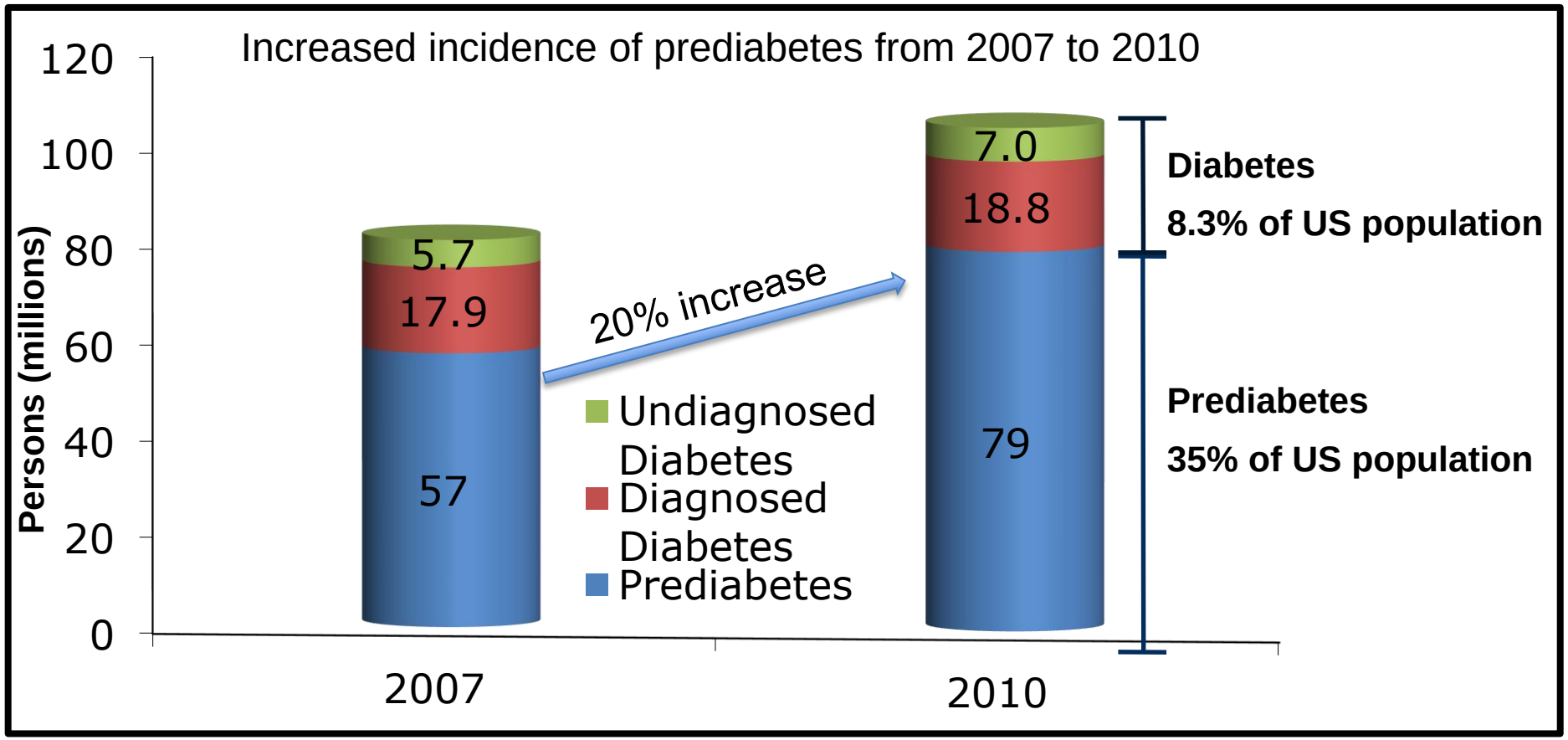
Why it is important to diagnose prediabetes???

- It is a major risk factor for development of diabetes.
- Prediabetes is associated with an increased risk of cardiovascular diseases.
- **It is treatable and curable !!!**

Background: **Progression to Type 2 Diabetes**



Prevalence of Diabetes and Prediabetes in the United States



CDC and Prevention. National diabetes fact sheet, 2007. http://www.**cdc**.gov/**Diabetes**/pubs/pdf/ndfs_2007.pdf.

