CPSL
Cyber-Physical Systems Laboratory

Assisting clinical decision making: individualized treatment effect estimation

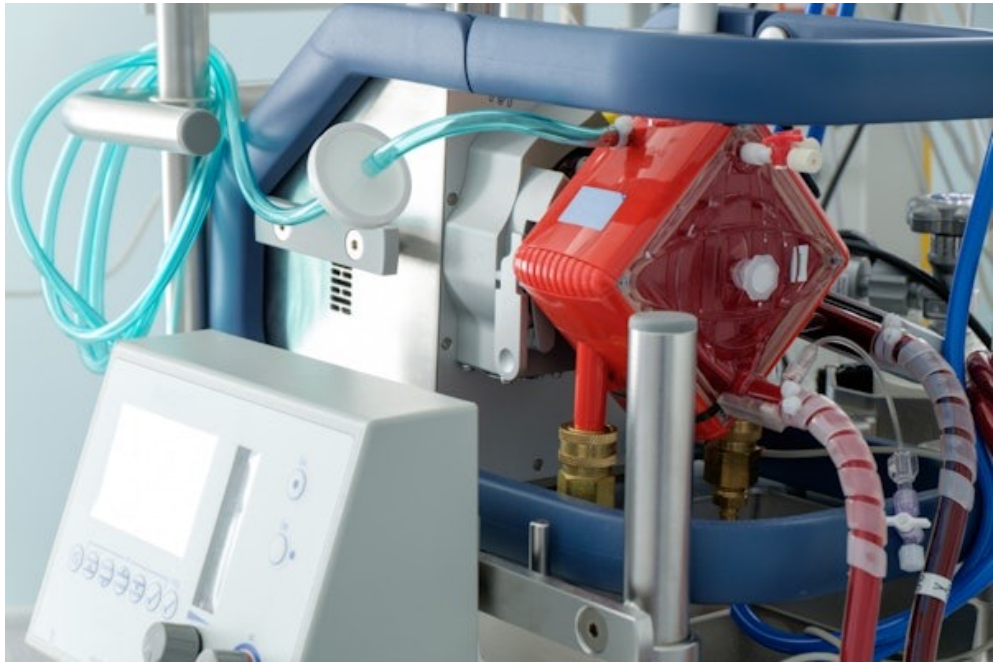
Bing Xue

AI for Health Institute
McKelvey School of Engineering

Washington University in St.Louis

Background

- What is ECMO?
  - ECMO: Extracorporeal Membrane Oxygenation
  - ECMO: used in critical care situations; allows the blood to "bypass" the heart and lungs, allowing these organs to rest and heal.
  - Critical: last-resort life support for severe COVID-19, ARDS and other infections.



An interdisciplinary team of researchers at Washington University in St. Louis has developed a new machine-learning model to assist clinicians in making the difficult decision on which patients should get ECMO treatment. (Credit: iStock photo)

Background

- Some characteristics of ECMO:
  - Resource constraint: limited availability;
    - Scarcity due to technical complexity, lack of expertise, etc.
    - 6.8M deaths,
    - Only 14K ECMO treatments
  - Label and resource extensive – hospitals and clinicians need time to prepare.
    - When treated with ECMO: 47% mortality rate
    - When treated without ECMO: 90% mortality rate



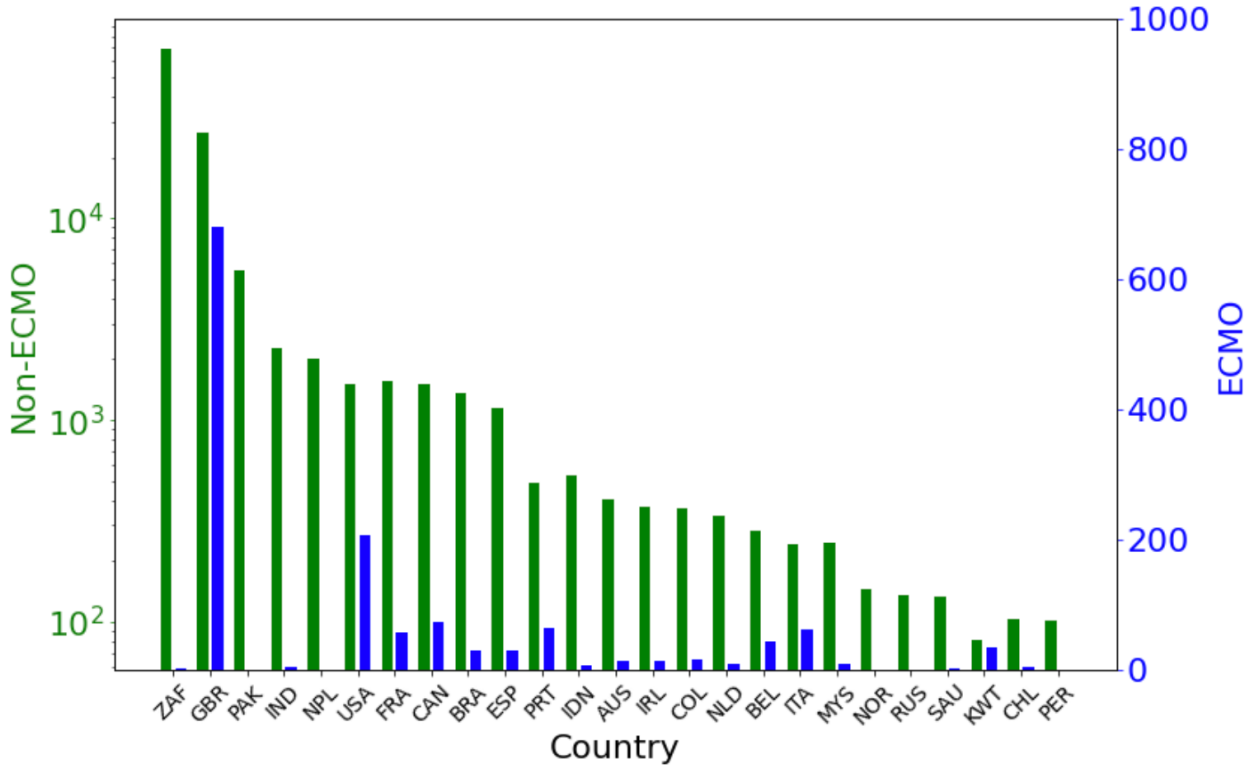
https://anesthesiology.wustl.edu/patient-care/perioperative-medicine/

Some reflections

- What we have done:
  - We developed a machine learning approach to:
    - Learn clinicians' decision making process
    - Provide lead time in identifying the patients-at-risk
- But:
  - Who should be eligible for ECMO? Do clinicians always make the right decisions?
  - Did we give the treatment resources to someone who could benefit more?
    - What would happen if we can't provide the treatment?

Who should be eligible for ECMO?

- There is no standard criterion to determine ECMO treatment;
  - Large variation between countries and hospitals
  - Which hospital's decision-making criteria is more reasonable?

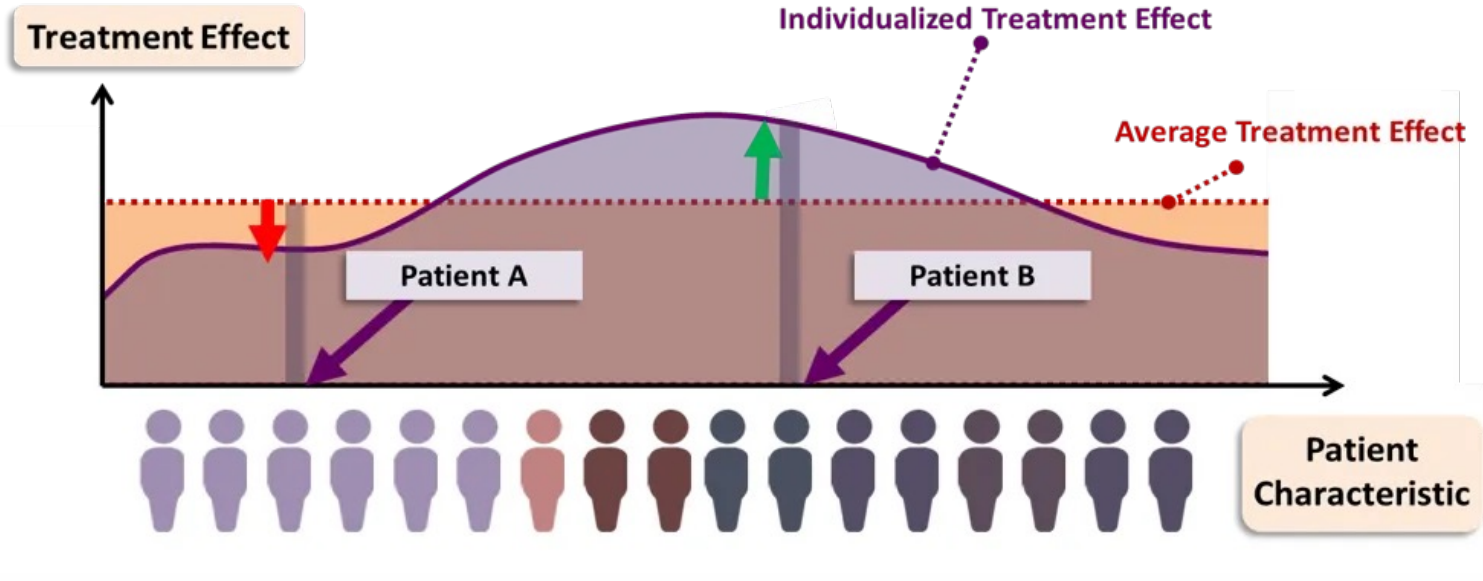


Did we give the treatment resources to someone who could benefit more?

