Fairness Analysis Report

Table of Contents:

1. Information about test dataset

Grouping by sex

Table 1.a

Category	Number of patients	Number of patients with event
F	1833	393
M	3253	780

Grouping by age_group

Table 1.b

Category	Number of patients	Number of patients with event
<50	766	127
50-65	1322	298
65-75	1418	350
75-85	1242	319
>85	338	79

Grouping by APACHE_group

Table 1.c

Category	Number of patients	Number of patients with event
Cardiovascular	1891	666
Neurological	1468	92
Gastrointestinal	522	151
Respiratory	471	102
Other	325	76
Trauma	279	61
Metabolic	98	19

Grouping by surgical_status

Table 1.d

Category	Number of patients	Number of patients with event
Surgical	2279	541
Non-surgical	2775	626

2. Model Performance Analysis

Goal: Comparing the model performance across cohorts of patients

Binary metrics computed with a threshold on score of 0.445.

2.1. Aggregated views

2.1.1. Summarized performance metrics per grouping

Grouping by sex

The minority category is F.

Table 2.1.1.a

Metric	Macro-average	Worst value (category)	For minority category
Recall ↑	0.201	0.198 (M)	0.204
Precision ↑	0.557	0.557 (M)	0.558
NPV ↑	0.965	0.962 (M)	0.967
FPR ↓	0.007	0.008 (M)	0.007
Corrected precision ↑	0.576	0.557 (M)	0.596
Corrected NPV ↑	0.967	0.967 (M)	0.967
Event-based recall ↑	0.805	0.793 (M)	0.816
Calibration error ↓	0.032	0.037 (F)	0.037
Avg. score on positive class	0.255	0.252 (M)	0.257
Avg. score on negative class	0.033	0.035 (M)	0.031
AUROC ↑	0.914	0.908 (M)	0.921
AUPRC ↑	0.39	0.385 (M)	0.396
Corrected AUPRC ↑	0.406	0.385 (M)	0.427
Event-based AUPRC ↑	0.694	0.674 (M)	0.715
Corrected event-based AUPRC ↑	0.707	0.674 (M)	0.741

Grouping by age_group

The minority category is >85.

Table 2.1.1.b

Metric	Macro-average	Worst value (category)	For minority category
Recall ↑	0.199	0.184 (<50)	0.204
Precision ↑	0.598	0.522 (50-65)	0.708
NPV ↑	0.963	0.953 (>85)	0.953
FPR ↓	0.007	0.01 (75-85)	0.006
Corrected precision ↑	0.67	0.564 (75-85)	0.719
Corrected NPV ↑	0.98	0.98 (50-65)	0.98
Event-based recall ↑	0.793	0.751 (<50)	0.795
Calibration error ↓	0.054	0.082 (>85)	0.082
Avg. score on positive class	0.257	0.253 (75-85)	0.266
Avg. score on negative class	0.034	0.041 (75-85)	0.038
AUROC ↑	0.915	0.885 (75-85)	0.916
AUPRC ↑	0.408	0.372 (75-85)	0.479
Corrected AUPRC ↑	0.475	0.393 (75-85)	0.489

Event-based AUPRC ↑	0.717	0.654 (75-85)	0.82
Corrected event-based AUPRC ↑	0.768	0.676 (75-85)	0.828

Grouping by APACHE_group

The minority category is Metabolic.

Table 2.1.1.c

Metric	Macro-average	Worst value (category)	For minority category
Recall ↑	0.18	0.058 (Neurological)	0.164
Precision ↑	0.552	0.386 (Neurological)	0.709
NPV ↑	0.959	0.944 (Cardiovascular)	0.952
FPR ↓	0.008	0.015 (Gastrointestinal)	0.004
Corrected precision ↑	0.679	0.58 (Gastrointestinal)	0.766
Corrected NPV ↑	0.99	0.989 (Neurological)	0.991
Event-based recall ↑	0.755	0.48 (Neurological)	0.716
Calibration error ↓	0.075	0.119 (Neurological)	0.113
Avg. score on positive class	0.24	0.132 (Neurological)	0.244
Avg. score on negative class	0.035	0.056 (Cardiovascular)	0.032
AUROC ↑	0.908	0.878 (Cardiovascular)	0.93
AUPRC ↑	0.383	0.172 (Neurological)	0.466
Corrected AUPRC ↑	0.491	0.415 (Cardiovascular)	0.524
Event-based AUPRC ↑	0.676	0.418 (Neurological)	0.808
Corrected event-based AUPRC ↑	0.77	0.707 (Cardiovascular)	0.843

Grouping by surgical_status

The minority category is Surgical.

