Learning "Healthy" Models for Healthcare

Marzyeh Ghassemi, PhD
University of Toronto, CS/Med
Vector Institute
Why Try To Work in Health?

• Improvements in health **improve lives**.

• Same **patient** $\rightarrow$ different **treatments** across hospitals, clinicians.

• Improving care requires **evidence**.
Why Try To Work in Health?

• Improvements in health **improve lives**.

• Same **patient** → **different treatments** across hospitals, clinicians.

• Improving care requires **evidence**.

What does it mean **mean** to be **healthy**?
Learning What Is Healthy?

Recruit a study population.
Learning What Is Healthy?

Learn a rule.
Learning What Is Healthy?

Does it generalize?
Learning What Is Healthy?

For whom does it generalize?
Evidence in Healthcare and Health?

Randomized Controlled Trials (RCTs) are rarely and expensive, and can encode structural biases that apply to very few people.
Evidence in Healthcare and Health?

Randomized Controlled Trials (RCTs) are rare and expensive

10 – 20% of Treatments are based on Randomized Controlled Trials (RCTs)

Evidence in Healthcare and Health?

Randomized Controlled Trials (RCTs) are rare and expensive, and can encode structural biases that apply to very few people.

- 10 – 20% of Treatments are based on Randomized Controlled Trials (RCTs)
- 6% of Asthmatics Would Have Been Eligible for Their Own Treatment RCTs.

Machine Learning What Is Healthy?

Can we use **data** to **learn** what is **healthy**?
Extracting Knowledge is Hard in Health

• Data are **not gathered** to answer your hypothesis.

• **Primary** case is to provide **care**.

• Secondary data are **hard** to work with.

<table>
<thead>
<tr>
<th>Heterogenous</th>
<th>Sparse</th>
<th>Uncertainty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling</td>
<td>Unmeasured</td>
<td>Labels</td>
</tr>
<tr>
<td>Data Type</td>
<td>Unreported</td>
<td>Bias</td>
</tr>
<tr>
<td>Time Scale</td>
<td>No Follow-up</td>
<td>Context</td>
</tr>
</tbody>
</table>
Lack of Expertise Is Challenging

- Media can create unrealistic expectations.
Be Careful What You Optimize For

- ML can be confidently wrong.\(^1,\ 2\)

\[\text{\begin{tabular}{c}
\text{king penguin} & \text{starfish} & \text{freight car} & \text{remote control}
\end{tabular}\] or

\[\text{\begin{tabular}{c}
\text{AllConv} & \text{NiN} & \text{VGG}
\end{tabular}\] or

\[\begin{array}{c}
\text{SHIP} \quad \text{CAR}(99.7\%) \\
\text{HORSE} \quad \text{FROG}(99.9\%) \\
\text{DEER} \quad \text{AIRPLANE}(85.3\%)
\end{array}\]


Natural Born Expertise Makes This Easier

• Humans are “natural” experts in NLP, ASR, Vision evaluation.¹

(a) Husky classified as wolf  
(b) Explanation

How Do We Know When We’re Wrong?

- Hyper-expertise makes attacks in clinical data harder to spot.¹

Healthy Models Require Domain Knowledge

- Learning without understanding is dangerous.¹

“…aggressive care received by asthmatic pneumonia patients (in the training set) was so effective that it lowered their risk of dying from pneumonia compared to the general population…”

“HasAsthma(x) ⇒ LowerRisk(x)”

Machine Learning For Health (ML4H)

1. **What Models are Healthy? Learning Good Representations.**
   - Unfolding Physiological State: Mortality Modelling in Intensive Care Unit (KDD 2014); A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU … (AAAI 2015);
   - Predicting Early Psychiatric Readmission with Natural Language Processing of Narrative … (Nature Trans Psych 2016);
   - Predicting Intervention Onset in the ICU with Switching State Space Models (AMIA-CRI 2017);
   - Clinical Intervention Prediction and Understanding using Deep Networks (MLHC 2017/JMLR W&C V68);
   - Semi-supervised Biomedical Translation with Cycle Wasserstein Regression GANs (AAAI 2018);

2. **What Healthcare is Healthy? Stratifying Human Risks.**
   - Continuous State-Space Models for Optimal Sepsis Treatment - Deep Reinforcement Learning … (MLHC/JMLR 2017);
   - Modeling Mistrust in End-of-Life Care (MLHC 2018/FATML 2018 Workshop);
   - The Disparate Impacts of Medical and Mental Health with AI. (In submission);

