Breakout Session 8: Track B

Enhancing Imputation for Clinical Trials: The Path for a Flexible Toolkit

Dr. Vida Abedi

Associate Professor, Penn State University

Enhancing Imputation for Clinical Trials: The Path for a Flexible Toolkit

Vida Abedi, PhD, Alireza Vafaei Sadr, PhD, and Vernon M. Chinchilli, PhD

Department of Public Health Sciences
College of Medicine, Penn State University

Type 1 Diabetes in Acute Pancreatitis Consortium (T1DAPC)

2024 NIH ODSS AI Supplement Program Virtual PI Meeting - FY23 NOT-OD-23-082 program

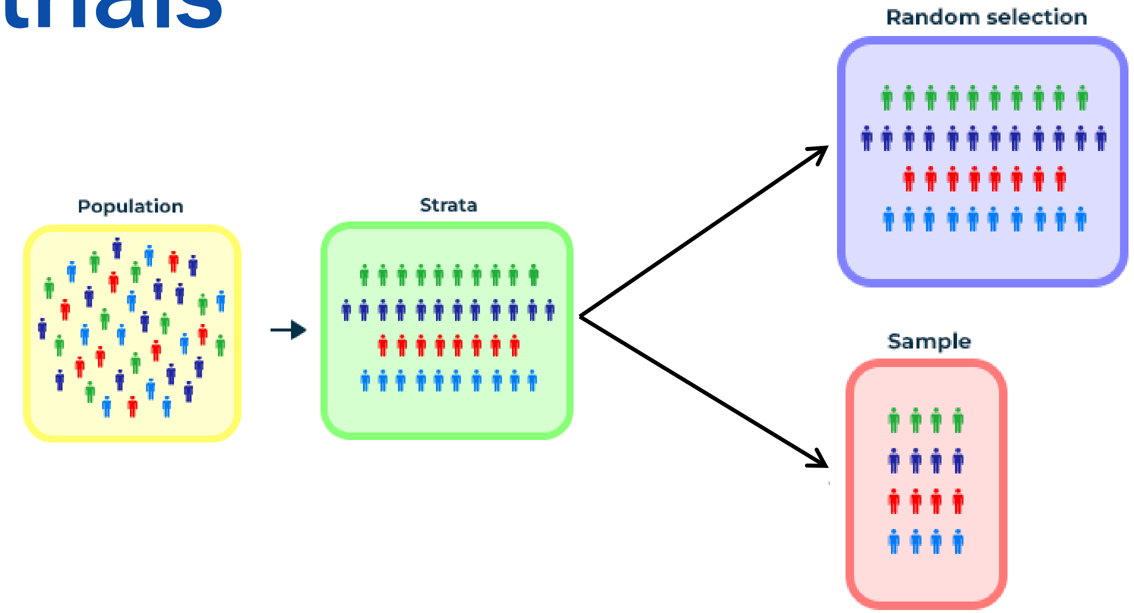
PennState

March 27-28, 2024

Outline

- Project Motivation
- Plan
- Expected outcome

Missing data in clinical trials



Randomization alone might not be enough.

Additional requirements for an unbiased study are:

1) **missing data from randomized patients do not bias the comparison of interventions** and
2) outcome assessments are obtained in a similar and unbiased manner for all patients.