CDC and Prevention. National diabetes fact sheet, 2011. http://www.cdc.gov/diabetes/pubs/pdf/ndfs_2011.pdf.

Methods: data source (NHANES)

National Health and Nutrition Examination Survey

- The survey examines a nationally representative sample of about 5,000 persons annually
- Unique: combination of interview and physical examination
- Interview: demographic, socioeconomic, dietary, and health-related questions
- Examination: medical, dental, and physiological measurement, Laboratory test
- Survey will be used to determine the prevalence and risk factors for major diseases
- Data will help develop sound public health policy, expand the health knowledge

Cycles 2005-2006 2007-2008 2009-2010 2011-2012 2013-2014

Component	Demographics	Examination	Laboratory	Questionnaire	Dietary
Component Data Files	Demographics (including survey design variables)	Audiometry Blood Pressure Body Measures Muscular Strength Oral health Vision Exam etc...	Urine Collection Hepatitis HIV Heavy metals Plasma Glucose Total Cholesterol Triglycerides etc...	Alcohol use Balance Blood Pressure Diabetes Drug Use Social Support Vision Weight History etc...	Dietary Interview Supplement Use etc...

Research questions/Hypotheses

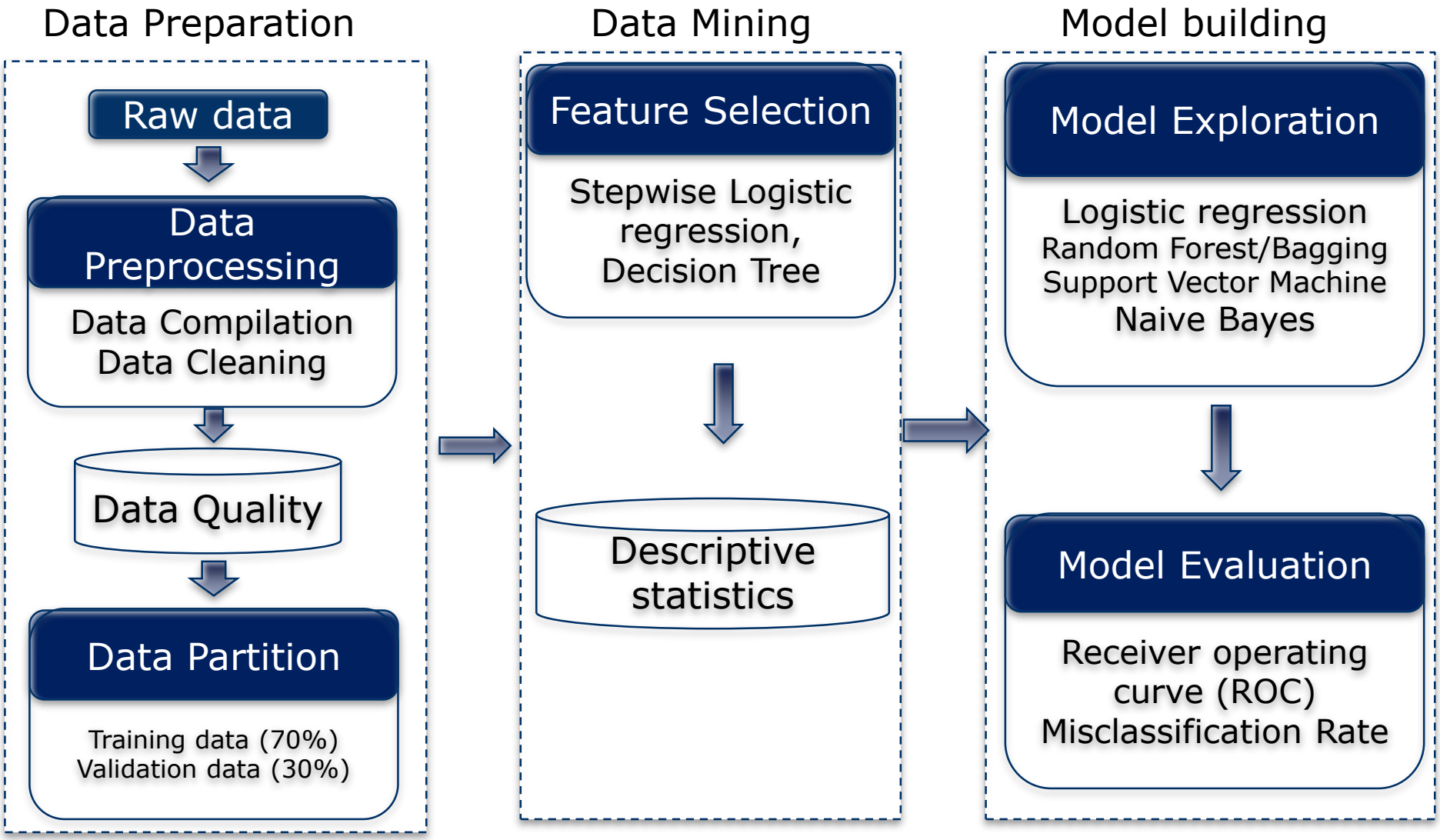
Research Questions:

- What's the prevalence of diabetes with continuous NHANES data compiled from 2005-2014?
- What's major risk factors that associated with prediabetes or undiagnosed diabetes with machine learning algorithm?

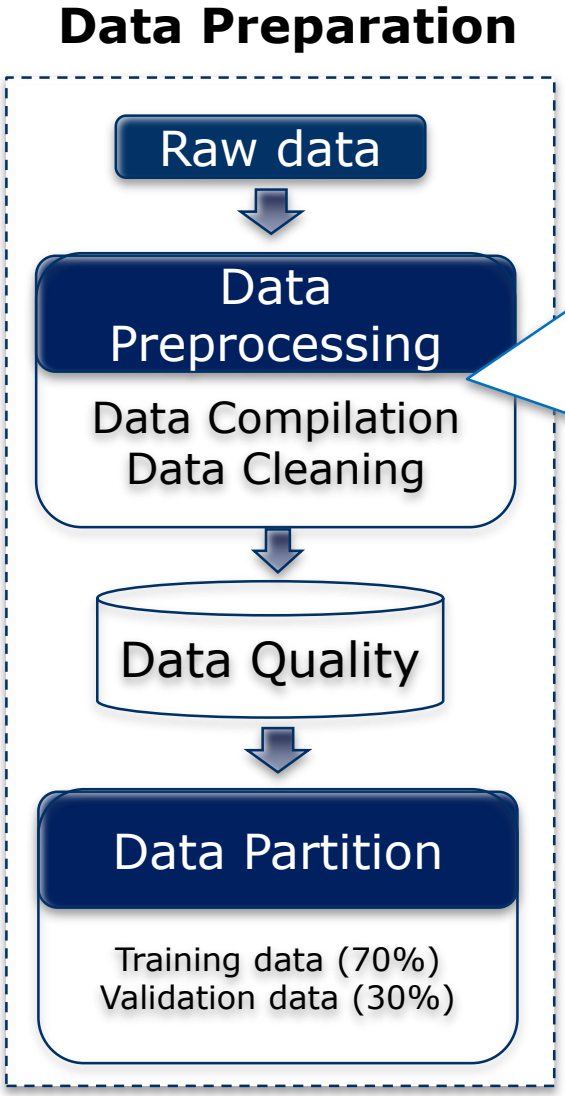
Research Hypotheses:

- The prevalence of diabetes is associated with demographic variables, such as age, gender, race/ethnicity, etc.
- Family history and other diabetes-related high risk diseases are key variables in predicting the early onset of diabetes

