Machine Learning for Healthcare: 
What’s next?

David Sontag
Clinical Machine Learning Group
MIT
@david_sontag
ML4H 2018: a workshop at NeurIPS 2018
Saturday December 08, 2018
Room 517 D, Palais des Congrès de Montréal,
Montreal, Canada

This workshop will bring together machine learning researchers, clinicians, and healthcare data experts. The program consists of invited talks, contributed posters and panel discussions.

Direct questions to:
ml4h.workshop.nips.2018@gmail.com
Hot topics in MLHC

- Interpretability
- Robustness to adversaries, dataset shift
- Fairness
- Reinforcement learning
Hot topics in MLHC

• **Interpretability**
• Robustness to adversaries, dataset shift
• Fairness
• Reinforcement learning
Interpretability

• Global interpretability – understand model as a whole
  – Will it work prospectively as intended?
  – What data was most useful?

• Local interpretability – understand predictions for individual patients
  – Build trust in predictions; recognize errors
  – Provide guidance to decision makers who may have additional information
CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

(a) Patient with multifocal community acquired pneumonia. The model correctly detects the airspace disease in the left lower and right upper lobes to arrive at the pneumonia diagnosis.

(b) Patient with a left lung nodule. The model identifies the left lower lobe lung nodule and correctly classifies the pathology.

(c) Patient with primary lung malignancy and two large masses, one in the left lower lobe and one in the right upper lobe adjacent to the mediastinum. The model correctly identifies both masses in the X-ray.

(d) Patient with a right-sided pneumothorax and chest tube. The model detects the abnormal lung to correctly predict the presence of pneumothorax (collapsed lung).

(e) Patient with a large right pleural effusion (fluid in the pleural space). The model correctly labels the effusion and focuses on the right lower chest.

(f) Patient with congestive heart failure and cardiomegaly (enlarged heart). The model correctly identifies the enlarged cardiac silhouette.

Figure 3. CheXNet localizes pathologies it identifies using Class Activation Maps, which highlight the areas of the X-ray that are most important for making a particular pathology classification. The captions for each image are provided by one of the practicing radiologists.

We identify the most important features used by the model in its prediction of the pathology by upscaling the map $M_c$ to the dimensions of the image and overlaying the image. Figure 3 shows several examples of CAMs on the pneumonia detection task as well as the 14-class pathology classification task.

7. Related Work

Recent advancements in deep learning and large datasets have enabled algorithms to surpass the performance of medical professionals in a wide variety of medical imaging tasks, including diabetic retinopathy detection (Gulshan et al., 2016), skin cancer classification (Esteva et al., 2017), arrhythmia detection (Rajpurkar et al., 2017), and hemorrhage identification (Grewal et al., 2017).

Automated diagnosis from chest radiographs has received increasing attention with algorithms for pulmonary tuberculosis classification (Lakhani & Sundaram, 2017) and lung nodule detection (Huang et al., 2017). Islam et al. (2017) studied the performance of various convolutional architectures on different abnormalities using the publicly available OpenI dataset (Demner-Fushman et al., 2015). Wang et al. (2017) released ChestX-ray-14, an order of magnitude larger than previous datasets of its kind, and also bench-
Hot topics in MLHC

• Interpretability
• Robustness to adversaries, dataset shift
• Fairness
• Reinforcement learning
Machine learning is brittle: adversarial perturbations

Correctly classified as a Dog

Machine learning is brittle: adversarial perturbations

Figure 5: Adversarial examples generated for AlexNet. Each of our models were trained with L-BFGS until convergence. The first three models are linear classifiers that work on the pixel level with various loss functions yet, but our first qualitative experiments with AlexNet gives us reason to believe that convolutional models may behave similarly as well. Each of our models are simple linear (softmax) classifier without hidden units (FC10). Two other models are a simple sigmoidal neural network with two hidden layers and a classifier. The last model, AE400-10, consists of a single layer with 400 nodes and a Softmax classifier. This network was trained until it got very high quality first layer filters and this layer was fine-tuned. The results presented here are consistent with those on a larger training set. According to our initial observations, adversarial examples for the higher layers seemed to be significantly more useful than those on the input or lower layers.

