Missing Values and Imputation in Healthcare Data:
Can Interpretable Machine Learning Help?

Written by Zhi Chen, Sarah Tan, Urszula Chajewska, Cynthia Rudin, Rich Caruana

Overview

- Motivation
- Related Work
- Definitions
- Testing for MCAR Using EBM
- Missing Values in Healthcare
- Imputation of Missing Values
- Conclusion

Motivation

- **Importance of Handling Missing Values in Healthcare Data**:
  - Common issue in healthcare datasets affecting model accuracy
  - Poor handling can lead to biased and unsafe decisions
- **Current Methods and Their Limitations**:
  - **Traditional**: Mean imputation, deletion assume MCAR—rarely true
  - **Advanced**: MissForest, KNN can create biases, reducing interpretability
- **Need for Interpretable Machine Learning**:
  - EBMs offer insights into missingness, unlike black-box models
  - Helps identify risks and biases from imputation methods
- **Research Objective**:
  - Explore how interpretable models improve handling of missing data and enhance transparency in healthcare

Related Work

- **Critique of Existing Imputation Methods**:
  - **Generative Methods**: Criticized for relying on untestable data distribution assumptions
  - **Discriminative Methods**: Performance varies with data type and missingness pattern (e.g., MissForest, KNN, MICE)
- **Connections Between Imputation and Causal Inference**:
  - **Assumptions**: Both "unconfoundedness" in causal inference and "missing at random" (MAR) in imputation are based on untestable assumptions
- **Use of Explainability Techniques**:
  - Studies use explainability to find dataset issues (e.g., spurious correlations, mislabeled data)
  - The paper used EBMs, not black-box models, for missing value issues

Related Work

- **Comparisons with Automated Data Cleaning Tools**:
  - Tools like Automatic Statistician and AlphaClean handle missing values automatically
  - The study focused on understanding and mitigating missing data impacts, not just automatic correction

Types of Missing Values

- **Missing Completely At Random (MCAR)**:
  - Missingness unrelated to any data (observed or unobserved)
  - Same probability of missing data for all cases
  - *Example*: Respondent accidentally skips a survey question
- **Missing At Random (MAR)**:
  - Missingness related to observed data, not missing data itself
  - Can be predicted from other variables
  - *Example*: Older respondents more likely to have missing income data, but not related to income level

Types of Missing Values

- **Missing Not At Random (MNAR):**
  - Missingness related to unobserved data
  - Caused by factors not captured in observed data
  - *Example*: Higher earners may not report income, making missingness dependent on income value

Missing Value Imputation

- **MissForest:**
  - Starts with mean/mode imputation
  - Uses random forest to iteratively predict missing features
  - Continues until values converge
  - Captures non-linear relationships and feature interactions
- **K-Nearest Neighbors (KNN) Imputation:**
  - Imputes based on the mean of the K nearest neighbors
  - Calculates distances using non-missing features
  - Fast and accurate but requires careful tuning of parameters
  - Effective when similar samples are expected to have similar missing values

Explainable Boosting Machines (EBMs)

- Based on Generalized Additive Models (GAMs)
- GAMs model the target as a sum of shape functions for each feature
- **Advantages of EBMs Over Traditional GAMs**:
  - Traditional GAMs use splines with smoothness constraints
  - EBMs use ensembles of boosted, depth-restricted trees, enhancing performance
  - Provide better representation and capture details more accurately

