Learning Healthy Models for Healthcare

Dr. Marzyeh Ghassemi, PhD MIT CSAIL
University of Toronto, Computer Science and Medicine
Canadian CIFAR AI Chair, Vector Institute
NeuroIPS Workshop Co-Chair

www.marzyehghassemi.com, @MarzyehGhassemi
Why Try To Work in Health?

• Improvements in health improve lives.

• Same patient → 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 rare 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.


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.

![Diagram]

- **Heterogenous**
  - Sampling
  - Data Type
  - Time Scale

- **Sparse**
  - Unmeasured
  - Unreported
  - No Follow-up

- **Uncertainty**
  - Labels
  - Bias
  - Context
Lack of Expertise Is Challenging

- Media can create unrealistic expectations.

+ ≠ Insight
Be Careful What You Optimize For

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

---


Natural Born Expertise Makes This Easier

- Humans are “natural” experts in NLP, ASR, Vision evaluation.\(^1\)

---

How Do We Know When We’re Wrong?

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

Learning Unintended Features

- CNN models can determine the hospital that the patient was admitted to with 95% accuracy... from the X-ray.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)”

What models are healthy?
What healthcare is healthy?
What behaviors are healthy?

Machine Learning For Health (ML4H)
Machine Learning For Health (ML4H)

What models are healthy?
What healthcare is healthy?
What behaviors are healthy?
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\(^1,2,3,4,5,6\)

---

\[\begin{align*}
1 & \text{ Unfolding Physiological State: Mortality Modelling in Intensive Care Unit (KDD 2014); A Multivariate Timeseries} \\
2 & \text{ Modeling Approach to Severity of Illness Assessment and Forecasting in ICU … (AAAI 2015);} \\
3 & \text{ Predicting Early Psychiatric Readmission with Natural Language Processing of Narrative … (Nature Trans Psych 2016);} \\
4 & \text{ Predicting Intervention Onset in the ICU with Switching State Space Models (AMIA-CRI 2017);} \\
5 & \text{ Clinical Intervention Prediction and Understanding using Deep Networks (MLHC 2017/JMLR W&C V68);} \\
6 & \text{ Semi-supervised Biomedical Translation with Cycle Wasserstein Regression GANs (AAAI 2018);} \\
\end{align*}\]
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>15</td>
<td>intubated vent ett secretions propofol abg respiratory resp care sedated</td>
<td>Respiratory failure</td>
</tr>
</tbody>
</table>

<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>
</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.
But Wait! 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?
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

![Graph showing AUC values for different variables.]

- **Static demographics**
- **Dynamic patient vitals @ t**
- **SSAM belief vector (10D) @ t**

### Variables

- **Vasopressor @ t+1 hr**
  - AUC values: 0.66, 0.77, 0.79, 0.66, 0.82 for s, x, s+x, b, b+s+x respectively.

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

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

- **Plasma @ t+4 hr**
  - AUC values: 0.55, 0.71, 0.72, 0.61, 0.75 for s, x, s+x, b, b+s+x respectively.
SSAM Post-hoc Interpretability

- Interpret classifier weights across interventions.

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

![Graphs showing weight distribution across states for vasopressor, ventilator, and ffp transfusion.]

![Average Emission Values chart showing mean values for various measurements.]
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 AUC</td>
<td>Baseline</td>
<td>0.50</td>
<td>0.79</td>
<td>0.55</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>LSTM Raw</td>
<td>0.96</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>Stay Off AUC</td>
<td>Baseline</td>
<td>0.94</td>
<td>0.71</td>
<td>0.93</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></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>0.95</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>
</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**
From Healthcare to Health

• Patients can be left on interventions longer than necessary.

• 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.
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

- Generated samples are realistic
- Account for missing samples (not just missing features)
- Ensure cycle/self-consistency\(^1\)

- Improved intervention response prediction
  - MLP MSE on paired intervention data
  - CWR-GAN augmented with data fails inclusion either pre-/post-intervention
  - \(~500\) paired, \(~3,000\) unpaired patients

<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>

---

Deploy Good Models To Forecast Response?

