Electronic Health Record Analytics: The Case of Optimal Diabetes Screening

Michael Hahsler\textsuperscript{1}, Farzad Kamalzadeh\textsuperscript{1}
Vishal Ahuja\textsuperscript{1}, and Michael Bowen\textsuperscript{2}

\textsuperscript{1} Southern Methodist University
\textsuperscript{2} UT Southwestern Medical Center and Parkland Health and Hospital System

EMIS Industry Advisory Board and Outreach Meeting
December 3, 2018
THE STAGGERING COSTS OF DIABETES IN AMERICA

Nearly 30 million Americans have diabetes. Medicare dollars are 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 amputation, 136 to enter end-stage kidney disease treatment and 1,795 to develop severe retinopathy that can lead to vision loss and blindness.

Diabetes Care 2013; 36:1033-1046.
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 Pre-diabetes

77 Million with Undiagnosed Pre-diabetes

Questions of Interest

Optimal screening decision under constraints and uncertainty

- Constraints on resources and patient availability. Population screening is not feasible.
- Individualize the decision based on cohort and patient characteristics.
- Focus on catching the disease (i.e., prevalence) at earlier stages.
Common Screening Strategies

1. Opportunistic Screening

2. Screening Guidelines
   - American Diabetes Association (ADA)
     All adults over age 45 OR any age if BMI ≥ 25 (or ≥ 23 in Asians) AND an additional risk factor
   - U.S. Preventive Services Task Force (USPSTF) 2015
     Adults 40-70 AND BMI≥25

3. Diabetes Risk Score
   - Incidence/prevalence risk score.
   - Not widely used in the US.

Setting and Data

- **Setting**: Parkland Health and Hospital System, a large integrated, safety-net healthcare system in North Texas.
- **Data Source**: Epic Electronic Medical Record (EHR)
- **Retrospective cohort** (N = 34,297 patients, 2012-2015)
- **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 Extracted from EHR

105 Features including

- **Demographic information**: Age, gender, ethnicity, etc.
- **Vitals**: Blood pressure, etc.
- **BMI**
- **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, fasting plasma glucose, oral glucose tolerance test

Only demographic information, BMI and vitals are widely available.

>20% of the data values are missing overall.

>50% of lab values missing.
# Health Analytics Framework

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Descriptive Analysis</th>
<th>Predictive / Prescriptive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease Progression Model</td>
<td>HMM</td>
<td>Screening Decision Model</td>
</tr>
</tbody>
</table>

| Individual Patient | Historical data | Risk factors | Predictive Risk Model |
Partially Observable Markov Decision Process

POMDP: Discrete Health Status States

Note: We only know if a patient has (pre)diabetes if we screen the patient.
POMDP: Actions, Transitions and Rewards

Actions:
- Screen/don’t screen

Rewards:
- Cost of screening, medical expenses, reduced quality of life, lost income

Transitions depend on actions and are associated with rewards.
POMDP: Observations and Belief States

Observations give us information about the unobservable health status → “Belief State”
A new observation results in a change of our “Belief State.”
Goal: **optimal policy**. I.e., optimal action for each state to maximize the expected future reward.
Health Analytics Framework

<table>
<thead>
<tr>
<th>Cohort Specific</th>
<th>Disease Progression Model</th>
<th>Screening Decision Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort</td>
<td>HMM</td>
<td>POMDP</td>
</tr>
<tr>
<td>Individual Patient</td>
<td>Historical data</td>
<td>Predictive Risk Model</td>
</tr>
</tbody>
</table>

Descriptive Analysis | Predictive / Prescriptive
HMM: Learn a Cohort-Specific Disease Progression Model

Health Analytics Framework

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Descriptive Analysis</th>
<th>Predictive / Prescriptive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease Progression Model</td>
<td>Disease Progression Model</td>
<td>Screening Decision Model</td>
</tr>
<tr>
<td>HMM</td>
<td>POMDP</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Individual Patient</th>
<th>Historical data</th>
<th>Risk factors</th>
<th>Predictive Risk Model</th>
<th>Personalized</th>
</tr>
</thead>
</table>
Predictive Risk Model

Collins et al., Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting, BMC Medicine 2011 9:103
Observations via Predictive Modeling

• **Idea:** Use predictive modeling (classification) to learn the relationship between clinical observations recorded in EHR and the unobservable health state. Predictions can be used as personalized observations resembling a “Virtual Screening.”

• **Our key questions are:**
  • How to we produce **simple predictive models** to guide screening using only already available data?
  • How do we deal with a large quantity of **missing data** and **data quality issues**?

• **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.
Comparison of Some Predictive Models

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASSO (best)</td>
<td>77%</td>
<td>100%</td>
</tr>
<tr>
<td>NB (select feat.)</td>
<td>76%</td>
<td>100%</td>
</tr>
<tr>
<td>NB (10)</td>
<td>74%</td>
<td>100%</td>
</tr>
<tr>
<td>LASSO (10)</td>
<td>73%</td>
<td>100%</td>
</tr>
<tr>
<td>RBG (avg)</td>
<td>76%</td>
<td>64%</td>
</tr>
<tr>
<td>BMI</td>
<td>67%</td>
<td>87%</td>
</tr>
<tr>
<td>RBG (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>

LASSO: Logistic Regression with Regularization
NB: Naïve Bayes Classifier
RBG: Random Blood Glucose Test
POMDP: Parameters

### Disease Progression (Transitions)

\[
P = \begin{pmatrix}
0.9438 & 0.048 & 0 & 0.0082 \\
0.0328 & 0.9242 & 0.0348 & 0.0082 \\
0 & 0 & 0.9916 & 0.0084 \\
0 & 0 & 0 & 1
\end{pmatrix}
\]

