Lecture 2: Risk stratification

Prof. David Sontag
MIT EECS, CSAIL, IMES

(Thanks to Narges Razavian for some of the slides)
Outline for today’s class

1. Case study for risk stratification: Early detection of Type 2 diabetes
2. Framing as supervised learning problem
3. Deriving labels
4. Evaluating risk stratification algorithms
5. Non-stationarity
Outline for today’s class

1. Case study for risk stratification: Early detection of Type 2 diabetes
2. Framing as supervised learning problem
3. Deriving labels
4. Evaluating risk stratification algorithms
5. Non-stationarity
Type 2 Diabetes: A Major public health challenge

$245$ billion: Total costs of diagnosed diabetes in the United States in 2012
$831$ billion: Total fiscal year federal budget for healthcare in the United States in 2014
Type 2 Diabetes Can Be Prevented *

Requirement for successful large scale prevention program

1. Detect/reach truly at risk population

2. Improve the interventions

3. Lower the cost of intervention

Type 2 Diabetes Can Be Prevented *

Requirement for successful large scale prevention program

1. Detect/reach truly at risk population

2. Improve the interventions

3. Lower the cost of intervention

Traditional Risk Prediction Models

- Successful Examples
  - ARIC
  - KORA
  - FRAMINGHAM
  - AUSDRISC
  - FINDRISC
  - San Antonio Model

- Easy to ask/measure in the office, or for patients to do online

- Simple model: can calculate scores by hand
Challenges of Traditional Risk Prediction Models

• A screening step needs to be done for every member in the population
  • Either in the physician’s office or as surveys
  • Costly and time-consuming
  • Infeasible for regular screening for millions of individuals

• Models not easy to adapt to multiple surrogates, when a variable is missing
  • Discovery of surrogates not straightforward
Population-Level Risk Stratification

• Key idea: Use readily available administrative, utilization, and clinical data

• Machine learning will find surrogates for risk factors that would otherwise be missing

• Perform risk stratification at the population level – millions of patients

[Razavian, Blecker, Schmidt, Smith-McLallen, Nigam, Sontag. Big Data. ‘16]
A Data-Driven approach on Longitudinal Data

- Looking at individuals who got diabetes today, (compared to those who didn’t)
  - Can we infer which variables in their record could have predicted their health outcome?

A Few Years Ago

Today
Reminder: Administrative & Clinical Data

Eligibility Record:
- Member ID
- Age/gender
- ID of subscriber
- Company code

Medical Claims:
- ICD9 diagnosis codes
- CPT code (procedure)
- Specialty
- Location of service
- Date of Service

Medications:
- NDC code (drug name)
- Days of supply
- Quantity
- Service Provider ID
- Date of fill

Lab Tests:
- LOINC code (urine or blood test name)
- Results (actual values)
- Lab ID
- Range high/low-Date
Top diagnosis codes

<table>
<thead>
<tr>
<th>Disease</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>4011 Benign hypertension</td>
<td>447017</td>
</tr>
<tr>
<td>2724 Hyperlipidemia NEC/NOS</td>
<td>382030</td>
</tr>
<tr>
<td>4019 Hypertension NOS</td>
<td>372477</td>
</tr>
<tr>
<td>25000 DMII wo cmp nt st uncnt</td>
<td>339522</td>
</tr>
<tr>
<td>2720 Pure hypercholesterolem</td>
<td>232671</td>
</tr>
<tr>
<td>2722 Mixed hyperlipidemia</td>
<td>180015</td>
</tr>
<tr>
<td>V7231 Routine gyn examination</td>
<td>178709</td>
</tr>
<tr>
<td>2449 Hypothyroidism NOS</td>
<td>169829</td>
</tr>
<tr>
<td>78079 Malaise and fatigue NEC</td>
<td>149797</td>
</tr>
<tr>
<td>V0481 Vaccin for influenza</td>
<td>147858</td>
</tr>
<tr>
<td>7242 Lumbago</td>
<td>137345</td>
</tr>
<tr>
<td>V7612 Screen mammogram NEC</td>
<td>129445</td>
</tr>
<tr>
<td>V700 Routine medical exam</td>
<td>127848</td>
</tr>
</tbody>
</table>

