Appendix D. Supplemental Tables and Figures

Table D.1: Predictor list and information from training dataset when patients were most acutely ill.

Feature	Type	Mean	Std. Dev.	Missing (%)	Missing (%) No Event*	Missing (%) With Event*
Age	Numeric	57.36	17.95	0.00	0.00	0.00
Albumin	Numeric	3.68	0.65	40.04	41.38	11.66
Anion Gap	Numeric	12.08	3.09	12.92	13.39	3.01
Blood Urea Nitrogen (BUN)	Numeric	20.44	15.85	13.06	13.57	2.27
Bilirubin	Numeric	1.09	2.58	41.20	42.54	12.64
CO2	Numeric	26.06	3.81	12.92	13.39	3.01
Calcium	Numeric	8.94	0.70	12.71	13.21	2.18
Chloride	Numeric	104.50	4.80	12.59	13.08	2.09
Creatinine	Numeric	1.13	1.19	12.50	12.99	2.12
Diastolic Blood Pressure	Numeric	66.75	13.96	0.03	0.03	0.06
Female	Indicator	0.49	0.50	0.00	0.00	0.00
Fluid Bolus Ordered	Indicator	0.10	0.30	0.00	0.00	0.00
Glasgow Coma Scale	Numeric	14.57	1.35	22.21	22.40	18.10
Glucose	Numeric	127.51	55.39	9.11	9.48	1.35
Heart Rate	Numeric	84.54	20.47	0.02	0.02	0.06
Height	Numeric	1.70	0.11	34.85	35.07	30.34
Hematocrit	Numeric	35.17	6.56	11.45	11.87	2.45
Hemoglobin	Numeric	11.70	2.37	12.20	12.63	2.91
International Normalized Ratio	Numeric	1.25	2.18	49.71	51.36	14.66
Lactate	Numeric	1.64	1.29	65.49	67.38	25.37
Magnesium	Numeric	1.93	0.34	50.82	52.51	15.09
Max O2 24hrs	Numeric	2.21	3.66	0.09	0.08	0.34

Mean Arterial Pressure	Numeric	85.33	16.81	0.03	0.03	0.06
Mean Corps. HGB	Numeric	29.60	2.68	12.21	12.65	2.91
Mean Corps. HGB Conc.	Numeric	33.19	1.52	12.21	12.65	2.91
Mean Corps. HGB Vol.	Numeric	89.13	6.63	12.20	12.64	2.85
Mean Platelet Vol.	Numeric	10.22	0.99	13.60	13.93	6.63
Min SPO2 24hrs	Numeric	92.80	5.85	0.09	0.09	0.09
Partial Thromboplastin	Numeric	28.53	8.34	58.78	60.24	27.67
Partial Thromboplastin Time	Numeric	12.84	6.32	50.47	52.12	15.34
Phosphorus	Numeric	3.61	1.07	58.50	60.17	23.22
Platelets	Numeric	227.23	105.65	12.31	12.76	2.91
Potassium	Numeric	4.21	0.55	11.55	12.02	1.63
Protein Level	Numeric	6.44	0.96	41.12	42.46	12.61
Pulse Pressure	Numeric	0.45	0.08	0.05	0.04	0.12
Race - Asian	Indicator	0.02	0.14	0.00	0.00	0.00
Race - Black or African American	Indicator	0.11	0.31	0.00	0.00	0.00
Race - Other	Indicator	0.04	0.20	0.00	0.00	0.00
Race - White or Caucasian	Indicator	0.83	0.38	0.00	0.00	0.00
Red Blood Cell Count	Numeric	3.97	0.79	12.20	12.64	2.85
Red Blood Cell Dist. Width	Numeric	14.74	2.44	12.25	12.68	3.01
Respiratory Device Used	Indicator	0.50	0.50	0.00	0.00	0.00
Respiratory Rate	Numeric	18.20	5.10	0.09	0.09	0.03
Shock Index	Numeric	0.71	0.50	0.08	0.08	0.15

Shock Index x Age	Numeric	40.18	33.14	0.08	0.08	0.15
Sodium	Numeric	138.43	3.78	11.56	12.03	1.63
Systolic Blood Pressure	Numeric	122.49	26.08	0.03	0.03	0.06
Temperature	Numeric	36.72	1.55	2.72	2.67	3.71
Time	Numeric	5.04	10.24	0.00	0.00	0.00
Urine 24hrs	Numeric	302.24	269.17	25.84	26.15	19.33
Weight	Numeric	84.25	24.90	15.08	15.49	6.29
White Blood Cell Count	Numeric	9.72	10.80	12.19	12.63	2.91

* We selected a single window per encounter. For patients with events, we took the window where the event occurred. For encounters without an event, we selected the windows with the highest NEWS score. Thus, for non-event encounters, the data used represents the time the patient was the most ill as determined by NEWS. Pulse-pressure was normalized by the systolic blood pressure.

Figure D.1: Dataset partitioning for training, tuning and testing our PICTURE model.