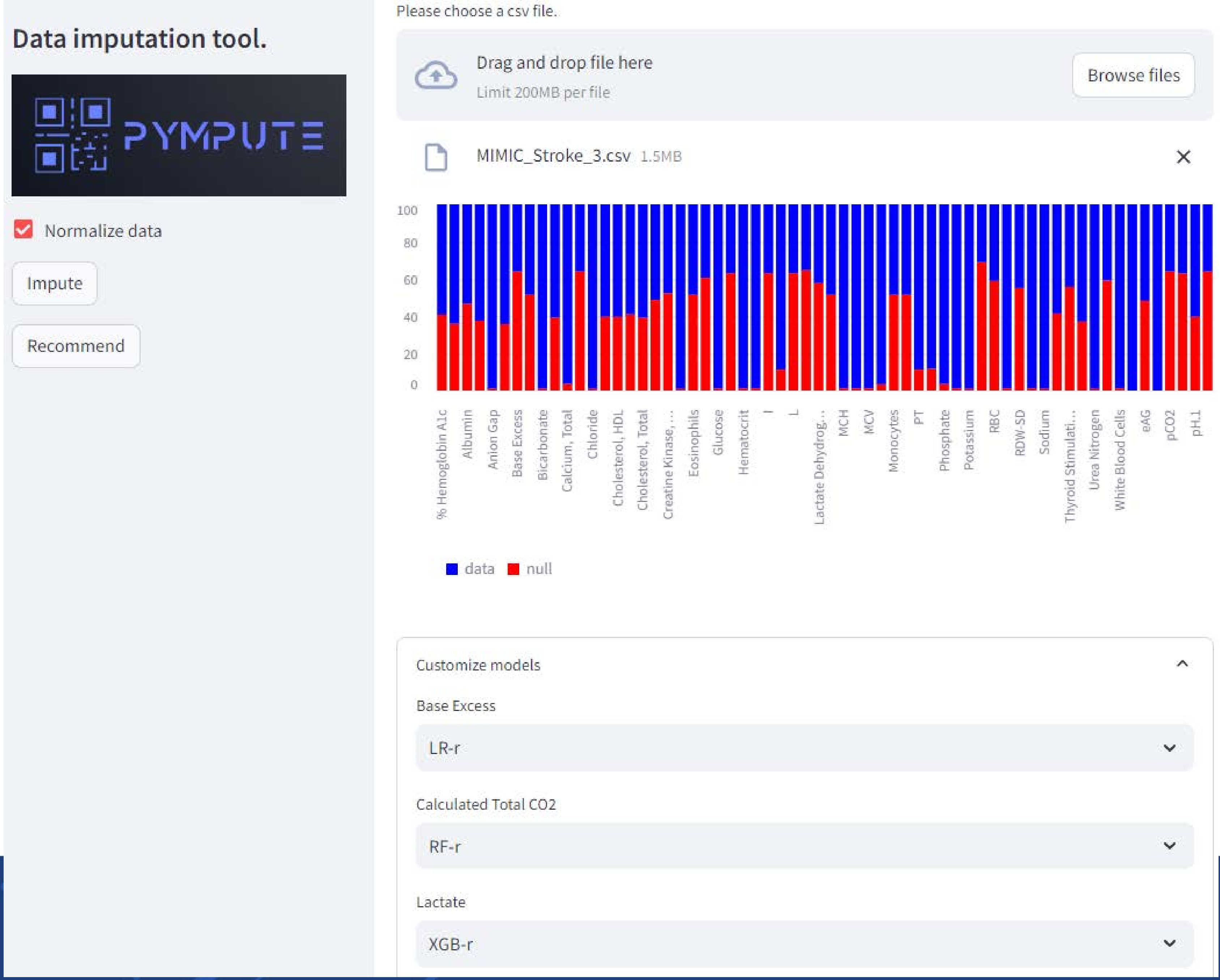
Missing data influences the Results

Various imputation techniques

- Replace the missing value by:
  - **Mean (Very common)**
  - **Median(Very common)**
  - Zero fill
- Performing multiple imputations (ex: by mean matching)
- Last observation carried forward
- Worst observation carried forward
- Likelihood estimation
- More advanced ML-based methods to estimate missing value

