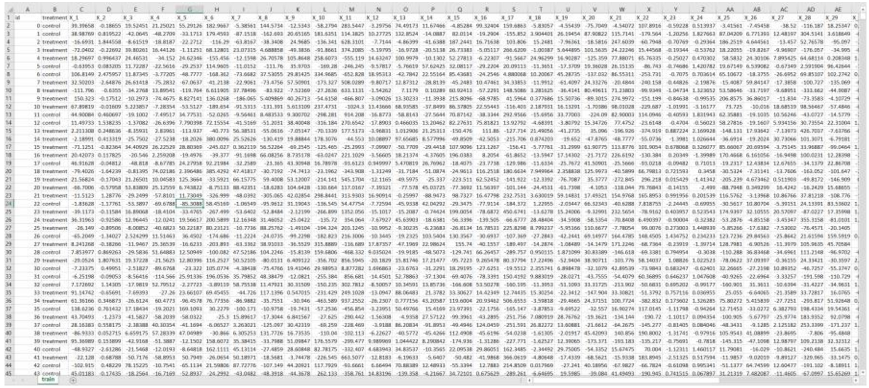
Tabular Data Prediction with Heterogeneous Features

Hedong Yan, CS, HKBU

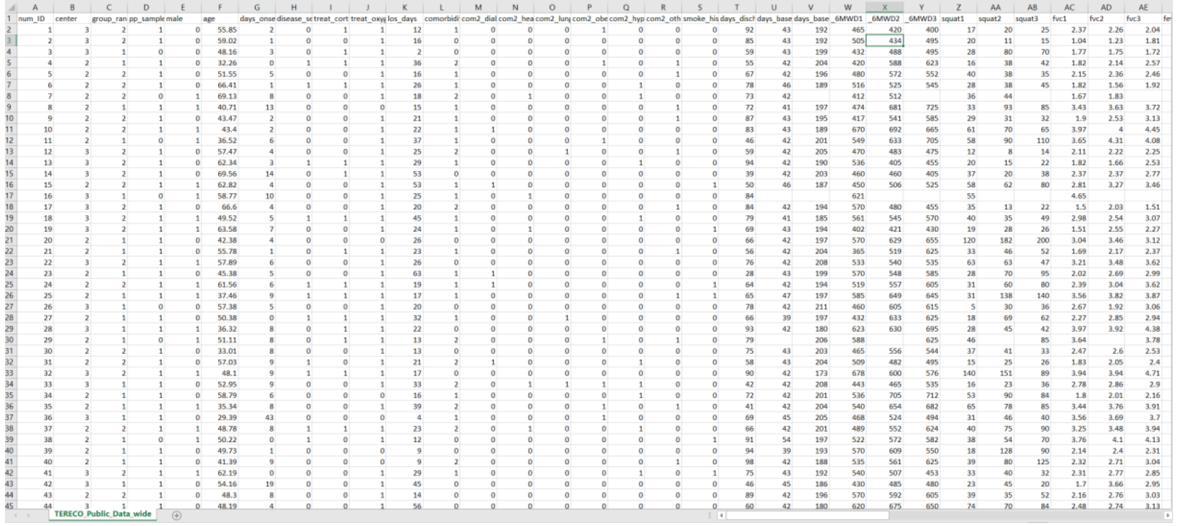
Supervisor: Prof. Yiuming Cheung

Background

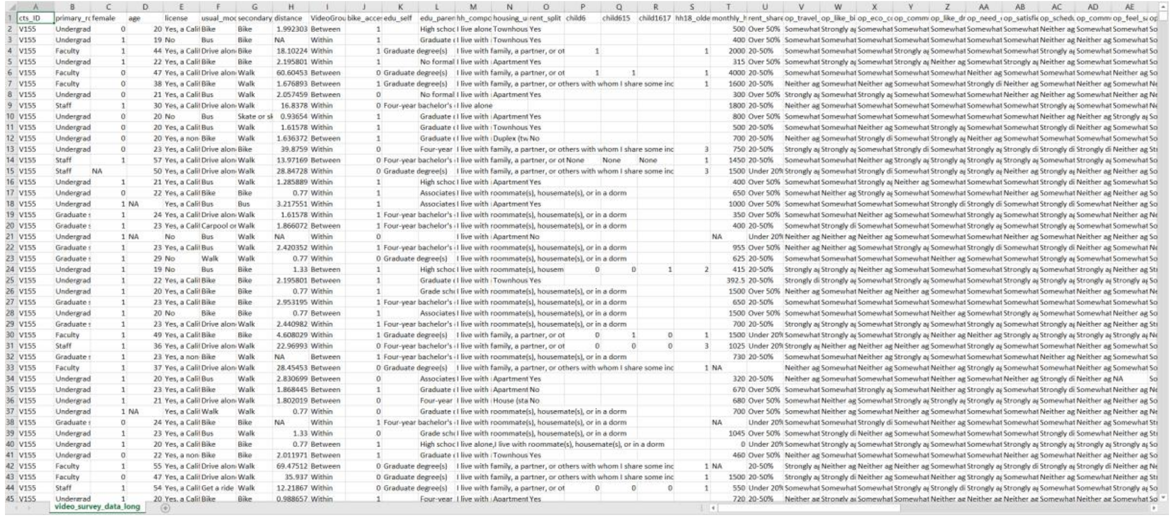
- Tabular data is widely existed in real world
  - Survey Data
  - Treatment Evaluation in Randomized Trials
  - Uplift Model (Improve User Conversion)



How to improve user conversion?