3. **What Behaviors are Healthy? Inferring Unseen Actions and States.**
   - Learning to Detect Vocal Hyperfunction from Ambulatory Necksurface Acceleration Features (IEEE TBME 2014);
   - Uncovering Voice Misuse Using Symbolic Mismatch (MLHC 2016/JMLR W&C V56);
   - Project BASELINE Mood Study with Alphabet’s Verily;
   - ClinicalVis Project with Google Brain. (*In submission);
1. **What Models are Healthy? Learning Good Representations.**
   Unfolding Physiological State: Mortality Modelling in Intensive Care Unit (KDD 2014); A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU … (AAAI 2015); Predicting Early Psychiatric Readmission with Natural Language Processing of Narrative … (Nature Trans Psych 2016); Predicting Intervention Onset in the ICU with Switching State Space Models (AMIA-CRI 2017); Clinical Intervention Prediction and Understanding using Deep Networks (MLHC 2017/JMLR W&C V68); Semi-supervised Biomedical Translation with Cycle Wasserstein Regression GANs (AAAI 2018);

2. **What Healthcare is Healthy? Stratifying Human Risks.**
   Continuous State-Space Models for Optimal Sepsis Treatment - Deep Reinforcement Learning … (MLHC/JMLR 2017); Modeling Mistrust in End-of-Life Care (MLHC 2018/FATML 2018 Workshop); The Disparate Impacts of Medical and Mental Health with AI. (In submission);

3. **What Behaviors are Healthy? Inferring Unseen Actions and States.**
   Learning to Detect Vocal Hyperfunction from Ambulatory Necksurface Acceleration Features (IEEE TBME 2014); Uncovering Voice Misuse Using Symbolic Mismatch (MLHC 2016/JMLR W&C V56); Project BASELINE Mood Study with Alphabet’s Verily; ClinicalVis Project with Google Brain. (“In submission);
MIMIC III ICU Data

- Learning with real patient data from the Beth Israel Deaconess Medical Center ICU.¹

Problem: Hospital Decision-Making / Care Planning

**Observe** Patient Data

“Real-time” **Prediction**

Of \{Drug/Mortality/Condition\}

By Gap Time

**Before** the Doctor Acted
Part 1: Predict Mortality With Clinical Notes

- **Acuity** (severity of illness) very important - use **mortality** as a proxy for **acuity**.¹

- Prior state-of-the-art focused on feature engineering in **labs/vitals** for target populations.²

- But **clinicians** rely on **notes**.

---


² Grady, Deborah, and Seth A. Berkowitz. "Why is a good clinical prediction rule so hard to find?." *Archives of internal medicine* 171.19 (2011): 1701-1702.
Patient Y, 12:45:00 EST

uneventful day. pt much improved. VS Stable nuero intact no compromise NSR BP stable Aline discontinued in afternoon. pt to transfer to floor awaiting bed. pt continues with nausea given anziment and started on reglan prn. small emesis in am. pt continues with ice chips. foley draining well adequate output. now replacing half cc for cc of urine. skin and surgical site unchanged, C/D/I. family (son and husband) at bedside for most of day. Plan: continue with current plan in progress, transfer to floor.
Represent Patients as Topic Vectors

• Model patient stays as an **aggregated set** of notes.

• Model notes as a **distribution** over topics.

• A “topic” is a **distribution** over words, that we learn.

• Use Latent Dirichlet Allocation (LDA)\(^1\) as an **unsupervised** way to **abstract** 473,000 notes from 19,000 patients into “topics”.\(^2\)

---


Correlation Between Average Topic Representation and Mortality

<table>
<thead>
<tr>
<th>Topic #</th>
<th>Top Ten Words</th>
<th>Possible Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cabg, pain, ct, artery, coronary, valve, post, wires, chest, sp</td>
<td>Cardiovascular surgery</td>
</tr>
<tr>
<td>15</td>
<td>intubated vent ett secretions propofol abg respiratory resp care sedated</td>
<td>Respiratory failure</td>
</tr>
</tbody>
</table>
Topic Representation Improves In-Hospital Mortality Prediction

• **First** to do forward-facing ICU mortality prediction with notes.

• **Latent** representations **add** predictive power.