Out of 135K patients who had laboratory data
## Top lab test results

<table>
<thead>
<tr>
<th>Lab test</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creatinine</td>
<td>1284737</td>
</tr>
<tr>
<td>Urea nitrogen</td>
<td>1282344</td>
</tr>
<tr>
<td>Potassium</td>
<td>1280812</td>
</tr>
<tr>
<td>Glucose</td>
<td>1282344</td>
</tr>
<tr>
<td>Alanine aminotransferase</td>
<td>1187809</td>
</tr>
<tr>
<td>Aspartate aminotransferase</td>
<td>1187965</td>
</tr>
<tr>
<td>Protein</td>
<td>1277338</td>
</tr>
<tr>
<td>Albumin</td>
<td>1274166</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>1268269</td>
</tr>
<tr>
<td>Triglyceride</td>
<td>1257751</td>
</tr>
<tr>
<td>Cholesterol.in LDL</td>
<td>1241208</td>
</tr>
<tr>
<td>Calcium</td>
<td>1165370</td>
</tr>
<tr>
<td>Sodium</td>
<td>1167675</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lab test</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cholesterol.in HDL</td>
<td>1155666</td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>1152726</td>
</tr>
<tr>
<td>Hemoglobin.total/Cholesterol.in HDL</td>
<td>1037730</td>
</tr>
<tr>
<td>Glomerular filtration rate/1.73 sq M.predicted</td>
<td>561309</td>
</tr>
<tr>
<td>Erythrocyte mean corpuscular hemoglobin</td>
<td>1070832</td>
</tr>
<tr>
<td>Leukocytes</td>
<td>1062980</td>
</tr>
<tr>
<td>Erythrocytes</td>
<td>1062445</td>
</tr>
<tr>
<td>Erythrocyte mean corpuscular volume</td>
<td>1063665</td>
</tr>
<tr>
<td>Thyrotropin</td>
<td>891807</td>
</tr>
</tbody>
</table>

### Count of people who have the test result (ever)

- Neutrophils/100 leukocytes: 952089
- Lymphocytes: 943918
- Basophils: 863448
- Eosinophils: 935710
- Monocytes/100 leukocytes: 943764
- Basophils/100 leukocytes: 863435
- Neutrophils: 943232
- Monocytes: 942978
- Eosinophils/100 leukocytes: 933929
- Thyrotropin: 891807
- Hemoglobin A1c/Hemoglobin.total: 527062
Outline for today’s class

1. Case study for risk stratification: Early detection of Type 2 diabetes
2. Framing as supervised learning problem
3. Deriving labels
4. Evaluating risk stratification algorithms
5. Non-stationarity
Framing for supervised machine learning

Align by absolute time

Gap is important to prevent label leakage
Alternative framings

• Align by relative time, e.g.
  – 2 hours into patient stay in ER
  – Every time patient sees PCP
  – When individual turns 40 yrs old

• Align by data availability

NOTE:

• If multiple data points per patient, make sure each patient in *only* train, validate, or test
Methods

• L1 Regularized Logistic Regression
  – Simultaneously optimizes predictive performance \textit{and}
  – Performs feature selection, choosing the subset of the features that are most predictive

• This prevents overfitting to the training data
Features used in models

Demographics (age, sex, etc.)

Health insurance coverage

Specialty of doctors seen (cardiology, rheumatology, ...)

Service place (urgent care, inpatient, outpatient, ...)

Medications taken (999 features) (laxatives, metformin, anti-arthritis, ...)

Procedures performed (457 features)

Laboratory indicators (7000 features)

For the 1000 most frequent lab tests:
• Was the test ever administered?
• Was the result ever low?
• Was the result ever high?
• Was the result ever normal?
• Is the value increasing?
• Is the value decreasing?
• Is the value fluctuating?
Features used in models

- **Demographics** (age, sex, etc.)
- **Health insurance coverage**
- **Service place** (urgent care, inpatient, outpatient, ...)
- **Specialty of doctors seen** (cardiology, rheumatology, ...)
- **Medications taken** (999 features) (laxatives, metformin, anti-arthritis, ...)
- **Procedures performed** (457 features)
- **Laboratory indicators** (7000 features)
- **16,000 ICD-9 diagnosis codes** (all history)

Total features per patient: 42,000
Outline for today’s class

1. Case study for risk stratification: Early detection of Type 2 diabetes
2. Framing as supervised learning problem
3. Deriving labels
4. Evaluating risk stratification algorithms
5. Non-stationarity
Where do the labels come from?

1. Manually label data by chart review
2. Electronic phenotyping from medical records
3. Use machine learning to get the labels themselves
Electronic phenotyping

PheKB

Access Validated Phenotype Algorithms

One-stop documentation and versioning of validated phenotype algorithms
Tailored searches for algorithms applicable to your EMR system

Validate existing phenotype algorithms on your EMR
Receive feedback and additional validation

Share Validated Phenotype Algorithms

Publicize your work to better find collaborators
Receive feedback and validation of your algorithm

Collaborate on Phenotype Algorithms
Electronic phenotyping

Figure 1: Algorithm for identifying T2DM cases in the EMR.
Visualization (looking at individual patients) is important to sanity check labeling method.
Getting the labels using the Anchor & Learn Framework

- Use a combination of domain expertise (simple rules) and vast amounts of data (machine learning)
- Method does not require any manual labeling
- Anchors are highly transferable between institutions

[Halpern et al., AMIA 2014]
What are anchors?