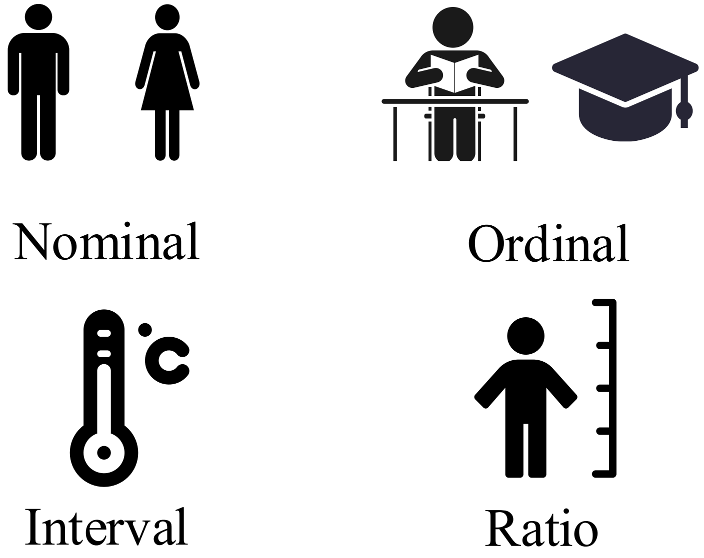
Alert can improve acute kidney injury (AKI)?



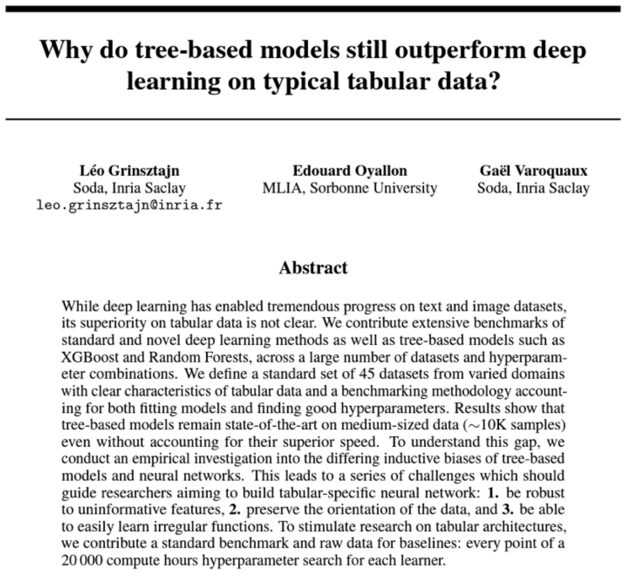
Bicycling comfort rate survey

Motivation

- Fact 1: Tabular data is often heterogenous.
- Fact 2: Neural network often NOT perform well on tabular data.



Heterogeneous Features



Why do tree-based models still outperform deep learning on typical tabular data?

Léo Grinsztajn
Soda, Inria Saclay
leo.grinsztajn@inria.fr

Edouard Oyallon
MLIA, Sorbonne University

Gaël Varoquaux
Soda, Inria Saclay

Abstract

While deep learning has enabled tremendous progress on text and image datasets, its superiority on tabular data is not clear. We contribute extensive benchmarks of standard and novel deep learning methods as well as tree-based models such as XGBoost and Random Forests, across a large number of datasets and hyperparameter combinations. We define a standard set of 45 datasets from varied domains with clear characteristics of tabular data and a benchmarking methodology accounting for both fitting models and finding good hyperparameters. Results show that tree-based models remain state-of-the-art on medium-sized data (~10K samples) even without accounting for their superior speed. To understand this gap, we conduct an empirical investigation into the differing inductive biases of tree-based models and neural networks. This leads to a series of challenges which should guide researchers aiming to build tabular-specific neural network: **1.** be robust to uninformative features, **2.** preserve the orientation of the data, and **3.** be able to easily learn irregular functions. To stimulate research on tabular architectures, we contribute a standard benchmark and raw data for baselines: every point of a 20 000 compute hours hyperparameter search for each learner.

Deep learning does NOT perform well

Problem: How to address on heterogeneous features effectively?

Related works (unsupervised)

One-hot

SEX
Male
Female

Male	Female
1	0
0	1

Ordinal

RATE
Bad
Neutral
Good

RATE
0
1
2

Rank-hot

RATE
Bad
Neutral
Good

Bad	Neutral	Good
1	0	0
1	1	0
1	1	1

Piece-wise linear

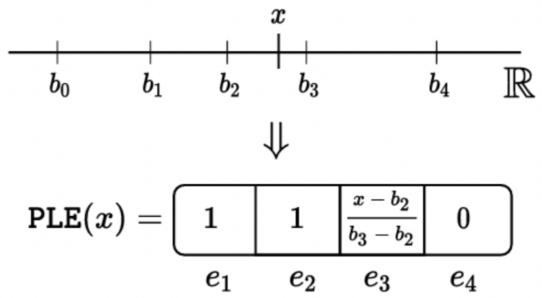


Figure 1. The piecewise linear encoding (PLE) in action, as defined in Equation 4. In the example, $T = 4$.

Related works (supervised)

Periodic

$$
\begin{aligned}
f_i(x) &= \texttt{Periodic}(x) = \texttt{concat}[\sin(v), \cos(v)], \\
v &= [2\pi c_1 x, \ldots, 2\pi c_k x]
\end{aligned} \tag{8}
$$