CPSL Cyber-Physical Systems Laboratory

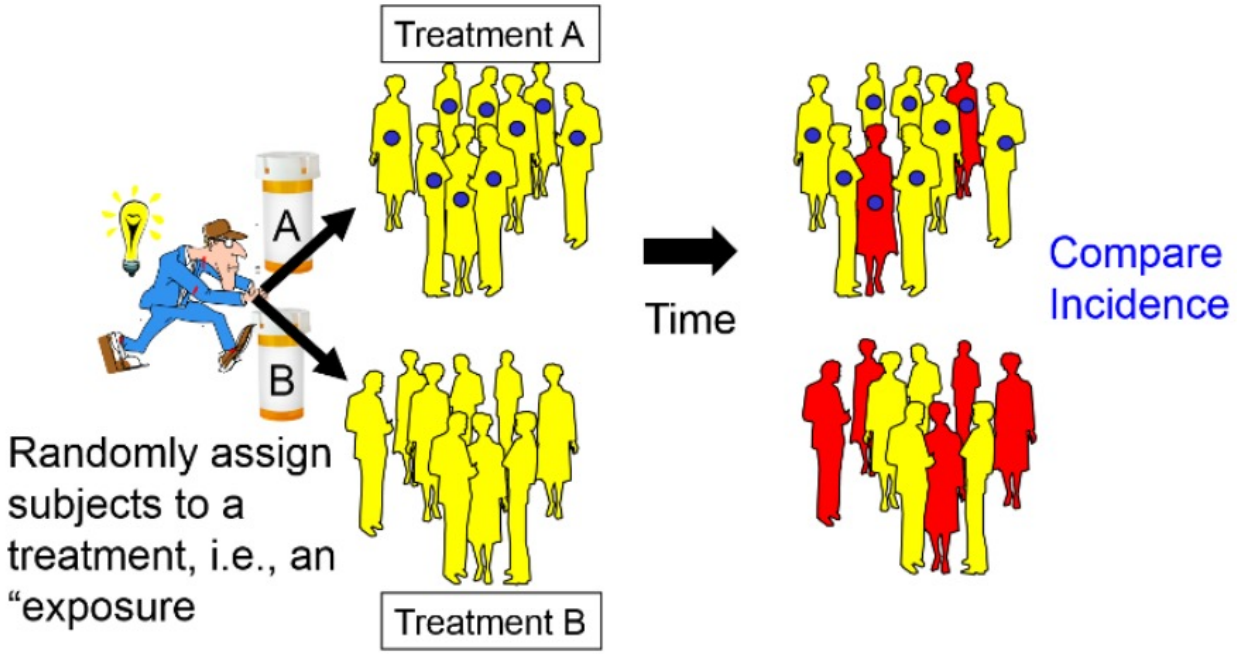
- Given resource constraint, we need to identify the patients who experience the greatest treatment effects using ECMO
  - We need to estimate each individual's treatment effect:
    - mortality risk with treatment – mortality risk without treatment

One of them is always unobserved!



Ideal case: randomized clinical trial

- Ideal case:
  - Randomly assign patients to ECMO and control group
  - Compare the treatment effects of similar/same patients in ECMO vs control



https://sphweb.bumc.bu.edu/otlt/mph-modules/ep/ep713_clinicaltrials/EP713_ClinicalTrials_print.html

However

- Real clinical decisions are not randomized clinical trials.
  - ECMO patients are generally sicker, older, etc.
  - Some patients are critically ill, but they are not suitable for ECMO treatment:
    - too old/young, too sick, BMI too high, etc.

However

- Real clinical decisions are not randomized clinical trials.
- Given the Electronic Health Records of patients, we are looking for a prediction model that solves 3 problems:
  1. Propensity score:
     - Is this patient critically ill?
     - What is the probability of the patient being assigned to ECMO treatment?
  2. Treatment effect:
     - What is the mortality risk if patient is assigned to ECMO;
  3. Non-treatment effect:
     - What is the mortality risk if patient is not assigned to ECMO.
  - The effect of ECMO treatment for this patient is :
    - Treatment effect - Non-treatment effect

Potential impact:

- Note that the treatment effect models are not limited to ECMO decision making
  - Other clinical decisions/treatments
    - Prescribe new medication as opposed to the standard medication?
    - One of the key topics in healthcare research
  - Non-clinical applications:
    - A/B test:
      - Online shopping: what if I offer free shipping / discount coupon to customers?
      - TikTok / IG Reels: what if I recommend these videos to users?
      - etc.
    - Key topic in industrial data science

Related works

- Individualized treatment effect estimation falls under causal inference
  - Used to be a statistical topic, but now more towards machine learning
- Two approaches:
  - Indirect estimation
  - Direct estimation
- Indirect estimation
  1. Model the treatment outcome:
     - What is the mortality risk if patient is assigned to ECMO;
  2. Model the non-treatment outcome:
     - What is the mortality risk if patient is not assigned to ECMO.
  3. The effect of ECMO treatment for this patient is :
     - Treatment effect - Non-treatment effect

Related works

- Direct estimation
  - Focus on treatment effect process.
  - Estimate the treatment effect without modeling the process that generates the observed outcomes.
  - Our intuition is that the outcome process is typically more complex than the treatment effect process.
    - For example, a customer's buying process is a complicated decision depending on awareness, context, money and shopping intention. A customer's susceptibility to a coupon also depends on money and shopping intention, but is intuitively less personal and much less complicated.
- Direct estimation vs indirect estimation: which one is preferred?

Related works

- Nowadays, indirect estimation is more favored. Why?
  - Clinical decision making:
    - Clinicians are not only interested in the treatment effects, but also the treatment and no-treatment outcomes

Related works

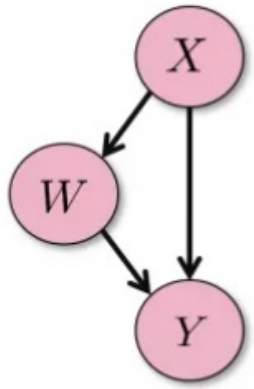
- Nowadays, indirect estimation is more favored. Why?
  - Easier ML modeling: we know the labels of factual outcomes, but we never know the labels of treatment effects!