### Risk Prediction Performance

\[
O(o|s) = P \begin{pmatrix}
0.8 & 0.15 & 0.05 \\
0.15 & 0.7 & 0.15 \\
0.05 & 0.25 & 0.7
\end{pmatrix}
\]

### Rewards (from Literature)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Source</th>
<th>Patient</th>
<th>Healthcare system</th>
<th>Society</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C_s)</td>
<td>Cost of a diabetes screening test</td>
<td>[55][56][57][16]</td>
<td>$134+$192</td>
<td>$8020</td>
<td>$8346</td>
</tr>
<tr>
<td>(Q)</td>
<td>Quality-Adjusted Life Year in U.S. dollars</td>
<td>[58]</td>
<td>$50,000</td>
<td></td>
<td>$50,000</td>
</tr>
<tr>
<td>(C_D)</td>
<td>Direct medical costs per year for new-onset diabetes</td>
<td>[55]</td>
<td>$4,174</td>
<td></td>
<td>$4,174</td>
</tr>
<tr>
<td>(C_P)</td>
<td>Incremental direct medical costs per year for a patient with prediabetes</td>
<td>[55]</td>
<td>$1,316</td>
<td></td>
<td>$1,316</td>
</tr>
<tr>
<td>(\alpha_P)</td>
<td>Annual utility decrease of living with prediabetes</td>
<td>[59][60]</td>
<td>0.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\alpha_{UD})</td>
<td>Annual utility decrease of living with undiagnosed diabetes</td>
<td>[59][61][62][63]</td>
<td>0.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\alpha_{DD})</td>
<td>Annual utility decrease of living with diagnosed diabetes</td>
<td>[59][61][62][63]</td>
<td>0.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(m_T)</td>
<td>Age-Adjusted mortality rate in U.S. in 2016</td>
<td>[53][64]</td>
<td>0.0084</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(m_D)</td>
<td>Age-adjusted mortality rate for Diabetes in 2016</td>
<td>[53][64]</td>
<td>0.00021</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(l_s)</td>
<td>Life expectancy for the U.S. population in 2016</td>
<td>[53]</td>
<td>78.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(l_d)</td>
<td>Lifespan decrement due to Diabetes</td>
<td>[65]</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(u_r)</td>
<td>Uptake rate of Diabetes screening</td>
<td>[66][67][68][69][70]</td>
<td>0.644</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
POMDP: Optimal Screening Policy

- We maintain for each patient a belief state.
- The belief state is updated with each new observation.
- The policy is a set of all considered belief states with the optimal action for each state.
POMDP: Optimal Screening Policy
# Effectiveness compared to Opportunistic Screening

<table>
<thead>
<tr>
<th>Screening Policy</th>
<th>ICER (incr. cost per QALY) (SD)</th>
<th>Years Gained (SD)</th>
<th>QALYs gained (SD)</th>
<th>Diagnosis lead time reduction (SD)</th>
<th>Macrovascular events prevented (SD)</th>
<th>Microvascular events prevented (SD)</th>
<th>Deaths prevented (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30+, every 3 years</td>
<td>$27,042 (1268)</td>
<td>0.75 (0.04)</td>
<td>2.04 (0.05)</td>
<td>19 (0.2)</td>
<td>22 (1.6)</td>
<td>207 (4)</td>
<td>48 (2)</td>
</tr>
<tr>
<td>45+, every year</td>
<td>$37,366 (1755)</td>
<td>0.62 (0.04)</td>
<td>1.18 (0.03)</td>
<td>14 (0.1)</td>
<td>21 (1.5)</td>
<td>178 (4)</td>
<td>45 (2)</td>
</tr>
<tr>
<td>45+, every 3 years</td>
<td>$31,155 (1791)</td>
<td>0.61 (0.04)</td>
<td>0.96 (0.03)</td>
<td>11 (0.1)</td>
<td>20 (1.4)</td>
<td>165 (4)</td>
<td>44 (2)</td>
</tr>
<tr>
<td>45+, every 5 years</td>
<td>$29,644 (2175)</td>
<td>0.60 (0.04)</td>
<td>0.86 (0.03)</td>
<td>9 (0.1)</td>
<td>20 (1.5)</td>
<td>157 (4)</td>
<td>44 (2)</td>
</tr>
<tr>
<td>60+, every 3 years</td>
<td>$32,201 (2966)</td>
<td>0.59 (0.04)</td>
<td>0.60 (0.03)</td>
<td>6 (0.1)</td>
<td>19 (1.4)</td>
<td>142 (4)</td>
<td>42 (2)</td>
</tr>
<tr>
<td>Maximum screening 30+</td>
<td>$36,801 (1233)</td>
<td>0.83 (0.05)</td>
<td>2.63 (0.05)</td>
<td>25 (0.2)</td>
<td>23 (1.5)</td>
<td>229 (4)</td>
<td>50 (2)</td>
</tr>
<tr>
<td>Proposed optimal policy</td>
<td>$20,426 (1839)</td>
<td>0.81 (0.04)</td>
<td>2.06 (0.05)</td>
<td>18 (0.2)</td>
<td>23 (1.5)</td>
<td>219 (5)</td>
<td>49 (2)</td>
</tr>
</tbody>
</table>

Limitations and Future Steps

- **HMM**: Estimation of transition probabilities may be biased because it is based on actually screened patients.
- **Predictive Model**: Missing data and data quality are a big issues.
- **POMDP**
  - Cost/reward structure in POMDP (e.g., real cost depends on time in a state)
  - Process is most likely not Markovian (more states can represent dependence on past information).
  - Other dimensions for the state space (E.g., age or BMI)? Make the model harder to solve due to an explosion of the number of belief states.
  - Set of possible/available actions (e.g., other interventions including diet and exercise).