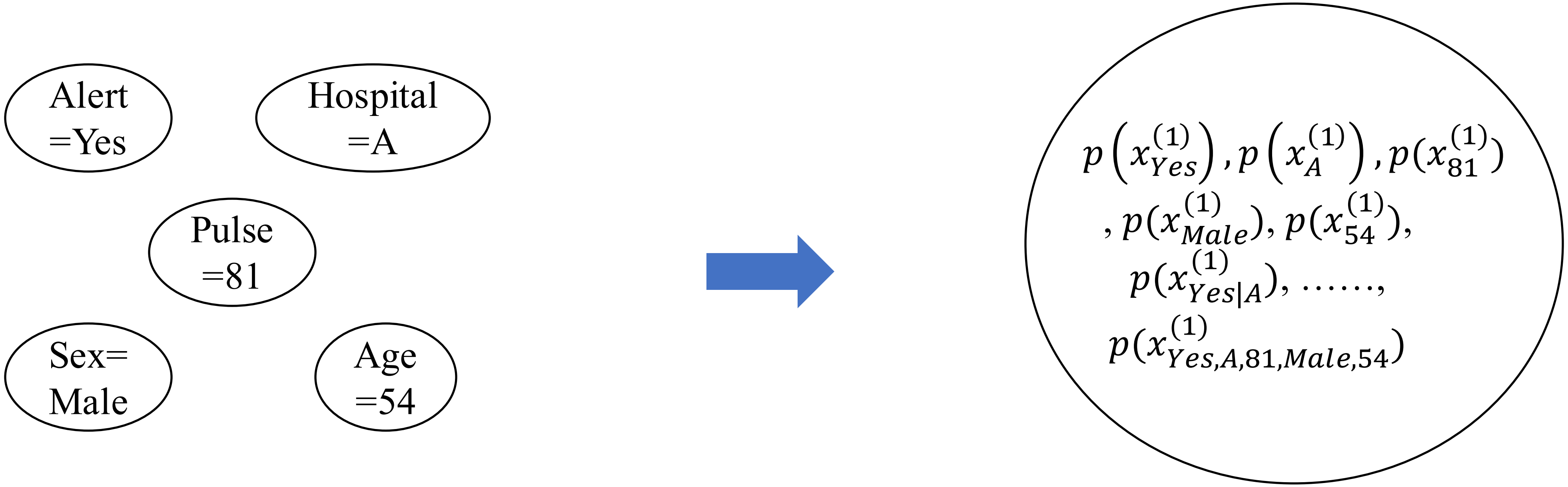
where c_i are trainable parameters initialized from $\mathcal{N}(0, \sigma)$. σ is an important hyperparameter that is tuned using validation sets.

Target Statistic

$$
x_k^i = \frac{\sum_{j=1}^{n} \mathbb{I}_{x_j^i = x_k^i} * y_j + ap}{\sum_{j=1}^{n} \mathbb{I}_{x_j^i = x_k^i} + a} \tag{1}
$$

Methodology

Idea



Instance in Heterogeneous Features Space H **Instance in Measurement Occurrence Space S**

Feature can be very heterogeneous. We list all potential measurements of an individual and use the measurement probabilities as new feature space where each axis is a potential measurement.

Methodology

$p\left(x_{Yes}^{(1)}\right), p\left(x_{A}^{(1)}\right), p(x_{81}^{(1)})$
$, p(x_{Male}^{(1)}), p(x_{54}^{(1)}),$
$p(x_{Yes|A}^{(1)}), \ldots\ldots,$
$p(x_{Yes,A,81,Male,54}^{(1)})$

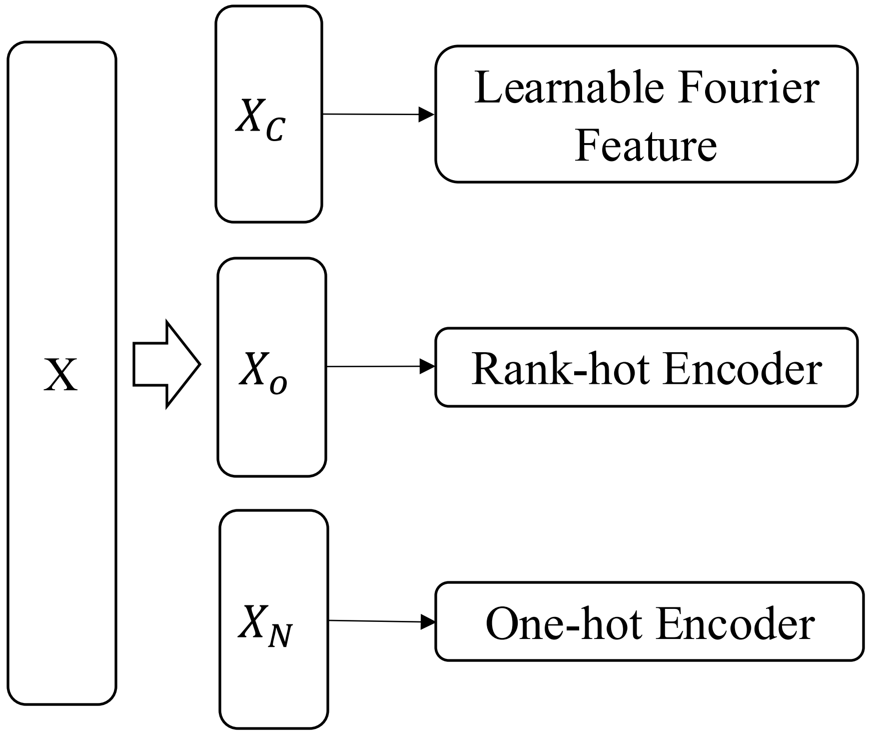
Measurement Occurrence Space S

Challenge 1: the dimensions of Space S is combinational.
Solution: We use subspaces S1,..., Sm that were separated by features and concatenate them together.