• Topics enable accurately **assess risk** from notes.
More Complex Models Haven’t Done Better

<table>
<thead>
<tr>
<th>Author</th>
<th>AUC</th>
<th>Method</th>
<th>Episodes</th>
<th>Hours</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ghassemi, 2014</td>
<td>0.84/0.85</td>
<td>LDA</td>
<td>19,308</td>
<td>24/48</td>
<td>53 - notes</td>
</tr>
<tr>
<td>Caballero, 2015</td>
<td>0.86</td>
<td>Text processing + medication</td>
<td>15,000</td>
<td>24</td>
<td>? - notes/meds</td>
</tr>
<tr>
<td>Che, 2015</td>
<td>0.8-0.82</td>
<td>Deep Learning (LSTM)</td>
<td>3,940</td>
<td>48</td>
<td>30 - vitals</td>
</tr>
<tr>
<td>Che, 2016</td>
<td>0.7/0.85</td>
<td>Deep Learning (GRU)</td>
<td>19,714</td>
<td>12/48</td>
<td>99 – vitals/meds</td>
</tr>
<tr>
<td>Che, 2018</td>
<td>0.85</td>
<td>Deep Learning (GRU-D)</td>
<td>19,714</td>
<td>48</td>
<td>99 – vitals/meds</td>
</tr>
</tbody>
</table>

Even When Complex and Clever

- Explicitly capture and use missing patterns in RNNs via systematically modified architectures.

- Performance bump is small, is MIMIC mortality our MNIST?

<table>
<thead>
<tr>
<th>Non-RNN Models</th>
<th>RNN Models</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mortality Prediction On MIMIC-III Dataset</strong></td>
<td><strong>LSTM-Mean</strong> 0.8142 ± 0.014</td>
</tr>
<tr>
<td>LR-Mean 0.7589 ± 0.015</td>
<td>SVM-Mean 0.7908 ± 0.006</td>
</tr>
<tr>
<td>LR-Forward 0.7792 ± 0.018</td>
<td>SVM-Forward 0.8010 ± 0.004</td>
</tr>
<tr>
<td>LR-Simple 0.7715 ± 0.015</td>
<td>SVM-Simple 0.8145 ± 0.008</td>
</tr>
<tr>
<td>LR-SoftImpute 0.7598 ± 0.017</td>
<td>SVM-SoftImpute 0.7540 ± 0.012</td>
</tr>
<tr>
<td>LR-KNN 0.6877 ± 0.011</td>
<td>SVM-KNN 0.7200 ± 0.004</td>
</tr>
<tr>
<td>LR-CubicSpline 0.7270 ± 0.005</td>
<td>SVM-CubicSpline 0.6376 ± 0.018</td>
</tr>
<tr>
<td>LR-MICE 0.6965 ± 0.019</td>
<td>SVM-MICE 0.7169 ± 0.012</td>
</tr>
<tr>
<td>LR-MF 0.7158 ± 0.018</td>
<td>SVM-MF 0.7266 ± 0.017</td>
</tr>
<tr>
<td>LR-PCA 0.7246 ± 0.014</td>
<td>SVM-PCA 0.7235 ± 0.012</td>
</tr>
<tr>
<td>LR-MissForest 0.7279 ± 0.016</td>
<td>SVM-MissForest 0.7482 ± 0.016</td>
</tr>
<tr>
<td><strong>Proposed GRU-D</strong> 0.8527 ± 0.003</td>
<td></td>
</tr>
</tbody>
</table>
Part 2: Predict **Interventions** With Clinical **Data**

- 34,148 ICU patients from MIMIC-III
- 5 static variables (gender, age, etc.)
- 29 time-varying vitals and labs (oxygen saturation, lactate, etc.)
- All clinical notes for each patient stay
Many Ways to Model, What Do We Learn?

SSAM

Learn model parameters over patients with variational EM.

Infer hourly distribution over hidden states with HMM DP (fwd alg.).