Pympute

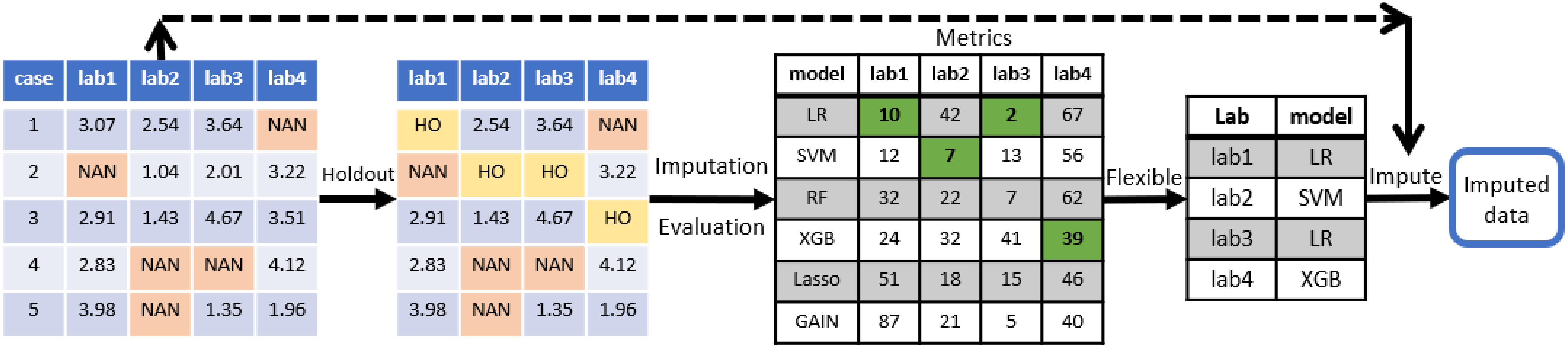
We have developed a web app designed specifically for clinical data from Electronic Health Records (EHR)



PennState

Imputation algorithm is recommended based on data distribution/observations

→ a FLEXIBLE algorithm



case	lab1	lab2	lab3	lab4
1	3.07	2.54	3.64	NAN
2	NAN	1.04	2.01	3.22
3	2.91	1.43	4.67	3.51
4	2.83	NAN	NAN	4.12
5	3.98	NAN	1.35	1.96

lab1	lab2	lab3	lab4
HO	2.54	3.64	NAN
NAN	HO	HO	3.22
2.91	1.43	4.67	HO
2.83	NAN	NAN	4.12
3.98	NAN	1.35	1.96

