• Diabetes: high blood glucose levels and insulin problems.
• There are two main types: Type 1 and Type 2.
• In 2015, 30.3 million Americans (9.4% of the population) had diabetes, and diabetes was the 7th leading cause of death in the US.¹
• 1.5 million Americans are diagnosed with diabetes every year.¹
• Diabetes is a prevalent disease that affects millions not just in the US, but worldwide.
What do these people have in common?
**Fast Facts on Diabetes**

29.1 million people or 9.3% of the U.S. population have diabetes.

**Diagnosed**
21.0 million people

**Undiagnosed**
8.1 million people
(27.8% of people with diabetes are undiagnosed).

---

**Estimated Diabetes Costs in the United States, 2012**

**Total (Direct and Indirect)**
$245 billion

**Direct Medical Costs**
$176 billion
After adjusting for population age and sex differences, average medical expenditures among people with diagnosed diabetes were 2.3 times higher than people without diabetes.

**Indirect Costs**
$69 billion
(disability, work loss, premature death).
The Diabetes Epidemic: 2010–2030

World
2011 = 366 million
2030 = 552 million
Increase = 51%

37.7
51.2
36%

52.8
64.2
22%

32.6
59.7
83%

14.7
28.0
90%

71.4
120.9
69%

131.9
187.9
42%

25.1
39.9
59%
Type 1 Diabetes
Poor Numeracy: The Elephant in the Diabetes Technology Room

David Kerr, M.D.

“one third of US adults have basic or below basic health literacy and would have difficulty managing common health-related task”

“large numbers of the adult population of England are functionally illiterate and innumerate”
Human Factors - how much insulin with a meal?

• To calculate an insulin dose for a meal
  – Carbs/ICR* + BG-Target/Correction factor
  – Minus the “Insulin on Board”

<table>
<thead>
<tr>
<th>Errors</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimating carbohydrates</td>
<td>15-25%</td>
</tr>
<tr>
<td>Constants for ICR and CF</td>
<td>15-25%</td>
</tr>
<tr>
<td>Insulin absorption variability</td>
<td>20-30%</td>
</tr>
</tbody>
</table>
Technology to the Rescue
A Failed mHealth Program Offers Lessons Learned For Future Projects

An project to have FQHC patients use an mHealth app to manage their diabetes and hypertension at home collapsed after a few weeks. But researchers say they learned valuable lessons.

- 1/3 of target patients did not download app.
- 12% of patients were regular users.
- Most patients stopped using the app after a few weeks.
- Staff used office visits to train patients in use of smartphones (!).
Artificial Pancreas Design Challenges

- Single versus multiple devices
- Device specific
- Battery life
- Connectivity
- Variable insulin sensitivity
- Glucagon formulation
- Human physiology
- Human psychology
- Affordability

Artificial Pancreas

Fully Automated Closed-Loop System

Doyle, Diabetes Care 2014
Shape Matters: Relationship between diabetes and obesity
Motivation

• The relationship between diabetes and BMI is well-known.
• Our sponsor Vitality observed that the correlation appeared to end around 2005/2006.
• BMI is only one of many factors. We want to explore the relationship between BMI and diabetes, controlling for other important factors, such as age, waist circumference, exercise, and education, and cholesterol.

This study was performed by students in my graduate class in Actuarial Research at the University of California, Santa Barbara, Fall-Winter 2017-8. Sponsored by a CAE Grant from the SOA.
Objective: To model the relationship between diabetes, BMI, and other health factors.

- Data is provided by The Vitality Group.
  - Data spans 4 years, i.e. 2012-2015; used data only for single year participants.
  - Includes demographic information and health measurements of over 300,000 people.
- Health factors and measurements include:

  - Age
  - Gender
  - Height
  - Weight
  - Body Mass Index (BMI)
  - Alcohol intake
  - Sleep
  - Daily nutritional intakes
  - Hemoglobin A1c level (HbA1c)
  - Fasting Plasma Glucose (FPG)
  - Kessler Stress Score (KSS)
  - Blood pressure levels (SBP, DBP)
  - Triglycerides level
  - Lipids (HDL, LDL, TC)
  - Count of workouts and activities
<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>HbA1c (percent)</th>
<th>FPG (mmol/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Below 5.7</td>
<td>Below 5.6</td>
</tr>
<tr>
<td>Pre-diabetes</td>
<td>5.7 to 6.5</td>
<td>5.6 to 6.9</td>
</tr>
<tr>
<td>Diabetes</td>
<td>6.5 or above</td>
<td>6.9 or above</td>
</tr>
</tbody>
</table>

- We looked for people who were classified as non-diabetic but had HbA1c and/or FPG levels indicating diabetes and re-classified them as having diabetes.
Missing Values Plot (Whole Data)
• Missing data are a problem for studies: different methods can be used to impute missing values.