In our future work, we plan to compare these effects in a systematic manner. Training set. Adversarial examples in this extreme setting as well. Two other models are a simple sigmoidal neural network with two hidden layers and a classifier. The last model, AE400-10, consists of a single layer with 400 nodes and a Softmax classifier. This network was trained until it got very high quality first layer filters and this layer was fine-tuned. The results presented here are consistent with those on a larger training set. According to our initial observations, adversarial examples for the higher layers seemed to be significantly more useful than those on the input or lower layers. In our future work, we plan to compare these effects in a systematic manner.

Training set. Adversarial examples in this extreme setting as well. Two other models are a simple sigmoidal neural network with two hidden layers and a classifier. The last model, AE400-10, consists of a single layer with 400 nodes and a Softmax classifier. This network was trained until it got very high quality first layer filters and this layer was fine-tuned. The results presented here are consistent with those on a larger training set. According to our initial observations, adversarial examples for the higher layers seemed to be significantly more useful than those on the input or lower layers. In our future work, we plan to compare these effects in a systematic manner.

Figure 6: Adversarial examples for QuocNet.

Machine learning is brittle: adversarial perturbations

Original image + Noise (not random)

[+]

Machine learning is brittle: adversarial perturbations

Original image + Noise (not random) = Classified as Ostrich!

Machine learning is brittle: adversarial perturbations

Dermoscopy

Nevus  Melanoma

0.0%  100.0%
100.0%  0.0%

Machine learning is brittle: natural changes in the data

Build population-level checks into deployment/transfer

[Figure adopted from Jen Gong and Tristan Naumann]
The top prognostic factors in the algorithm were all clinically relevant and included the terms heart failure in problem list, an inpatient loop diuretic, or a BNP level of more than 120 pg/mL. Of these hospitalizations, 33.8% had heart failure listed on the problem list (algorithm sensitivity of 67.0% and PPV of 52.5%). Among patients with heart failure listed on the problem list, 96.9% had heart failure on the problem list, followed by mention of heart failure in problem list (algorithm sensitivity of 84.2% and PPV of 58.8%).

### Table 1. Characteristics of 47 119 Hospitalized Patients

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Finding*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, mean (SE), y</td>
<td>60.9 (18.15)</td>
</tr>
<tr>
<td>Female</td>
<td>23 952 (50.8)</td>
</tr>
<tr>
<td>Black/African American race</td>
<td>5258 (11.2)</td>
</tr>
<tr>
<td>Hispanic/Latino ethnicity</td>
<td>3667 (7.8)</td>
</tr>
<tr>
<td>Medicaid</td>
<td>8303 (17.6)</td>
</tr>
<tr>
<td>Heart failure in problem list</td>
<td>3630 (7.7)</td>
</tr>
<tr>
<td>Prior diagnosis of any heart failure</td>
<td>2985 (6.3)</td>
</tr>
<tr>
<td>Prior diagnosis of primary heart failure</td>
<td>615 (1.3)</td>
</tr>
<tr>
<td>Prior echocardiography</td>
<td>15 938 (33.8)</td>
</tr>
<tr>
<td>Loop diuretics</td>
<td></td>
</tr>
<tr>
<td>Inpatient</td>
<td>6837 (14.5)</td>
</tr>
<tr>
<td>Outpatient</td>
<td>6427 (13.6)</td>
</tr>
<tr>
<td>ACE inhibitors or ARB</td>
<td></td>
</tr>
<tr>
<td>Inpatient</td>
<td>13 166 (27.9)</td>
</tr>
<tr>
<td>Outpatient</td>
<td>14 797 (31.4)</td>
</tr>
<tr>
<td>β-Blockers</td>
<td></td>
</tr>
<tr>
<td>Inpatient</td>
<td>19 748 (41.9)</td>
</tr>
<tr>
<td>Outpatient</td>
<td>14 870 (31.6)</td>
</tr>
<tr>
<td>Heart failure with β-blockers</td>
<td></td>
</tr>
<tr>
<td>Inpatient</td>
<td>6310 (13.4)</td>
</tr>
<tr>
<td>Outpatient</td>
<td>8644 (18.4)</td>
</tr>
</tbody>
</table>