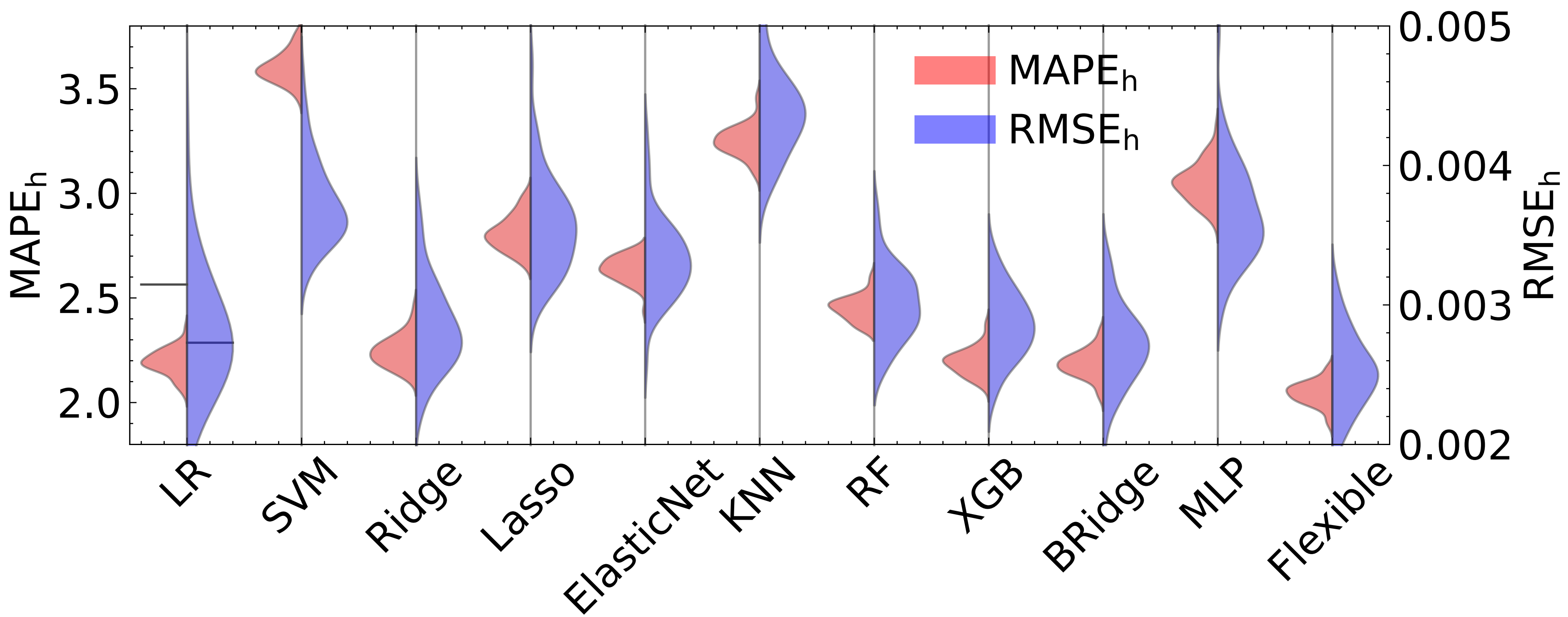
Metrics

model	lab1	lab2	lab3	lab4
LR	**10**	42	**2**	67
SVM	12	**7**	13	56
RF	32	22	7	62
XGB	24	32	41	**39**
Lasso	51	18	15	46
GAIN	87	21	5	40

Lab	model
lab1	LR
lab2	SVM
lab3	LR
lab4	XGB

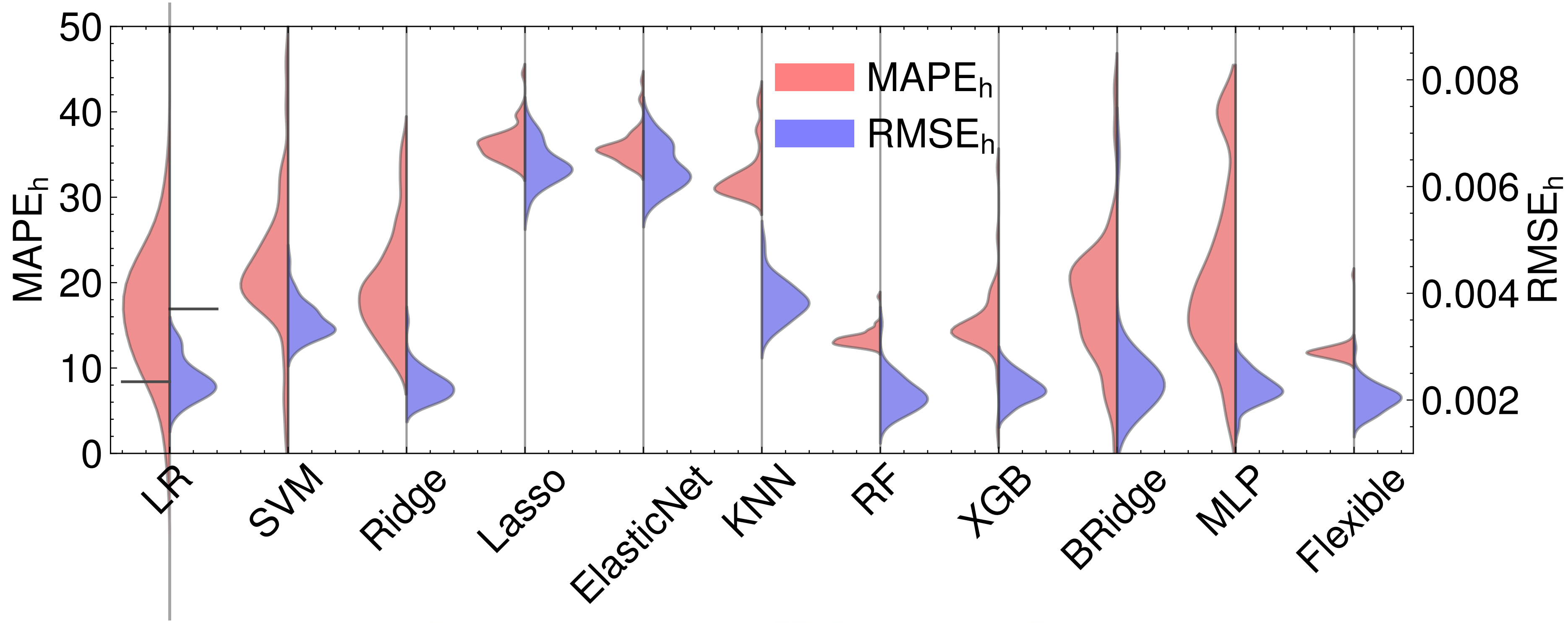
As expected, a FLEXIBLE algorithm outperforms any other algorithm (based on two error metrics)

Using clinical data from MIMIC dataset.