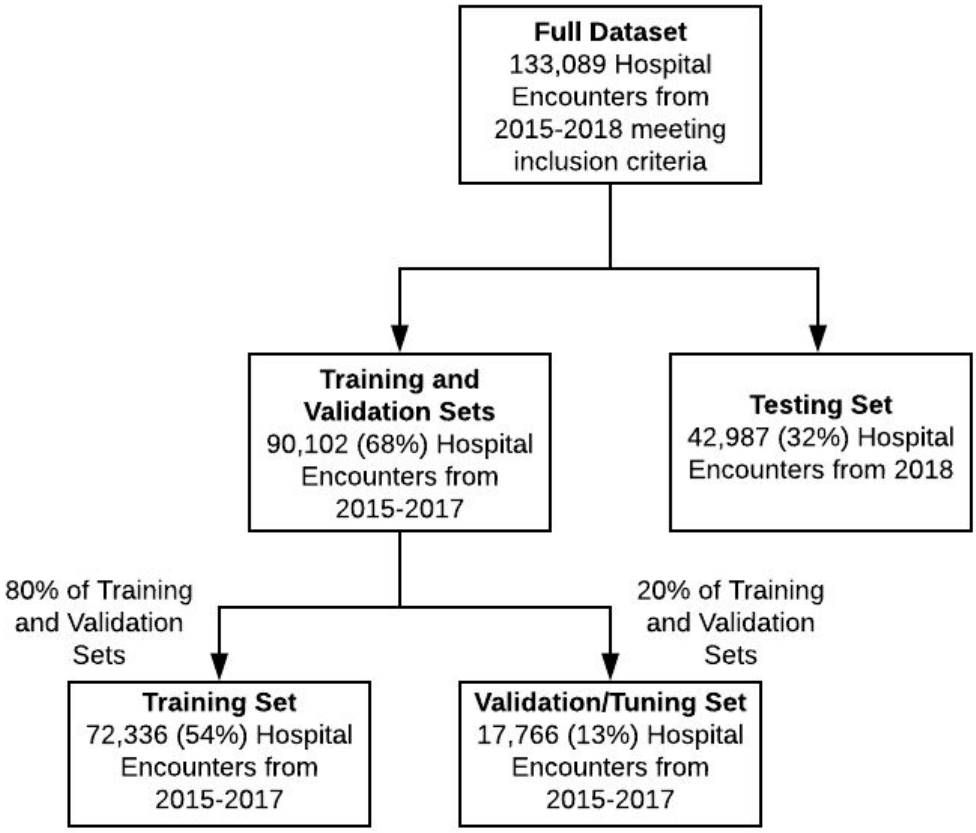

Figure D.2: Different granularity levels for making predictions. A prediction can be made at every unique observation for a hospital encounter. This is the lowest level of granularity. An entire hospital encounter can be summarized into a single score by taking the maximum score 24 before hours some adverse event or discharge. For training, we partitioned a hospital encounter into discrete 8-hour non-overlapping blocks. The most recent data up to 1 hour (for missing features) into the window was used to represent the current data for a window. This windowing step normalizes the number of training observations across patients so that two patients in the hospital for the same amount of time will have the same number of training examples.

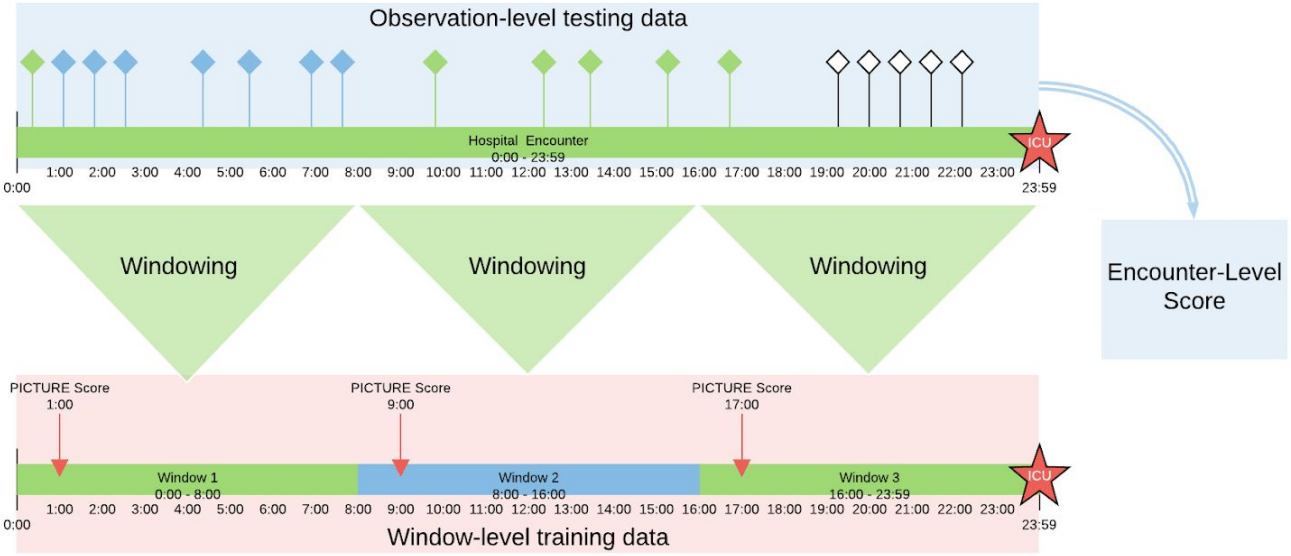

Figure D.3: Simulation results when training on data with half the missing rate of University of Michigan. We compared the test performance of XGBoost, a Random forest classifier, logistic regression, and logistic regression with missing value indicators as features across different test setting missingness patterns when training on half the missing data. The same imputation method that was used in training was also used during testing. In Scenario 2, the training data and the test dataset followed the same missingness pattern: the half the missing rates of those in **Table D.1**. In Scenario 1, there was twice as much missing data in the testing dataset as in the training dataset. In Scenario 3, the missingness pattern in the testing dataset was fixed across classes, and it was different in the training dataset. In Scenario 4, the testing dataset was fully observed. These different missing patterns could represent different hospitals or data collection changes. Each panel represents a different classification model and each color within a panel represents a different imputation method. The learning of the missingness pattern is clearly illustrated in the XGBoost classification model with the "Extreme Value" imputation method, where the AUPR was 0.62 in Scenario 2 and 0.2 in Scenario 3. The Random Forest classifiers AUPR dropped from 0.45 in Scenario 2 to 0.25 in Scenario 3. Note that these are the same simulated test patients with a different missingness pattern. Bayesian imputation strategies eliminated the learning of the missingness pattern. Since the tree models learned the missingness pattern, the classification performance increased when the missing rate doubled from Scenario 2 to Scenario 1.