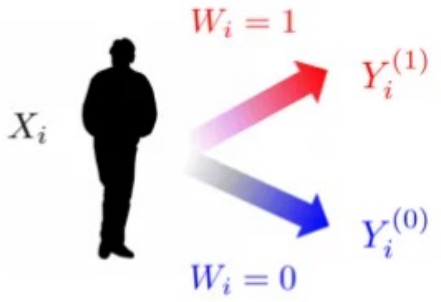
Observational data (X_i, W_i, Y_i)

Each patient i has features $X_i \in \mathcal{X} \subset \mathbb{R}^d$

Two potential outcomes $Y_i^{(1)}, Y_i^{(0)} \in \mathbb{R}$

Treatment assignment $W_i \in \{0, 1\}$





Factual outcomes

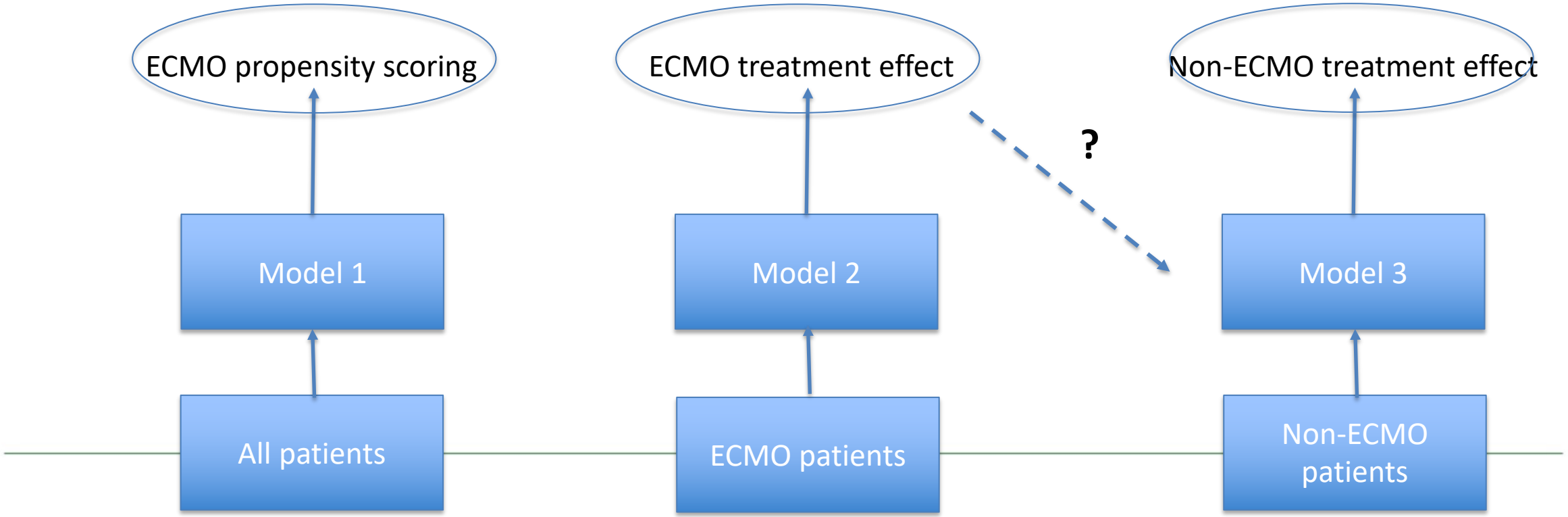
$$Y_i = W_i\, Y_i^{(1)} + (1 - W_i)\, Y_i^{(0)}$$

Causal effects

$$T(x) = \mathbb{E}\left[Y_i^{(1)} - Y_i^{(0)} \,\middle|\, X_i = x \right]$$

Challenges:

- Building single models for each problem is flexible, but it does not fully address our concerns.
  - Selection bias: each model is built on different groups of patients



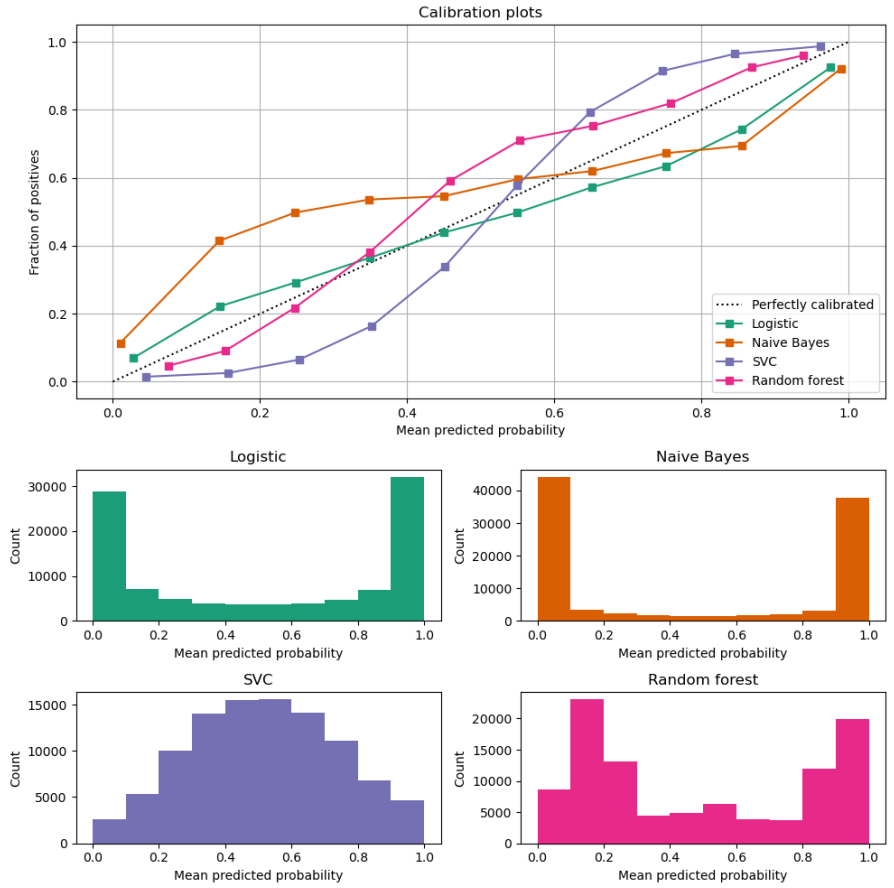
Challenges:

- Model 2 never learned the characteristics of Non-ECMO patients, can it make a good prediction for Non-ECMO patients?
- Model 3 never learned the characteristics of ECMO patients, can it make a good prediction for ECMO patients?
- The effect of ECMO treatment for this patient is :
  - Treatment effect - Non-treatment effect

If one of them is not reliable, does the treatment effect still make sense?

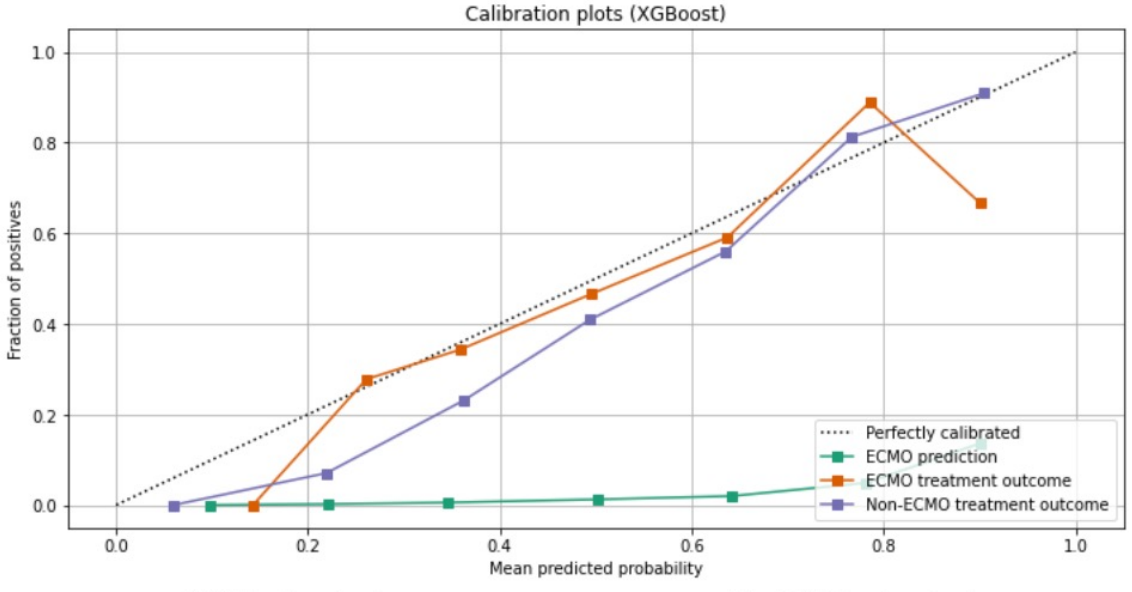
Challenges:

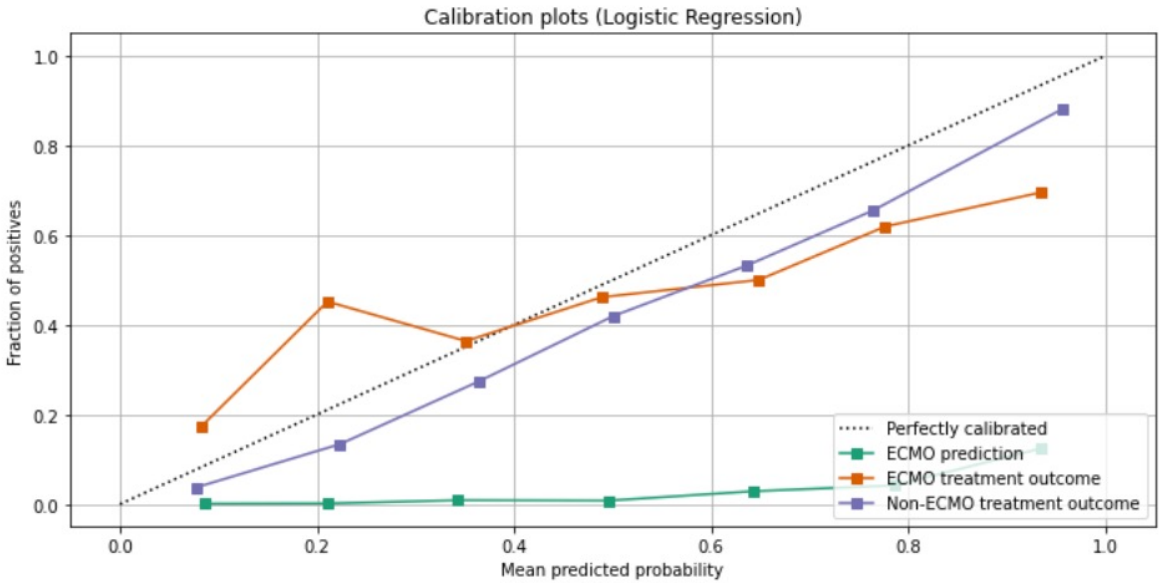
- Even if the single models are generalizable to other patients, are their probability/risk scores comparable?
  - Probability Calibration:
    - If model A is used to predict ECMO outcome, and
    - model B is used to predict non-ECMO outcome, then
    - Treatment effect is NOT equal to $p_A - p_B$!



Challenges:

- Model calibration: A perfect model is along the dashed diagonal.





Problem formulation:

- We need a solution with the following features:
  - Indirect estimation:
    - Predict treatment assignment (propensity score)
    - Predict treatment outcome
    - Predict no-treatment outcome
  - Selection **debiase**:
    - Can predict treatment outcome for control patients
    - Can predict no-treatment outcome for ECMO patients
  - Calibrated probability:
    - The treatment outcome and non-treatment outcome can be directly compared

Approach 1

- **Single-Model-Approach (aka S-Learner):**
  - Examples: BART, Causal Forest, etc.
  - We include the treatment indicator as a covariate into the model
  - To predict the ITE, we would set the treatment variable once to 1 and once to 0 for the same X. The difference in predictions is the treatment effect estimate.

BART consists of two pieces: a sum-of-trees model and a regularization prior. The model lets

$$Y = g(z, x; T_1, M_1) + g(z, x; T_2, M_2) + \cdots + g(z, x; T_m, M_m) + \epsilon,$$

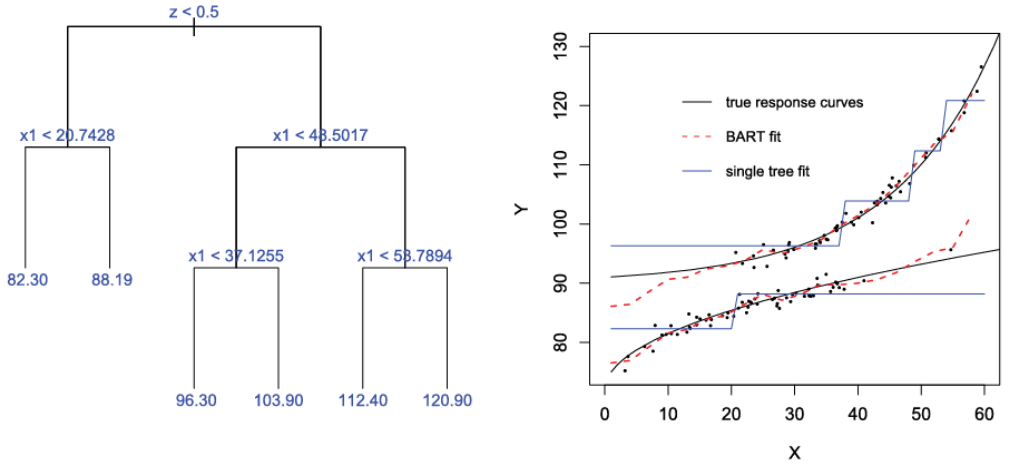


Figure 2. Left panel: the binary tree fit to the data from Figure 1. Right panel: single-tree fits (solid lines) and BART fits (dashed lines). A color version of this figure is available in the electronic version of this article.

Approach 2

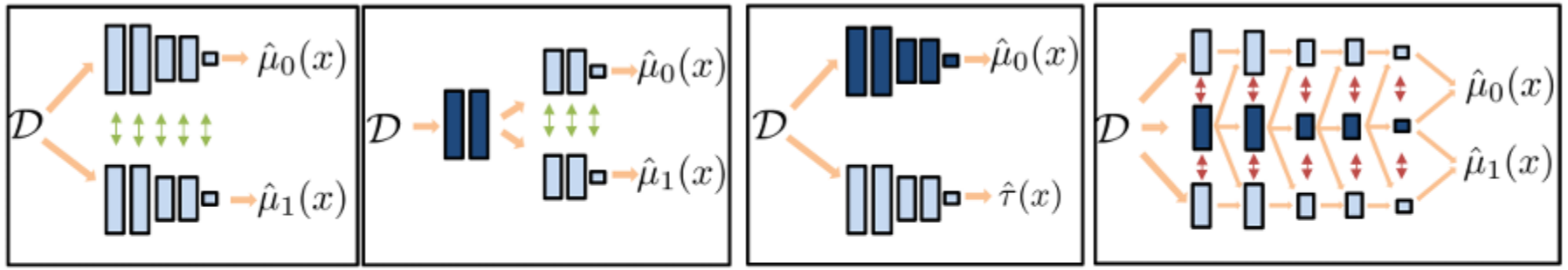
CPSL Cyber-Physical Systems Laboratory

Difference in Conditional Means

- aka K-Model approach, Two-Model Approach, T-Learner, Conditional Mean Regressions, etc
- We estimate an outcome model for each treatment group separately and calculate the treatment effect as the difference between the estimated outcomes.
- The outcome models (*base learners*) can take any form, so we could use neural networks or boosted trees to do the heavy lifting.
- Selection debiase & probability calibration:
  - Option 1: 2 NN models with parameter sharing in the lower hidden layers
  - Option 2: single NN model with multi-task architecture
  - Option 3: 3 NN models:
    - A NN model for shared information between 2 groups
    - 2 NN models for private information in each group

Approach 2

- The key is how to regularize the two outcome predictions.
  - By sharing representation/information:

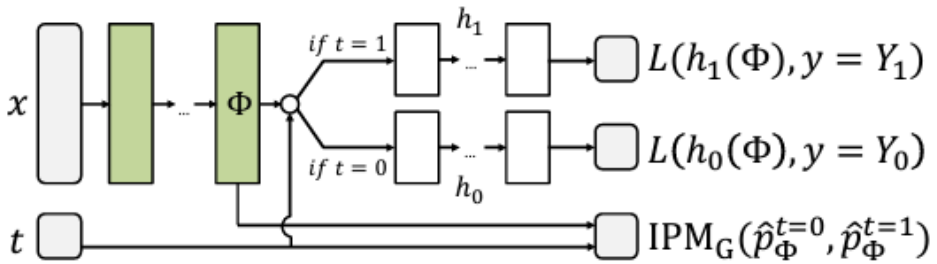


(1) Regularization for TNet (left) and TARNet (right) (2) Reparametrization (3) FlexTENet

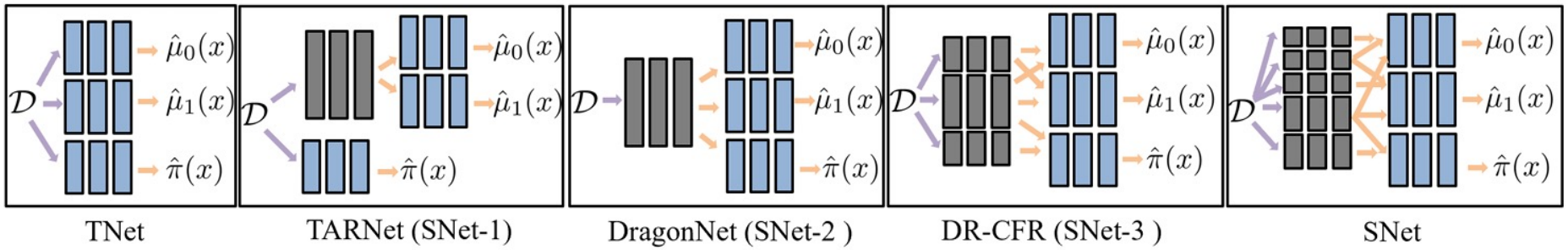
Figure 2: The three approaches under investigation. Dark layers indicate parameters shared between POs, light layers indicate private parameters. Green arrows indicate regularization encouraging parameters to be similar, red arrows indicate regularization that encourages orthogonalization.