Methods: processing flowchart



Methods: processing flowchart



Selected Explanatory Variables

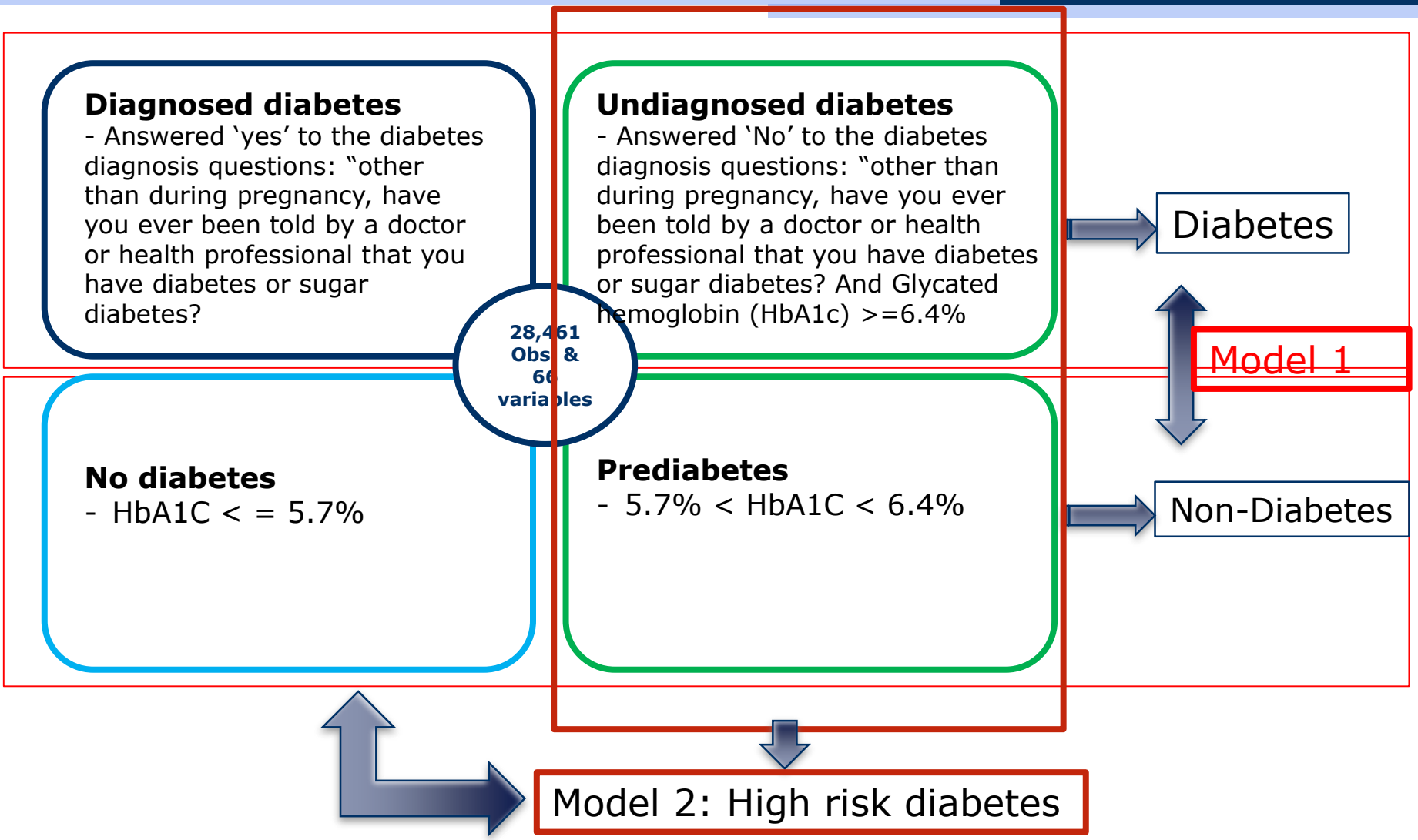
Modifiable risk factors:

- Overweight/obesity
- Physical activity
- Smoking status/history
- Alcohol status/history
- Diet status/history
- Laboratory tests

Non-modifiable risk factors:

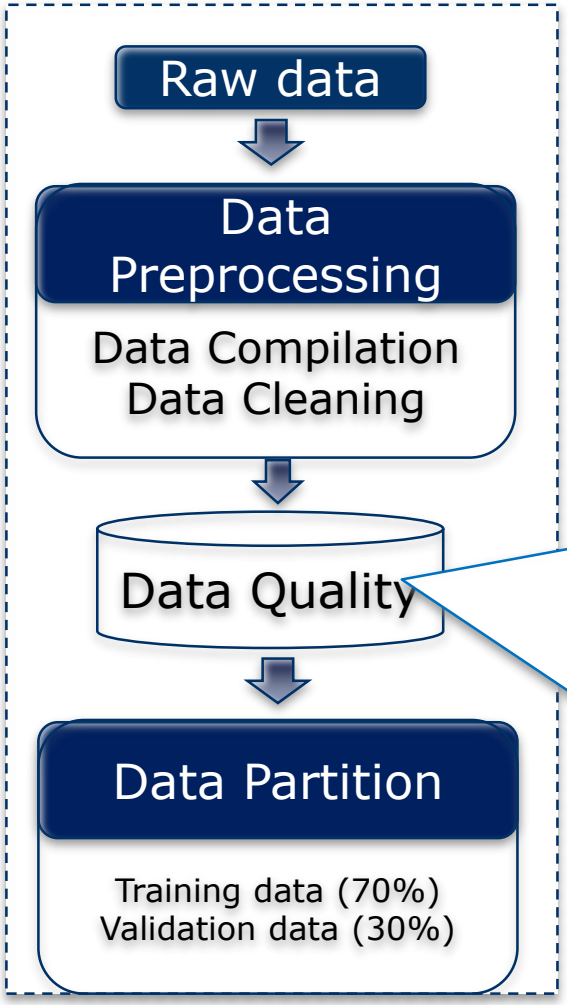
- Socioeconomic status
- Age
- Gender
- Race/ethnicity
- Education
- Family disease history
- Family income

Methods: Target Variables



Methods: processing flowchart

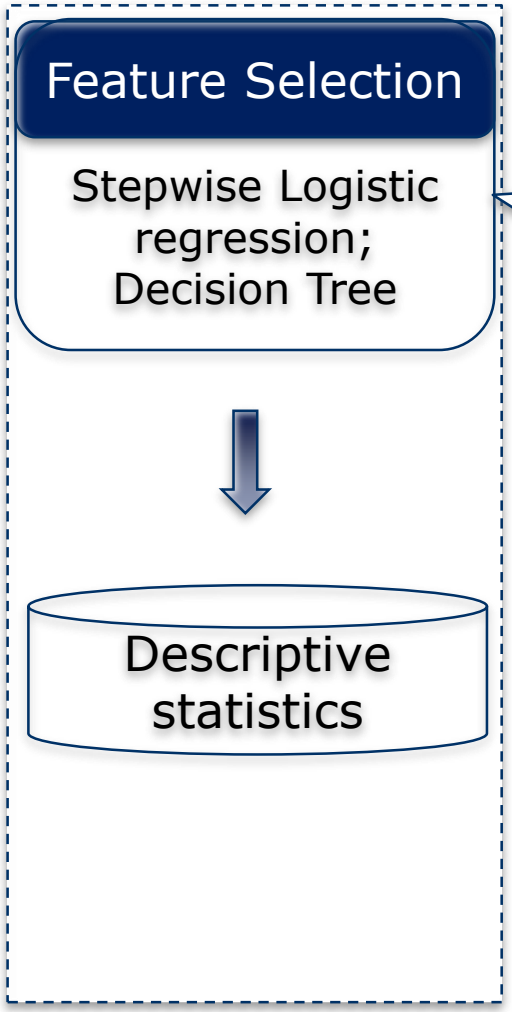
Data Preparation



Data Quality

- Remove the variables that with more than 50% missing values
- Check variables distribution and correlation
  - binning the demographic variables into categorical variables
  - Converting clinical test-related variables into categorical variables based on clinical criteria
- Handling missing values
  - adding dummy (1/0) variables to the analytical model to indicate whether the value for that variable is missing

Feature Selection



Personal/family history:
- Ever told have health risk for diabetes
- Ever told you have prediabetes
- Close relative had diabetes
- Doctor told you – high cholesterol level
- Close relative had asthma
- Ever told you had high blood pressure
- Doctor ever said you were overweight