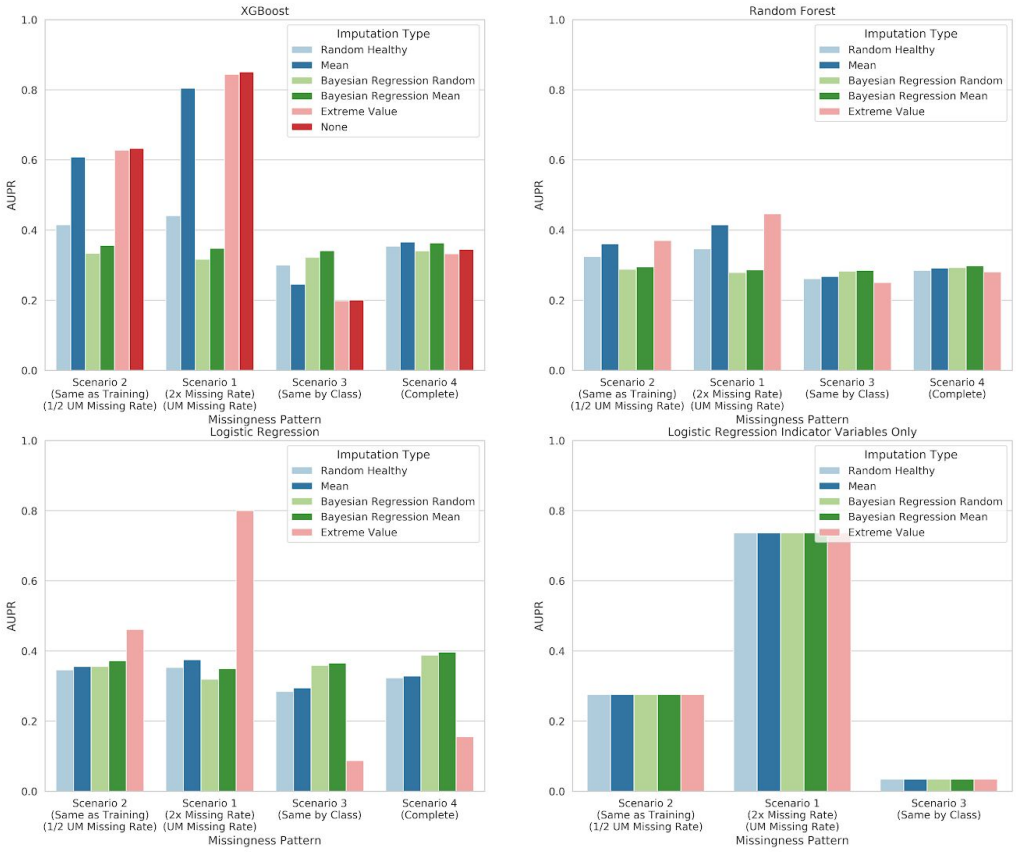

Figure D.4: Simulation results when training on data with four times lower missing rate than University of Michigan. We compared the test performance of XGBoost, a Random forest classifier, logistic regression, and logistic regression with missing value indicators as features across different test setting missingness patterns when training on four times less missing data. The same imputation method that was used in training was also used during testing. In Scenario 5, the training data and the test dataset followed the same missingness pattern: the four times lower missing rates of those in **Table D.1**. In Scenario 1, there was twice as much missing data in the testing dataset as in the training dataset. In Scenario 3, the missingness pattern in the testing dataset was fixed across classes, and it was different in the training dataset. In Scenario 4, the testing dataset was fully observed. These different missing patterns could represent different hospitals or data collection changes. Each panel represents a different classification model and each color within a panel represents a different imputation method. The learning of the missingness pattern is now less pronounced in the XGBoost classification model with the "Extreme Value" imputation method, where the AUPR was 0.5 in Scenario 2 and 0.23 in Scenario 3. The Random Forest classifiers AUPR dropped from 0.31 in Scenario 2 to 0.26 in Scenario 3. Therefore, in this training testing situation, the Random Forest classifier did not learn the missingness pattern as much as XGBoost. Since XGBoost still learned the missingness pattern (when not using an appropriate imputation method), its performance dropped when the missing rate increased.