Approach 2

- Extract/distill the selection bias



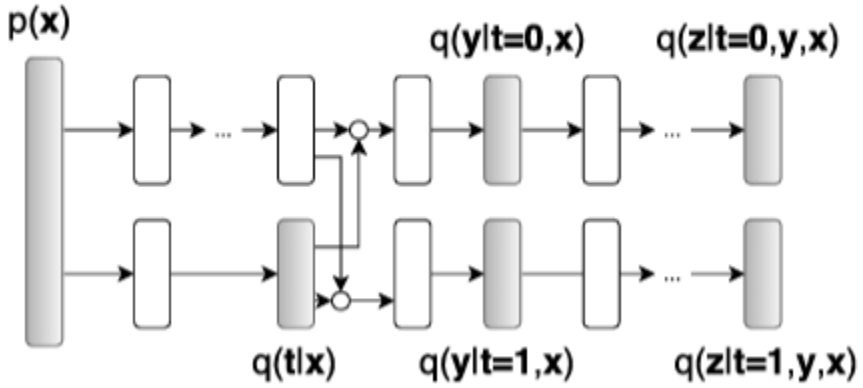
- Revolutions of NNs in sharing representations and distilling selection bias



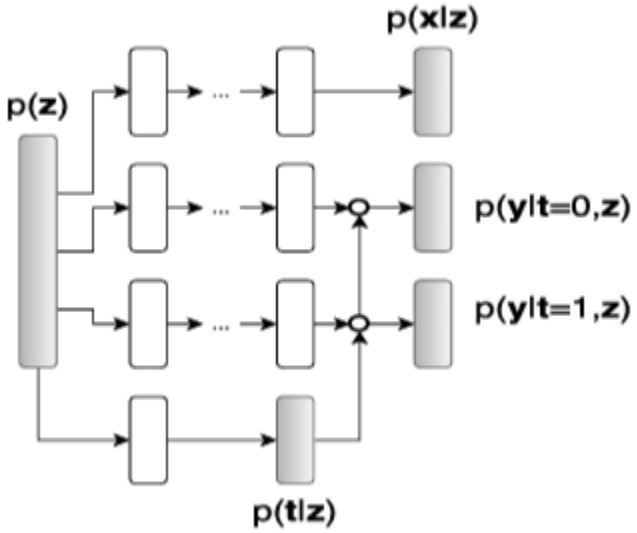
Note: Grey boxes are shared layers, blue boxes are task-specific layers.

Approach 3

- GenAI: generative modeling of patients characteristics
  - The concept was GenAI was not *so* popular until this year
  - Key concept: a generative encoder-decoder framework
    - Encoder: a new representation of patients
    - Decoder: transform patient characteristics into other forms of information



(a) Inference network, $q(\mathbf{z}, t, y|\mathbf{x})$.



(b) Model network, $p(\mathbf{x}, \mathbf{z}, t, y)$.

Approach 3

- GenAI: why do we need generative modeling?
  - Think about OpenAI's ChatGPT-4:
    - Inputs contain various information
      - Some are redundant/useless to the task
      - Inputs have different formats (images) than outputs (texts)
    - Outputs vary depend on tasks/commands
  - In our case:
    - Encoder: Extracts the useful information from inputs
    - Decoder: Transform the useful information into desired outputs (treatment/control outcomes)

User What is funny about this image? Describe it panel by panel.



Source: hmmm (Reddit)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

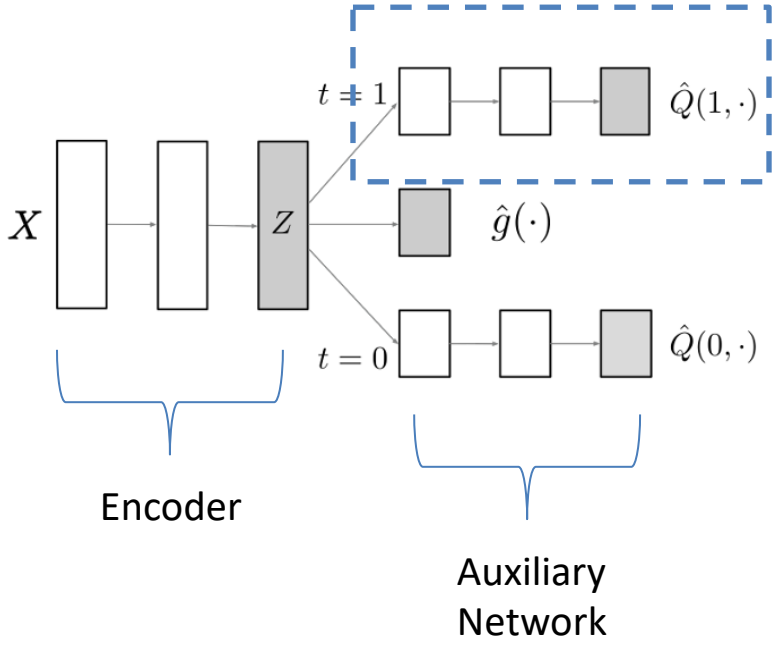
Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Why do we need a new model?

- Summary of the modeling challenges in ECMO treatment effect estimation:
  1. Counterfactual: Only one outcome is observed
  2. Strong selection bias between two groups
  3. Scarcity of treatment resources: ECMO cases are rare
  4. Characteristics of Electronic Health Records:
     - Always partially observable
     - High-dimensional: thousands of lab tests / measurements / medications

Why do we need a new model?

- The problems come from the auxiliary networks
  - Compared to ChatGPT or other GenAI use cases, our dataset is highly imbalanced



This is not well-trained

- when treatment cases are rare;
- when there is strong selection bias.

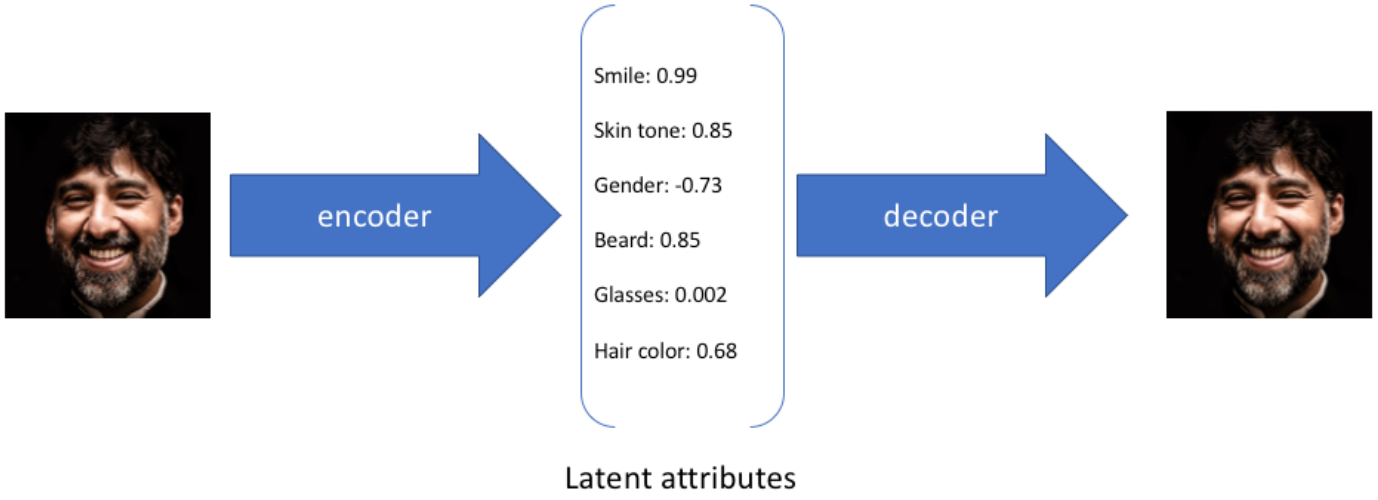
Our modeling strategy

A *different* perspective from existing studies:

- Instead of removing selection bias, can we make use of the selection bias?
- Can we get rid of auxiliary networks (task-specific layers)?
- Can we regularize counterfactual predictions by the intrinsic characteristics of data?