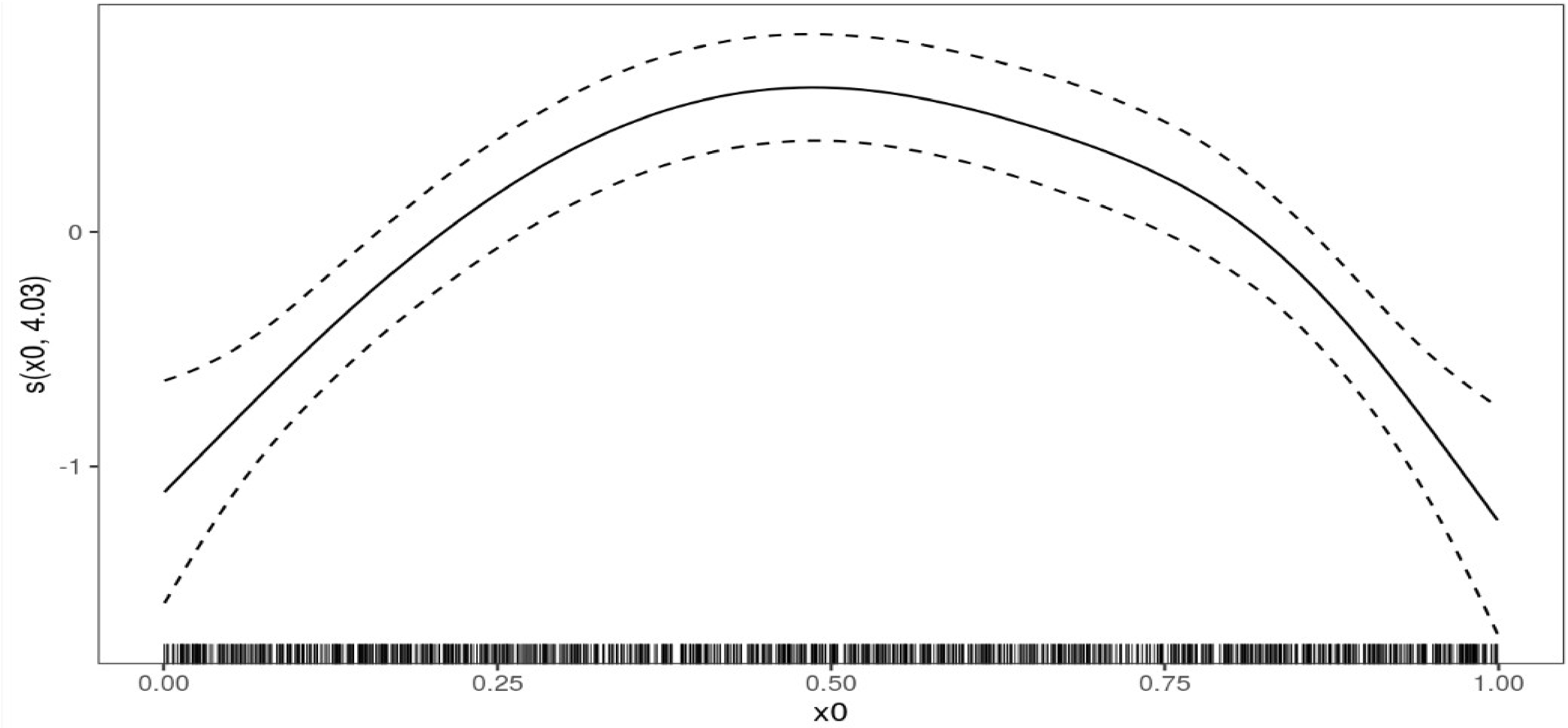


Image Source: https://mfasiolo.github.io/mgcViz/reference/plot.gamViz.html

Explainable Boosting Machines (EBMs)