Table 2.1.1.d

Metric	Macro-average	Worst value (category)	For minority category
Recall ↑	0.205	0.194 (Non-surgical)	0.217
Precision ↑	0.554	0.553 (Non-surgical)	0.555
NPV ↑	0.963	0.961 (Surgical)	0.961
FPR ↓	0.008	0.009 (Surgical)	0.009
Corrected precision ↑	0.571	0.557 (Surgical)	0.557
Corrected NPV ↑	0.966	0.965 (Non-surgical)	0.966
Event-based recall ↑	0.802	0.799 (Non-surgical)	0.804
Calibration error ↓	0.033	0.033 (Surgical)	0.033
Avg. score on positive class	0.258	0.248 (Non-surgical)	0.268
Avg. score on negative class	0.035	0.04 (Surgical)	0.04
AUROC ↑	0.912	0.911 (Surgical)	0.911
AUPRC ↑	0.388	0.385 (Non-surgical)	0.39
Corrected AUPRC ↑	0.4	0.391 (Surgical)	0.391
Event-based AUPRC ↑	0.687	0.682 (Surgical)	0.682
Corrected event-based AUPRC ↑	0.698	0.682 (Surgical)	0.682

2.1.2. Summary view based on the ratio of significantly worse metrics

We first show an overview of this analysis over all groupings.
Worst ratio: 73.3% for category 75-85 (age_group) with the biggest delta 0.053 on Corrected event-based AUPRC.
Worst delta: 0.346 on Event-based recall for category Neurological (APACHE_group).

In the following tables, we display the ratio of significantly worse metrics (over the total number of analysed performance metrics) for each category of patients.

Grouping by sex

Worst ratio: 60.0% for category M with the biggest delta 0.068 on Corrected event-based AUPRC.
Worst delta is the same as above.

Table 2.1.2.a

F	M
6.7%	60.0%

Grouping by age_group

Worst ratio: 73.3% for category 75-85 with the biggest delta 0.053 on Corrected event-based AUPRC.
Worst delta: 0.056 on Event-based recall for category <50.

Table 2.1.2.b

<50	50-65	65-75	75-85	>85
20.0%	33.3%	33.3%	73.3%	20.0%

Grouping by APACHE_group

Worst ratio: 46.7% for category Cardiovascular with the biggest delta 0.104 on Corrected precision.
Worst delta: 0.346 on Event-based recall for category Neurological.

Table 2.1.2.c

Cardiovascular	Neurological	Gastrointestinal	Respiratory	Other	Trauma	Metabolic
46.7%	46.7%	40.0%	40.0%	33.3%	6.7%	26.7%

Grouping by surgical_status

Worst ratio: 40.0% for category Surgical with the biggest delta 0.031 on Corrected event-based AUPRC.
Worst delta is the same as above.

Table 2.1.2.d

Surgical	Non-surgical
40.0%	13.3%

2.1.3. Top 3 cohorts with the biggest performance metric discrepancies

In the following table, we show for each performance metric the 3 cohorts with the biggest delta that are significantly worse off than the rest of the patients. If some cells are empty, this means that there are less than 3 cohorts, possibly none, that are significantly worse than the rest of the patients for this particular metric.