Modeling strategy

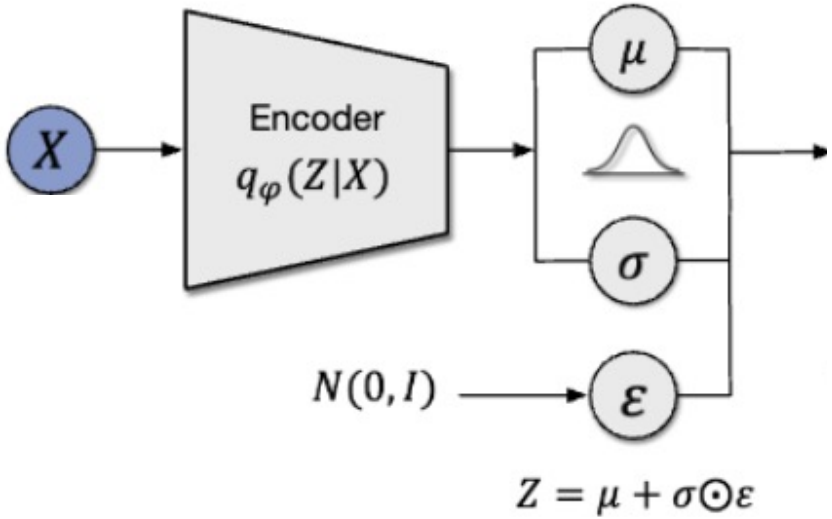
- There are many possible GenAI networks, here we use variational autoencoder (VAE):



Modeling strategy

- We start with the encoder:
  - Transform Electronic Health Records to posterior latent distribution.
- Why?
  - Always partially observable
  - High-dimensional: thousands of lab tests / measurements / medications

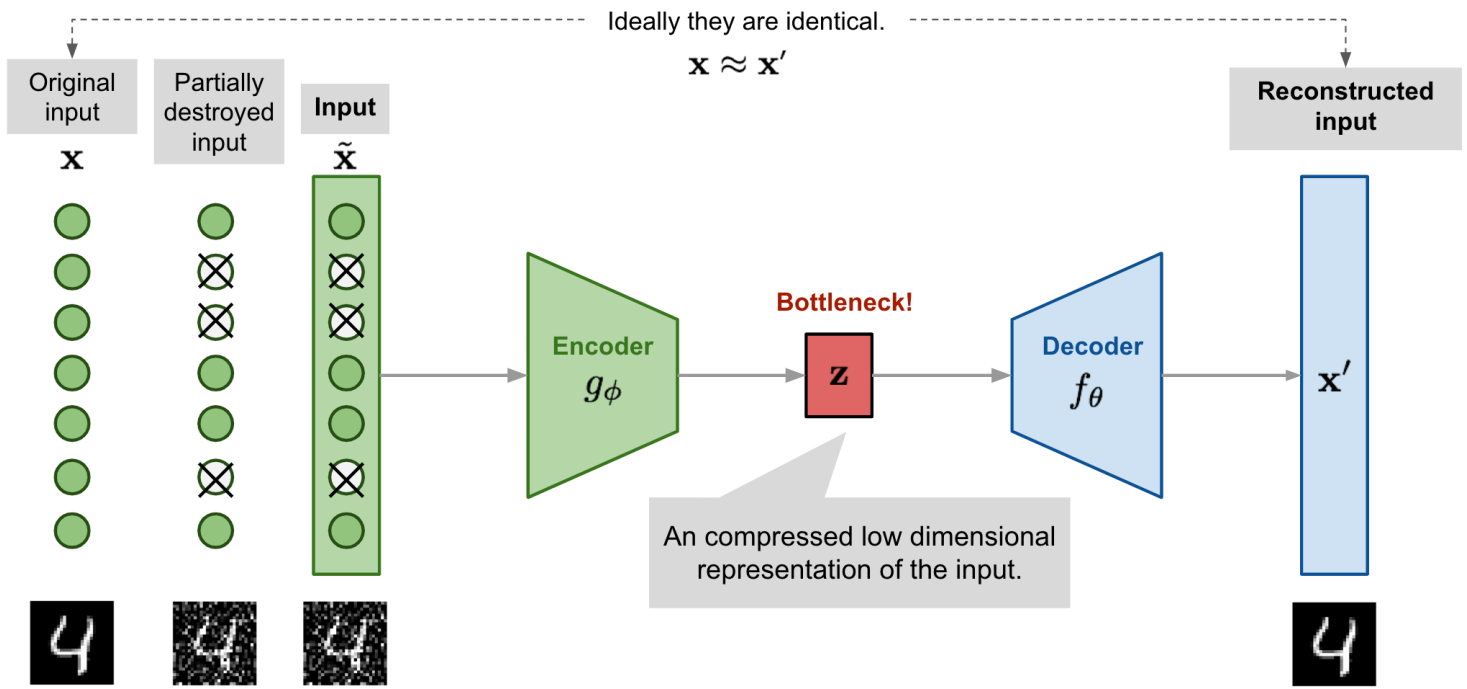
Input Features:
- Redundant/correlated
- High-dimensional
- Multi-modal
- Partially observable



Latent Embedding:
- Lower-dimensional
- No missingness
- Regularizable: disentangled, semi-supervised, clustered, etc.