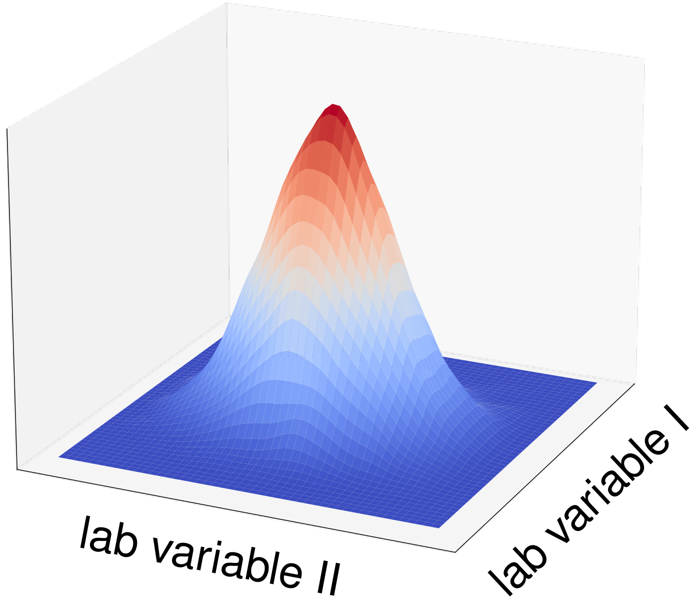
As expected, a FLEXIBLE algorithm outperforms any other algorithm (based on two error metrics)

Using clinical data from Penn State EHR.

