Why Does Interpretability Matter in Health Care?

David Sontag
Clinical Machine Learning Group, MIT

Department of Electrical Engineering and Computer Science (EECS),
Computer Science and Artificial Intelligence Laboratory (CSAIL)
Institute for Medical Engineering & Science (IMES)
Adoption of Electronic Health Records (EHR) has increased 9x since 2008

Percentage of hospitals in the US

9.4% 12.2% 15.6% 27.6%* 44.4%* 59.4%* 75.5%* 83.8%*  

[Henry et al., ONC Data Brief, May 2016]
Wealth of digital health data available

- genomics
- imaging
- lab tests
- proteomics
- social media
- phone
- vital signs
- devices
- genomics
Where can machine learning help?

Computational pathologist
(Beck et al., Sci Transl Med, 2011)

Finding undiagnosed Type 2 diabetics
(my lab: Razavian et al., Big Data 2016)

Improving EHR documentation
(my lab: Jernite et al., 2013)

Differential diagnosis
(INTERNIST-1/Quick Medical Reference, 1980’s)

Symptoms | Differential diagnosis
---|---
Cough | 1. Common cold
Fever | 2. Flu
Headache | 3. Strep throat
Sore throat | 4. Meningitis

In-silico models for precision medicine

For this specific individual, which medication is better, A or B?
Outline for today’s talk

• **Reasons to want interpretable models:** trust, causality, transferability, informativeness

• **Case studies:**
  - Early detection of Type 2 diabetes
  - Framingham Coronary Heart Disease Risk Score
  - Improving clinical documentation with “auto-complete”
Typical ICML paper: we get SOTA on ____ benchmark
Real world ML: “it’s complicated”

When humans are consumer of ML, often we want something the metric doesn’t capture. But, what?

(Slide credit: Zachary Lipton)
Trust

• Does the model know when it’s uncertain?

• Does the model make same mistakes as human? (e.g., would we be happy delegating decision making authority?)

• Are we comfortable with the model?

(Slide credit: Zachary Lipton)
Causality

- We may want models to tell us something about the natural world
- Supervised models are trained simply to make predictions, but often used to take actions
- Naïve interpretations can be misleading
Transferability

• The idealized training setups often differ from the real world
  • E.g., data leakage, errors in outcome definition from observational data

• Real problem may be non-stationary, noisier, etc.

• Want sanity-checks that the model doesn’t depend on weaknesses in setup

(Slide credit: Zachary Lipton)
Informativeness

• We may train a model to make a decision

• But it’s real purpose is usually to aid a person in making a decision

• Thus an interpretation may be valuable for the extra bits it carries

  I.e., ability to integrate model output with human prior beliefs

(Slide credit: Zachary Lipton)
CASE STUDY 1

Early detection of Type 2 diabetes and its complications

Work led by Narges Razavian
Early Detection of Type 2 Diabetes

- Global prevalence will go from 171 million in 2000 to 366 million in 2030
- 25% of people in the US with diabetes are undiagnosed
- Leads to complications of cardiovascular, cerebrovascular, renal, and vision systems
Traditional risk assessment

- Use small number of risk factors (e.g. ~20)
- Easy to ask/measure in the office
- Simple model: can calculate scores by hand
Population-Level Risk Stratification

• Key idea: Use automatically collected administrative, utilization, and clinical data

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

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

[Razavian, Blecker, Schmidt, Smith-McLallen, Nigam, Sontag. Big Data. ‘16]
ML task formulation
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, ...)

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

Procedures performed (457 features)

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

Laboratory indicators (7000 features)

16,000 ICD-9 diagnosis codes (all history)

Medications taken (999 features)

Total features per patient: 42,000
Positive predictive value (PPV)

<table>
<thead>
<tr>
<th></th>
<th>Traditional risk factors</th>
<th>Full model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 100 Predictions</td>
<td>0.06</td>
<td>0.15</td>
</tr>
<tr>
<td>Top 1000 Predictions</td>
<td>0.07</td>
<td>0.17</td>
</tr>
<tr>
<td>Top 10000 Predictions</td>
<td>0.06</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Diabetes 1-year gap
Questions

1. Did we set up the prediction task correctly so that it mimics how we would apply it prospectively?
   
   Transferability

2. Do our models suggest any causal hypotheses of the mechanism in which patients become diabetic?
   
   Causality

3. Is anything fundamentally new discovered?
   
   Trust, Informativeness

4. Is the model likely to be useful in new settings?
   
   Transferability
What are the Discovered Risk Factors?

- 769 variables have non-zero weight
- No time to look at all 769. Instead, we do a regression with much higher regularization *just for visualization purposes*
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
### Top History of Disease Risk Factors

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

• 769 variables have non-zero weight
Questions

1. Did we set up the prediction task correctly so that it mimics how we would apply it prospectively?

2. Do our models suggest any causal hypotheses of the mechanism in which patients become diabetic?

3. Is anything fundamentally new discovered?

4. Is the model likely to be useful in new settings?
1. Did we set up the prediction task correctly so that it mimics how we would apply it prospectively?