Table 2.1.3.a

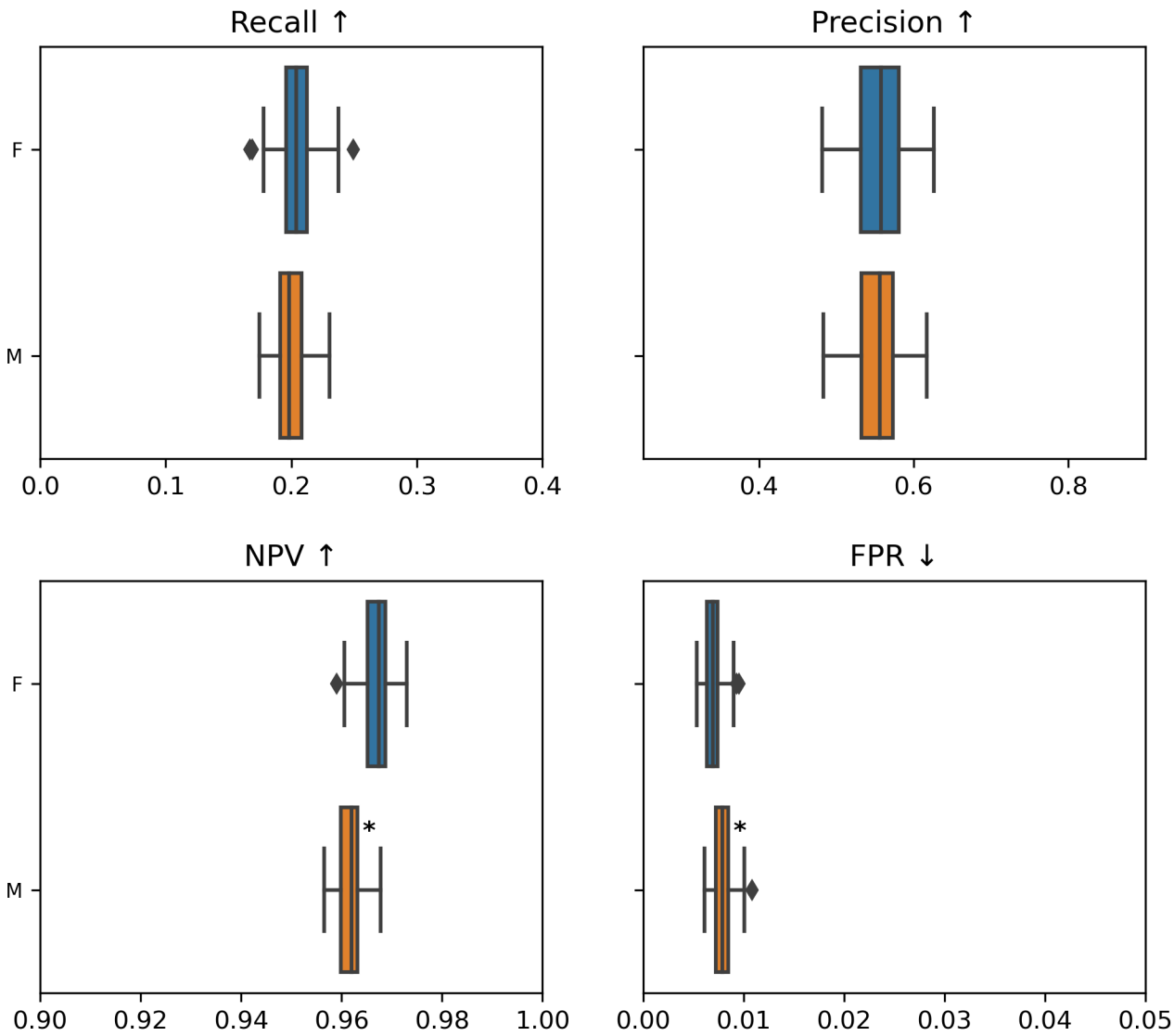
Metric	Cohort 1 (Δ)	Cohort 2 (Δ)	Cohort 3 (Δ)
Recall ↑	Neurological (0.156)	Respiratory (0.047)	Metabolic (0.038)
Precision ↑	Neurological (0.176)	50-65 (0.048)	75-85 (0.03)
NPV ↑	Cardiovascular (0.028)	Gastrointestinal (0.016)	Other (0.014)
FPR ↓	Gastrointestinal (0.008)	Cardiovascular (0.007)	75-85 (0.004)
Corrected precision ↑	Cardiovascular (0.104)	Gastrointestinal (0.067)	75-85 (0.04)
Corrected NPV ↑	-	-	-
Event-based recall ↑	Neurological (0.346)	Metabolic (0.088)	<50 (0.056)
Calibration error ↓	Neurological (0.096)	Metabolic (0.09)	>85 (0.056)
Avg. score on positive class	Neurological (0.133)	Respiratory (0.026)	Non-surgical (0.02)
Avg. score on negative class	Cardiovascular (0.031)	Gastrointestinal (0.013)	Surgical (0.01)
AUROC ↑	Cardiovascular (0.045)	75-85 (0.034)	Respiratory (0.027)
AUPRC ↑	Neurological (0.233)	75-85 (0.023)	50-65 (0.016)
Corrected AUPRC ↑	Cardiovascular (0.084)	M (0.042)	75-85 (0.035)
Event-based AUPRC ↑	Neurological (0.281)	75-85 (0.046)	M (0.041)
Corrected event-based AUPRC ↑	Cardiovascular (0.079)	M (0.068)	75-85 (0.053)