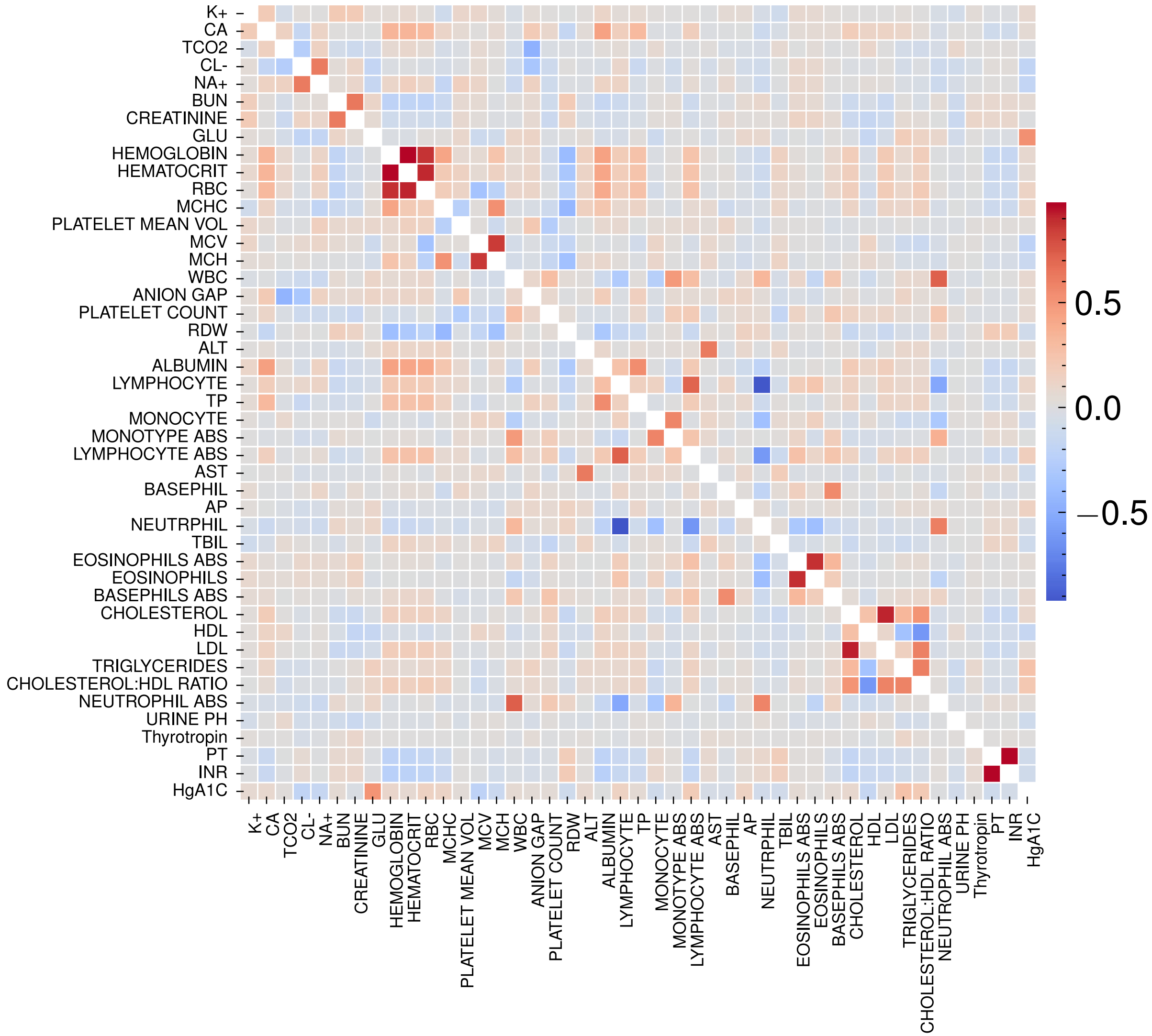


Simulate data based on EHR data from Geisinger

Multivariate normal distribution



