A Road Map for Robust AI

Gilmer Valdes PhD, DABR
Department of Radiation Oncology
Department of Epidemiology and Biostatistics
University of California, San Francisco.
Outline

1. Causal Inference.
2. Risks of Models built using correlation.
3. Interpretability.
4. Expert Augmented Machine Learning
Causal Inference Framework

\[ Y_A \rightarrow \text{survival time when treatment } t = A \]

\[ Y_B \rightarrow \text{survival time when treatment } t = B \]

\[ x \rightarrow \text{patient's characteristics.} \]

Treatment Effect

\[ \text{Effect} = E_x [ (Y_A - Y_B) | x ] \]

\[ \text{Effect} = E_x [ Y_A | x ] - E_x [ Y_B | x ] \]

Where the expectation is taken over all patients
Causal Inference from Training Data

\[
\text{Effect} = E_x [Y_A | x] - E_x [Y_B | x]
\]

**Problem:** We never observe \( Y_A \) and \( Y_B \) for a patient because they either receive treatment A or B.

**Approximation:**

\[
E_x [Y_A | x] \approx E_x [Y | x, t = A]
\]

No hidden confounders

\[
(Y_A, Y_B) \perp t | x
\]

\[
E_x [Y_B | x] \approx E_x [Y | x, t = B]
\]
No Hidden Confounder

\[(Y_A, Y_B) \perp t \mid x\]

**When does it break?**

\[t = A\] is only given to very sick patients. Both \( Y_A, Y_B \) are very small.

\[t = B\] is only given to healthy patients. Both \( Y_A, Y_B \) are big.

**Hidden Confounder: patient selection mechanism**
Hidden Confounder in Prediction Settings

\[ \text{Prediction} = E \left[ Y \mid x \right] \]
Risks of Models built using correlation

88% ranking. The CNN was learning the hospital type.

https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98
Risks of Models built using correlation

Example: Predicting Risk of dying of Pneumonia for In-hospital patients

Most accurate model trained: Multi-purpose neural net....

Rule Based Model

Asthmatic ➔ Lower Risk

- Harmful to patients
- High Risk of Liability

**Risks of Models built using correlation**

**Example:** *Predicting Risk of stroke for Emergency Department patients*

<table>
<thead>
<tr>
<th>Table 1—Predicting and Mispredicting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Stroke</strong></td>
</tr>
<tr>
<td>--------------------------------------</td>
</tr>
<tr>
<td>Prior stroke</td>
</tr>
<tr>
<td>(0.012)</td>
</tr>
<tr>
<td>Prior accidental injury</td>
</tr>
<tr>
<td>(0.095)</td>
</tr>
<tr>
<td>Abnormal breast finding</td>
</tr>
<tr>
<td>(0.092)</td>
</tr>
<tr>
<td>Cardiovascular disease history</td>
</tr>
<tr>
<td>(0.029)</td>
</tr>
<tr>
<td>Colon cancer screening</td>
</tr>
<tr>
<td>(0.178)</td>
</tr>
<tr>
<td>Acute sinusitis</td>
</tr>
<tr>
<td>(0.155)</td>
</tr>
</tbody>
</table>

*Notes:* Logistic regression on demographics and prior diagnoses in EHR data. Sample: 177,825 ED visits in 2010–2012 to a large academic hospital.

Context is everything

That was surprisingly easy. How come the robotic uprising used spears and rocks instead of missiles and lasers?

If you look to historical data, the vast majority of battle-winners used pre-modern weaponry.