### Table 2. Blood Pressure, Creatinine, Sodium, BNP, and Problem List Elements

<table>
<thead>
<tr>
<th>Blood pressure, mean (SE), mm Hg</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Systolic</td>
<td>123.3 (18.3)</td>
</tr>
<tr>
<td>Diastolic</td>
<td>67.8 (12.8)</td>
</tr>
<tr>
<td>Creatinine, mean (SE), mg/dL</td>
<td>1.01 (1.1)</td>
</tr>
<tr>
<td>Sodium, mean (SE), mEq/L</td>
<td>138.4 (3.7)</td>
</tr>
<tr>
<td>BNP, pg/mL</td>
<td></td>
</tr>
<tr>
<td>&lt;500</td>
<td>1721 (23.4)</td>
</tr>
<tr>
<td>500-999</td>
<td>878 (12.0)</td>
</tr>
<tr>
<td>1000-4999</td>
<td>2498 (34.0)</td>
</tr>
<tr>
<td>5000-9999</td>
<td>931 (12.7)</td>
</tr>
<tr>
<td>10 000-19 999</td>
<td>652 (8.9)</td>
</tr>
<tr>
<td>≥20 000</td>
<td>667 (9.1)</td>
</tr>
</tbody>
</table>

| Blood pressure                          |                      |
| Any systolic                            | 46 982 (99.7)        |
| Any diastolic                           | 46 982 (99.7)        |
| Any creatinine                          | 46 598 (98.9)        |
| Any sodium                              | 46 613 (98.9)        |
| Any BNP                                 | 7347 (15.6)          |

| Problem list                            |                      |
| Acute MI                                | 952 (2.0)            |
| Atherosclerosis                         | 6147 (13.0)          |

| Final discharge diagnosis of heart failure |                      |
| Any diagnosis                            | 6549 (13.9)          |
| Principal diagnosis                      | 1214 (2.6)           |

Abbreviations: ACE, angiotensin-converting enzyme; ARB, angiotensin receptor blockers; BNP, brain natriuretic peptide.

[Sources: Blecker et al., Comparison of Approaches for Heart Failure Case Identification From Electronic Health Record Data, JAMA Cardiology 2016]
Hot topics in MLHC

- Interpretability
- Robustness to adversaries, dataset shift
- Fairness
- Reinforcement learning
Machine Bias

There’s software used across the country to predict future criminals. And it’s biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016
Fair Regression for Health Care Spending

Anna Zink
Harvard University
and
Sherri Rose
Harvard Medical School*

January 31, 2019

Abstract

The distribution of health care payments to insurance plans has substantial consequences for social policy. Risk adjustment formulas predict spending in health insurance markets in order to provide fair benefits and health care coverage for all enrollees, regardless of their health status. Unfortunately, current risk adjustment formulas are known to undercompensate payments to health insurers for specific groups of enrollees (by underpredicting their spending). Much of the existing algorithmic fairness literature for group fairness to date has focused on classifiers and binary outcomes. To improve risk adjustment formulas for undercompensated groups, we expand on concepts from the statistics, computer science, and health economics literature to develop new fair regression methods for continuous outcomes by building fairness considerations directly into the objective function. We additionally propose a novel measure of fairness while asserting that a suite of metrics is necessary in order to evaluate risk adjustment formulas more fully. Our data application using the IBM MarketScan Research Databases and simulation studies demonstrate that these new fair regression methods may lead to massive improvements in group fairness with only small reductions in overall fit.

Keywords: Constrained regression, Penalized regression, Risk adjustment, Fairness
(a) Using Tukey’s range test, we can find the 95%-significance level for the zero-one loss for each group over 5-fold cross validation.

(b) As training set size increases, zero-one loss over 50 trials decreases over all groups and appears to converge to an asymptote.

(c) Topic modeling reveals subpopulations with high differences in zero-one loss, for example cancer patients and cardiac patients.

Figure 3: Mortality prediction from clinical notes using logistic regression. Best viewed in color.