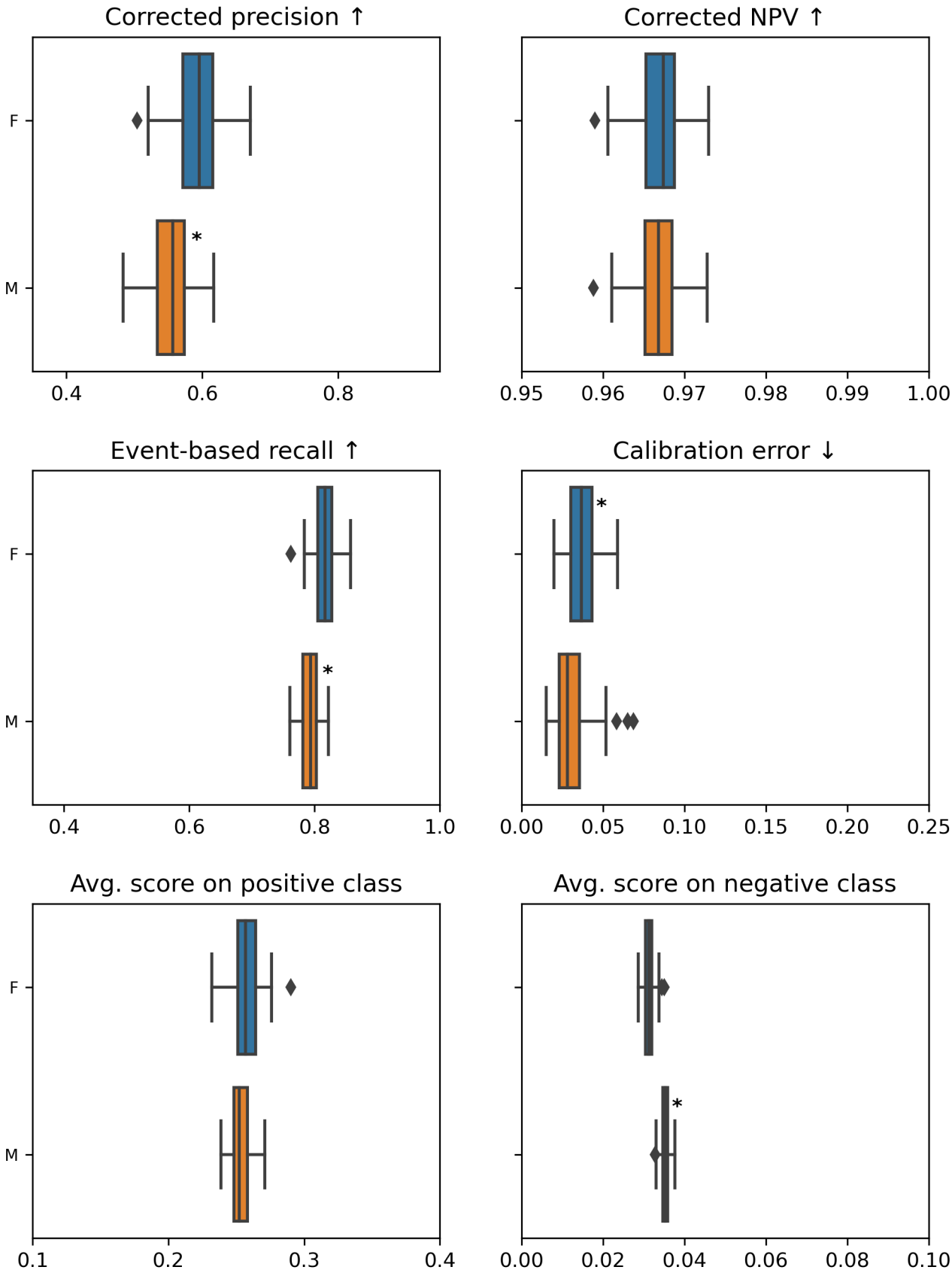
2.2. Grouping by

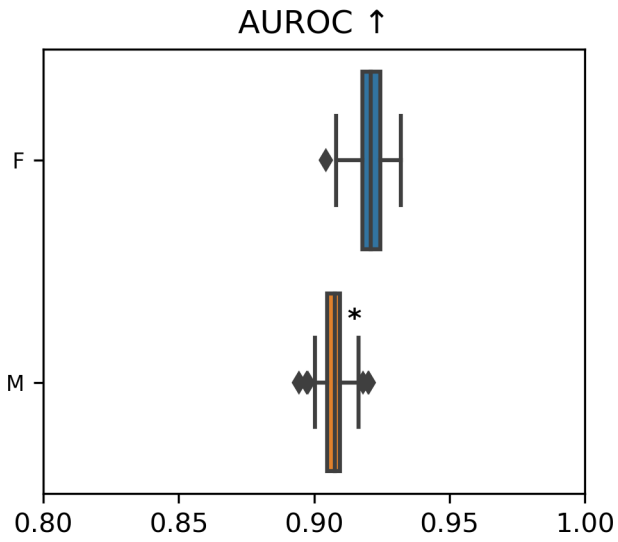
For each grouping, we display box plots that show the performance metrics' distributions for the different categories of patients. For each metric, we emphasize with a black star the cohorts that are significantly worse off compared to the rest of the patients and with a red star the cohorts that appear in the table **Top 3 cohorts with the biggest performance metric discrepancies**.
For each grouping, we propose a table that presents the results of the statistical analysis: comparing the different performance metrics for a cohort against the rest of