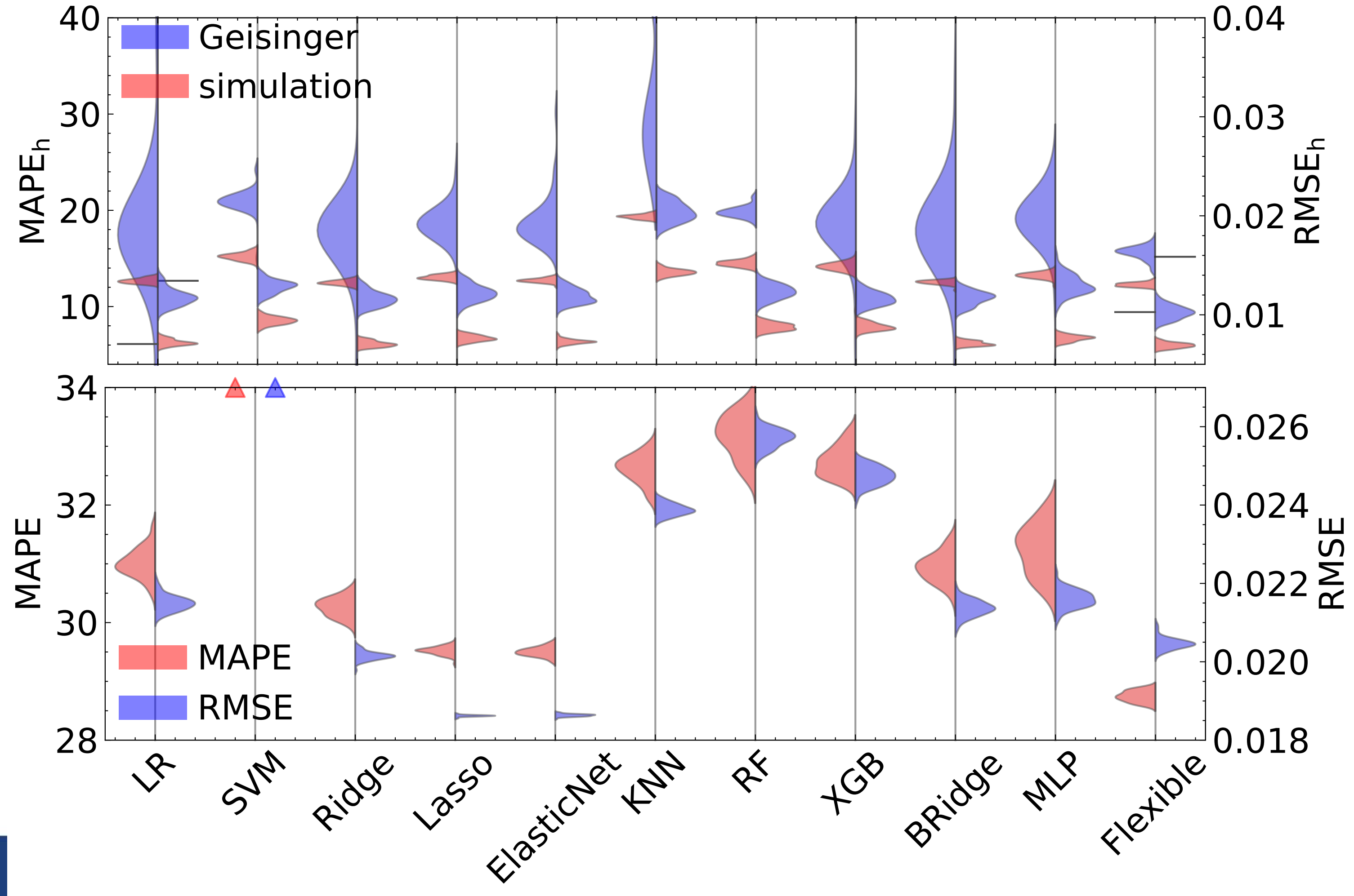
$$N(x|\mu,\Sigma) \triangleq \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$