• Rather than provide gold-standard labels, construct a simple rule that can catch some positive cases.

• Examples:

<table>
<thead>
<tr>
<th>Clin. state var</th>
<th>Possible Anchor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetic</td>
<td>gsn:016313 (insulin) in Medications</td>
</tr>
<tr>
<td>Cardiac</td>
<td>ICD9:428.X (heart failure) in Diagnoses</td>
</tr>
<tr>
<td>Nursing home</td>
<td>“from nursing home” in text</td>
</tr>
<tr>
<td>Social work</td>
<td>“social work consulted” in text</td>
</tr>
</tbody>
</table>
What are anchors?

• Rather than provide gold-standard labels, construct a simple rule that can catch some positive cases. **Low sensitivity here is ok!**

• Examples:

<table>
<thead>
<tr>
<th>Clin. state var</th>
<th>Possible Anchor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetic</td>
<td>gsn:016313 (insulin) in Medications</td>
</tr>
<tr>
<td>Cardiac</td>
<td>ICD9:428.X (heart failure) in Diagnoses</td>
</tr>
<tr>
<td>Nursing home</td>
<td>“from nursing home” in text</td>
</tr>
<tr>
<td>Social work</td>
<td>“social work consulted” in text</td>
</tr>
</tbody>
</table>
Learning with Anchors

- Identify anchors
- Learn to predict the anchors (anchor as pseudo-labels)
- Account for the difference between anchors and labels

<table>
<thead>
<tr>
<th>LOINC</th>
<th>UMLS CUID</th>
<th>RXnorm</th>
<th>ICD9</th>
<th>Unstructured Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Transform

Predict anchor

Predict label
Theoretical basis for anchors

• Unobserved variable: Y, Observation: A
• A is an anchor for Y if conditioning on $A=1$ gives uniform samples from the set of positive cases.
Theoretical basis for anchors

• Unobserved variable: Y, Observation: A
• A is an anchor for Y if conditioning on A=1 gives uniform samples from the set of positive cases.
• Alternative formulation – two necessary conditions:

\[ P(Y = 1|A = 1) = 1 \quad \text{AND} \quad A \perp \chi|Y \]

- Positive condition
- Conditional independence

\( \chi \) represents all other observations.
Theoretical basis for anchors

• Unobserved variable: Y, Observation: A

• A is an anchor for Y if conditioning on A=1 gives uniform samples from the set of positive cases.

• Alternative formulation – two necessary conditions:

\[ P(Y = 1|A = 1) = 1 \quad \text{AND} \quad A \perp \mathcal{X}|Y \]

Positive condition

\begin{itemize}
  \item e.g. If patient is taking *insulin*, the patient is surely *diabetic*.
\end{itemize}

Conditional independence

\begin{itemize}
  \item e.g. If we know the patient had *heart failure*, knowing whether the *diagnosis code* appears does not inform us about the rest of the record.
\end{itemize}
Theoretical basis for anchors

• Unobserved variable: Y, Observation: A

• A is an anchor for Y if conditioning on $A=1$ gives uniform samples from the set of positive cases.

• Theorem [Elkan & Noto 2008]:

  *In the above setting, a function to predict A can be transformed to predict Y*

• Can also use more recent advances on *learning with noisy labels* (e.g., Natarajan et al., NIPS ‘13)
# Learning with anchors

**Input:** anchor A
- unlabeled patients

**Output:** prediction rule

1. Learn a calibrated classifier (e.g. logistic regression) to predict:
   \[
   \Pr(A = 1 \mid \mathcal{X})
   \]
2. Using a validate set, let \( P \) be the patients with \( A=1 \). Compute:
   \[
   C = \frac{1}{|P|} \sum_{k \in P} \Pr(A = 1 \mid \mathcal{X}^{(k)})
   \]
3. For a previously unseen patient \( t \), predict:
   \[
   \frac{1}{C} \Pr(A = 1 \mid \mathcal{X}^{(t)}) \quad \text{if } A^{(t)} = 0
   
   1 \quad \text{if } A^{(t)} = 1
   \]

[Elkan & Noto 2008]

---

**Learning**
- Learn to predict \( A \) from the other variables.

**Calibration**
- \( C \) is the average model prediction for patients with anchors.

**Transformation**
- If no anchor present, according to a scaled version of the anchor-prediction model.
Outline for today’s class

1. Case study for risk stratification: Early detection of Type 2 diabetes
2. Framing as supervised learning problem
3. Deriving labels
4. **Evaluating risk stratification algorithms**
5. Non-stationarity
What are the Discovered Risk Factors?