• 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?

What models are healthy?

What healthcare is healthy?

What behaviors are healthy?
Robust ML is Crucial in Healthcare

- Machine learning in healthcare requires robustness.
  - Technical replicability
  - Statistical replicability
  - Conceptual replicability

Evaluation Metrics:
A. Technical replicability
   1. Code available
   2. Public dataset
B. Statistical replicability
   1. Variance reported
C. Conceptual replicability
   1. Multiple datasets

[1] Reproducibility in Machine Learning for Health; ICLR Reproducibility Workshop 2018 (under review); Matthew B. A. McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Marzyeh Ghassemi, Luca Foschini
Large Data Is Not The Failsafe You Imagine

In a combined EHR/claims dataset from 11 sources/4 countries/250 million patients, how many followed a unique treatment pathway?

- Diabetes:
- Depression:
- Hypertension:

Large Data Is Not The Failsafe You Imagine

In a combined EHR/claims dataset from 11 sources/4 countries/250 million patients, how many followed a unique treatment pathway?

- **Diabetes:** 10% of patients
- **Depression:**
- **Hypertension:**

Large Data Is Not The Failsafe You Imagine

In a combined EHR/claims dataset from 11 sources/4 countries/250 million patients, how many followed a unique treatment pathway?

- Diabetes: 10% of patients
- Depression: 11% of patients
- Hypertension:

Large Data Is Not The Failsafe You Imagine

In a combined EHR/claims dataset from 11 sources/4 countries/250 million patients, how many followed a unique treatment pathway?

- Diabetes: 10% of patients
- Depression: 11% of patients
- Hypertension: 24% of patients

“In an underlying population of 250 million, based on my 3-y treatment pathway, what patients are like me?”

Large Data Is Not The Failsafe You Imagine

“In an underlying population of 250 million, based on my 3-y treatment pathway, what patients are like me?”

For 24% of hypertension patients, “No one.”

Bias Is Part of the Clinical Landscape Already

- How does/should ML interact with fairness/health$^{1,2,3,4,5}$?
How Can We Improve Health Care For All?

- Patient populations have high variance.

- Differences in treatment by race, sex, and socioeconomic status

  treatment A or B

- Calculating differences in prediction accuracy by group can give us additional insight
We can predict ICU mortality and 30-day psychiatric readmission, but notes have group-specific heterogeneity.
Unfair Accuracies in Medical and Mental Health

- Significant differences in model accuracy for race, sex, and insurance type in **ICU notes** and insurance type in **psychiatric notes**.
Supporting Task-Focused Model 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>
Moving Forward; Opportunities and Challenges

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
Challenges are Secret Opportunities!

Opportunities in Machine Learning for Healthcare

Marzyeh Ghassemi
University of Toronto, Vector Institute
Toronto, Canada
marzyeh@cs.toronto.edu

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 School of Public Health
Boston, MA 02115
andrew_beam@hms.harvard.edu

Irene Y. Chen
Massachusetts Institute of Technology
Cambridge, MA 02139
iychen@mit.edu

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

Abstract

Modern electronic health records (EHRs) provide data to answer clinically meaningful questions. The growing data in EHRs makes healthcare ripe for the use of machine learning. However, clinical data presents unique challenges that complicate the use of common machine learning methodologies. For example, these challenges include disease labels in EHRs, encompassing multiple underlying phenotypes, and the under representation of healthy individuals. This article serves as a primer to illuminate these challenges and highlights opportunities for members of the machine learning and data science communities to contribute to this domain.
UofT / Vector Team

PhD Students

Bret Nestor
Denny Wu
Amy Lu
Matthew McDermott

Technical Collaborators

Dr. Anna Goldenburg
Dr. Shalmali Joshi

Clinical Collaborators

Dr. Amol Verma
Dr. Fahad Razak
Dr. Muhammad Mamdani
Complex Data Challenges

- We know that **Data Quality Matters**, but **Disease Data is Imbalanced**, and restrictive access makes **Data Only for Few researchers**.