• One method that we use at the University that is not well-known (but deserves to be) is MICE.
  – Use MICE package in R to impute missing values
    (MICE = Multivariate Imputation by Chained Equations)
    https://cran.r-project.org/web/packages/mice/mice.pdf

How it Works:
MICE uses observed values to predict missing values on a variable by variable basis using logistic regression.

• Compare models with and without the imputed data to determine the value of imputation.

Interpretation from: Analytics Vidhya https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/
• Using MICE, Imputation models are built for each variable to predict the missing values.
• Variables we imputed were BMI, Hours of Sleep, Waist Circumference, Total Cholesterol,… etc.
• After MICE:

One Year
217,254 members

Two or more years
144,083 members
<table>
<thead>
<tr>
<th>Gender</th>
<th>Count</th>
<th>Mean Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>98,399</td>
<td>43.4</td>
</tr>
<tr>
<td>Female</td>
<td>118,855</td>
<td>43.4</td>
</tr>
</tbody>
</table>
One-Year File

- 217,254 members
  - Diabetes Group: 17,554 members
  - Non-Diabetes Group: 199,700 members
BMI Median Values

<table>
<thead>
<tr>
<th>Diabetes Status</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetics</td>
<td>30.9</td>
<td>32.5</td>
</tr>
<tr>
<td>No Diabetics</td>
<td>27.9</td>
<td>26.7</td>
</tr>
</tbody>
</table>
Modeling

- **Step 1**
  - Logistic regression model (all variables, based on whole data set)
  - Random forests model (all variables, based on whole data set)

- **Step 2 Sub-model**
  - Logistic regression model (important variables, based on whole data set)

- **Step 3 Balanced Data**
  - Logistic regression model (important variables, based on balanced data set)
  - Random forests model (all variables, based on balanced data set)
Random forest

Logistic regression model

<table>
<thead>
<tr>
<th>Important variable</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>Waist_circumference</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>Count of other program activities</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>Education 11</td>
<td>0.00012</td>
</tr>
<tr>
<td>Education 6</td>
<td>8.83E-13</td>
</tr>
<tr>
<td>Average alcohol drinks per week</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>BMI</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>Kessler stress score</td>
<td>3.19E-15</td>
</tr>
<tr>
<td>SBP</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>HDL</td>
<td>8.18E-14</td>
</tr>
<tr>
<td>DBP</td>
<td>1.23E-15</td>
</tr>
<tr>
<td>LDL</td>
<td>2.34E-07</td>
</tr>
<tr>
<td>TC</td>
<td>4.85E-05</td>
</tr>
<tr>
<td>Heart disease</td>
<td>4.03E-16</td>
</tr>
<tr>
<td>daily sedentary hours</td>
<td>4.35E-14</td>
</tr>
<tr>
<td>chronic lung disease</td>
<td>6.85E-06</td>
</tr>
</tbody>
</table>
### Validation of Predicted Values

1. Derivations from confusion matrix
2. AUC

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset Used</th>
<th>Predictors</th>
<th>Threshold</th>
<th>AUC</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>CV Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>Whole dataset</td>
<td>All variables</td>
<td>0.08</td>
<td>0.7903</td>
<td>0.7127</td>
<td><strong>0.7380</strong></td>
<td>0.7359</td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>Balanced dataset</td>
<td>All variables</td>
<td>0.50</td>
<td><strong>0.8000</strong></td>
<td><strong>0.8486</strong></td>
<td>0.5753</td>
<td><strong>0.7465</strong></td>
<td>0.7279</td>
</tr>
<tr>
<td>Random Forests</td>
<td>Whole dataset</td>
<td>All variables</td>
<td>0.08</td>
<td>0.7982</td>
<td>0.7514</td>
<td>0.7041</td>
<td>0.7078</td>
<td></td>
</tr>
<tr>
<td>Random Forests</td>
<td>Balanced dataset</td>
<td>All variables</td>
<td>0.50</td>
<td>0.7973</td>
<td>0.7678</td>
<td>0.6908</td>
<td>0.7445</td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>Whole dataset</td>
<td>Important variables</td>
<td>0.08</td>
<td>0.7897</td>
<td>0.7094</td>
<td><strong>0.7382</strong></td>
<td>0.7358</td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>Balanced dataset</td>
<td>Important variables</td>
<td>0.50</td>
<td><strong>0.7999</strong></td>
<td><strong>0.8486</strong></td>
<td>0.5757</td>
<td><strong>0.7466</strong></td>
<td>0.7262</td>
</tr>
</tbody>
</table>