- 769 variables have non-zero weight

<table>
<thead>
<tr>
<th>Top History of Disease</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impaired Fasting Glucose (Code 790.21)</td>
<td>4.17 (3.87 4.49)</td>
</tr>
<tr>
<td>Abnormal Glucose NEC (790.29)</td>
<td>4.07 (3.76 4.41)</td>
</tr>
<tr>
<td>Hypertension (401)</td>
<td>3.28 (3.17 3.39)</td>
</tr>
<tr>
<td>Obstructive Sleep Apnea (327.23)</td>
<td>2.98 (2.78 3.20)</td>
</tr>
<tr>
<td>Obesity (278)</td>
<td>2.88 (2.75 3.02)</td>
</tr>
<tr>
<td>Abnormal Blood Chemistry (790.6)</td>
<td>2.49 (2.36 2.62)</td>
</tr>
<tr>
<td>Hyperlipidemia (272.4)</td>
<td>2.45 (2.37 2.53)</td>
</tr>
<tr>
<td>Shortness Of Breath (786.05)</td>
<td>2.09 (1.99 2.19)</td>
</tr>
<tr>
<td>Esophageal Reflux (530.81)</td>
<td>1.85 (1.78 1.93)</td>
</tr>
</tbody>
</table>

Diabetes
1-year gap
What are the Discovered Risk Factors?

- 769 variables have non-zero weight

<table>
<thead>
<tr>
<th>Top History of Disease</th>
<th>Odds Ratio</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impaired Fasting Glucose (Code 790.21)</td>
<td>4.17</td>
<td>(3.87 4.49)</td>
</tr>
<tr>
<td>Abnormal Glucose NEC (790.29)</td>
<td>4.07</td>
<td>(3.76 4.41)</td>
</tr>
<tr>
<td>Hypertension (401)</td>
<td>3.28</td>
<td>(3.17 3.39)</td>
</tr>
<tr>
<td>Obstructive Sleep Apnea (327.23)</td>
<td>2.98</td>
<td>(2.78 3.20)</td>
</tr>
<tr>
<td>Obesity (278)</td>
<td>2.88</td>
<td>(2.75 3.02)</td>
</tr>
<tr>
<td>Abnormal Blood Chemistry (790.6)</td>
<td>2.49</td>
<td>(2.36 2.62)</td>
</tr>
<tr>
<td>Hyperlipidemia (272.4)</td>
<td>2.45</td>
<td>(2.37 2.53)</td>
</tr>
<tr>
<td>Shortness Of Breath (786.05)</td>
<td>2.09</td>
<td>(1.99 2.19)</td>
</tr>
<tr>
<td>Esophageal Reflux (530.81)</td>
<td>1.85</td>
<td>(1.78 1.93)</td>
</tr>
</tbody>
</table>

Additional Disease Risk Factors Include:
- Pituitary dwarfism (253.3),
- Hepatomegaly (789.1), Chronic Hepatitis C (070.54), Hepatitis (573.3), Calcaneal Spur (726.73), Thyrotoxicosis without mention of goiter (242.90), Sinoatrial Node dysfunction (427.81), Acute frontal sinusitis (461.1), Hypertrophic and atrophic conditions of skin (701.9), Irregular menstruation (626.4), ...

Diabetes 1-year gap
What are the Discovered Risk Factors?

- 769 variables have non-zero weight

<table>
<thead>
<tr>
<th>Top Lab Factors</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemoglobin A1c /Hemoglobin.Total (High - past 2 years)</td>
<td>5.75 (5.42 6.10)</td>
</tr>
<tr>
<td>Glucose (High- Past 6 months)</td>
<td>4.05 (3.89 4.21)</td>
</tr>
<tr>
<td>Cholesterol.In VLDL (Increasing - Past 2 years)</td>
<td>3.88 (3.53 4.27)</td>
</tr>
<tr>
<td>Potassium (Low - Entire History)</td>
<td>2.58 (2.24 2.98)</td>
</tr>
<tr>
<td>Cholesterol.Total/Cholesterol.In HDL (High - Entire History)</td>
<td>2.29 (2.19 2.40)</td>
</tr>
<tr>
<td>Erythrocyte mean corpuscular hemoglobin concentration -(Low - Entire History)</td>
<td>2.25 (1.92 2.64)</td>
</tr>
<tr>
<td>Eosinophils (High - Entire History)</td>
<td>2.11 (1.82 2.44)</td>
</tr>
<tr>
<td>Glomerular filtration rate/1.73 sq M.Predicted (Low -Entire History)</td>
<td>2.07 (1.92 2.24)</td>
</tr>
<tr>
<td>Alanine aminotransferase (High Entire History)</td>
<td>2.04 (1.89 2.19)</td>
</tr>
</tbody>
</table>

Diabetes
1-year gap
What are the Discovered Risk Factors?