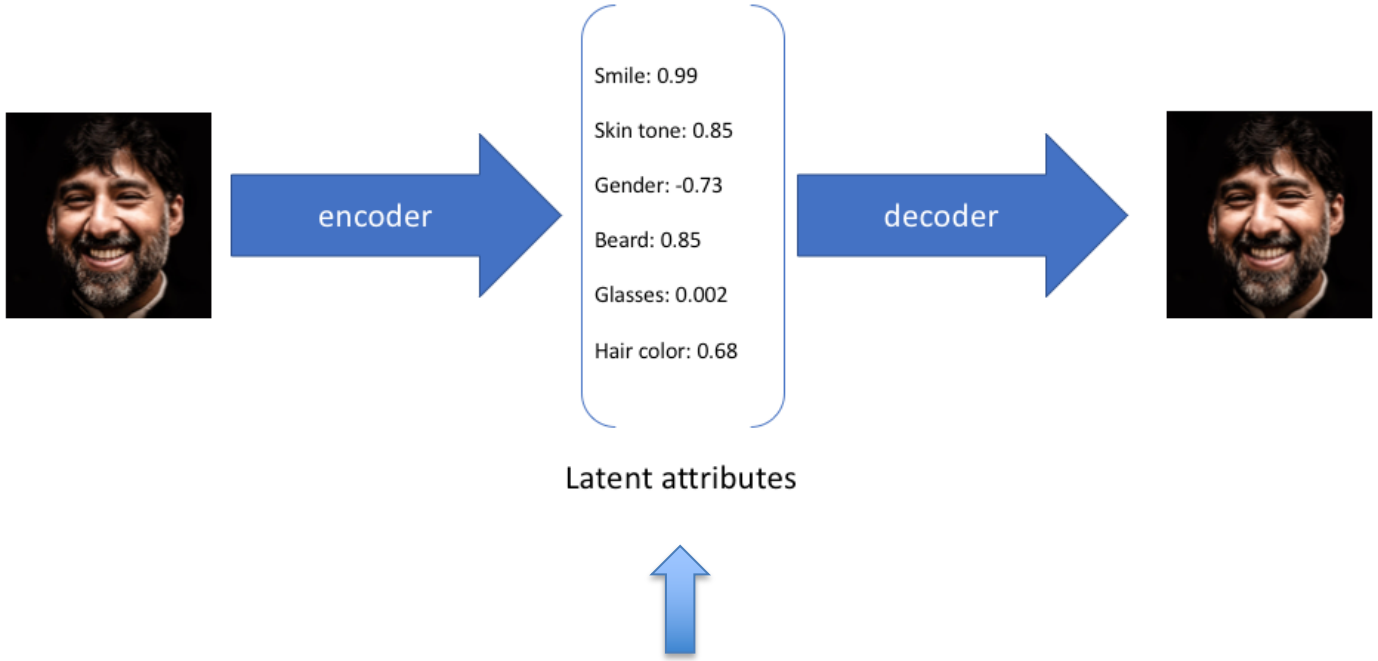
Modeling strategy

- Similar idea has been validated in image reconstruction



Modeling strategy

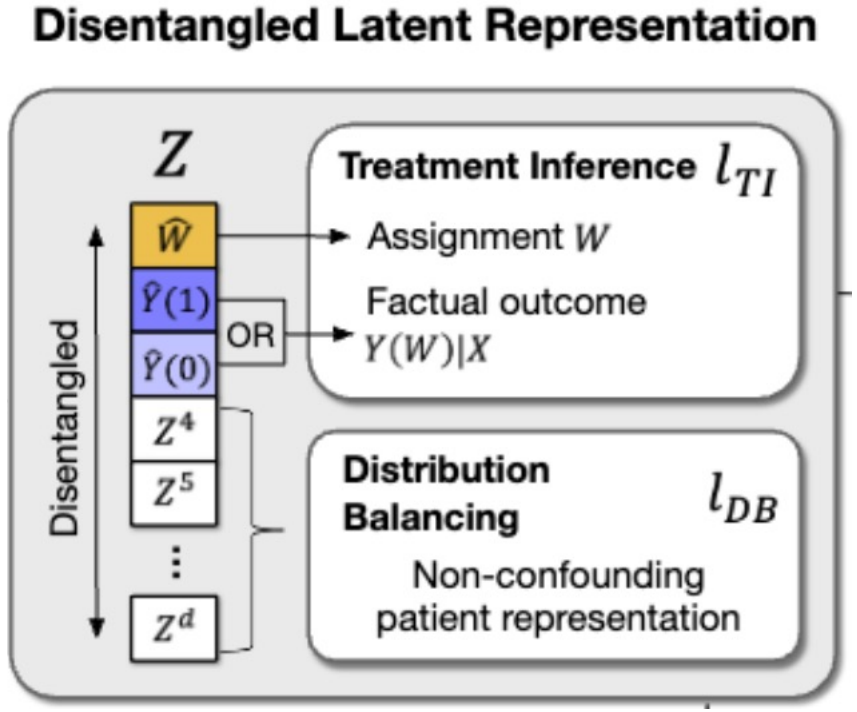
- Before we move to decoder



How to regularize/control the latent attributes?

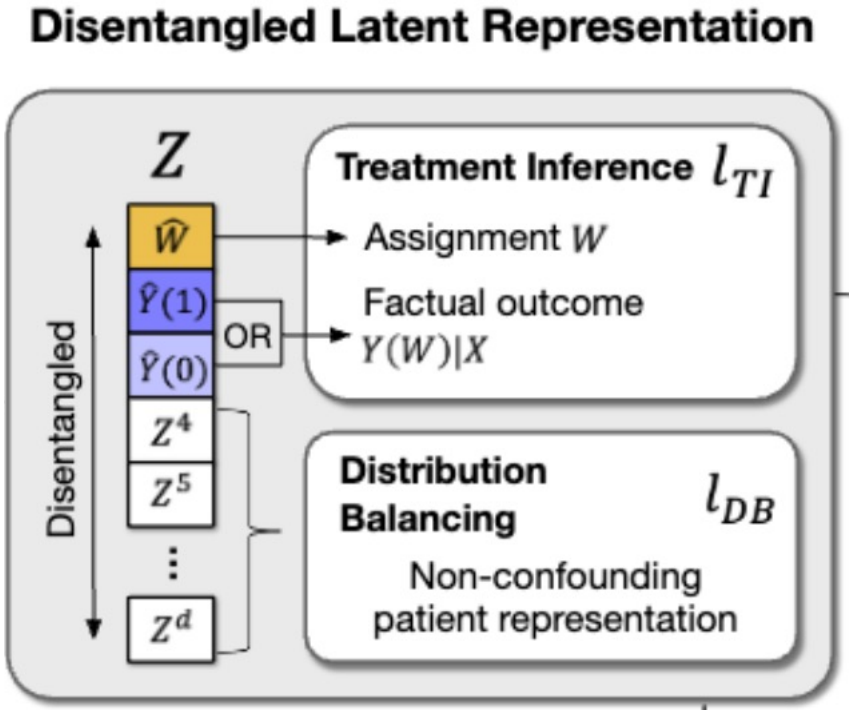
Modeling strategy

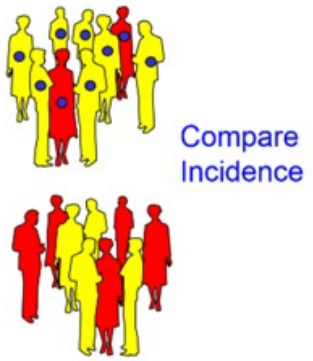
- Next, we constructed a regularized latent space:
  - Disentangled latent space:
    - Each dimension represents a single aspect of information
    - Isolate shared representation from group-specific representations
    - Extract selection bias into treatment assignment dimension
  - Semi-supervised latent space:
    - Direct predictions from designated dimensions
    - Learn from factual predictions



Modeling strategy

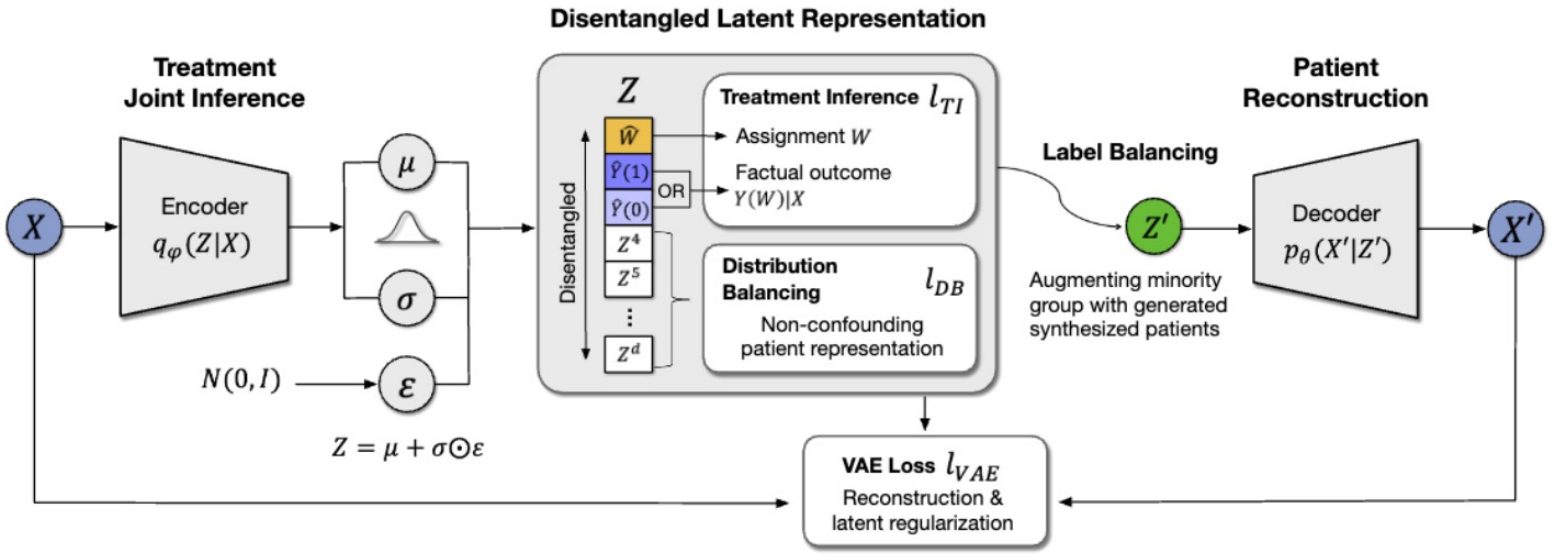
- Next, we constructed a regularized latent space:
  - Distribution balancing in shared latent dimensions
    - Minority group: learn from majority group
    - Mitigate overfitting
  - Clustering effects in latent space:
    - Similar patients are close to each other
      - Easy to infer their counterfactual outcomes