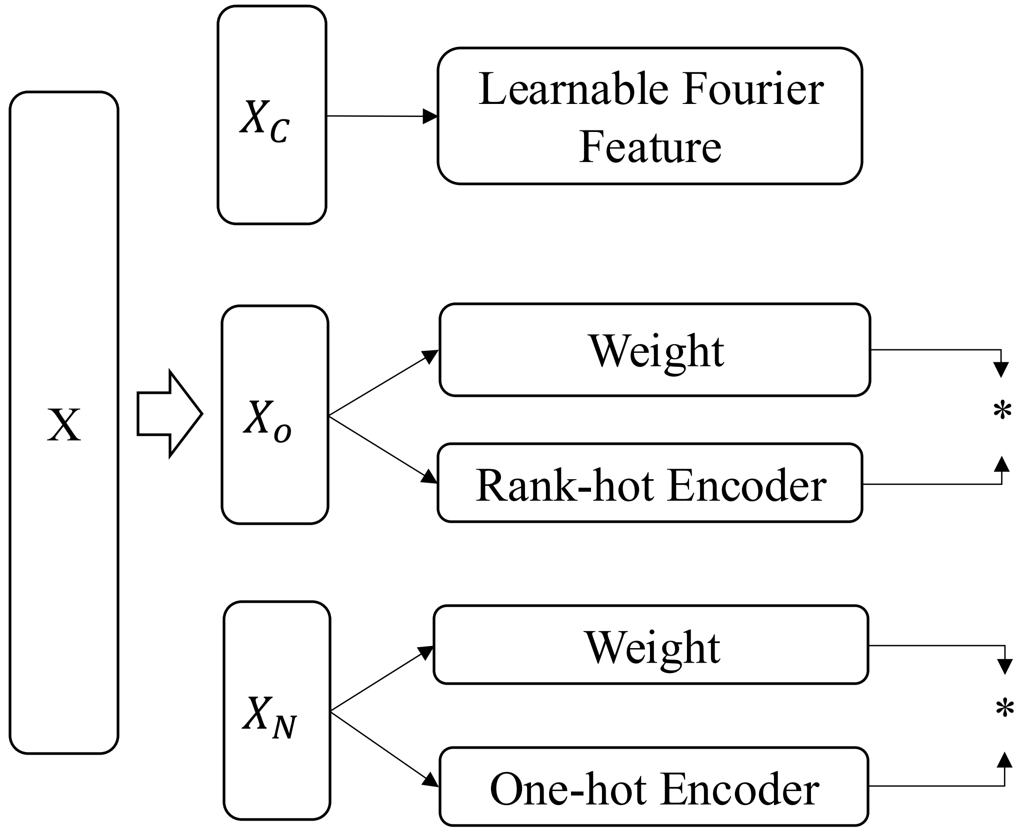
Challenge 2: the number of probabilities is combinational.
Solution: We use zero-order probabilities $p\left(x_{Yes}^{(1)}\right), p\left(x_{A}^{(1)}\right), p(x_{81}^{(1)}), p(x_{Male}^{(1)}), p(x_{54}^{(1)})$ as approximation.

Methodology

Encoder of A Trivial MLP



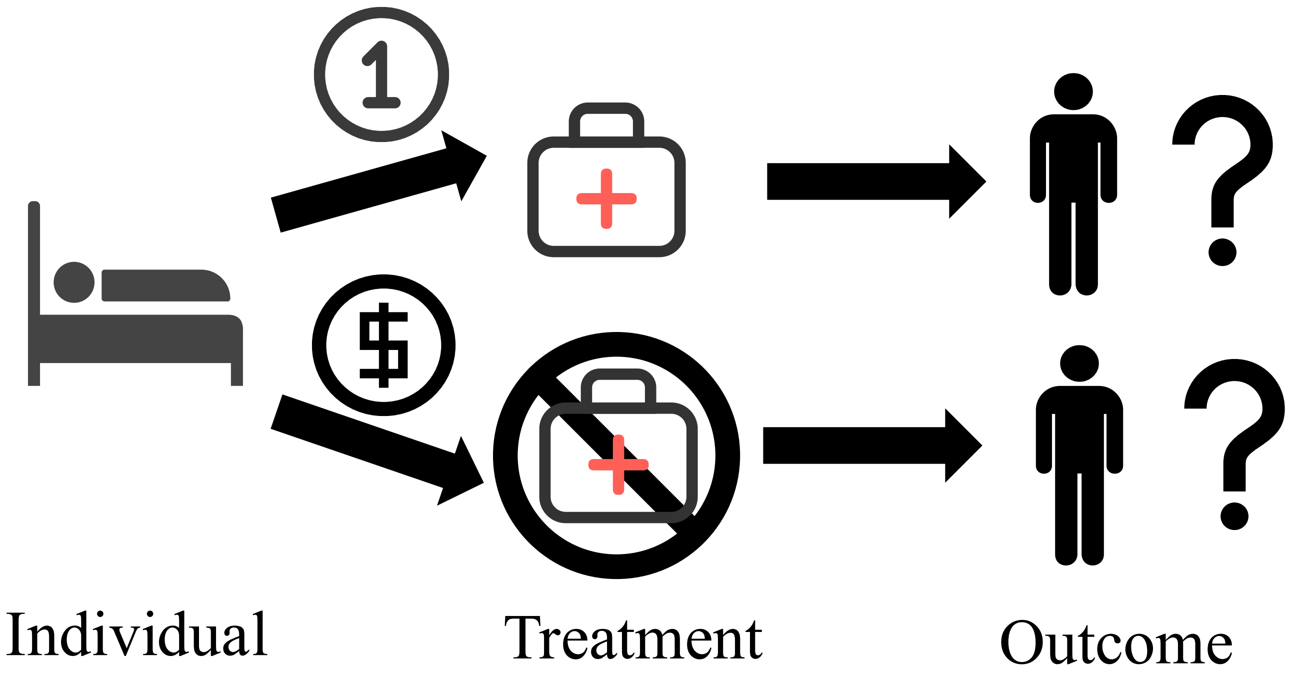
Encoder of HetMLP



Finally, we get a weighted encoder for heterogeneous features where weight is occurrence's inverse probability. It can better deal with rare event.