[Chen, Johansson, Sontag, Why is my classifier discriminatory?, NeurIPS, 2018]
Hot topics in MLHC

• Interpretability
• Robustness to adversaries, dataset shift
• Fairness
• Reinforcement learning
Learning to play Atari games

Watch video: https://www.youtube.com/watch?v=V1eYniJ0Rnk

Could we use such reinforcement learning algorithms in health care?
(Off-Policy) Reinforcement Learning

• **Goal:** Find a dynamic treatment regime (policy) \( \pi(A_t \mid H_t) \)
  
  – that selects actions \( A_t \)
  
  – which optimize outcomes \( Y_{t:T} \) (i.e., future rewards)
  
  – given the history \( H_t = \{(S_0, A_0, Y_0), \ldots, (S_{t-1}, A_{t-1}, Y_{t-1}), S_t\} \)
    of states \( S_t \), actions and outcomes

• **Given:** samples of past histories (no exploration possible)

• **Algorithms:** e.g., deep Q-learning
Example: Managing sepsis in the ICU

$S_t$: Heart rate, blood oxygenation, etc.

$A_t$: Mechanical ventilation? Sedation? Vasopressors?

$Y_t$: Observed (e.g., patient dies)

Unobserved

Off-policy RL has to be done with care\(^1\)

- In performing and evaluating observational studies of sequential decision making, we must ask:
  1. Do we have access to the information currently used in decision making?
  2. Are we optimizing the right reward/outcome?
  3. Is our data large enough to compare our proposed policy to existing ones?

\(^1\)Guidelines for reinforcement learning in healthcare. Gottesman, O; Johansson, F; Komorowski, M; Faisal, A; Sontag, D; Doshi-Velez, F; and Celi, L. *Nature Medicine*, 25(1): 16–18. 2019
Off-policy RL guidelines: confounding

1. Do we have access to the information used by doctors in making this choice?

If not, our estimate will likely be **confounded**
Off-policy RL guidelines: outcome label

2. What **reward** are we optimizing? Does it capture long-term effects?
Off-policy RL guidelines: sample size

• Standard to make use only of patient trajectories that agree with the proposed policy—small effective sample size

3. How large is the effective sample size?
Opportunities in Machine Learning for Healthcare

Marzyeh Ghassemi
Massachusetts Institute of Technology, Verily
Cambridge, MA 02139
mghassem@mit.edu, marzyeh@google.com

Tristan Naumann
Massachusetts Institute of Technology
Cambridge, MA 02139
tjn@mit.edu

Peter Schulam
Johns Hopkins University
Baltimore, MD 21218
pschulam@cs.jhu.edu

Andrew L. Beam
Harvard Medical School
Boston, MA 02115
andrew Beam@hms.harvard.edu

Rajesh Ranganath
New York University
New York, NY 10011
rajesh@cims.nyu.edu

Abstract

Healthcare is a natural arena for the application of machine learning, especially as modern electronic health records (EHRs) provide increasingly large amounts of data to answer clinically meaningful questions. However, clinical data and practice present unique challenges that complicate the use of common methodologies. This article serves as a primer on addressing these challenges and highlights opportunities for members of the machine learning and data science communities to contribute to this growing domain.
And that’s a wrap!

• Thanks for a great two days
• Keep in touch:

E-mail: dsontag@csail.mit.edu
Twitter: david_sontag
LinkedIn: https://www.linkedin.com/in/david-sontag/
Readings
References for risk stratification

- Predicting the Risk and Trajectory of Intensive Care Patients Using Survival Models Caleb Hug, Master's thesis at MIT, 2006 [https://dspace.mit.edu/handle/1721.1/38326]
- Passive Detection of Atrial Fibrillation Using a Commercially Available Smartwatch Tison et al., JAMA Cardiology 2018 [https://jamanetwork.com/journals/jamacardiology/article-abstract/2675364]
- Scalable and accurate deep learning with electronic health records Rajkomar et al., Nature Digital Medicine, 2018 [https://www.nature.com/articles/s41746-018-0029-1]
  Supplementary: [https://static-content.springer.com/esm/art%3A10.1038%2Fs41746-018-0029-1/MediaObjects/41746_2018_29_MOESM1_ESM.pdf]
References for causal inference

• Miguel Hernan’s causal inference book

• Paul Rosenbaum’s *Design of Observational Studies*

• High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Schneeweiss et al., Epidemiology 2009
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3077219/

• Estimation and Inference of Heterogeneous Treatment Effects using Random Forests
  Stefan Wager, Susan Athey, JASA ‘18
  https://arxiv.org/abs/1510.04342