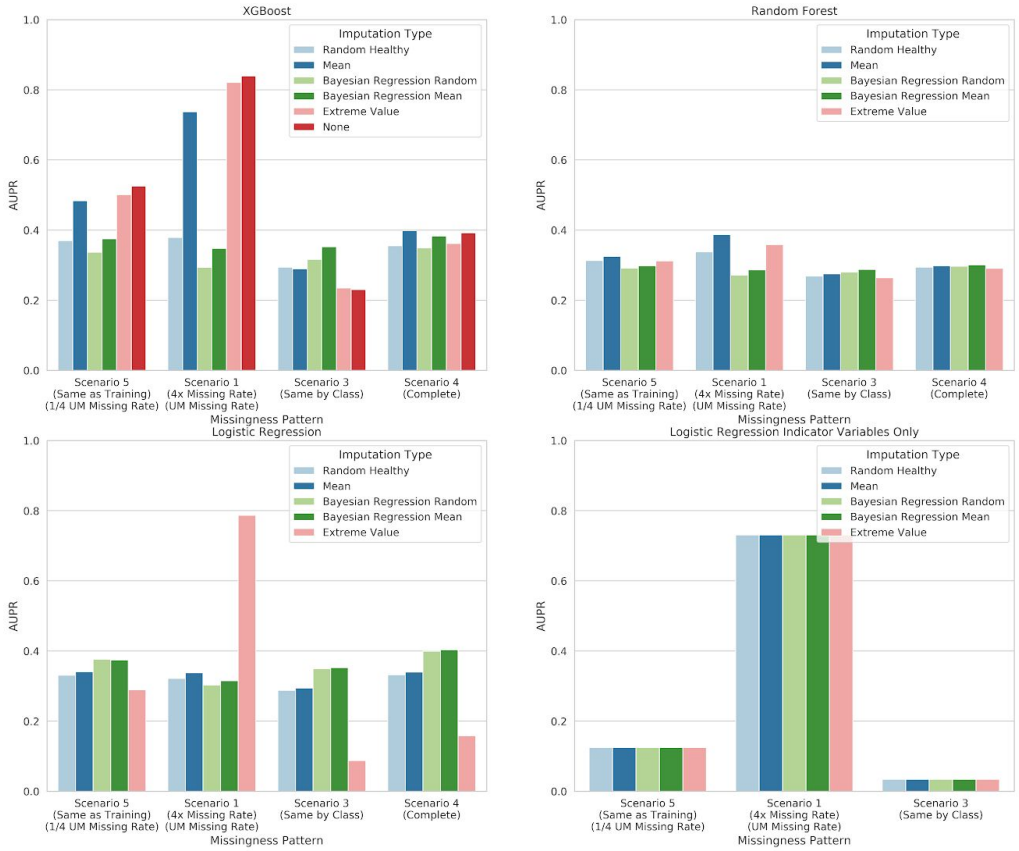

Figure D.5: Receiver Operating Characteristic curve for PICTURE, logistic regression, NEWS and SOFA at the observation-level.

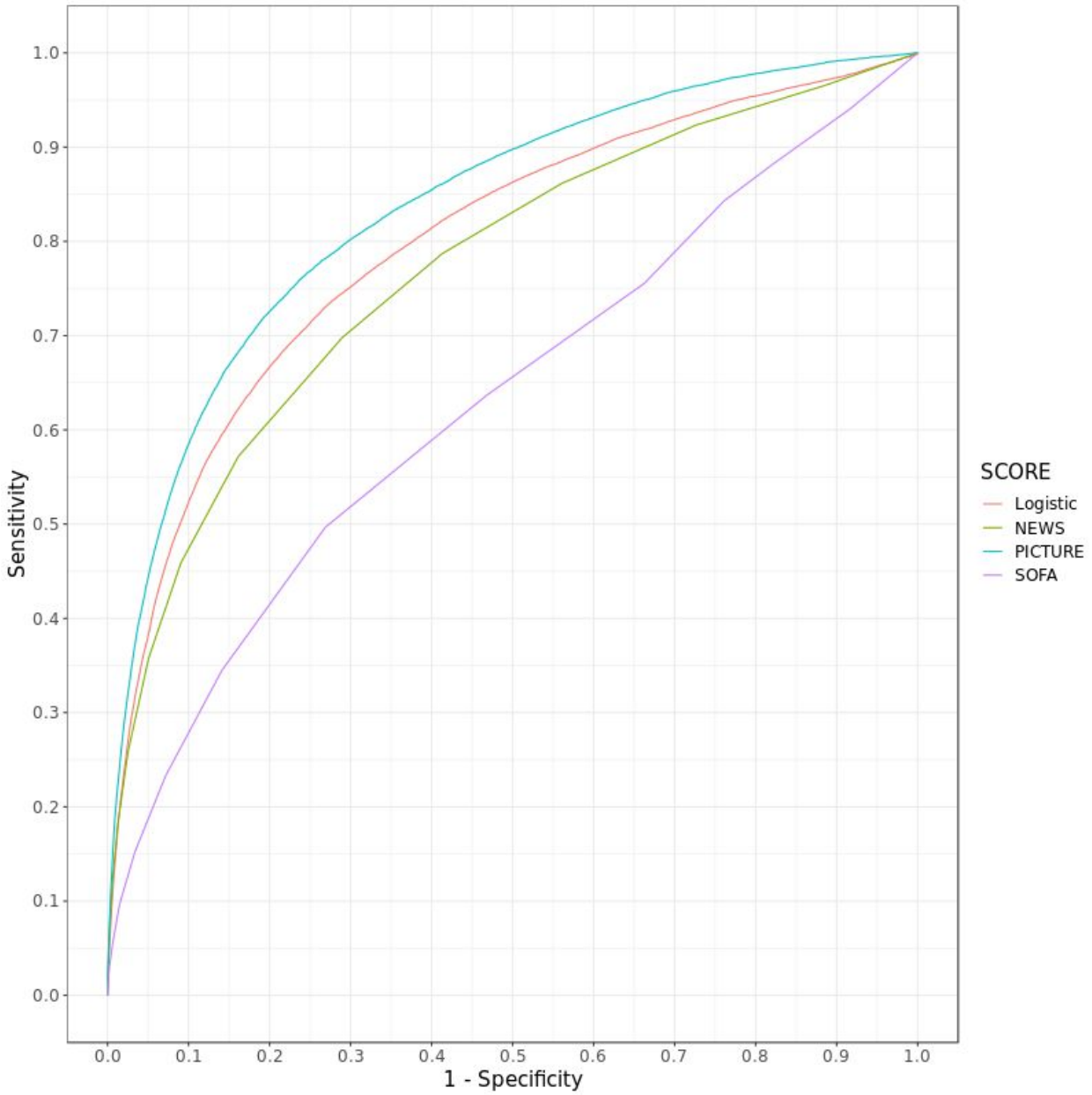

Figure D.6: Precision-Recall Curve for PICTURE, logistic regression, NEWS, and SOFA at the observation-level. The event rate is 1.0%.