Logistic regression (with label-balanced cost function)

Predict onset in advance

LSTM

2 Layer/512 node LSTM with sequential hourly data; at end of window, use the final hidden state to predict output.

CNN

CNN for temporal convolutions at 3/4/5 hours, max-pool, combine the outputs, and run through 2 fully connected layers for prediction.
State Space Beliefs Improve Prediction

- **Vasopressor @ t+1 hr**
  - AUC values for different conditions:
    - s: 0.66
    - x: 0.77
    - s+x: 0.79
    - b: 0.66
    - b+s+x: 0.82

- **Vasopressor @ t+4 hr**
  - AUC values for different conditions:
    - s: 0.66
    - x: 0.70
    - s+x: 0.74
    - b: 0.64
    - b+s+x: 0.78

- **Ventilator @ t+4 hr**
  - AUC values for different conditions:
    - s: 0.61
    - x: 0.62
    - s+x: 0.65
    - b: 0.56
    - b+s+x: 0.67

- **Plasma @ t+4 hr**
  - AUC values for different conditions:
    - s: 0.55
    - x: 0.71
    - s+x: 0.72
    - b: 0.61
    - b+s+x: 0.75

**Additional Notes**
- Static demographics
- Dynamic patient vitals @ t
- SSAM belief vector (10D) @ t
SSAM Post-hoc Interpretability

- Interpret classifier weights across interventions.

- Investigate data associated with vasopressor onset state (9).

![Average Emission Values](image-url)
NNs Do Well; Improved Representation Helps

<table>
<thead>
<tr>
<th>Task</th>
<th>Model</th>
<th>VENT</th>
<th>NI-VENT</th>
<th>VASO</th>
<th>COL BOL</th>
<th>CRYS BOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Onset AUC</td>
<td>Baseline</td>
<td>0.60</td>
<td>0.66</td>
<td>0.43</td>
<td>0.65</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>LSTM Raw</td>
<td>0.61</td>
<td>0.75</td>
<td>0.77</td>
<td>0.52</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>LSTM Words</td>
<td>0.75</td>
<td>0.76</td>
<td>0.76</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>0.62</td>
<td>0.73</td>
<td>0.77</td>
<td>0.70</td>
<td>0.69</td>
</tr>
<tr>
<td>Wean AUC</td>
<td>Baseline</td>
<td>0.83</td>
<td>0.71</td>
<td>0.74</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>LSTM Raw</td>
<td>0.90</td>
<td>0.80</td>
<td>0.91</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>LSTM Words</td>
<td>0.90</td>
<td>0.81</td>
<td>0.91</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>0.91</td>
<td>0.80</td>
<td>0.91</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Stay On</td>
<td>Baseline</td>
<td>0.50</td>
<td>0.79</td>
<td>0.55</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Stay Off</td>
<td>Baseline</td>
<td>0.94</td>
<td>0.71</td>
<td>0.93</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Stay AUC</td>
<td>LSTM Raw</td>
<td>0.95</td>
<td>0.86</td>
<td>0.96</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>LSTM Words</td>
<td>0.97</td>
<td>0.86</td>
<td>0.95</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>0.96</td>
<td>0.86</td>
<td>0.96</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Macro AUC</td>
<td>Baseline</td>
<td>0.72</td>
<td>0.72</td>
<td>0.66</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>LSTM Raw</td>
<td>0.86</td>
<td>0.82</td>
<td>0.90</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>LSTM Words</td>
<td>0.90</td>
<td>0.82</td>
<td>0.89</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>0.86</td>
<td>0.81</td>
<td>0.90</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Representations with “physiological words” for missingness significantly increased AUC for interventions with the lowest proportion of examples.

Deep models perform well in general, but words are important for ventilation tasks.
NN Post-hoc Interpretability

- Feature-level occlusions identify important per-class features.

- Convolutional filters target known short-term trajectories.

Physiological data were more important for the more invasive interventions.

Higher diastolic blood pressure, respiratory rate, and heart rate, and lower oxygen saturation: **Hyperventilation**

Decreased systolic blood pressure, heart rate and oxygen saturation rate: **Altered peripheral perfusion or stress hyperglycemia**

Decreased creatinine, phosphate, oxygen saturation and blood urea nitrogen: **Neuromuscular respiratory failure**
ML for Healthcare, or ML for Health?

- Patients can be left on interventions longer than necessary.

![Histogram of Extra Intervention Time](image)

- Extended interventions can be costly and detrimental to patient health.¹,²

---


Finding Where We “Could” Wean Early?

- One example of a 62-year-old male patient with a cardiac catheterization.

- More complexity/higher misclassification penalty don’t solve this!
Part 3: Forecast **Response to An Intervention**

- Fully paired biomedical datasets are
  - Privacy sensitive
  - Expensive and difficult to collect
  - Often homogenous

- Sufficiently large, heterogeneous paired datasets are rare.
Using Adversarial Training To Overcome Missingness

- GANs are used for data augmentation\(^1\), imputation\(^2\).

- We use adversarial learning techniques to learn distributional signals from additional, unpaired data to augment predictions on a limited training set.