<table>
<thead>
<tr>
<th>Problem</th>
<th>Short-Term Solutions</th>
<th>Long-term Outlook</th>
</tr>
</thead>
</table>
| **Data Quality Matters** | - Sparsity  
- Missingness  
- Biased sampling | - Data aggregation  
- Imputation techniques  
- Synthetic data | - High-quality research data  
- Data generation documentation |
| **Disease Data Imbalances** | - Sporadic diseases  
- Unbalanced observation | - Modified loss functions  
- Important class targeting  
- Data subsampling | - Patient self-reporting  
- Passive data collection  
- “Normal” baselines |
| **Data Only For The Few** | - Limited access  
- Few datasets | - Standardized metrics  
- Anonymized learning  
- Privacy-preserving models | - Patient-shared data pools  
- Well-defined anonymization regulation |
Rethinking Clinical Prediction

Out of sample generalization is particularly important in clinical settings.

Three training paradigms for mortality prediction in MIMIC III. Item-ID and Clinically Aggregated representations are trained on

A) 2001-2002 data only,
B) previous year only,
C) all previous years.

Dashed line is year-agnostic model performance, aka what most papers report for performance.

Only models trained on all previous data using clinically aggregated features **generalise across hospital policy changes and year of care.**
Robustness to The Unseen

- As devices and practices change the **Same Name maybe a Different Measure**, while novel $x, y, x|y$ require **Anticipating New Data** and **Handling the Next Zika**.

<table>
<thead>
<tr>
<th>Problem</th>
<th>Short-Term Solutions</th>
<th>Long-term Outlook</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same Name, Different Measure</td>
<td>- Measurement drift</td>
<td>- Better devices</td>
</tr>
<tr>
<td></td>
<td>- Changing equipment</td>
<td>- Self-calibration</td>
</tr>
<tr>
<td>Anticipating New Data</td>
<td>- Unseen people</td>
<td>- Regulatory incentives</td>
</tr>
<tr>
<td></td>
<td>- Drifting populations</td>
<td>- Fund data generalizability</td>
</tr>
<tr>
<td>Handling the Next Zika</td>
<td>- Unseen conditions</td>
<td>- Expedited clinical capture</td>
</tr>
<tr>
<td></td>
<td>- Drifting diseases</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Abnormality detection</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Human-in-the loop models</td>
<td></td>
</tr>
</tbody>
</table>
Modeling the Biological Pathology Continuum with HSIC-regularized Wasserstein Auto-encoders

- Create latent representations that reflect side information with WAE to model pathology continuum, and HSIC to enforce dependency between certain latent features and the provided side information

Training loss and HSIC loss vs. training steps + malignancy score of the nearest neighbors of generated samples vs. dependant axis; the trend of malignancy correlates with the dependent axis. Lung Image Data of thoracic scans from 1018 patient cases with 2670 images.

- Regularized generative model constructs interpretable latent features, and models continuous morphological change corresponding to provided side information

Scatter plot of test images on latent space of ~10,000 images from leukemia cell line K562 with dilutions of adriamycin. Class separation is obvious on x (dependant axis), but not on y (1st PC of independent axes. Generated images sampled from the dependent axis and the 1st PC of all other axes; generated cells vary in shape.
Unknown Knowns

- Fundamental research is needed in healthcare to understand **Difficult Disease Endotyping**, which may require that researchers work with clinicians to **Create Common Ground**.

<table>
<thead>
<tr>
<th>Problem</th>
<th>Short-Term Solutions</th>
<th>Long-term Outlook</th>
</tr>
</thead>
</table>
| Difficult Disease Endotyping | - Underlying disease heterogeneity  
- Unknown subtypes | - Generative modeling  
- Unsupervised clustering | - Additional data  
- Fundamental research  
- Robust clinical endotype |
| Creating Common Ground | - No consensus  
- Meaningful targets | - Causal inference  
- Diagnostic baselines | - Patient self-report  
- Combining labels |
The Effect of Heterogeneous Data for Alzheimer’s Disease Detection from Speech

NeurIPS 2018 ML4H Workshop Aparna Balagopalan, Jekaterina Novikova, Frank Rudzicz, Marzyeh Ghassemi

- Augment AD with multi-task healthy data + analyze class boundaries.