PennState

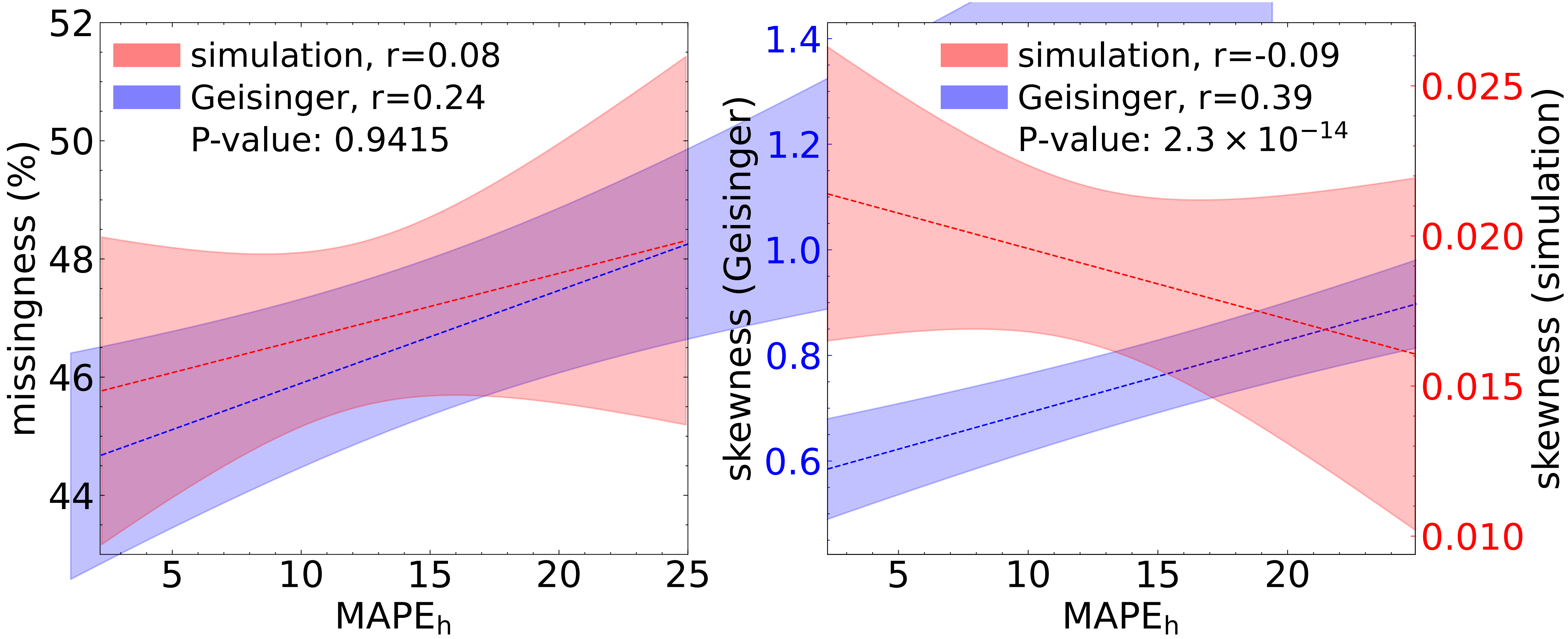
Comparing Geisinger vs. Simulated data

- Flexible finds best options for both Geisinger and Simulated data
- **Results are much better when using simulated data**→caution when studies only report results using simulated data



PennState

Missingness and skewness impact on performance



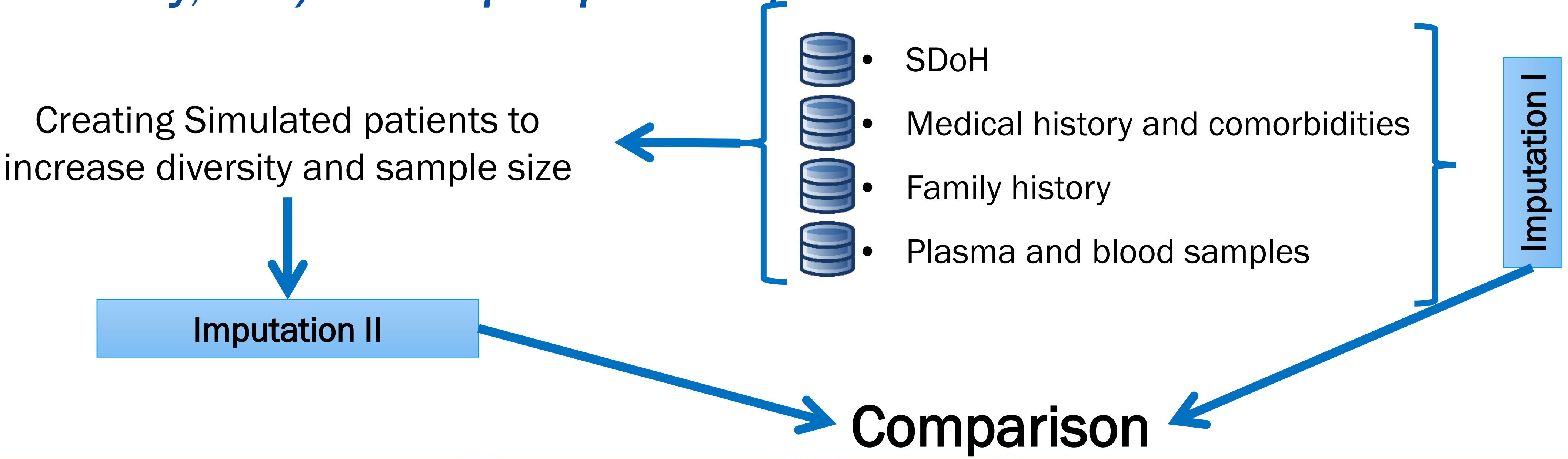
PennState

PLAN

- *Evaluating various imputation strategies*

> ***Evaluating if imputation results can be improved when clinical trial data is augmented/enriched with simulated patient data***

> ***Evaluating if inclusion of other variables (such as SDoH, past medical history, etc.) can help improve imputation of clinical trial data***



Expected Outcomes

- Missing of certain features/variables will not be at random
- Certain features/variables are expected to be missing in a specific group of patient population
- Improving imputation will improve prognosis/diagnosis prediction
- Simulated data can aid in improving imputation results
- A user-friendly tool to help impute clinical trial data

Questions