Experiment

Outcome Prediction Task For RCT



- **Task Goal**
  Predict Outcome in Randomized Control Trial
- **Meaning**
  We can compare the predicted outcome difference between different treatment for an individual to decide whether a patient should accept the treatment.
- **Metric**
  Mean Average Precision for Classification
  Mean Squared Error for Regression

Assumption: for an individual, if a unbiased model can predict its factual outcome in RCT better, then it can predict its counterfactual outcome in RCT better.

Experiment

Our Gathered Dataset

TABLE IV: Heterogeneous datasets for outcome prediction

Dataset	Instance	Outcome	Treatment
Safety and Preliminary Efficacy of Intranasal Insulin for Cognitive Impairment in Parkinson Disease and Multiple System Atrophy	16	Parkinson disease	Intranasal insulin
	https://physionet.org/content/inipdmsa/1.0/		
Tai Chi, Physiological Complexity, and Healthy Aging - Gait	60	Gait and EMG data	Tai Chi
	https://physionet.org/content/taichidb/1.0.2/		
ECG Effects of Dofetilide, Moxifloxacin, Dofetilide+Mexiletine, Dofetilide+Lidocaine and Moxifloxacin+Diltiazem	22	ECG	Dofetilide, Moxifloxacin, Dofetilide+Mexiletine, Dofetilide+Lidocaine and Moxifloxacin+Diltiazem
	https://physionet.org/content/ecgdmmld/1.0.0/		
ECG Effects of Ranolazine, Dofetilide, Verapamil, and Quinidine	22	ECG	Ranolazine, Dofetilide, Verapamil, and Quinidine
	https://physionet.org/content/ecgrdvq/1.0.0/		
CAST RR Interval Sub-Study Database	734	Cardiac arrhythmia suppression	Encainide, flecainide, moricizine (antiarrhythmic drugs) or a placebo
	https://physionet.org/content/crisdb/1.0.0/		
Randomized trial of AKI alerts in hospitalized patients	6030	Acute Kidney Injury	Electronic AKI alert versus usual care
	https://datadryad.org/stash/dataset/doi:10.5061%2Fdryad.59zw3r27n		
Telerehabilitation program for COVID-19 survivors (TERECO) - Randomized controlled trial	120	Exercise capacity, lower-limb muscle strength (LMS), pulmonary function, health-related quality of life (HRQOL), and dyspnoea	Telerehabilitation program for COVID-19 survivors
	https://datadryad.org/stash/dataset/doi:10.5061%2Fdryad.59zw3r27n		
Bicycling comfort video experiment	15289	Bicycle rating	Video Type
	https://datadryad.org/stash/dataset/doi:10.25338%2FB8KG77		
Megafon uplift competition	1.5 million	User conversion	Exposure
	https://ods.ai/tracks/df21-megafon/competitions/megafon-df21-comp/data		
Infant Health and Development Program	1090	Cognitive development, Behavior problems, Health status	Home visits, attendance at a special child development center
	https://www.icpsr.umich.edu/web/HMCA/studies/9795		
National Supported Work Evaluation Study	6600	effects of the Supported Work Program	Offered a job in supported work
	https://www.icpsr.umich.edu/web/ICPSR/studies/7865		
CPAP Pressure and Flow Data from a Local Trial of 30 Adults at the University of Canterbury	30	Breathing	Continuous positive airway pressure
	https://physionet.org/content/cpap-data-canterbury/1.0.1/		

- Most of those tabular datasets are heterogeneous features.
- More detail can be seen at: https://github.com/herdonyan/RandomizedTrialDataset

Experiment

Alert2AKI Dataset

Intervention	AKI Alert or Not
Main-outcome	AKI Progression in 14 Days
Pre-treatment	EHR Records
Patients Num	6030 in 5 Hospitals (5082/948)

SCALE	**NUM**
Nominal	9
Ordinal	19
Interval	3
Ratio	20

PR-AUC (5 Random Splits)

HetMLP	Trivial MLP	MLP	Random
.2117±.0009	.2087±.0164	.2009±.0329	.1568±.0089

Our HetMLP got **1.43%** performance up compared with Trivial MLP.

AKI: Acute Kidney Injury

Will the patients be benefited from the alert?

Splitting 1

Metrics	Num
Patients Num	3536
Benefited: AKI=1→AKI=0	15
Harmful: AKI=0→AKI=1	14

Splitting 2

Metrics	Num
Patients Num	3552
Benefited: AKI=1→AKI=0	8
Harmful: AKI=0→AKI=1	2

Splitting 3

Metrics	Num
Patients Num	3504
Benefited: AKI=1→AKI=0	26
Harmful: AKI=0→AKI=1	4

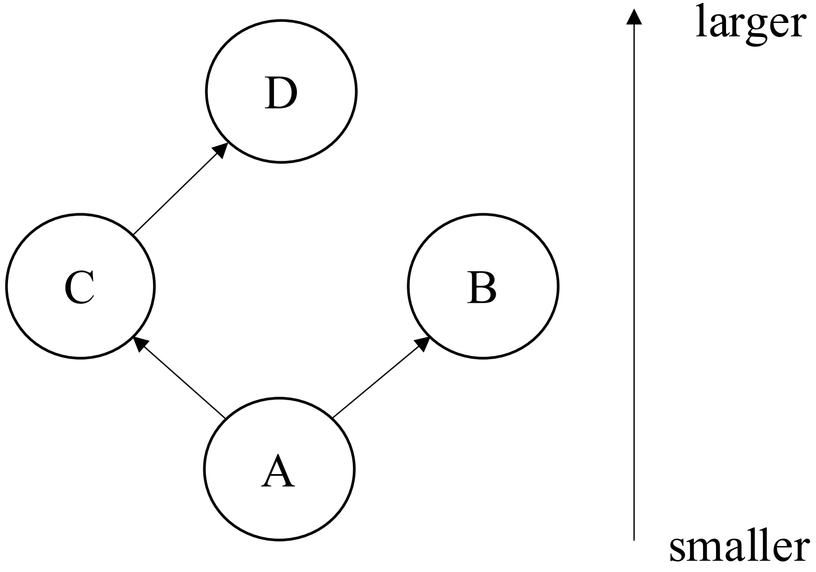
Splitting 4

Metrics	Num
Patients Num	3536
Benefited: AKI=1→AKI=0	9
Harmful: AKI=0→AKI=1	9

Futural Plan

1. Add more models and datasets for further detailed comparison in experiment
2. Consistency constrain
3. Extend to time-series data