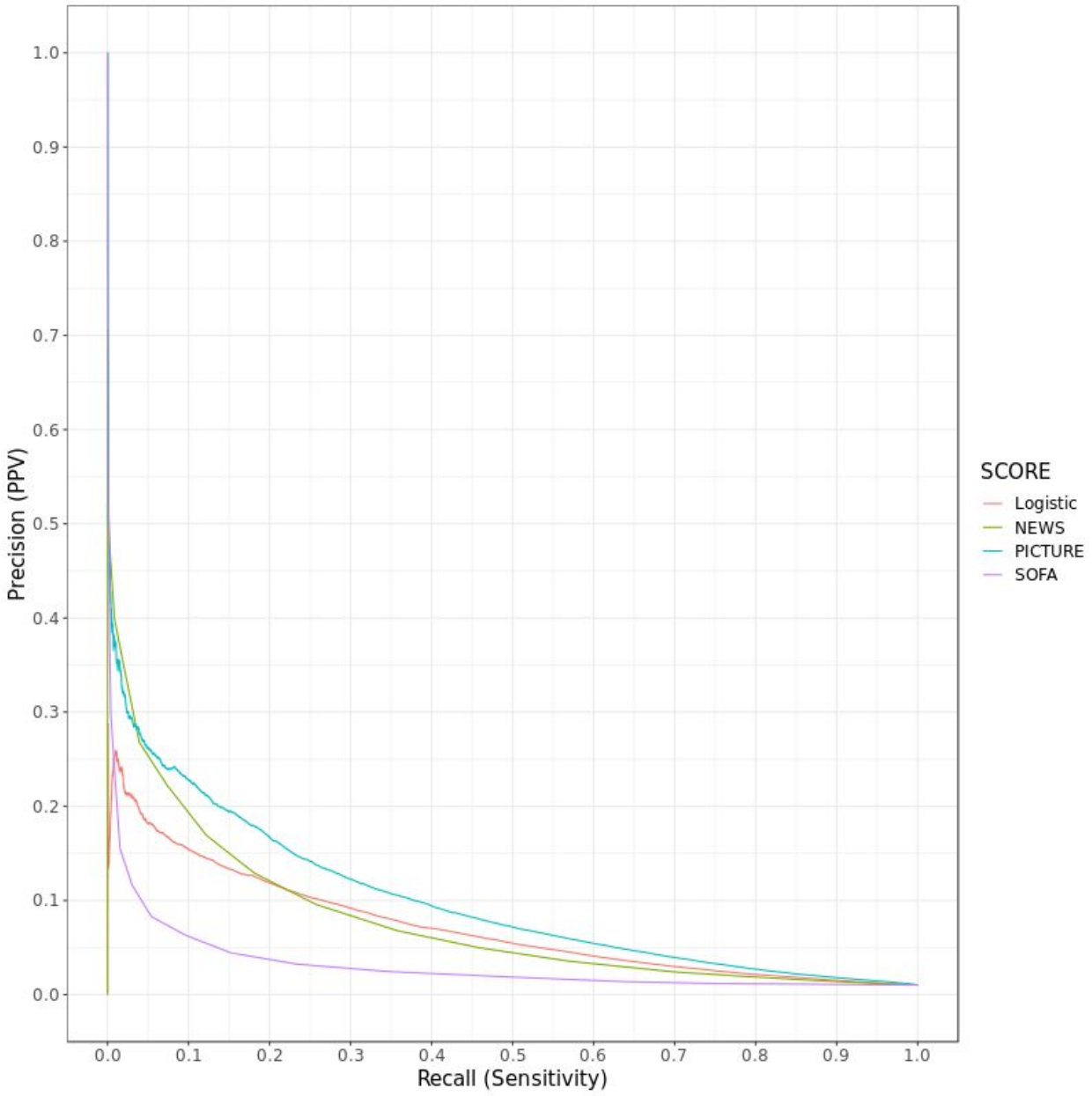

Figure D.7: The adjusted Precision-Recall Curve for PICTURE, logistic regression, NEWS, and SOFA at the observation-level. The event rate is 4%.