---


Model Learns on Unpaired Data, $G_X$ Used to Eval

- Ensure generated samples are realistic, account for missing samples (not just missing features), and ensure cycle/self-consistency.¹

![Diagram](image)

# Improved Intervention Response Prediction

<table>
<thead>
<tr>
<th>Intervention Type</th>
<th>VENT</th>
<th>NOREP</th>
<th>DOP</th>
<th>PHEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model MSE</td>
<td>3.780</td>
<td>2.829</td>
<td>2.719</td>
<td>3.186</td>
</tr>
<tr>
<td>Baseline MLP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CWR-GAN (% Delta)</td>
<td>-0.5%</td>
<td>-7.4%</td>
<td>+2.7%</td>
<td>-4.5%</td>
</tr>
</tbody>
</table>

- Mean-squared-error of a traditional MLP on only paired intervention data vs. the CWR-GAN augmented with data that failed to meet inclusion criteria on either the pre-intervention side or post-intervention side (~500 paired, ~3,000 unpaired patients).
The Problem With Models That Learn...

- Exciting work on to be done on learning what treatments are best for individuals based on environment and context!

- But there are other factors...
Health Questions Beyond The Obvious

- Across these use cases, a number of ethical, social, and political challenges are raised and the 10 most important are:

  01 What effect will AI have on human relationships in health and care?

  02 How is the use, storage and sharing of medical data impacted by AI?

  03 What are the implications of issues around algorithmic transparency/explainability on health?

  04 Will these technologies help eradicate or exacerbate existing health inequalities?

  05 What is the difference between an algorithmic decision and a human decision?

  06 What do patients and members of the public want from AI and related technologies?

  07 How should these technologies be regulated?

  08 Just because these technologies could enable access to new information, should we always use it?

  09 What makes algorithms, and the entities that create them, trustworthy?

  10 What are the implications of collaboration between public and private sector organisations in the development of these tools?
1. What Models are Healthy? Learning Good Representations.
   Unfolding Physiological State: Mortality Modelling in Intensive Care Unit (KDD 2014);  
   A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU … (AAAI 2015);  
   Predicting Early Psychiatric Readmission with Natural Language Processing of Narrative … (Nature Trans Psych 2016);  
   Predicting Intervention Onset in the ICU with Switching State Space Models (AMIA-CRI 2017);  
   Clinical Intervention Prediction and Understanding using Deep Networks (MLHC 2017/JMLR W&C V68);  
   Semi-supervised Biomedical Translation with Cycle Wasserstein Regression GANs (AAAI 2018);

   Continuous State-Space Models for Optimal Sepsis Treatment - Deep Reinforcement Learning … (MLHC/JMLR 2017);  
   Modeling Mistrust in End-of-Life Care (MLHC 2018/FATML 2018 Workshop);  
   The Disparate Impacts of Medical and Mental Health with AI. (In submission);  
   ClinicalVis Project with Google Brain. (*In submission);

3. What Behaviors are Healthy? Inferring Unseen Actions and States.
   Learning to Detect Vocal Hyperfunction from Ambulatory Necksurface Acceleration Features (IEEE TBME 2014);  
   Uncovering Voice Misuse Using Symbolic Mismatch (MLHC 2016/JMLR W&C V56);  
   Project BASELINE Mood Study with Alphabet’s Verily;
Modelling Mistrust in EOL Care

- Replicate documented racial disparities in open databases.

(a) MIMIC Mechanical Ventilation
White: 4810 patients
Black: 510 patients
p = 0.005

(b) eICU Mechanical Ventilation
White: 4911 patients
Black: 655 patients
p < 0.001

- Algorithmically mistrust demonstrates treatment disparity > than race, even with acuity factored in.

<table>
<thead>
<tr>
<th></th>
<th>OASIS</th>
<th>SAPS II</th>
<th>Noncompliance</th>
<th>Autopsy</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>OASIS</td>
<td>1.0</td>
<td>0.679</td>
<td>0.050</td>
<td>-0.012</td>
<td>0.075</td>
</tr>
<tr>
<td>SAPS II</td>
<td>0.679</td>
<td>1.0</td>
<td>0.013</td>
<td>-0.013</td>
<td>0.086</td>
</tr>
<tr>
<td>Noncompliance</td>
<td>0.050</td>
<td>0.013</td>
<td>1.0</td>
<td>0.262</td>
<td>0.058</td>
</tr>
<tr>
<td>Autopsy</td>
<td>-0.012</td>
<td>-0.013</td>
<td>0.262</td>
<td>1.0</td>
<td>0.044</td>
</tr>
<tr>
<td>Sentiment</td>
<td>0.075</td>
<td>0.086</td>
<td>0.058</td>
<td>0.044</td>
<td>1.0</td>
</tr>
</tbody>
</table>
We can predict ICU mortality and 30-day psychiatric readmission, but notes have group-specific heterogeneity.