Adding in same task healthy data (122 samples). Pic. descriptions (PD); 28.6% out of task error

Adding in different structured task healthy data (327 samples) PD + structured tasks; 17.8% out of task error

Adding in general speech healthy data (231 samples) PD + general speech; 3.6% out of task error

- Adding in datasets with general, unstructured conversations improves models trained using structured tasks!
Machine Learning For Health (ML4H)

What models are healthy?

What healthcare is healthy?

What behaviors are healthy?

What is Canada doing?
Health Data Is A Resource

- All data is valuable; health data particularly so.
- Robust algorithms require large scale datasets for research use.

AWS Machine Learning Blog
Improving Patient Care with Machine Learning At Beth Israel Deaconess Medical Center

Beth Israel Deaconess Medical Center has launched a multi-year, innovative research program on how machine learning can improve patient care, supported by an academic research sponsorship grant from AWS. The Harvard Medical School-affiliated teaching hospital will use a broad array of AWS machine learning services to uncover new ways that machine learning technology can enhance clinical care, streamline operations, and eliminate waste, with the goal of improving patient care and quality of life.

Improving patient care with machine learning

Inefficiencies in hospital management and operations are not only extremely costly to providers, insurers, patients, and taxpayers, but they can result in precious resources being diverted away from patient care. These inefficiencies drive healthcare costs up and can contribute to life-threatening medical conditions.

Amazon Comprehend Medical
Extract information from unstructured medical text accurately and quickly
No machine learning experience required

Amazon Comprehend Medical is a natural language processing service that makes it easy to use machine learning to extract relevant medical information from unstructured text. Using Amazon Comprehend Medical, you can quickly and accurately gather information, such as medical condition, medication dosage, strength, and frequency from a variety of sources like doctors’ notes, clinical trial reports, and patient health records.

Google Tries to Patent Healthcare Deep Learning, EHR Analytics

Google has applied for a sweeping patent including the fundamentals of deep learning and EHR analytics in the healthcare industry.

Source: Google
Create Research with a Resource

• ML4H is currently defined by ONE dataset - MIMIC from the Beth Israel Deaconess Medical Center ICU.¹

A Decade of Vetted Access to De-identified Data

• MIMIC has been around for over a decade.

• No lawsuits or newspaper headlines regarding privacy failures.

• Vetted access to de-identified data demonstrably safe, even for a single source in a small city.

IRB Approval

This study was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). Requirement for individual patient consent was waived as the study did not impact clinical care and all data were de-identified.

The MIMIC II database was collected as part of a Bioengineering Research Partnership (BRP) grant from the National Institute of Biomedical Imaging and Bioengineering entitled, “Integrating Data, Models and Reasoning in Intensive Care” (RO1-EB001659). The project was established in October 2003 and included an interdisciplinary team from academia (MIT), industry (Philips Medical Systems) and clinical medicine (Beth Israel Deaconess Medical Center). The objective of the BRP is to develop and evaluate advanced Intensive Care Unit (ICU) patient monitoring systems that will substantially improve the efficiency, accuracy and timeliness of clinical decision making in intensive care.
The MIMIC Model Works - ICES/GEMINI Options

- Openly accessible, de-identified clinical dataset
- Privacy risks mitigated with vetted users under EULA
- Streamlined access to data
- Enabling collaboration, benchmarking, reproducibility

New researchers approved for MIMIC:

Machine Learning in Health overfits models to MIMIC:

Funded NIH Grants based on MIMIC (~$1.3M in 2018):
Created a large community!

And many more!
Planning an ML4H Unconference

- May 28/29 in Toronto; ML researchers and clinicians.

[ML4H Unconference website link]
What models are healthy?

What healthcare is healthy?

What behaviors are healthy?