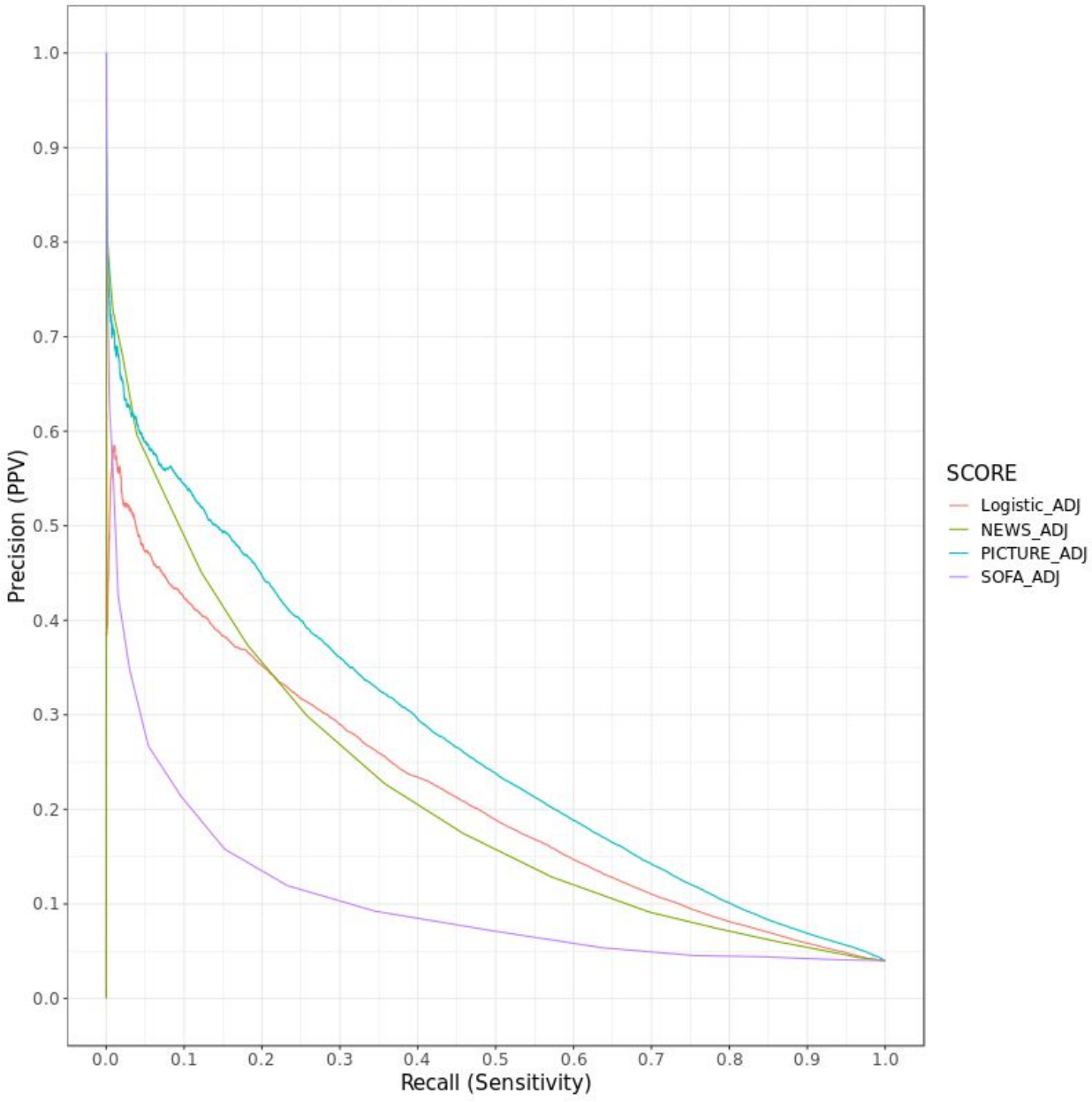

Table D.2: Example Individual prediction explanations for a patient who had an adverse event. We present clinicians with a ranked list of the features with the highest attribution and their corresponding observed values when the patient's risk score passes the alarm threshold.

Feature	Feature Attribution	Observed Feature Value
Glasgow Coma Scale	1.95	3.0
Respiratory Rate	1.76	33.0
Lactate	1.25	4.1
Shock Index	0.9	1.24
Hematocrit	0.86	18.3

Figure D.8: The top 20 features as determined by the mean absolute Shapley value computed across all encounters using the maximum PICTURE score within an encounter.

