Predicting Unplanned Medical Visits among Patients with Diabetes Using Machine Learning

Arielle Selya, PhD, Sanford Research
Eric L. Johnson, MD, University of North Dakota
Diabetes

• High blood glucose
  – Type 1: no insulin production
  – Type 2: insulin resistance

• 9% of U.S. population

• $245 billion annual costs [1]
  – $176 billion in direct medical costs [1]

Effects of Diabetes

- Hyper/hypoglycemia [2]
- Vascular complications [2]
  - Cardiovascular disease (CVD)
  - Nerve damage
  - Kidney damage
  - Eye damage
- Infections [3]
  - Soft tissue
  - Respiratory tract
  - Urinary tract
- Patients have many unplanned medical visits [4]
Diabetes and Smoking

• Smoking exacerbates the complications of diabetes:
  – Decreases glycemic control [1]
  – Increases risk of infection [5]
  – Amplifies CVD risk [6]
Can we predict unplanned visits?

- LACE index for 30-day readmissions
Risk Prediction Challenges

Hypothetical: Perfect prediction

Hypothetical: Random prediction

Unplanned Visits

LDL level (mg/dL)

Participant Number

Unplanned Visits

1+ Unplanned Visits

LDL level (mg/dL)

Participant Number
Risk Prediction Challenges

Actual Data

- Small but statistically significant difference (88.3 vs. 87.5, \( p = .005 \))
- Statistical differences do not necessarily indicate predictive ability!
Machine Learning for Risk Prediction

• Classification task (any vs. no unplanned visits)
  – Linear and quadratic discriminant analysis
  – Support vector machines (SVM)
  – Artificial neural nets (NN)

• Relative to status quo
  – Logistic regression analysis
  – LACE index
Sanford Data Collaborative 2017

- EMR data 2014-16
- N=63,245 patients:
  - Age 18 or over
  - Diabetes diagnosis
  - Zip codes in MN, ND, SD
- Unplanned visits
  - 4 separate types
  - 54.7% had ≥1 unplanned visit
- Predictors:
  - Age
  - Blood pressure
  - Number on “problem list”
  - Number of prescriptions
  - Body mass index (BMI)
  - Cholesterol (HDL, LDL)
  - A1C
  - Ranked smoking status
Unplanned Visits by Smoking Status

<table>
<thead>
<tr>
<th></th>
<th>Smokers</th>
<th>Nonsmokers</th>
<th>$p$</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minnesota</td>
<td>58.6%</td>
<td>50.4%</td>
<td>&lt;.0001</td>
<td>+ 8.2%</td>
</tr>
<tr>
<td>North Dakota</td>
<td>59.6%</td>
<td>57.5%</td>
<td>.0400</td>
<td>+ 2.1%</td>
</tr>
<tr>
<td>South Dakota</td>
<td>59.5%</td>
<td>55.5%</td>
<td>.0003</td>
<td>+ 4.0%</td>
</tr>
</tbody>
</table>

- Patients with diabetes who smoke are more likely to have at least one unplanned visit
# Most Common Diagnoses in Unplanned Visits

## Smokers with diabetes

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>R10.xx: Abdominal and pelvic pain</td>
<td>6.6% (N=4109)</td>
</tr>
<tr>
<td>M54.xx: Dorsalgia</td>
<td>6.1% (N=3750)</td>
</tr>
<tr>
<td>R07.xx: Pain in throat and chest</td>
<td>3.6% (N=2252)</td>
</tr>
<tr>
<td>M25.xx: Other joint disorder, not elsewhere classified</td>
<td>3.4% (N=2114)</td>
</tr>
<tr>
<td>M79.xx: Other and unspecified tissue disorders</td>
<td>3.4% (N=2085)</td>
</tr>
<tr>
<td>L03.xx: Cellulitis and acute lymphangitis</td>
<td>2.3% (N=1453)</td>
</tr>
<tr>
<td>E11.xx: Type 2 Diabetes Mellitus</td>
<td>2.1% (N=1314)</td>
</tr>
<tr>
<td>R05.xx: Cough</td>
<td>2.1% (N=1298)</td>
</tr>
<tr>
<td>J40.xx: Bronchitis, not specified as acute or chronic</td>
<td>2.0% (N=1265)</td>
</tr>
<tr>
<td>G43.xx: Migraine</td>
<td>1.9% (N=1197)</td>
</tr>
</tbody>
</table>

