Predictive Models for Making Patient Screening Decisions

MICHAEL HAHSLER¹, VISHAL AHUJA¹, MICHAEL BOWEN², AND FARZAD KAMALZADEH¹

¹ Southern Methodist University,
² UT Southwestern Medical Center and Parkland Health and Hospital System
The staggering costs of diabetes in America

Nearly 30 million Americans have diabetes.

$1 in $3 Medicare dollars is spent caring for people with diabetes.

Diabetes and prediabetes cost America $322 billion per year.

86 million Americans have prediabetes.

$1 in $5 health care dollars is spent caring for people with diabetes.

Today, 3,835 Americans will be diagnosed with diabetes. Today, diabetes will cause 200 Americans to undergo an amputation, 136 to enter end-stage kidney disease treatment and 1,795 to develop severe retinopathy that can lead to vision loss and blindness.

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)

8.1 Million Undiagnosed
77 Million with Undiagnosed Pre-diabetes

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)
Predictive Problem: Initial Screening Decision

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

Follow-up period >12 month

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
- Female: 69%
- Male: 31%

Race
- Hispanic: 45%
- Black: 36%
- White: 14%
- Asian: 5%
- Other: 0%

Payment
- Charity: 40%
- Medicaid-Medicare: 26%
- Self-pay: 22%
- Commercial: 12%
- Other: 0%

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

AUC = 0.67

Available for 100% of patients
Single-Factor Threshold Models
Usual risk factors: Age and BMI

AUC = 0.58

AUC = 0.67

1- False Alarm Rate

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

AUC = 0.67

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

AUC = 0.58
Single-Factor Threshold Models
Usual risk factors: Age and BMI

AUC = 0.67

Sensitivity : 0.817   Specificity : 0.377

= USPSTF 2015 (Age>40, BMI>25)
Single-Factor Threshold Models

Uncommon risk factor: Random Blood Glucose

Available for 64% of patients

Available for 15% of patients

AUC = 0.73

AUC = 0.65

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., overweighted 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
  - Just ignore the missing value (pairwise deletion) – needs to be supported by the method
  - Imputation

Not practical with many missing values
And introduce bias for all but MCAR

Naïve Bayes Classifier

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

\[
p(C_k | x) = \frac{p(C_k) \prod_{i=1}^{n} p(x_i | C_k)}{p(x)}
\]

• \(C_k\) is the class, \(x\) is a feature vector. We use a threshold on \(p(C_{diabetes} | x)\) 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.
Multi Factor Model
Naive Bayes Model (NB) - All 105 Predictors

Available for 100% of patients

AUC = 0.73

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

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

AUC = 0.78

Available for 100% of patients

Backward Feature Selection

5 of 10 predictors are not in current guidelines
Generalized Linear Model with LASSO

- GLM with binomial response and L1 regularization.

\[
\min_{\beta} \left\{ \frac{1}{N} \left\| y - \frac{1}{1 - \exp(-X\beta)} \right\|^2 \right\} \quad \text{s.t. } \|\beta\|_1 \leq t
\]

- **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

### Regularization Path

<table>
<thead>
<tr>
<th>Feature</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>LABRANDOM GLUCOSE_MEAN</td>
<td>1.49</td>
</tr>
<tr>
<td>BMI</td>
<td>1.26</td>
</tr>
<tr>
<td>BP_SYSTOLIC</td>
<td>1.10</td>
</tr>
<tr>
<td>COMORB HYPERTENSION</td>
<td>1.03</td>
</tr>
</tbody>
</table>
GLM – LASSO
Cross Validated selection of lambda (43 features)

AUC = 0.77

Available for 100% of patients

Feature                  | Odds Ratio
LAB_NON_HDL_CHOLESTEROL_NA | 1.686548773596
LAB_RANDOM_GLUCOSE_MEAN   | 1.6687952164283
BMI                       | 1.3945863570232
LAB_PLATELET_NA           | 1.2186934500033
COMORB_FAMILY_HIST        | 1.18536725238753
AGE                       | 1.17605474716379
BP_SYSTOLIC               | 1.14286625129544
LAB_CHOLESTEROL_NA        | 1.1131974786007
LAB_RED_BLOOD_COUNT       | 1.10245828551814
MED_DM_biguanide          | 1.08600204482025
GLM – LASSO
22 Features

Available for 100% of patients
### Comparison of Predictive Models

<table>
<thead>
<tr>
<th>Feature</th>
<th>AUC</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB (select feat.)</td>
<td>78%</td>
<td>100%</td>
</tr>
<tr>
<td>LASSO (best)</td>
<td>77%</td>
<td>100%</td>
</tr>
<tr>
<td>NB (all features)</td>
<td>76%</td>
<td>100%</td>
</tr>
<tr>
<td>LASSO (25)</td>
<td>76%</td>
<td>100%</td>
</tr>
<tr>
<td>RGB (avg)</td>
<td>76%</td>
<td>64%</td>
</tr>
<tr>
<td>BMI</td>
<td>67%</td>
<td>87%</td>
</tr>
<tr>
<td>RGB (std. dev.)</td>
<td>65%</td>
<td>15%</td>
</tr>
<tr>
<td>BP</td>
<td>63%</td>
<td>99%</td>
</tr>
<tr>
<td>HDL Ratio</td>
<td>61%</td>
<td>50%</td>
</tr>
<tr>
<td>Age</td>
<td>58%</td>
<td>100%</td>
</tr>
</tbody>
</table>
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.
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.