Physical activity:
- Moderate recreational activities

Demographic factors:
- Age
- Education level
- Race/ethnicity

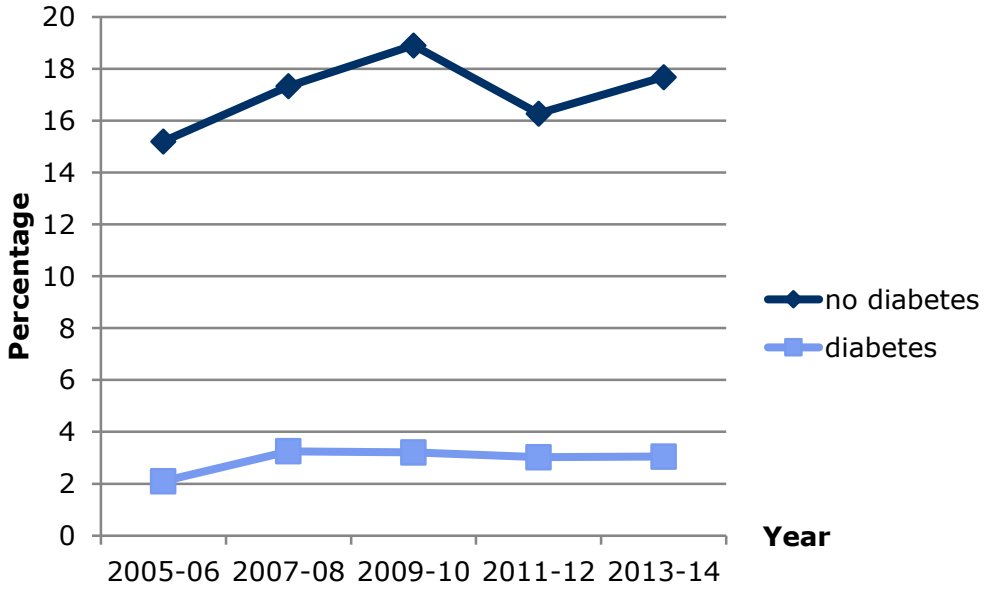
Body examination:
- Waist circumference
- Trouble seeing even with glass/contacts
- Had blood tested past three years

Smoke/alcohol status/history:
- Smoked at least 100 cigarettes in life
- How often drink alcohol over past 12 mos

Lab test:
- Blood cotinine level

Results: Prevalence of Diabetes

- Diabetes prevalence across the survey periods has increasing substantially over the past few years
- 3.05% in 2013-14 compared with 2.1% in 2005-06, which corresponds to approximately 10 million U.S adults with total confirmed diabetes



Results: Demographic variables are related with diabetes prevalence

Demographic Variables		no diabetes	diabetes	Percentage of diabetes (%)	p value
Age	20-39	8256	263	3.09	<.0001
	40-59	6985	1183	14.50	
	≥60	5969	2168	26.63	
Gender	male	10171	1829	15.25	0.005
	female	11039	1785	13.92	
Race/Ethnicity	non-hispanic white	10223	1333	11.54	<.0001
	non-hispanic black	4135	1003	19.52	
	Mexican American	3197	645	16.80	
	others	3655	633	14.77	
Education	below college	9903	2191	18.12	<.0001
	some college	6251	920	12.84	
	above college	5056	503	9.07	
Annual Family Income	<$25,000	6163	1428	18.80	<.0001
	$25,000-$75,000	10435	1714	14.10	
	>$75,000	4612	472	9.28	
Annual Household Income	<$25,000	6929	1504	17.84	<.0001
	$25,000-$75,000	10020	1656	14.18	
	>$75,000	4261	454	9.64	
family PIR	<=1	4330	845	16.32	<.0001
	>1	16189	2604	13.86	

Diabetes Prevalence was associated with

- Increasing age
- Males
- Population with low income, low education level
- Non-Hispanic Black

Chi-square test: test for independence to determine whether the demographic variables were related to diabetes prevalence

Null hypotheses: there was no association between these demographic variables and prevalence of diabetes