## Nonsmokers with diabetes

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>R10.xx: Abdominal and pelvic pain</td>
<td>5.4% (N=8856)</td>
</tr>
<tr>
<td>M54.xx: Dorsalgia</td>
<td>4.3% (N=7099)</td>
</tr>
<tr>
<td>R07.xx: Pain in throat and chest</td>
<td>4.3% (N=6971)</td>
</tr>
<tr>
<td>M25.xx: Other joint disorder, not elsewhere classified</td>
<td>3.2% (N=5175)</td>
</tr>
<tr>
<td>M79.xx: Other and unspecified tissue disorders</td>
<td>3.1% (N=5092)</td>
</tr>
<tr>
<td>R05.xx: Cough</td>
<td>2.8% (N=4547)</td>
</tr>
<tr>
<td>L03.xx: Cellulitis and acute lymphangitis</td>
<td>2.2% (N=3526)</td>
</tr>
<tr>
<td>J40.xx: Bronchitis, not specified as acute or chronic</td>
<td>2.1% (N=3422)</td>
</tr>
<tr>
<td>J02.xx: Acute pharyngitis</td>
<td>2.0% (N=3222)</td>
</tr>
<tr>
<td>R51.xx: Headache</td>
<td>1.8% (N=3.14)</td>
</tr>
</tbody>
</table>
Evaluating Classifiers

• Cross-validation testing
  – How well did the classifier learn patterns that are truly diagnostic of a category/outcome?

• Confusion matrices

<table>
<thead>
<tr>
<th>Actual class: 0</th>
<th>Predicted class: 0</th>
<th>Correct rejection</th>
<th>False alarm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual class: 1</td>
<td>Miss</td>
<td>Hit</td>
<td></td>
</tr>
</tbody>
</table>

• Average prediction accuracy: mean of correct rejection and hit rates
### Risk Prediction Results

#### Logistic regression

<table>
<thead>
<tr>
<th>Predicted: No visits</th>
<th>Predicted: 1+ visit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual: No visits</td>
<td>60.5%</td>
</tr>
<tr>
<td>Actual: 1+ visit</td>
<td>29.8%</td>
</tr>
</tbody>
</table>

**Average:** 65.4%

#### Best-case classifier (radial-basis support vector machine)

<table>
<thead>
<tr>
<th>Predicted: No visits</th>
<th>Predicted: 1+ visit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual: No visits</td>
<td>67.8%</td>
</tr>
<tr>
<td>Actual: 1+ visit</td>
<td>34.1%</td>
</tr>
</tbody>
</table>

**Average:** 66.9% (+1.5% points)

LACE index for 30-day readmissions:
66.3% hit rate; 53.3% false rejection rate = average 59.8%
Impact of More Accurate Prediction

• For the broader population (not restricted to patients with diabetes):
  – N=379,870 people with 1+ unplanned visit
  – Using SVM over regression correctly identifies N≈3039 people at risk (≈10,000 visits)

• Analyses of cost were not feasible
Clinical Implications

• Can’t conclude causality (from classifiers or regressions)
• Separate treatment from prediction?
• How to extract clinical implications?
  – E.g. what predictor variables, if modified, would lower unplanned visits?
  – Remove patients with certain ranges on modifiable variables, and re-run models
Variables’ Impact on Prediction Accuracy

<table>
<thead>
<tr>
<th>Restricted variable range</th>
<th>New prediction accuracy</th>
<th>Change in accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI &lt; 30 (N=15,885)</td>
<td>65.9%</td>
<td>-1.0%</td>
</tr>
<tr>
<td>BP &lt; 120/80 (N=11,996)</td>
<td>65.0%</td>
<td>-1.9%</td>
</tr>
<tr>
<td>No current smoking (N=38,370)</td>
<td>65.8%</td>
<td>-1.1%</td>
</tr>
<tr>
<td>LDL &lt; 130 (N=39,384)</td>
<td>65.8%</td>
<td>-1.1%</td>
</tr>
<tr>
<td>HDL &gt; 50 (N=30,058)</td>
<td>65.1%</td>
<td>-1.8%</td>
</tr>
<tr>
<td>A1C &lt; 6.5 (N=13,857)</td>
<td>66.2%</td>
<td>-0.7%</td>
</tr>
</tbody>
</table>

High levels of BP and HDL were most informative for predicting unplanned visits
Next Steps

• Data Analysis of Existing Data
  – Adding more variables and refining the model
  – Validate the model with forthcoming data
  – Generate recommendations for clinical targets

• Clinical Research at Sanford
  – Target strong predictors (BP, HDL, smoking) and prospectively look at unplanned visits
  – Identify causal relationships to leverage
  – Automated system for flagging high-risk patients