Significant differences in model accuracy for race, sex, and insurance type in ICU notes and insurance type in psychiatric notes.
ClinicalVis: Supporting Clinical Task-Focused Design Evaluation

1. Present real patient data to HCPs using open-source prototype.

2. Ask HCPs to plan care for two interventions in an eICU simulation.

3. Evaluate the confidence, accuracy and time-to-task under different visual prototypes.

<table>
<thead>
<tr>
<th></th>
<th>Vasopressor Positive (VP+)</th>
<th>Ventilator Positive (VE+)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy (%)</strong></td>
<td>Baseline 50.00 %</td>
<td>56.25 %</td>
</tr>
<tr>
<td></td>
<td>ClinicalVis 68.83 %</td>
<td>62.79 %</td>
</tr>
<tr>
<td><strong>Confidence Score</strong></td>
<td>Baseline 0.68</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>ClinicalVis 1.41</td>
<td>1.27</td>
</tr>
<tr>
<td><strong>Average Time to Task (seconds)</strong></td>
<td>Baseline 92.31 s</td>
<td>92.73 s</td>
</tr>
<tr>
<td></td>
<td>ClinicalVis 84.43 s</td>
<td>86.86 s</td>
</tr>
</tbody>
</table>
Future of Machine Learning For Health (ML4H)

1. What Models are Healthy? Learning Good Representations.
   - Balancing multi-target output learning
   - Finding useful abstractions
   - “Explaining” decisions in case/controls

   - Providing meaningful, calibrated notions of uncertainty
   - Finding causes and establishing causality
   - Defining and targeting fairness

3. What Behaviors are Healthy? Inferring Unseen Actions and States.
   - Data quality and availability
   - Real-time decision making
   - Robustness in the face of unexpected data
**ML4H - What Can You Do?**

**Opportunities in Machine Learning for Healthcare**

- **Maryam Ghavamzadeh**
  Massachusetts Institute of Technology, Verizon
  Cambridge, MA 02139
  mg3waan@mit.edu, mail@mg3waan.com

- **Tribina Namjou**
  Massachusetts Institute of Technology
  Cambridge, MA 02139
  tj@mit.edu

- **Peter Schmidt**
  Johns Hopkins University
  Baltimore, MD 21218
  pschmidt@jhu.edu

- **Andrew L. Beam**
  Harvard Medical School
  Boston, MA 02115
  andrewbeam@hms.harvard.edu

- **Rajesh Ranganath**
  New York University
  New York, NY 10011
  rajesh@cs.nyu.edu

**Abstract**

Healthcare is a natural area 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.

---

**Current Opportunities in Healthcare**

- **CAUSALITY**

- **MISSINGNESS**

- **PATIENT**

- **HIGH FREQ. SIGNALS**

- **LABS**

- **NOTES**

---

**High Freq. Signals**
Bedside equipment record real time patient data.

**Labs**
Vitals and labs measure biomarkers of patient state.

**Notes**
Notes record the interaction between patient and a healthcare team.

**Causality**
Difficult to establish causality from observational data.

**Missingness**
Data are often missing in ways that break common ML assumptions.

**Defining Outcomes**
Identifying the right problems requires multiple data sources and domain expertise.

**Automation**
Routine processes can be automated; e.g. triage.

**Support & Augmentation**
Standardizing processes and integrating fragmented records.

**Expanding Clinical Capacities**

**Practicing Non-Stationarity**
Models that can continue learning as new data are available.

**Interpretable Models**
Machine learning systems facilitate interaction and collaboration with human experts.

**Representations in Networks**
Integrating predictions from multi-source high dimensional data.
Technical Challenges!

ML4H - What Can You Do?
ML4H - What Can You Do?

Technical Challenges!

Health Opportunities!
ML4H - What Can You Do?

Technical Challenges!
Health Opportunities!
ML Work Needed!
What models are healthy?

What healthcare is healthy?

What behaviors are healthy?