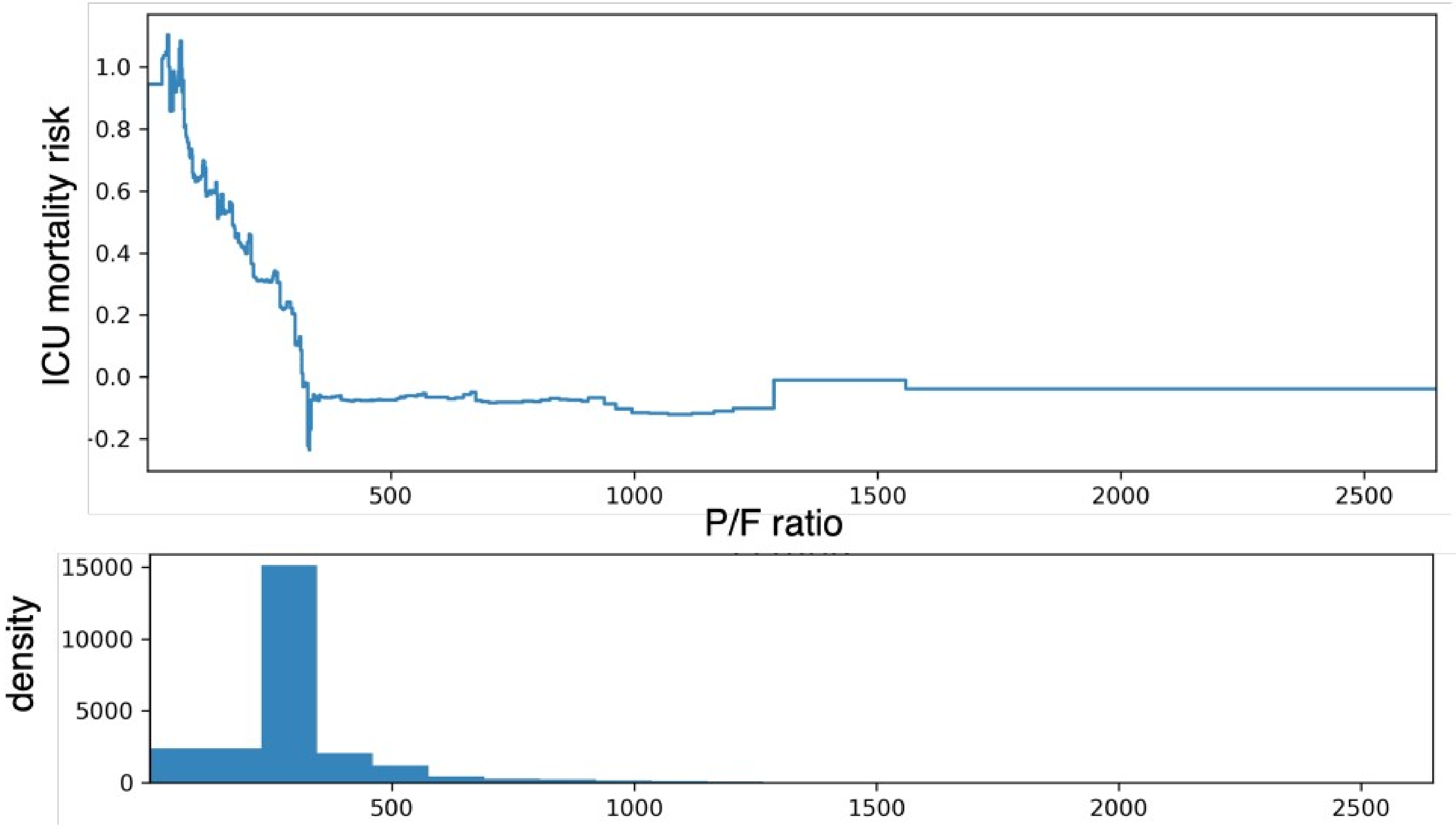


Image source:arXiv:2304.11749v1 [cs.LG] 23 Apr 2023

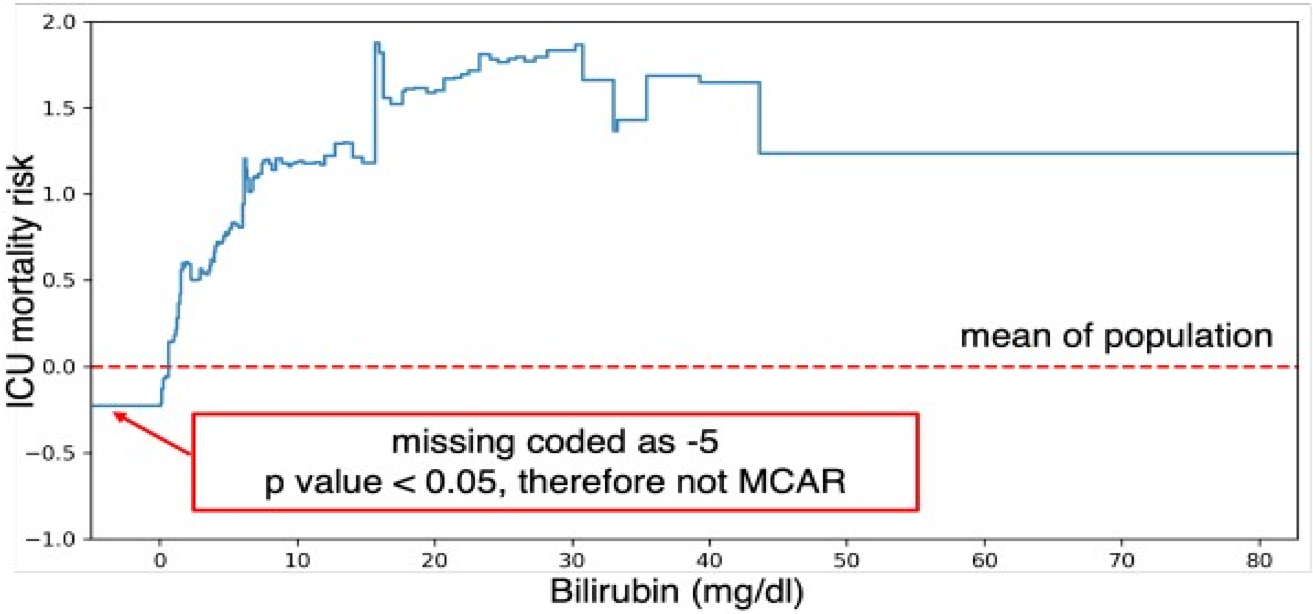
Testing for Missing Completely At Random (MCAR) Using EBM

- **Standard Tests for MCAR**:
  - Common tests include Little's test
  - Provide a statistical basis to determine if data is missing completely at random (MCAR)
- **Proposed Method Using EBMs**:
  - New method to test for MCAR using EBM shape functions
  - Utilizes EBM's visual interpretability to detect MCAR patterns from shape function plots
- **Benefits of the EBM Approach**:
  - Improves interpretability and understanding of missingness
  - Detects subtle patterns and interactions indicating if data is MCAR or otherwise

Testing for Missing Completely At Random (MCAR) Using EBM

- **How the EBM Approach Works:**
  - Assign a unique value to missing data (e.g., -1 or separate category)
  - EBM shape functions split values into bins, each with a prediction score
  - EBM shape function shows contribution of feature values, including missing data, to predictions
  - If missingness is MCAR all samples are missing with the same probability
  - Expected score of the bins should be 0
  - Wald test is used for the p-value

Testing for Missing Completely At Random (MCAR) Using EBM



Type	MCAR datasets↓			MAR datasets↑		
p_m	0.1	0.2	0.3	0.1	0.2	0.3
Little's	**0.035**	0.070	0.055	**1.000**	**1.000**	**1.000**
Ours	0.080	**0.005**	**0.005**	0.910	0.885	0.890

Image source:arXiv:2304.11749v1 [cs.LG] 23 Apr 2023

Missing Values in Healthcare

- **Common Missing Data Patterns in Healthcare**:
  - Lab results may not be recorded if considered "normal“
  - Measurements within normal range might be omitted, focusing on abnormal findings