For gap=0, possibly not!

We see that the diabetic medication Metformin (a first-line diabetic treatment) is predictive
Preventing diabetic onset or progression

• Our goal ultimately is to use the models’ predictions to prioritize *interventions* to prevent progression of diabetes

• What should these interventions be?
  1. **Approach 1**: Look for existing interventions in the data that might show promise (but perhaps are inconsistently applied)

     “Gastric bypass surgery” is highest negative weight (9th most predictive feature)

     **Does gastric bypass surgery prevent onset of diabetes?**

  2. **Approach 2**: Characterize the patient population that we can predict well for, and use clinical expertise
CASE STUDY 2
Framingham Coronary Heart Disease (CHD) Risk Score
Transferability: non-stationary

Data created during health care is from a non-stationary process due to changes in:

- Medical science
- Incentives & regulations
- Business processes
Transferability: non-stationary

Top 100 lab measurements over time

Time (in months, from 1/2005 up to 1/2014)
Transferability: non-stationary

- Testing for covariate shift (wound healing):

  - Fit a model to distinguish 2013 vs pre-2013 samples
    - 0.98 AUC on test set

  - Using just data from 2013
    - Train a model from first two-thirds of 2013 to predict on last third
    - 29k train, 14k test (1/3 data)
    - AUC of 0.863

(Difference credit: Ken Jung)
Case study on transferability: Framingham CHD risk score

• Many ML models are trained in one place and deployed more broadly

• **Example:** Framingham coronary heart disease (CHD) risk score
  
  • Model based on 6 major risk factors: age, BP, smoking, diabetes, total cholesterol (TC), and high-density lipoprotein cholesterol (HDL-C)

[Wilson et al., Circulation, 1998]
CHD score sheet for men using TC or LDL-C categories.

Case study on transferability: Framingham CHD risk score

• Many ML models are trained in one place and deployed more broadly

• **Example**: Framingham coronary heart disease (CHD) risk score

---

**Prediction of coronary heart disease using risk factor categories**

**Authors**: Peter WF Wilson, Ralph B D'Agostino, Daniel Levy, Albert M Belanger, Halit Silbershatz, William B Kannel

**Publication date**: 1998/5/1

**Journal**: Circulation

**Volume**: 97

**Issue**: 18

**Pages**: 1837-1847

**Publisher**: Lippincott Williams & Wilkins

**Description**: Background—The objective of this study was to examine the association of Joint National Committee (JNC-V) blood pressure and National Cholesterol Education Program (NCEP) cholesterol categories with coronary heart disease (CHD) risk, to incorporate them into coronary prediction algorithms, and to compare the discrimination properties of this approach with other noncategorical prediction functions. Methods and Results—This work was designed as a prospective, single-center study in the setting of a community-based...
Case study on transferability: Framingham CHD risk score

• Many ML models are trained in one place and deployed more broadly

• **Example:** Framingham coronary heart disease (CHD) risk score
  - 99% of Framingham participants are of European descent
  - How well does it generalize to a Chinese population?

• C-statistic (=AUC on censored data) 0.705/0.742 (M/F)
• Re-fit using local data only slightly improves C-statistic (=AUC on censored data), to 0.736/0.759 (M/F)

[Liu et al., JAMA ‘04]
Could we say the same about our more complex machine learning models?

What would we need to look at to get confidence that they would transfer as well?
CASE STUDY 3

Improving clinical documentation with “auto-complete”
Improving Quality of Structured Data

• Much of the valuable data in EHRs is in the form of free text notes

• Collecting structured data is slow and error-prone

• We can make the process faster and more accurate:
  1. Automatically keeping problem lists up to date
  2. Assigning diagnosis codes and other documentation
  3. Assigning chief complaints by leveraging the text data in a patient’s EHR
Example: Chief complaints

Changed workflow to have chief complaints assigned *last*. Predict them.

Using for all 55,000 patients/year that present at BIDMC ED
Example: Chief complaints

Percentage of standardized chief complaints (per week)
Example: Chief complaints

Changed workflow to have chief complaints assigned last. Predict them.

Challenge: trust / face validity

If clinician writes, “does not have chest pain”, then “chest pain” had better not be a suggested chief complaint

Using for all 55,000 patients/year that present at BIDMC ED
Conclusions

• Model introspection seems to be essential for using machine learning in health care
• Fertile terrain to develop new ML methods for directly tackling these issues where “interpretability” arises:
  • Building trust
  • Checking that prediction task is set up properly
  • Identifying causal hypotheses
  • Getting the gist of what is predictive
  • Assessing transferability