- 769 variables have non-zero weight

### Top Lab Factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>Odds Ratio</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemoglobin A1c /Hemoglobin.Total (High - past 2 years)</td>
<td>5.75</td>
<td>(5.42 6.10)</td>
</tr>
<tr>
<td>Glucose (High - past 6 months)</td>
<td>4.05</td>
<td>(3.89 4.21)</td>
</tr>
<tr>
<td>Cholesterol.In VLDL (Increasing - past 2 years)</td>
<td>3.88</td>
<td>(3.53 4.27)</td>
</tr>
<tr>
<td>Potassium (Low - entire history)</td>
<td>2.58</td>
<td>(2.24 2.98)</td>
</tr>
<tr>
<td>Cholesterol.Total/Cholesterol.In HDL (High - entire history)</td>
<td>2.29</td>
<td>(2.19 2.40)</td>
</tr>
<tr>
<td>Erythrocyte mean corpuscular hemoglobin concentration -(Low - entire history)</td>
<td>2.25</td>
<td>(1.92 2.64)</td>
</tr>
<tr>
<td>Eosinophils (High - entire history)</td>
<td>2.11</td>
<td>(1.82 2.44)</td>
</tr>
<tr>
<td>Glomerular filtration rate/1.73 sq M.Predicted (Low - entire history)</td>
<td>2.07</td>
<td>(1.92 2.24)</td>
</tr>
<tr>
<td>Alanine aminotransferase (High - entire history)</td>
<td>2.04</td>
<td>(1.89 2.19)</td>
</tr>
</tbody>
</table>

### Additional Lab Test Risk Factors Include:

- Albumin/Globulin (Increasing - entire history),
- Urea nitrogen/Creatinine -(high - entire history),
- Specific gravity (Increasing, past 2 years),
- Bilirubin (high - past 2 years),

---

Diabetes 1-year gap
Receiver-operator characteristic curve

Obtained by varying prediction threshold

Want to be here

True positive rate

False positive rate

Diabetes 1-year gap

Full model

Traditional risk factors
Receiver-operator characteristic curve

Area under the ROC curve (AUC)

AUC = Probability that algorithm ranks a positive patient over a negative patient

Invariant to amount of class imbalance

Diabetes 1-year gap
Receiver-operator characteristic curve

Risk stratification usually focuses on just this region (because of the cost of interventions)

Full model \( AUC = 0.78 \)

Traditional risk factors \( AUC = 0.74 \)

Random \( AUC = 0.5 \)

Diabetes 1-year gap

True positive rate

False positive rate
Positive predictive value (PPV)

- Top 100 Predictions: 0.15
  - Traditional risk factors: 0.06
  - Full model: 0.17
- Top 1000 Predictions: 0.07
  - Traditional risk factors: 0.07
  - Full model: 0.17
- Top 10000 Predictions: 0.1
  - Traditional risk factors: 0.06
  - Full model: 0.1

Diabetes 1-year gap
Calibration (*note: different dataset*)

Predicting infection in the ER

Model
- BoW
- CC
- Topics
- Vitals

Fraction of patients the model predicts to have this probability of infection
Outline for today’s class

1. Case study for risk stratification: Early detection of Type 2 diabetes
2. Framing as supervised learning problem
3. Deriving labels
4. Evaluating risk stratification algorithms
5. Non-stationarity
Major challenge: non-stationarity

• ICD10 rolled out in 2015: predictive models learned using ICD9 features are no longer useful!
• Logistical issues => some features may not be available!
• Prevalence and significance of features may change over time
• Automatically derived labels may change meaning
Top 100 lab measurements over time

Time (in months, from 1/2005 up to 1/2014)
A slight downward trend of new diagnoses
(Among patients who are enrolled between 2009 and 2013, how many newly diagnosed diabetics do we have each month?)
Diabetes Onset after 2009

External validity

• Motivates multi-institution evaluations
• Good practice is to let the test data be from a future year