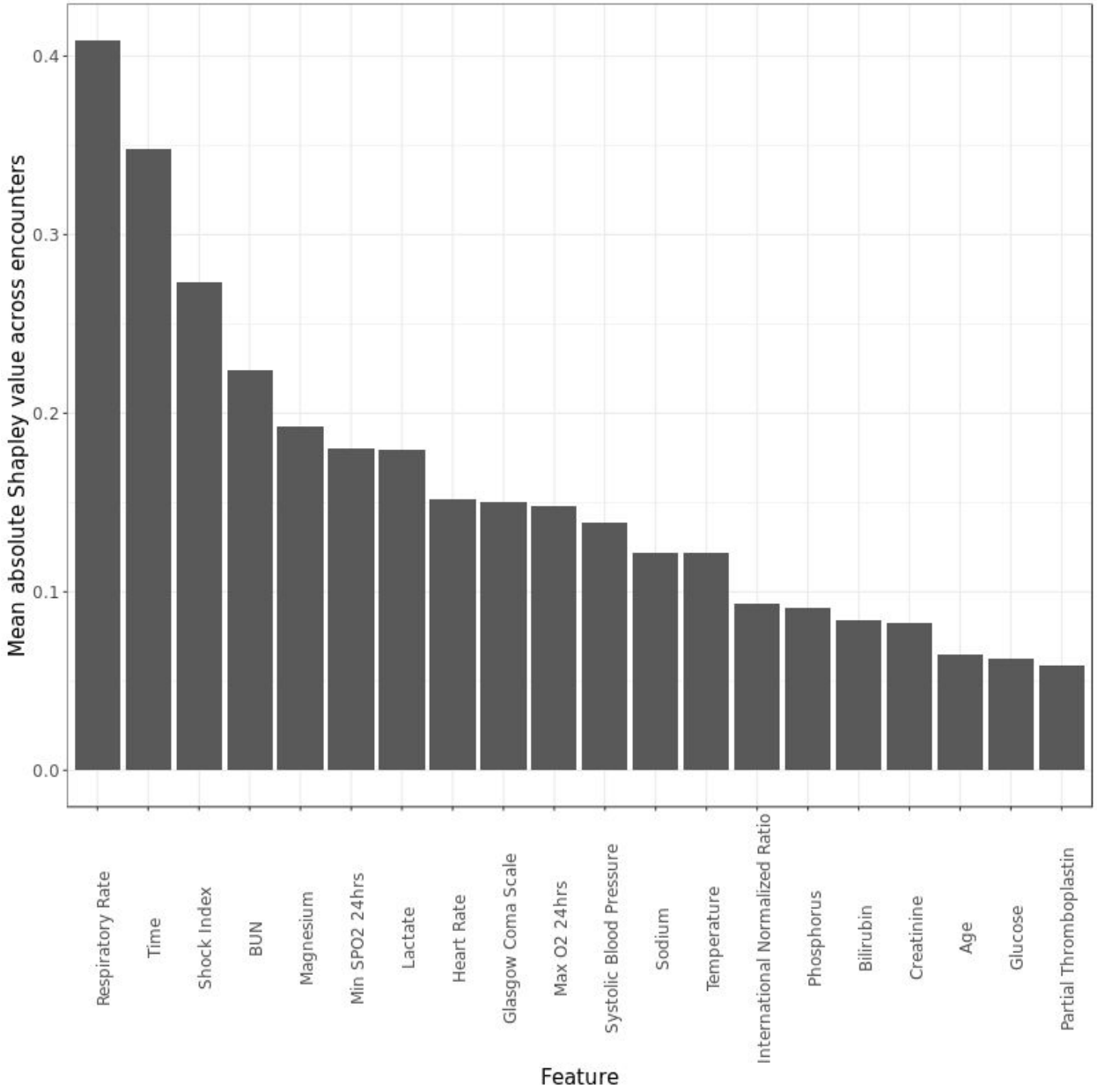

Table D.3: Simulation performance of imputation methods on testing data. In this analysis, we set the difference in missing rates between classes (i.e., those with vs. without adverse events) to be 0.8. Extreme Value Imputation and Healthy Mean Imputation experience a significant drop in accuracy and PPV when the missingness pattern changes from the same as training to completely observed data. The Random Healthy Imputation method did not learn the missingness pattern and did not show a performance drop.

Model	Same Missingness Pattern Accuracy	Complete Data Accuray	Accuracy Drop	Same Missingness Pattern PPV	Complete Data PPV	PPV Drop
Fully Observed	0.77	0.77	**0**	0.69	0.69	**0**
Healthy Mean Imputation	0.90	0.59	**0.31**	0.83	0.37	**0.47**
Extreme Value Imputation	0.90	0.74	**0.15**	0.75	0.25	**0.50**
Random Healthy Imputation	0.77	0.78	**0**	0.64	0.65	**0.01**

Figure D.9: The accuracy difference due to the missingness pattern change increases as the missing rate difference increases between classes (i.e., those with adverse events vs. those without adverse events). The y-axis shows the decrease in accuracy on the test dataset when using fully observed data instead of data with the same missingness pattern as the training dataset. The x-axis shows the missing rate difference between classes. The larger the missing rate difference between classes, the larger the performance drop for extreme value imputation and healthy mean imputation.