Thanks to machine-learning algorithms, the robot apocalypse was short-lived.
Possible Solutions

Prediction = \( E [Y|x] \)

Interpretability

1. The model is interpretable in a global sense

2. The model is interpretable locally.
   Post-hoc justifications or explanations.

Variable Importance (salient map), Use a simpler model to explain a more complex one, visualizations, etc
Possible Solutions

Post-hoc interpretations are rarely faithful

Salient Map

https://www.nature.com/articles/s42256-019-0048-x

LIME

Possible Solutions

**Expert-Augmented Machine Learning**

- Train a state-of-the-art predictive model using RuleFit
  *This represents the best machine-learned model to predict the outcome of interest given the training data*

- Extract human expert knowledge from panel of domain experts using MediForest.com
  *This provides an automated way to extract problem-specific human expert priors*

- Combine ML model with expert priors to build a robust, efficient, and interpretable EAML model
  *This represents the merging of human expert knowledge with a machine-learned model for best-of-both performance*
Expert-Augmented Machine Learning incorporates human expertise into ML models

1. Build Trees using Gradient Boosting
   ![Gradient Boosting Diagram]

2. Convert Trees to Rules
   ![Converted Trees Diagram]

3. Select Rules using LASSO
   ![Selected Rules Diagram]

4. Extract experts’ assessment of each rule on MediForest.com and rank
   ![Experts' Assessment Table]

5. Compare empirical and expert rule ranking
   ![Comparison Table]

6. Build EAML model by combining rules & expert assessments
   
   \[
   \hat{C} = \text{argmin}_C \sum_{i=1}^{N} [y_i - CR(i, :)]^2 \\
   + \lambda \sum_{k=1}^{K} f(\Delta \text{Ranking}_k, \text{STDEV}_k) \| C_k \|_2
   \]
   
   Gennatas et al 2019
   
   n cases; k rules
   
   arxiv.org/abs/1903.09731

Experts assess a few simple rules instead of a vast number of individual cases
Model- vs. Expert-derived variable importance

**Model variable importance based on correlational structure of data**

![Graph A](image)

**Variable importance estimated from clinicians’ responses based on their causal & correlational knowledge**

![Graph B](image)

**PaO₂/FiO₂** is the most important feature for both the model and clinicians but for different reasons
EAML allows to train with less data

MIMIC2–trained model on MIMIC3 data

Mean AUC

0.78
0.76
0.74
0.72
0.70

Training N Cases

200 400 800 1600 3200 6400

Rank difference

<5 (All Rules)
<4
<3
<2
<1

\[ \hat{C} = \text{argmin}_{\hat{C}} \sum_{i=1}^{N} [y_i - C \ast R(i,:)^2] + \lambda \sum_{k=1}^{K} f(\Delta \text{Ranking}_k, \text{STDV}_k) \ast C_k^2 \]
“...There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot...”

https://www.wired.com/2008/06/pb-theory/
EAML Project: The team

We are a multidisciplinary team with extensive clinical and quantitative expertise, and a shared goal of developing advanced Machine Learning algorithms for accurate, interpretable, safe and fair clinical predictive modeling.

In alphabetical order:

Andrew Auerbach, MD, MPH
University of California, San Francisco, Hospital Medicine

Eliier Delgado, MSc
Innova Montreal, Web Application Engineering

Eric Eaton, PhD
University of Pennsylvania, Machine Learning, Deep Learning

Jerome H. Friedman, PhD
Stanford University, Statistics, Machine Learning

Efstathios D. Gennatas, MBBS, PhD
University of Pennsylvania, Neuroscience, Machine Learning, Biomedical Data Science

Yannet Interian, PhD
University of San Francisco, Data Science

Mark J. van der Laan, PhD
University of California, Berkeley, Biostatistics

Jose Marcio Luna, PhD
University of Pennsylvania, Radiation Oncology, Machine Learning

Romain Pirracchio, MD, PhD
University of California, San Francisco, Anesthesia, Biostatistics

Lara G. Reichmann, PhD
University of San Francisco, Data Science

Charles B. Simone II, MD
NY Proton, Radiation Oncology

Timothy D. Solberg, PhD
University of California, San Francisco, Radiation Oncology

Lyle H. Ungar, PhD
University of Pennsylvania, Machine Learning, Biomedical Data Science, Natural Language Processing

Gilmer Valdes, PhD
University of California, San Francisco, Radiation Oncology, Machine Learning