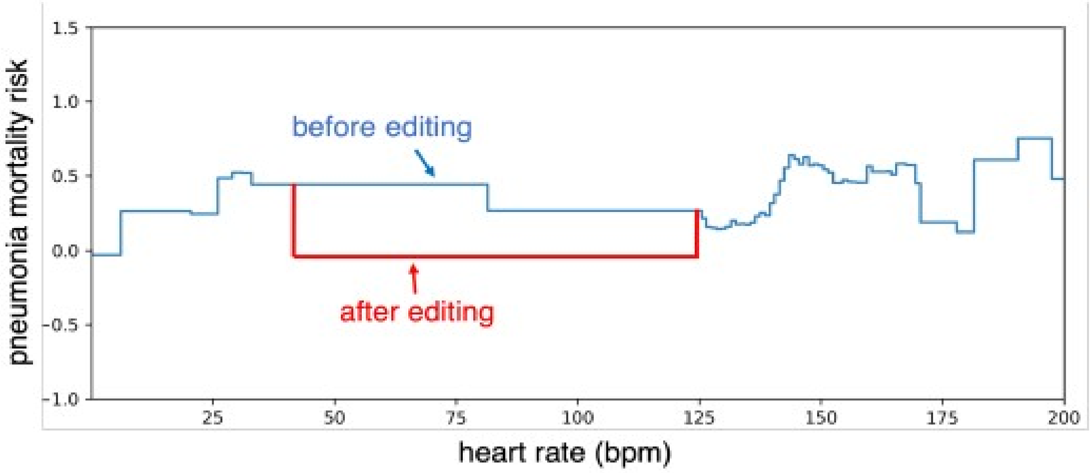


Image source:arXiv:2304.11749v1 [cs.LG] 23 Apr 2023

Predicting Missingness

- **Understanding Missing Data Beyond MCAR**:
  - Most missing values are not Missing Completely At Random (MCAR)
  - Distinguishing between MNAR (Missing Not At Random) and MAR (Missing At Random) is key for effective handling
- **Using EBMs to Predict Missingness:**
  - EBMs predict missingness by using observed variables to infer the missingness of another variable
  - Uses a missingness indicator (0-1) as the target, with other features as inputs (including the target)

MAR

model	p_m	linear	curvilinear	quadratic
LR	0.1	0.954±0.014	0.902±0.016	**0.883±0.02**
RF		0.943±0.014	0.946±0.013	**0.883±0.02**
KNN		0.895±0.013	0.894±0.009	0.881±0.021
EBM		**0.956±0.015**	**0.959±0.013**	0.881±0.02
LR	0.2	0.928±0.019	0.839±0.034	0.815±0.013
RF		0.911±0.019	0.928±0.019	**0.831±0.017**
KNN		0.813±0.024	0.81±0.022	0.812±0.008
EBM		**0.930±0.019**	**0.946±0.02**	0.822±0.016
LR	0.3	0.906±0.022	0.809±0.054	0.710±0.025
RF		0.887±0.021	0.926±0.019	**0.812±0.03**
KNN		0.744±0.032	0.752±0.042	0.711±0.016
EBM		**0.908±0.022**	**0.946±0.02**	0.795±0.03

Image source:arXiv:2304.11749v1 [cs.LG] 23 Apr 2023

MNAR

model	p_m	linear	curvilinear	quadratic
LR	0.1	0.957±0.013	0.901±0.013	**0.886±0.017**
RF		0.944±0.013	0.948±0.011	**0.886±0.017**
KNN		0.899±0.012	0.898±0.01	0.885±0.018
EBM		**0.959±0.012**	**0.963±0.011**	0.885±0.017
LR	0.2	0.928±0.018	0.847±0.035	0.817±0.010
RF		0.910±0.016	0.933±0.016	**0.828±0.012**
KNN		0.816±0.024	0.82±0.025	0.813±0.008
EBM		**0.931±0.017**	**0.953±0.016**	0.819±0.012
LR	0.3	0.914±0.016	0.805±0.048	0.706±0.024
RF		0.891±0.015	0.925±0.015	**0.811±0.028**
KNN		0.760±0.035	0.764±0.039	0.711±0.017
EBM		**0.916±0.016**	**0.949±0.015**	0.789±0.03

Image source:arXiv:2304.11749v1 [cs.LG] 23 Apr 2023

Missing Values in Healthcare

- **Visualization of Missingness Contributions**:
  - EBMs allow visualization of how different feature values contribute to the missingness of a specific variable