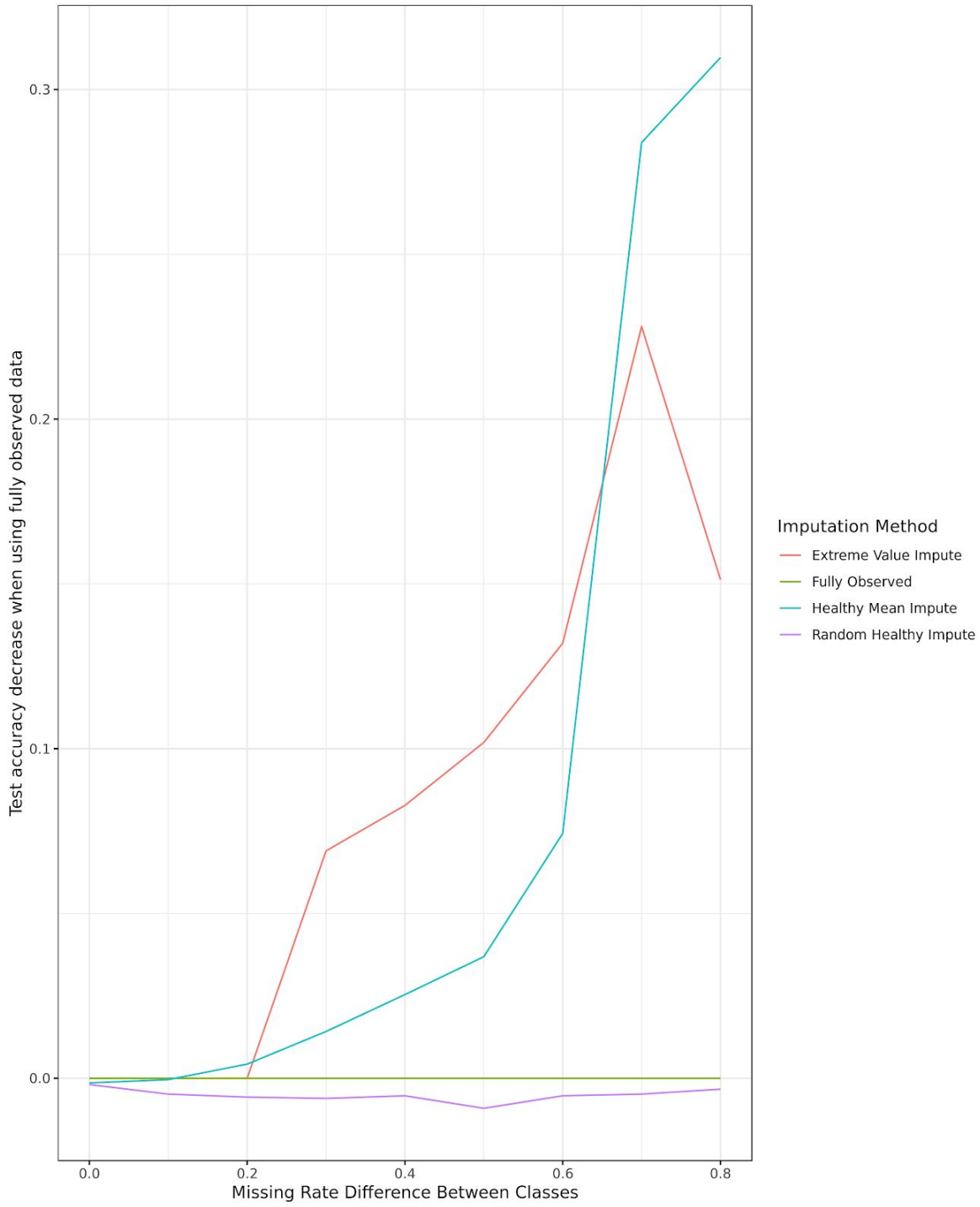

Figure D.10: The PPV difference due to the missingness pattern change increases as the missing rate difference increases between classes (i.e., those with adverse events vs. those without adverse events). The y-axis shows the decrease in PPV on the test dataset when using fully observed data instead of data with the same missingness pattern as the training dataset. The x-axis shows the missing rate difference between classes. The larger the missing rate difference between classes, the larger the performance drop for extreme value imputation and healthy mean imputation.

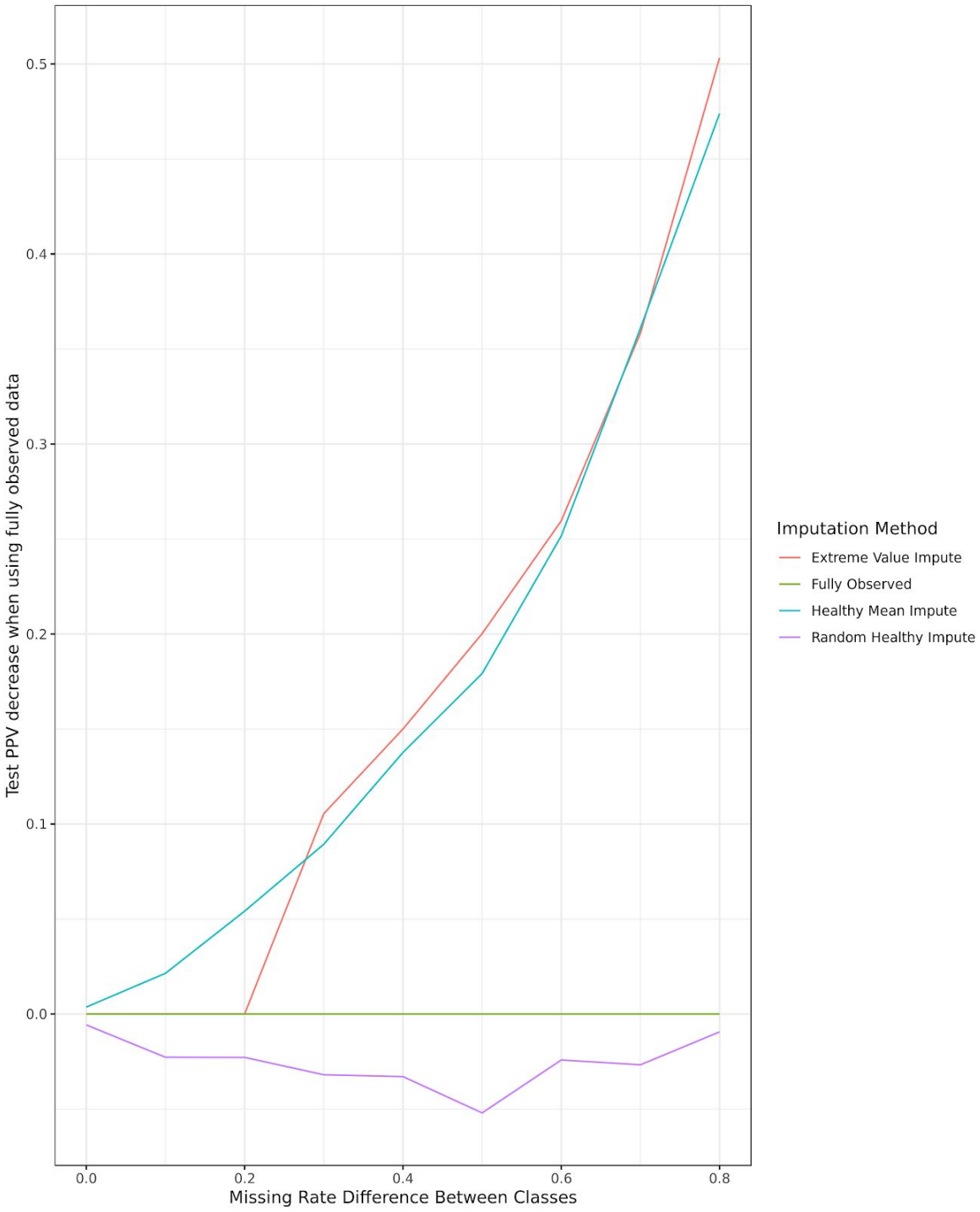

Table D.4: PICTURE subgroup performance. Event rates for AUPR were standardized to 4% or the raw encounter-level event rate. All performance statistics were computed at the encounter-level.

Subgroup	Raw Encounter-Level Event Rate	AUROC	Raw AUPR	Standardized AUPR
All Events	3.4%	0.855	0.286	0.314
All Events Sex = Female	3.0%	0.860	0.279	0.349
All Events Race = Black	3.6%	0.845	0.257	0.273
ICU Transfer vs. No Event	2.4%	0.856	0.198	0.283
Death vs. No Event	0.7%	0.945	0.208	0.546
Vasoactive medications vs. No Event	1.1%	0.820	0.07	0.253
Cardiac Arrest vs. No Event	0.4%	0.825	0.004	0.157