• Estimating individual treatment effect: generalization bounds and algorithms

• Postsurgical prescriptions for opioid naive patients and association with overdose and misuse: retrospective cohort study
  Gabriel Brat et al., BMJ 2017
  https://www.bmj.com/content/360/bmj.j5790
References for time-series

- Factorial Switching Linear Dynamical Systems applied to Physiological Condition Monitoring
  Quinn et al., TPAMI 2008
  https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4586385

- Integration of Early Physiological Responses Predicts Later Illness Severity in Preterm Infants
  Saria et al., Science Translational Medicine 2010
  http://stm.sciencemag.org/content/2/48/48ra65

- Clifford et al. AF Classification from a Short Single Lead ECG Recording: the PhysioNet/Computing in Cardiology Challenge, Computing in Cardiology 2017
  https://www.physionet.org/challenge/2017/

- Abductive reasoning as the basis to reproduce expert criteria in ECG Atrial Fibrillation identification
  Teijeiro, Garcia, Castro, Felix. arXiv:1802.05998, 2018
  https://arxiv.org/abs/1802.05998

- Cardiologist-Level Arrhythmia Detection With Convolutional Neural Networks

- Modeling Disease Progression via Fused Sparse Group Lasso
  Zhou et al., KDD ’12
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4191837/
References for causal inference

• Personalized Diabetes Management Using Electronic Medical Records
  Dimitris Bertsimas, Nathan Kallus, Alexander M. Weinstein, and Ying Daisy Zhuo
  Diabetes Care, 2016
  http://care.diabetesjournals.org/content/early/2016/12/01/dc16-0826.full-text.pdf

• Medical Homes and Cost and Utilization Among High-Risk Patients
  Susannah Higgins; Ravi Chawla; Christine Colombo; Richard Snyder; and Somesh Nigam
  American Journal of Managed Care, 2014
References for interpretability

- Implications of non-stationarity on predictive modeling using EHRs
  Kenneth Jung, Nigam Shah. JBI 2015

- Intriguing properties of neural networks
  Szegedy et al. 2014
  https://arxiv.org/abs/1312.6199

- “Why Should I Trust You?” Explaining the Predictions of Any Classifier
  Ribeiro et al., KDD ’16

- Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission
  Caruana et al., KDD 2015

- The Mythos of Model Interpretability
  Zachary C. Lipton
  https://arxiv.org/abs/1606.03490
References for disease subtyping

- Phenomapping for Novel Classification of Heart Failure with Preserved Ejection Fraction
  Shah et al., Circulation 2015

- Subtyping: What It Is and Its Role in Precision Medicine
  Saria & Goldberg, IEEE Intelligent Systems 2015

- Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis

- Cluster Analysis and Clinical Asthma Phenotypes
  Haldar et al., Am J Respir Crit Care Med. 2008.
References on MLHC policy

• **Real-World Evidence In Support Of Precision Medicine: Clinico-Genomic Cancer Data As A Case Study**
  Vineeta Agarwala, Sean Khozin, Gaurav Singal, Claire O’Connell, Deborah Kuk, Gerald Li, Anala Gossai, Vincent Miller, and Amy P. Abernethy
  Health Affairs, 2018

• **FDA Regulation of Mobile Medical Apps**
  Shuren, Patel, Gottlieb. JAMA, 2018
  [https://jamanetwork.com/journals/jama/fullarticle/2687221](https://jamanetwork.com/journals/jama/fullarticle/2687221)

• **How Tech Can Turn Doctors Into Clerical Workers: The Threat That Electronic Health Records and Machine Learning Pose to Physicians' Clinical Judgement -- and their Well-Being**
  Abraham Verghese

• **Predictive modeling of U.S. health care spending in late life**
  Einav et al., Science 2018
  [http://science.sciencemag.org/content/360/6396/1462](http://science.sciencemag.org/content/360/6396/1462)

• **Hacking Healthcare: A Guide to Standards, Workflows, and Meaningful Use**
  Trotter & Uhlman. O'Reilly Media, 2011

• **Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients**

• **Predicting the Future — Big Data, Machine Learning, and Clinical Medicine**
  Ziad Obermeyer, M.D., and Ezekiel J. Emanuel, M.D., Ph.D.
References for fairness

• Why is My Classifier Discriminatory?
  Irene Chen, Fredrik Johansson, David Sontag
  NeurIPS 2018