Sensitivity: True positive rate
Specificity: True negative rate
<table>
<thead>
<tr>
<th>Predictors</th>
<th>Variable Description</th>
<th>Odds Ratio (95% Conf. Int.)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>In years; range is 18 - 80</td>
<td>1.0502 (1.0475, 1.0529)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Education (11)</td>
<td>College degree</td>
<td>0.9402 (0.8806, 1.0038)</td>
<td>0.0648</td>
</tr>
<tr>
<td>Education (6)</td>
<td>Postgraduate degrees</td>
<td>0.8214 (0.7515, 0.8978)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Daily Sedentary Hours</td>
<td>In hours</td>
<td>1.0137 (1.0065, 1.0210)</td>
<td>0.0002</td>
</tr>
<tr>
<td>Kessler Stress Score</td>
<td>Range is 1 - 50</td>
<td>1.0140 (1.0073, 1.0207)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Average Alcohol Drinks per Week</td>
<td>Average drinks</td>
<td>0.9661 (0.9586, 0.9735)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Chronic Lung Disease (2)</td>
<td>Yes versus No; does not have disease</td>
<td>0.5944 (0.3592, 0.9549)</td>
<td>0.0365</td>
</tr>
<tr>
<td>Heart Disease (2)</td>
<td>Yes versus No; does not have disease</td>
<td>0.6696 (0.5467, 0.8161)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Body Mass Index</td>
<td>In kilograms per meter squared</td>
<td>1.0343 (1.0268, 1.0417)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Waist Circumference</td>
<td>In centimeters</td>
<td>1.0217 (1.0186, 1.0249)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Systolic Blood Pressure</td>
<td>In mmHg</td>
<td>1.0169 (1.0142, 1.0195)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Diastolic Blood Pressure</td>
<td>In mmHg</td>
<td>0.9908 (0.9870, 0.9946)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Total Cholesterol</td>
<td>In mmol/L</td>
<td>0.7860 (0.6946, 0.8890)</td>
<td>0.0001</td>
</tr>
<tr>
<td>High Density Lipoprotein</td>
<td>In mmol/L</td>
<td>0.8379 (0.7286, 0.9640)</td>
<td>0.0132</td>
</tr>
<tr>
<td>Low Density Lipoprotein</td>
<td>In mmol/L</td>
<td>0.9243 (0.8164, 1.0468)</td>
<td>0.214</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>In mmol/L</td>
<td>1.5019 (1.4119, 1.5988)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Count of Other Program Activities</td>
<td>Count per week, self-reported</td>
<td>0.9307 (0.9207, 0.9408)</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
• Want to investigate the data to see any interesting subpopulations.
• Clustering data of mixed type:
  – Gower distance
  – Partition around medoids (PAM)
  – Silhouette width
• 10,683 observations used (5% of data)
• t-Distributed Stochastic Neighbor Embedding (t-SNE)
Cluster based on gender

- Cluster 1: Male
- Cluster 2: Female

The 2-cluster model is not too interesting.
Clusters

1. Male with post graduate degree
2. Male college graduate; AA alcohol consumption
3. Female college dropout; high stress levels
4. Female college graduate; smoking history
5. Mostly female high school graduate; BA alcohol consumption
6. Male college dropout; AA BMI, AA waist circumference
7. Female college graduate; no smoking;
8. Female with post graduate degree; BA BMI, BA waist circumference
9. Female college dropout; high depression rate with a smoking history
10. Male college graduate; no smoking, AA waist circumference

* AA – above average; BA- below average
• Logistic Regression and Random Forest Models
• Other important factors other than BMI such as:
  • Age
  • Waist circumference
  • Cholesterol levels
  • Education
• Future work: two or more years longitudinal data
Thank you