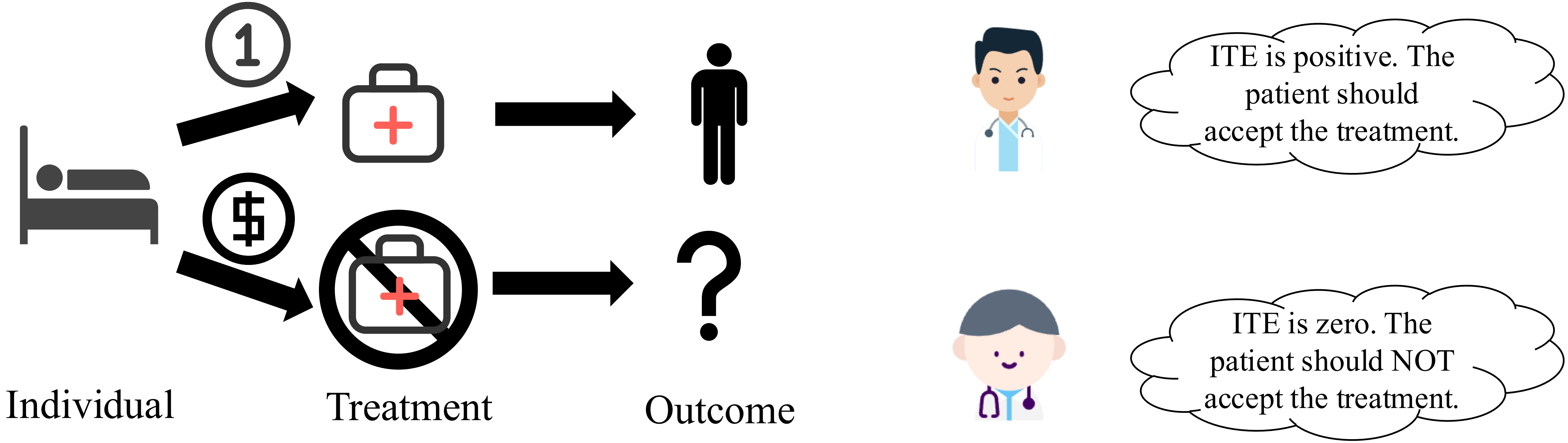
Motivation: Evaluate Individual Treatment Effect
Fundamental Problem: counterfactual is unknown
Methodology:



M1,M2对factual的预测是无偏的，对counterfactual的预测是无偏的，所有预测误差服从高斯分布，其中M1和M2对于counterfactual的预测方差相比factual的预测方差较大，则在左图假设下，在factual上具有较低MSE的模型以接近1的概率具有较低的ITE误差,
因此，可以通过Factual上的MSE来评估ITE。

A = MAE of M1 for factual
B = MAE of M1 for counterfactual
C = MAE for M2 for factual
D = MAE for M2 for counterfactual

Individual Treatment Effect Evaluation



How to evaluate causal models for ITE estimation task?

Fundamental Problem: only one clinical ending can be measured

Problem Formulation

- Population Evaluation: Given $M_1: (A, X) \to Y$, $M_2: (A, X) \to Y$, test dataset (A,X,Y,F=1) where A is randomized, compare $MSE\left(M(1, X) - M(0, X) - \left(Y(1) - Y(0)\right)\right)$ of M_1 and M_2 where only one of $Y_i(1), Y_i(0)$ is given.
- Individual Evaluation: Given $M_1: (A, X) \to Y$, $M_2: (A, X) \to Y$, individual (1,x,y,F=1) where A is randomized, compare $|M(1, x) - M(0, x) - \left(y - Y(0)\right)|$ of M_1 and M_2 where $Y(0)$ is not given.