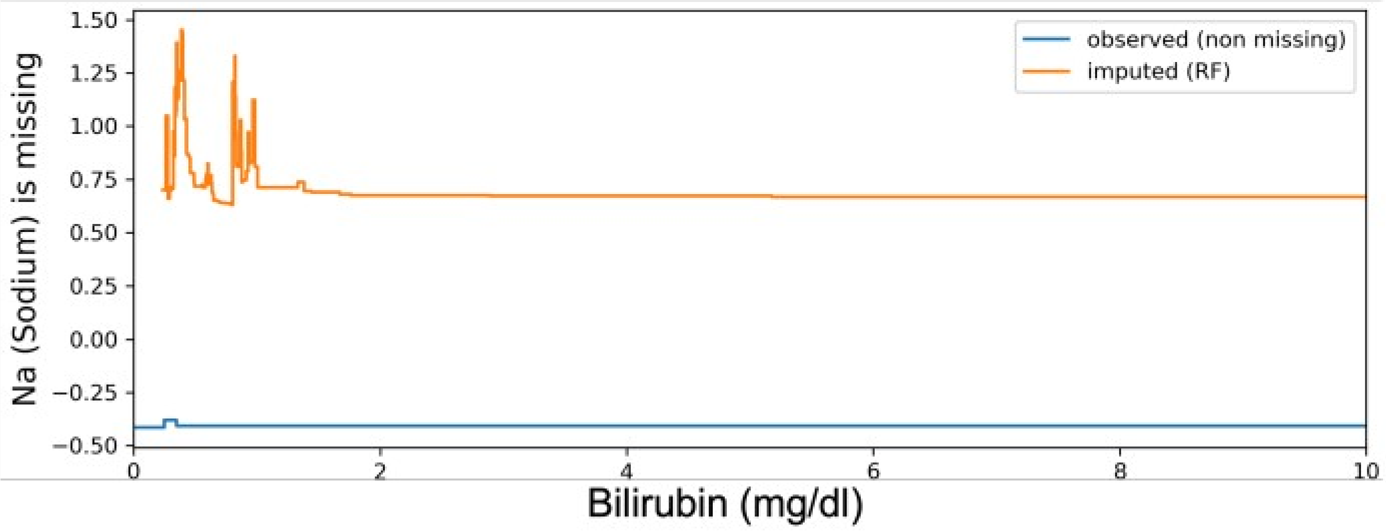


Image source:arXiv:2304.11749v1 [cs.LG] 23 Apr 2023

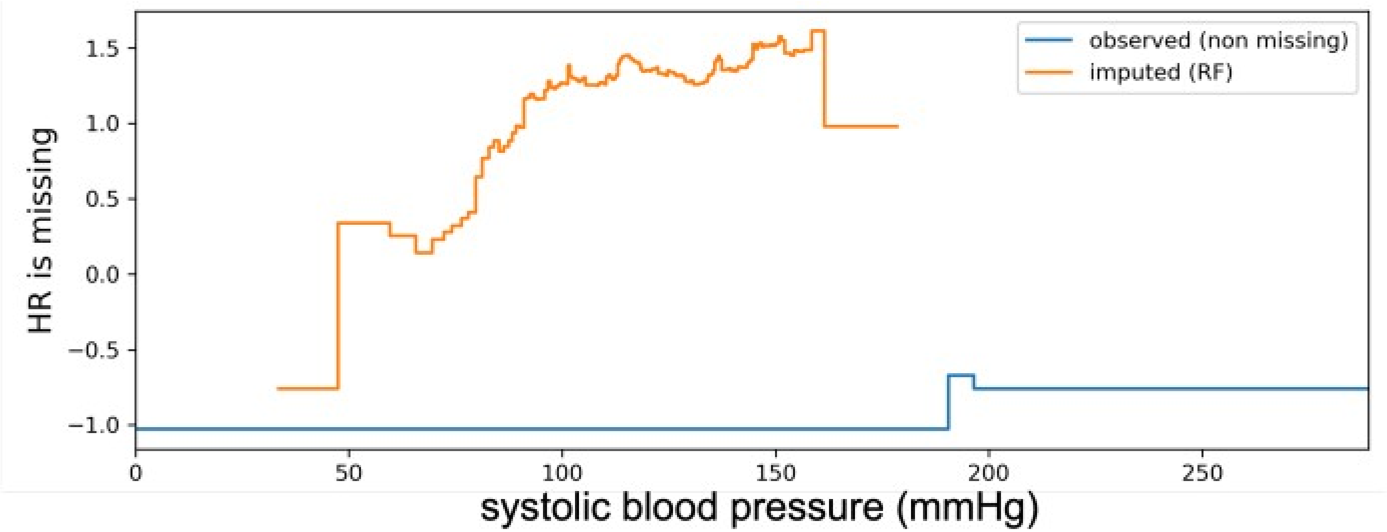


Image source:arXiv:2304.11749v1 [cs.LG] 23 Apr 2023

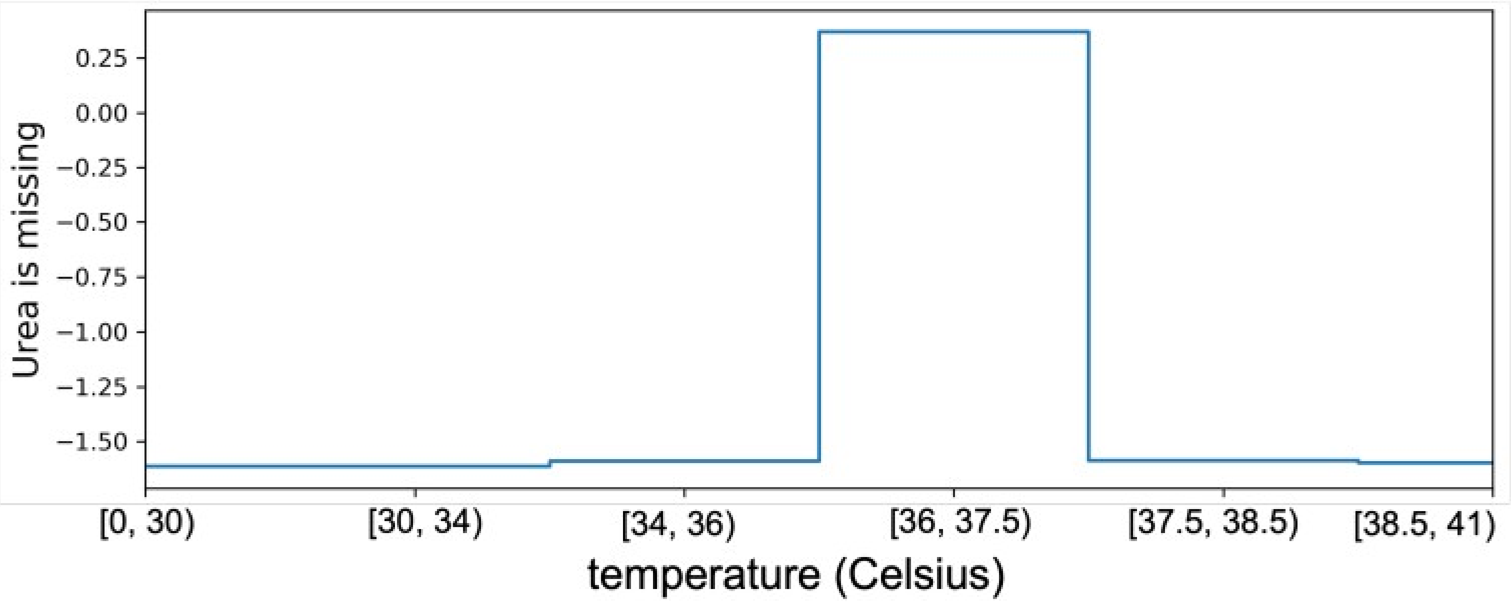


Image source:arXiv:2304.11749v1 [cs.LG] 23 Apr 2023

Detecting and avoiding risks

- **Common Practice of Missing Value Imputation:**
  - Widely used due to models' inability to handle missing data
  - Techniques: mean, median imputation, unique values (e.g., 0, -99), advanced methods like MissForest
- **Risks Associated with Mean Imputation**:
  - Mean imputation is one of the most common methods
  - problematic if missing data differs from non-missing data
  - does not significantly affect model accuracy but poses a risk of underestimating the risk for patients

Detecting and avoiding risks

- **Challenges with Mean Imputation:**
  - Mean imputation can obscure key data distinctions, leading to misleading predictions
  - Difficult to edit models effectively as it aligns low-risk (missing) and high-risk (actual) groups at the same point