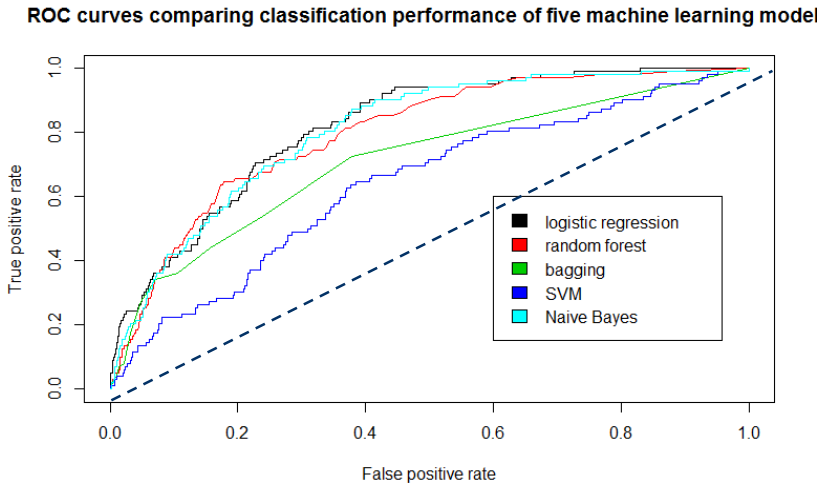
Results: Model Exploration

Model Exploration

Logistic regression
Random Forest/Bagging
Support Vector Machine
Naive Bayes

Model Evaluation

Receiver operating curve (ROC)
Misclassification Rate



	logistic regression	random forest	bagging	SVM	Naïve Bayes
AUC	0.816	0.798	0.711	0.639	0.806
misclassification rate	0.0452	0.0452	0.0474	0.0452	0.0661

Logistic regression performs exceptionally well with overall **lowest** misclassification rate and **highest** AUC (area under curve)

Results: Risk Factors for Diabetes (model 1)

Parameter		DF	Estimate
Intercept		1	10.6443
DIQ160	1	1	-1.4408
DIQ160	2	0	0
MCQ300B	1	1	-0.4257
MCQ300B	2	0	0
PAQ665	1	1	0.4718
PAQ665	2	0	0
ALQ120Q	1	1	-9.4561
ALQ120Q	2	1	-9.3465
ALQ120Q	3	1	-13.3278
ALQ120Q	4	0	0
MCQ300C	1	1	-0.5022
MCQ300C	2	0	0
RIDAGEYR	1	1	1.3617
RIDAGEYR	2	1	0.4271
RIDAGEYR	3	0	0
BMXWAIST	1	1	2.1830
BMXWAIST	2	1	1.4646
BMXWAIST	3	1	0.5130
BMXWAIST	4	0	0
LBXTC	1	1	0.5294
LBXTC	2	1	0.6184
LBXTC	3	0	0

Model 1:

- Binary outcome: diabetes vs. non-diabetes
- Eight variables were significant predictors of diabetes
  - high waist circumference
  - prediabetes history
  - older age
  - active alcohol status
  - close relative diabetes history
  - less physical activities
  - high total cholesterol level
  - close relative asthma history

Calculate corresponding odds ratio:

- Young aged and middle aged diabetes population were **3.903** and **1.533** times less in risk to develop diabetes than that of older aged population
- High/extreme high waist circumference had higher risk to develop diabetes

Results: Risk Factors for high-risk diabetes (model 2)

	Model 1	Model 2	
Target Variables	**Diabetes vs. non-diabetes**	**Prediabetes & undiagnosed diabetes vs. no-diabetes**	**Labels**
Important Variables	DIQ160	DIQ160	Ever told you have prediabetes
	MCQ300B		close relative had asthma
	PAQ665	PAQ665	moderate recreational activities
	ALQ120Q	ALQ120Q	how often drink alcohol over past 12 mos
	MCQ300C	MCQ300C	Close relative had diabetes
	RIDAGEYR	RIDAGEYR	age
	BMXWAIST	BMXWAIST	waist circumference
	LBXTC	LBXTC	total cholesterol
		LBXCOT	blood cotinine level
		BPQ080	doctor told you had high cholesterol level
		BPQ020	ever told you had high blood pressure
		DIQ180	had blood tested past three years

Model 2:

- Prediabetes & undiagnosed diabetes vs. no-diabetes
- **11 variables** were significant predictors of prediabetes & undiagnosed diabetes
- **4 variables** were the key factors only for model 2

Key factors for predication of high-risk diabetes

Conclusions

Diabetes prediction model: Logistic regression

- Prevalence prediction model (logistic regression) might be a promising diabetes surveillance instrument for the community-based population