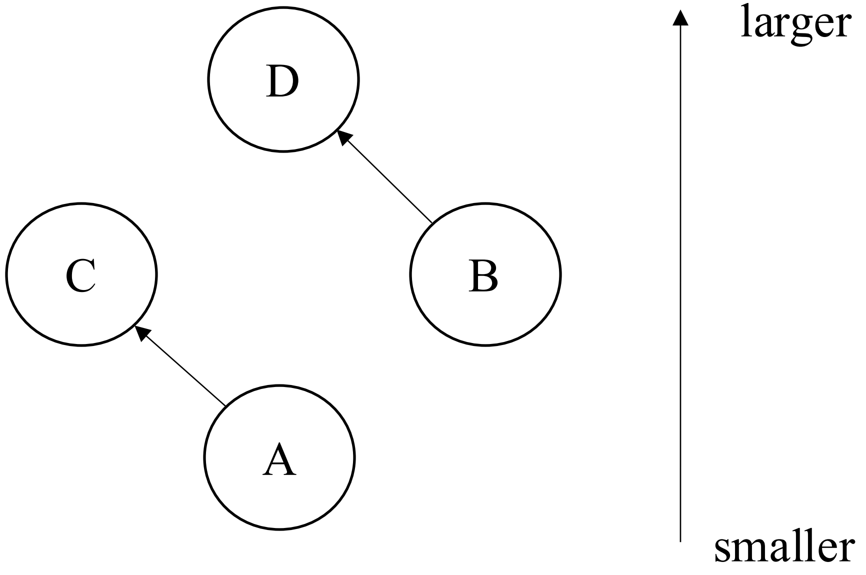
Methodology

- M_1, M_2 is unbiased on factual dataset and counterfactual dataset
- Factual individual (Y can be measured) and counterfactual individual (Y can not be measured) with same randomized treatment A follows identical distribution
  - $MSE(M_1,[F]) < MSE(M_1,[F]) \rightarrow MSE(M_1,[F,CF]) \underset{\sim}{<} MSE(M_2,[F,CF])$ with high probability (≥95%) as n increase $n \geq \frac{16}{((\frac{MSE(M_2)}{MSE(M_1)})^2-1)^2}$, $n = 77$ if ratio is 1.1

- So, we can use MSE on randomized factual data to evaluate ITE

Thanks!

Methodology



A = Testing MSE of M1 for factual
B = Testing MSE of M1 for counterfactual
C = Testing MSE for M2 for factual
D = Testing MSE for M2 for counterfactual

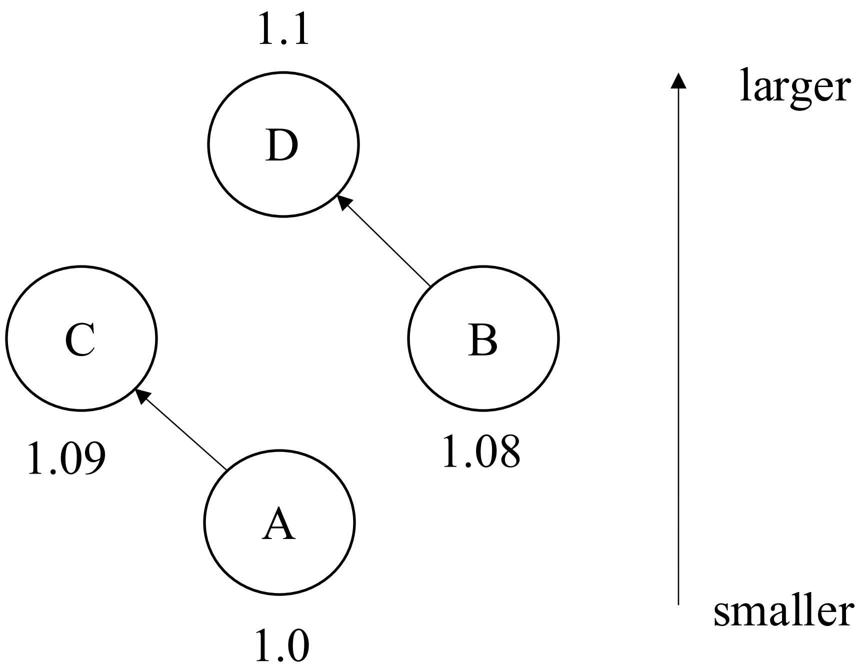
Assumption

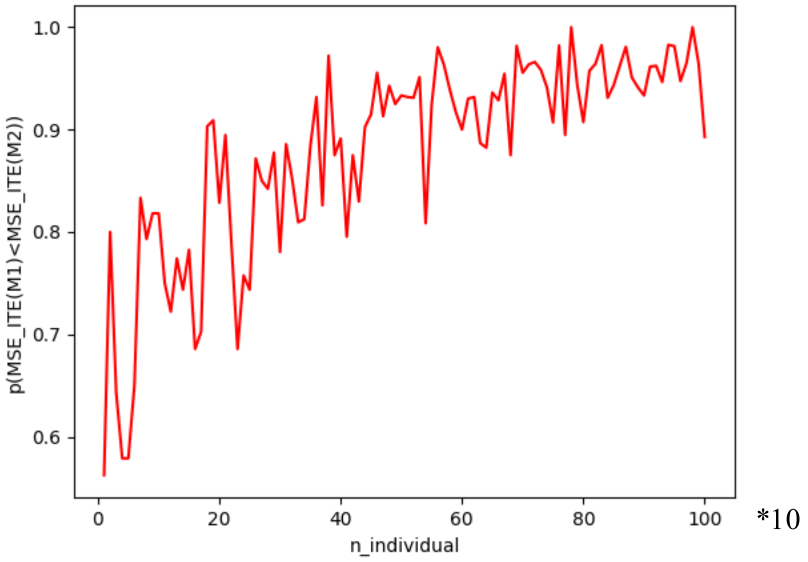
M1, M2 is unbiased on factual data

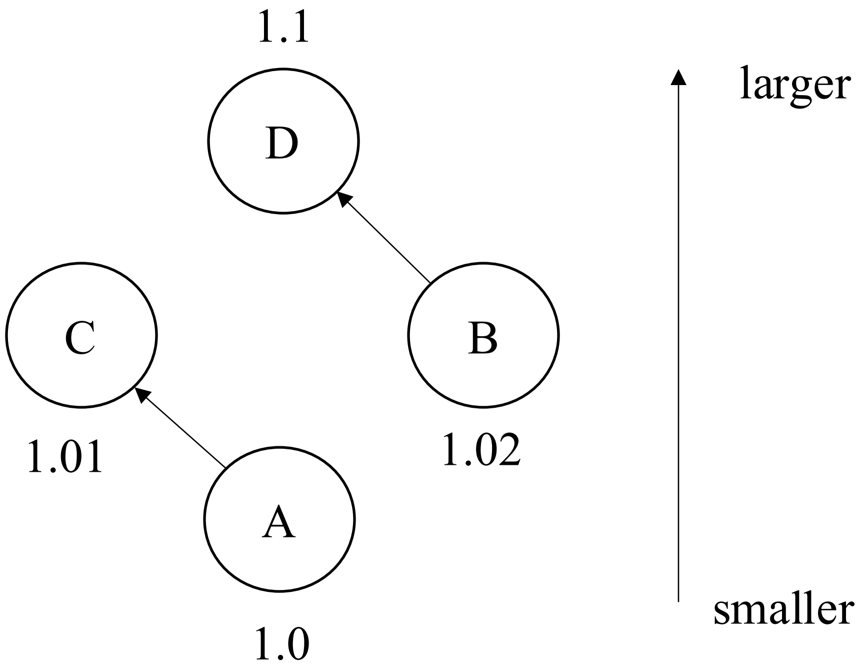
M1, M2 is unbiased on counterfactual data