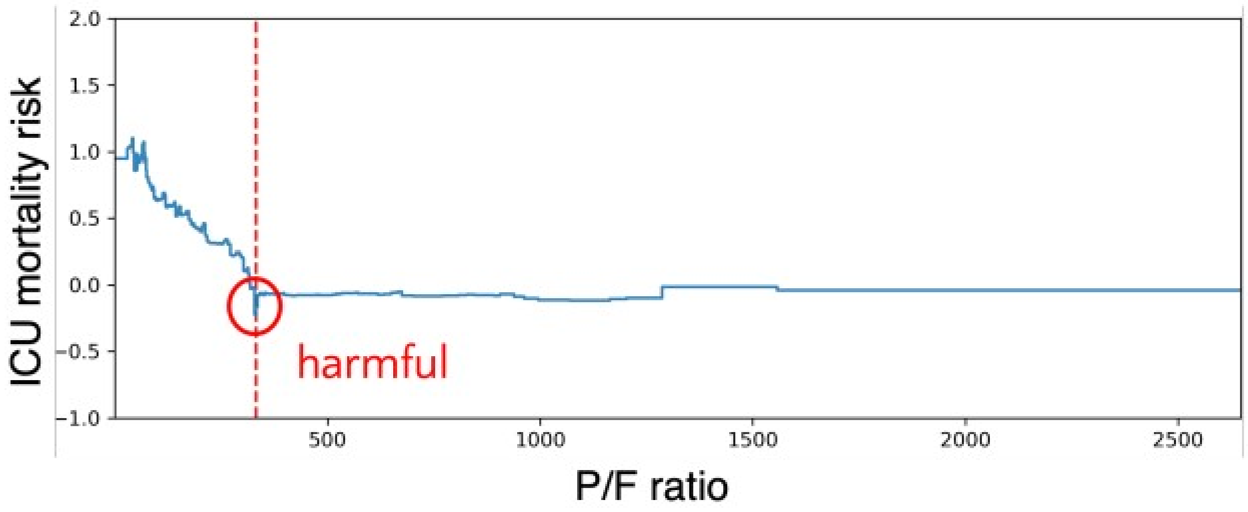


Image source:arXiv:2304.11749v1 [cs.LG] 23 Apr 2023

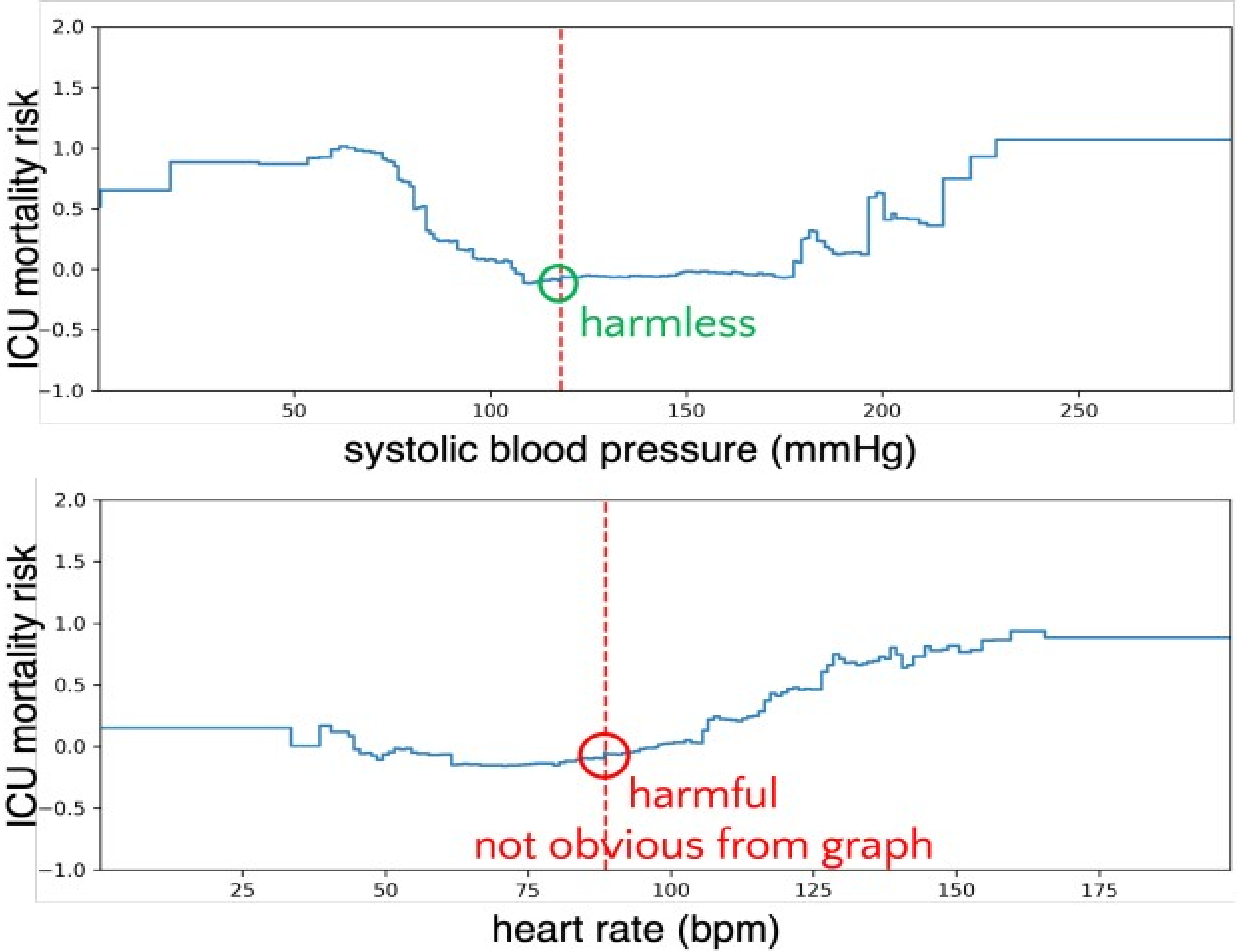
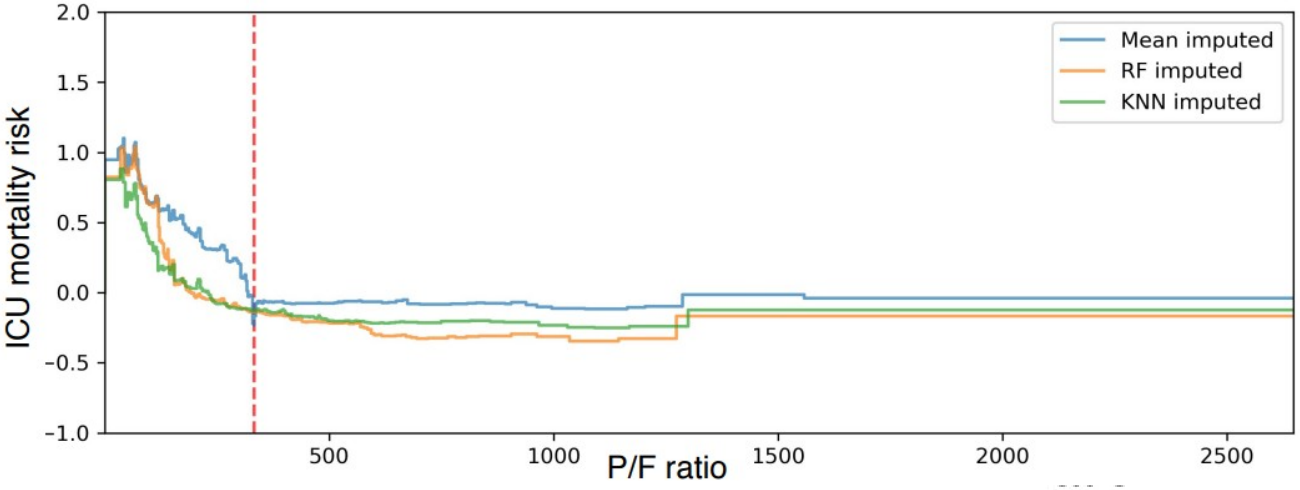


