Transfer Learning using Electronic Health Record data

ICME Expo 2017

Ken Jung, Shah Lab
Stanford University School of Medicine
Thanks Sebastien & Nathanael!

- Sebastien Dubois (ICME)
- Nathanael Romano (MS&E)
Familial Hypercholesterolemia

- Genetic disorder causing very high LDL ("bad" cholesterol)
- 1/250 - 1/400 prevalence
- 90% undiagnosed
- Treatment by latest generation of statins can help (PCSK9 inhibitors)
Cost structure

- Assume 1/100 patients in a lipid clinic are undiagnosed cases
- $300 for genetic test
- $10,000 for PCSK9 inhibitor, 1 year
- Net $40,000 to find and treat one patient, 75% to find
How can we bring cost down?

- Target genetic testing to enrich cases
- Assume a classifier with 50% PPV
- 1/2 tested patients are cases
- Net cost to find and treat one patient: $10,600
Electronic health records

- Where would the data come from?
- Electronic Health Records (EHRs) have been in use at Stanford Hospital for many years
- Increasing adoption in US (75%)
Data from Stanford Hospital’s EHR

- > 2 million patients
- Structured data (diagnosis, procedure, and medication codes) - 43 million records
- Unstructured data (free text clinical notes) - 42 million notes
But...

- 93 confirmed cases
Does this seem right?

❖ “Fruit” appearing in the clinical notes is a significant predictor...
Explicit incorporation of domain knowledge

❖ “automate” existing diagnosis guideline
❖ Extract features from EHR that approximate guideline inputs
❖ Fit a linear model on ~20 features

Table 4. Dutch Lipid Clinic Network diagnostic criteria for Familial Hypercholesterolemia^3

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Family history</strong></td>
<td></td>
</tr>
<tr>
<td>First-degree relative with known premature* coronary and vascular disease, OR</td>
<td>1</td>
</tr>
<tr>
<td>First-degree relative with known LDL-C level above the 95th percentile</td>
<td></td>
</tr>
<tr>
<td>First-degree relative with tendinous xanthomata and/or arcus cornealis, OR</td>
<td>2</td>
</tr>
<tr>
<td>Children aged less than 18 years with LDL-C level above the 95th percentile</td>
<td></td>
</tr>
<tr>
<td><strong>Clinical history</strong></td>
<td></td>
</tr>
<tr>
<td>Patient with premature* coronary artery disease</td>
<td>2</td>
</tr>
<tr>
<td>Patient with premature* cerebral or peripheral vascular disease</td>
<td>1</td>
</tr>
<tr>
<td><strong>Physical examination</strong></td>
<td></td>
</tr>
<tr>
<td>Tendinous xanthomata</td>
<td>6</td>
</tr>
<tr>
<td>Arcus cornealis prior to age 45 years</td>
<td>4</td>
</tr>
<tr>
<td><strong>Cholesterol levels mg/dl (mmol/liter)</strong></td>
<td></td>
</tr>
<tr>
<td>LDL-C &gt;= 330 mg/dL (≥8.5)</td>
<td>8</td>
</tr>
<tr>
<td>LDL-C 250 – 329 mg/dL (6.5–8.4)</td>
<td>5</td>
</tr>
<tr>
<td>LDL-C 190 – 249 mg/dL (5.0–6.4)</td>
<td>3</td>
</tr>
<tr>
<td>LDL-C 155 – 189 mg/dL (4.0–4.9)</td>
<td>1</td>
</tr>
<tr>
<td><strong>DNA analysis</strong></td>
<td></td>
</tr>
<tr>
<td>Functional mutation in the LDLR, apo B or PCSK9 gene</td>
<td>8</td>
</tr>
<tr>
<td><strong>Diagnosis (diagnosis is based on the total number of points obtained)</strong></td>
<td></td>
</tr>
<tr>
<td>Definite Familial Hypercholesterolemia</td>
<td>&gt;8</td>
</tr>
<tr>
<td>Probable Familial Hypercholesterolemia</td>
<td>6 – 8</td>
</tr>
<tr>
<td>Possible Familial Hypercholesterolemia</td>
<td>3 – 5</td>
</tr>
<tr>
<td>Unlikely Familial Hypercholesterolemia</td>
<td>&lt;3</td>
</tr>
</tbody>
</table>

* Premature = < 55 years in men; < 60 years in women
LDL-C = low density lipoprotein cholesterol; FH, familial hypercholesterolemia.
LDLR = low density lipoprotein receptor
Apo B = apolipoprotein B
PCSK9 = Proprotein convertase subtilisin/kexin type 9

Transfer learning using EHR data

- Despite “millions of patients”, very few labels of things we care about.
- Opportunity for transfer learning?
Transfer Learning

- Models trained on ImageNet ILSVRC are very useful for other tasks with less data.
- Learn a model on a source task with lots of labels
- Use for a target task with few labels
Clinical text

- Focus on transfer learning using clinical text
- Clinical notes are most complete source of information
  - but harder to use than medical codes
- Most applications of ML to EHR data uses structured data
Representing clinical notes

- Each patient has a sequence of notes
- Sebastien & Nathanael explored ways of summarizing this sequence
Representing clinical notes

- Baselines: bag of words, LDA
- Compare against:
  - Embed and aggregate
  - RNN Sequence models
Predicting complex clinical events

![Graph showing AUROC for different models](chart.png)
Acknowledgements

- Sebastien Dubois and Nathanael Romano
- ICME NVIDIA seed grant program
- Familial Hypercholesterolemia Foundation
- Josh Knowles MD, et al
- Nigam Shah