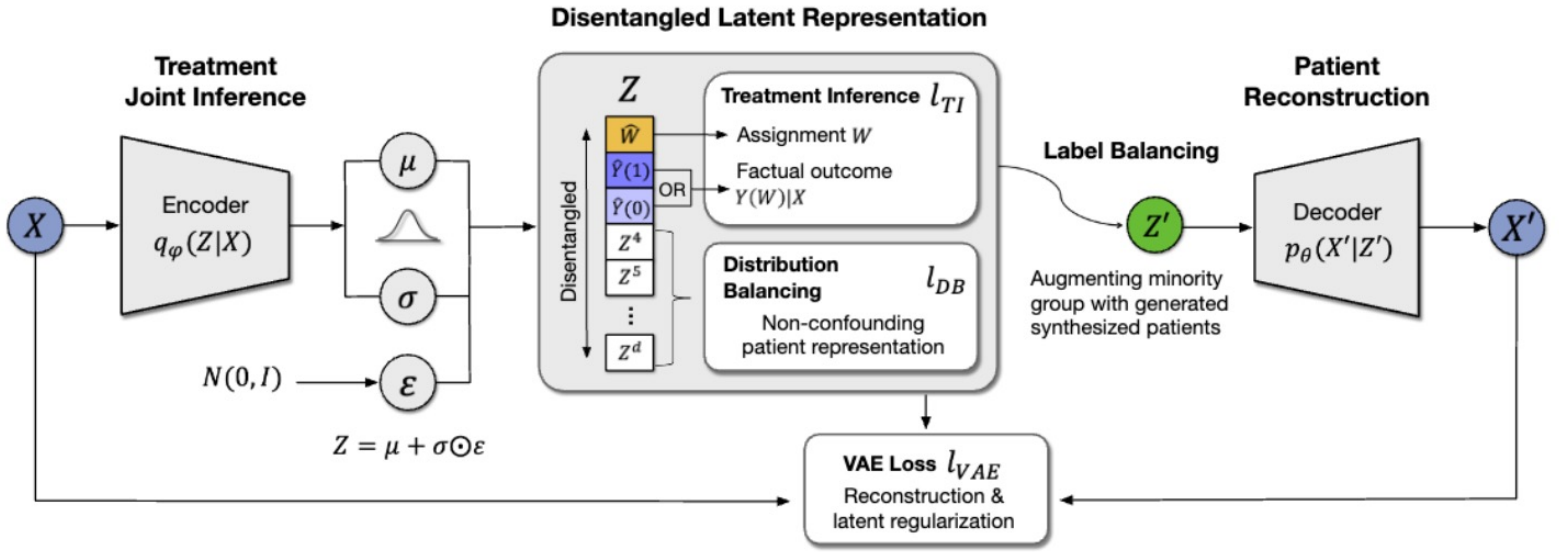
Modeling strategy

- Last, we add a generative decoder:
  - Regularize counterfactual predictions by minimizing reconstruction loss
  - Clustering effect: counterfactual predictions are pushed towards similar patients
  - Data augmentation: enrich treatment group from posterior distributions



Modeling strategy

- Key features:
  1. Generative encoder-decoder approach;
  2. Encoder: disentangled, semi-supervised latent space;
     - No auxiliary network is needed;
     - Selection bias is disentangled from other information;
  3. Decoder: reconstructs the inputs
     - Regularizes the counterfactual prediction;
     - Augment the rare ECMO cases



Modeling strategy

- Is our proposed model specific to ECMO treatment effect estimation?
  - The model features are to tackle the generic problems in healthcare
    - Partially observed inputs;
    - High-dimensional, correlated features;
    - Highly imbalanced datasets.
  - Foundational modeling approach: we can apply similar framework to many tasks/clinical decision making problems

Experiments

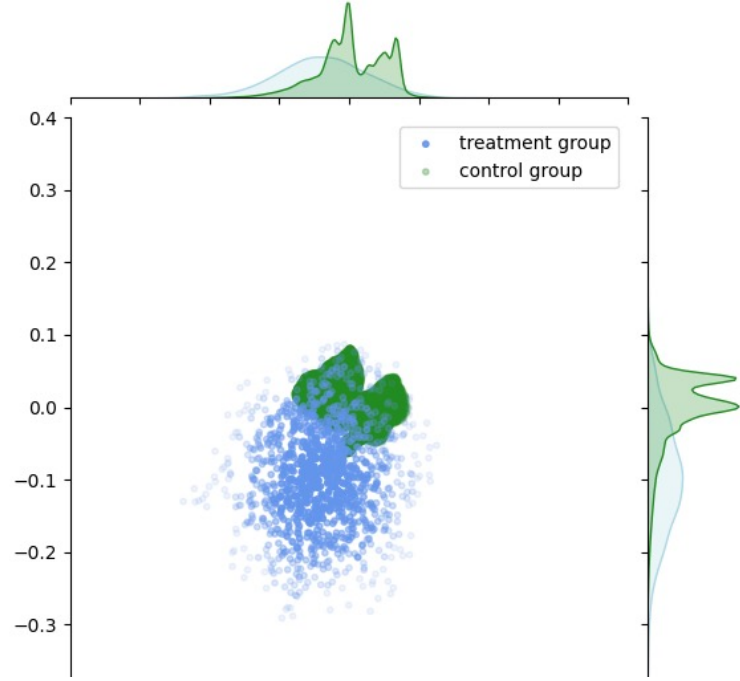
- **Dataset 1: ISARIC public dataset**
  - 118,801 ICU patients from 1651 hospitals in 63 countries
  - 1,451 ECMO treatment cases (1.2%) with 40% mortality
- **Dataset 2: Barnes Jewish HealthCare dataset**
  - 6,016 ICU patients from 15 hospitals
  - 134 ECMO treatment cases (2.2%) with 47% mortality
- **Dataset 3: IHDP public dataset**
  - Synthetic dataset of 747 subjects
  - 139 treatment cases with continuous treatment responses

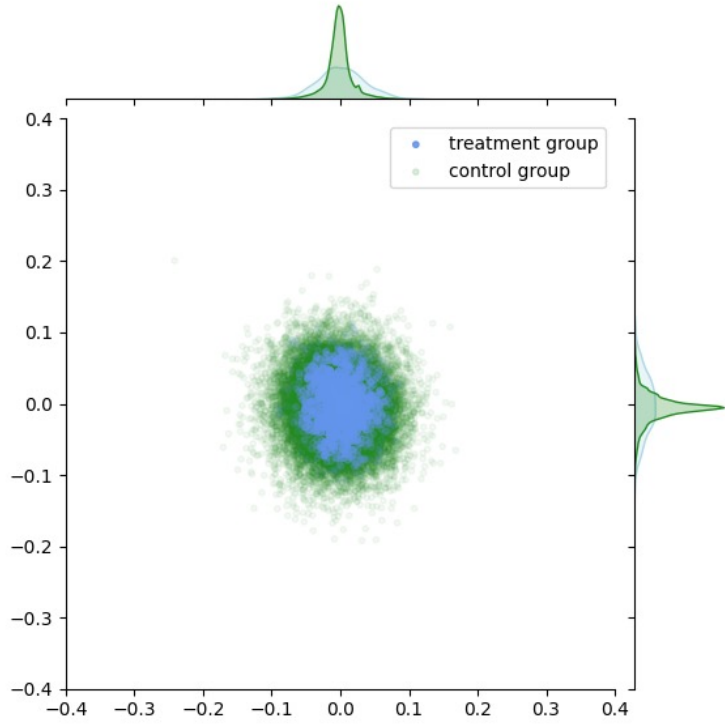
Dataset	Treatment Assignment	Treatment Response	Counterfactual Outcome	Treatment cases	Best Model
ISARIC	Binary	Binary	N.A	Rare	**Ours**
BJC	Binary	Binary	N.A	Rare	**Ours**
IHDP	Binary	Continuous	Synthesized	Not rare but imbalanced	**Ours**

Case study

How does distribution balancing construct the latent space?

- More balanced distribution;
- Death cases are sicker and have higher mortality risks even with ECMO treatment





Conclusion

- Some key features of our proposed generative modeling approach:
  - To assist with decision making under treatment resource constraints;
  - To express propensity score, factual and counterfactual responses as transformed representation of covariates;
  - To regularize counterfactual prediction through the clustering and self-reconstruction;
  - Majority help minority: to learn a better representation of minority group from distribution balancing.

Future work

- Multiple treatment options :
  - Problem:
    - What is the treatment effect for each treatment option?
    - Which treatment should be recommended?
  - Modeling direction:
    - Treatment effect estimation + recommendation system
      - Treatment effect estimation: for each treatment option, maximize the prediction accuracy of treatment outcomes
      - Recommendation system: for each patient, maximize the recall of correct treatment assignment

Future work

- Continuous treatment effect estimation: FAST project
  - Problem:
    - The patient can be discharged on any day, but we only have outcome for one day
  - Modeling direction:
    - Develop a monotonically decreasing probability curve
    - Minimze the binary cross entropy loss