Image source:arXiv:2304.11749v1 [cs.LG] 23 Apr 2023

Imputation with Advanced Methods



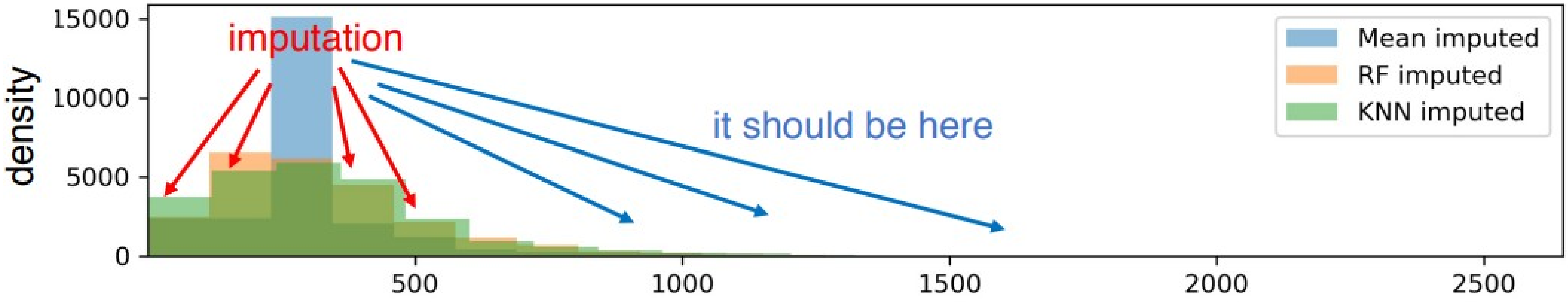


Image source:arXiv:2304.11749v1 [cs.LG] 23 Apr 2023

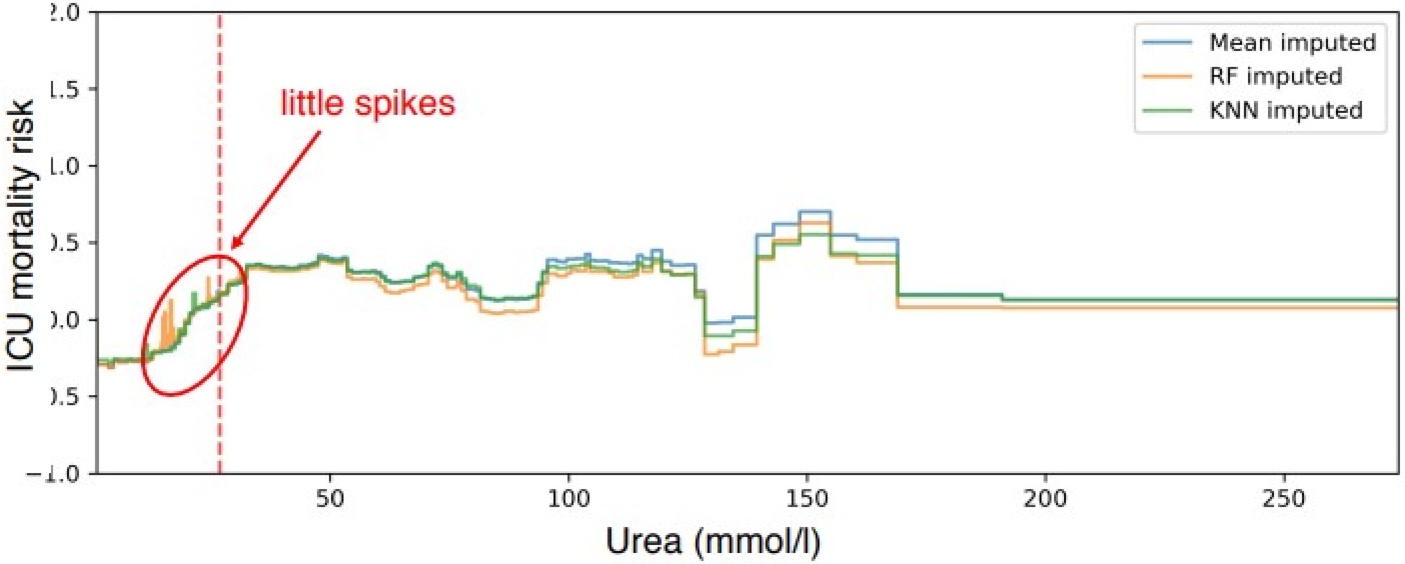


Image source:arXiv:2304.11749v1 [cs.LG] 23 Apr 2023

Conclusion

- **Key Contributions**:
  - **Testing for MCAR**: Developed an EBM-based method to determine if data is Missing Completely At Random (MCAR)
  - **Identifying Assumed Normal Values**: EBM shape functions detect missing values due to normality assumptions, clarifying missingness mechanisms
  - **Predicting Missingness**: EBMs predict missingness of features using observed data, enhancing interpretability
  - **Automatic Detection of Harmful Imputations**: EBMs identify harmful imputations (e.g., mean, median)
  - **Advanced Imputation Methods**: Visualization with EBMs assesses the impact of advanced methods (e.g., MissForest, KNN) on performance and reveals subtle issues

Conclusion

- **Impact and Future Directions**:
  - Provides a robust framework for handling missing data in healthcare, enhancing model reliability and safety
  - Future research may explore more interpretability techniques to improve data handling and model performance across domains

Questions?