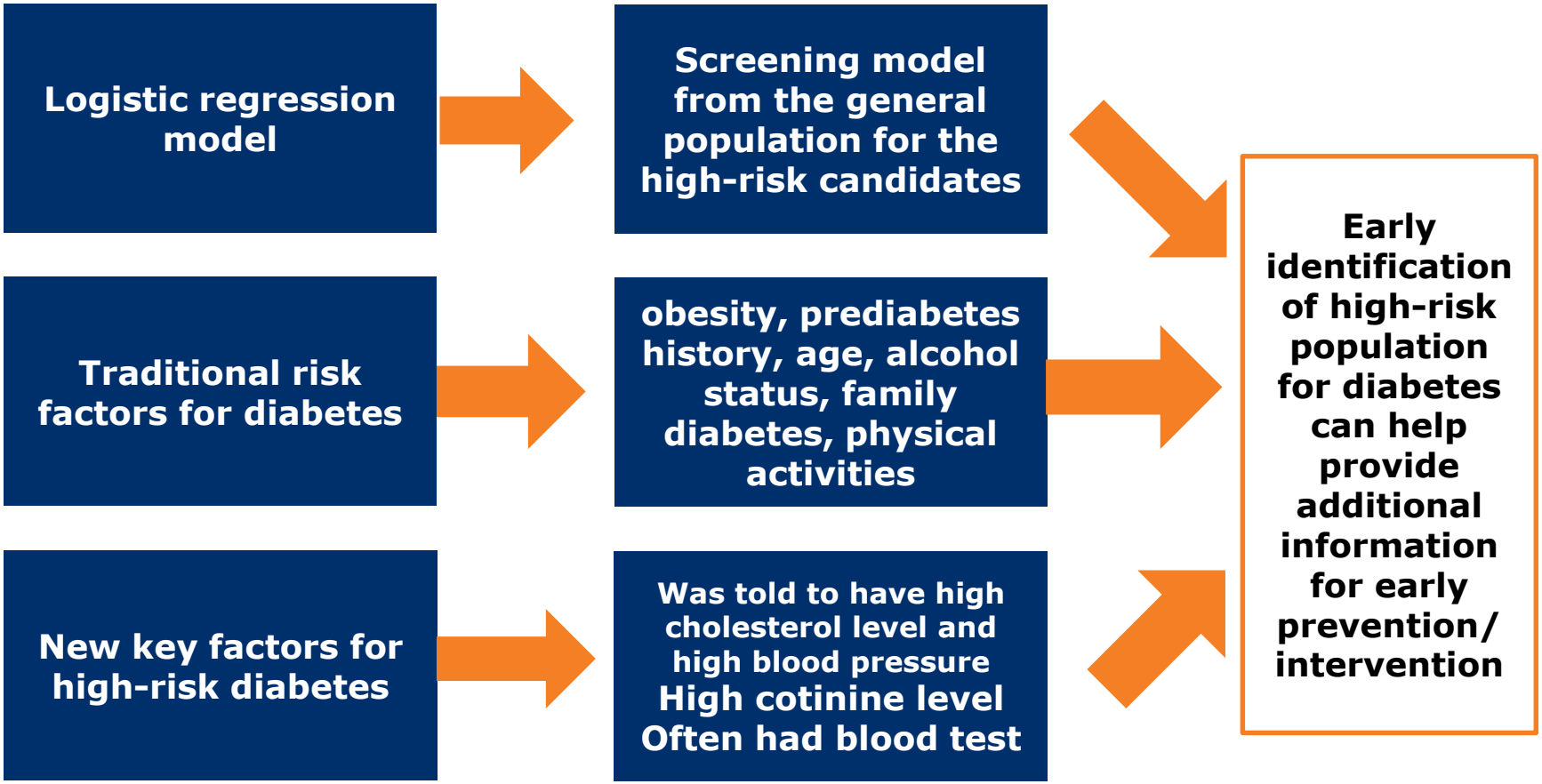
High risk factors for prediabetes/undiagnosed diabetes

- Blood Cotinine level, Doctor told high cholesterol level, Doctor told high blood pressure and has blood tested past three years might play critical roles in prediction of high-risk diabetes (pre-diabetes and undiagnosed diabetes)

Research Limitations

- No consider the sampling weights in the analysis, therefore the results cannot be extrapolated to generalize the US population
- Additional community-based survey data used for validation/test could provide robust evidence for the model performance
- Enable the complete records for survey data to be analyzed without regard for the missing data is challenging

Recommendations



Thank you!

Q & A