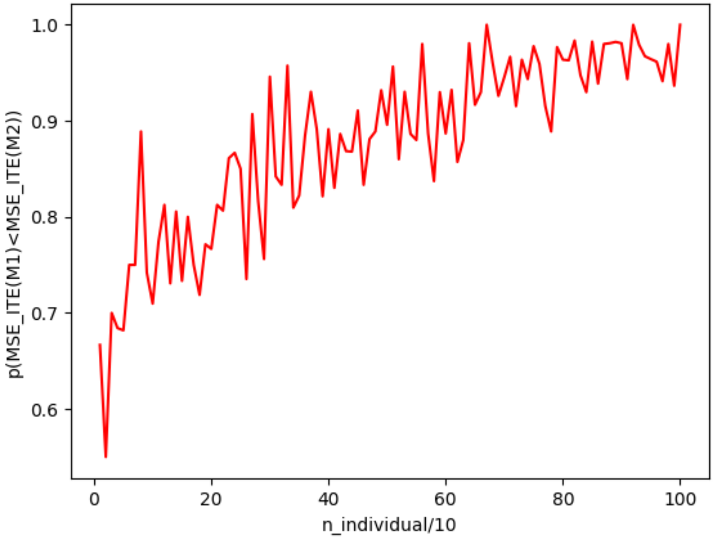
A < C -> A+B < C+D

Errors are both sampling from Gaussian


1.1
D
C
B
1.09
1.08
A
1.0
larger
smaller


p(MSE_ITE(M1)<MSE_ITE(M2))
1.0
0.9
0.8
0.7
0.6
0
20
40
60
80
100
*10
n_individual


1.1
D
C
B
1.01
1.02
A
1.0
larger
smaller


p(MSE_ITE(M1)<MSE_ITE(M2))
1.0
0.9
0.8
0.7
0.6
0
20
40
60
80
100
n_individual/10

Conclusion

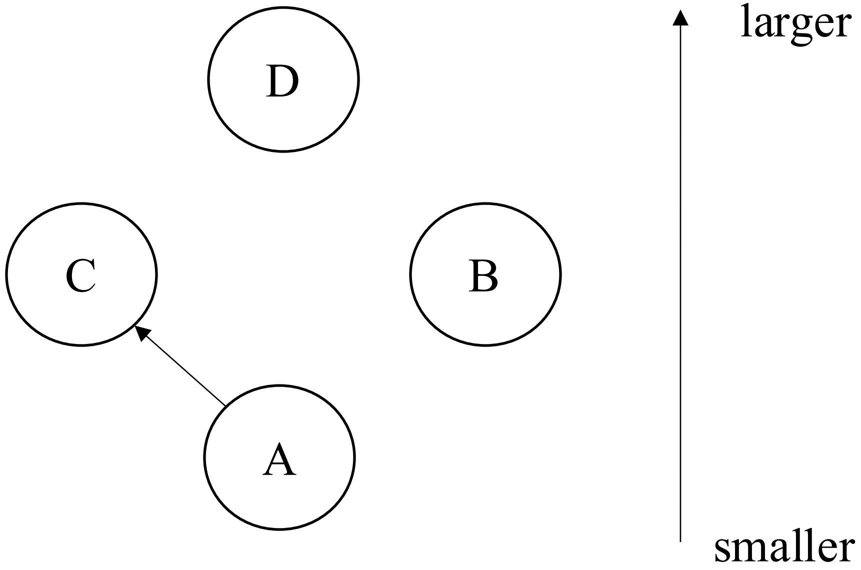
容易证明

当n趋向无穷，
MAE(ITE_M1)<MAE(ITE_M2)的概率是1，
收敛速率和sqrt(A+B)/(C+D)线性相关

因此，评估模型时不需要知道反事实输出，只需要计算事实数据上的预测的MSE误差对A和C进行ite估计进行评估

- N>=16/((C+D)/(A+B))
- $N \geq \frac{16}{\frac{C+D}{A+B}^{2}-1^{2}}$，95%置信度
- 如果比值为1.1，仅需要77个样本，比值越大需要的样本越少

进一步解释和弱化假设



A = Testing MSE of M1 for factual
B = Testing MSE of M1 for counterfactual
C = Testing MSE for M2 for factual
D = Testing MSE for M2 for counterfactual

A < C -> A+B < C+D

- 对于任意模型，接受治疗和不治疗的人数相同，对于测试集中**全部个体估计treatment=1时的结局时，假设治疗组个体和不治疗组个体的误差为同分布**
- 这是因为
  - 1、test中的treatment是被随机分配的，因此对于测试中treatment相同的个体以及反事实个体(A=1,X,Y)可以认为在同一个分布
  - 2、test的factual数据(A,X,Y)是训练时没见过的，而对应的counterfactual数据(!A,X,Y)也是训练时没见过的，所以他们的均方误差相同

因此，对于测试集，A < C 推出 A+B < C+D是可靠的假设