the patients. P-values are obtained by running the Mann-Whitney U test with Bonferroni correction. We display only metrics and cohorts with a significant p-value (smaller than 0.001/number of comparisons) and whose delta is bigger than 0.
For binary grouping, we display the category with the worst distribution for each metric. While for multicategorical grouping, we display whether the distribution for the category is better or worse than for the rest of patients
We also display the calibration curve for each grouping's categories as well as the curves corresponding to each score-based metrics.

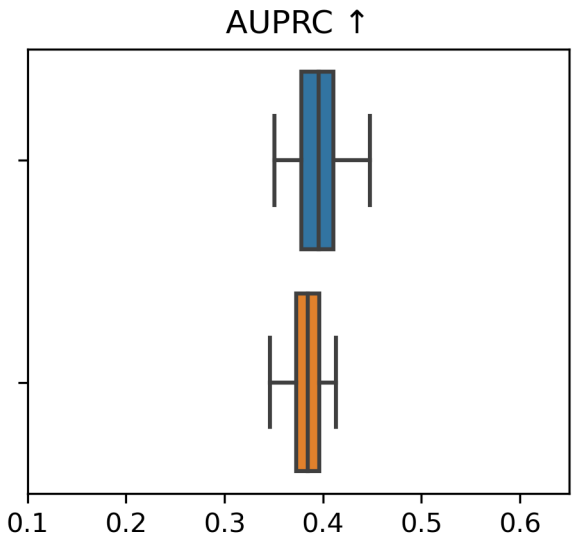
2.2.1. ... sex

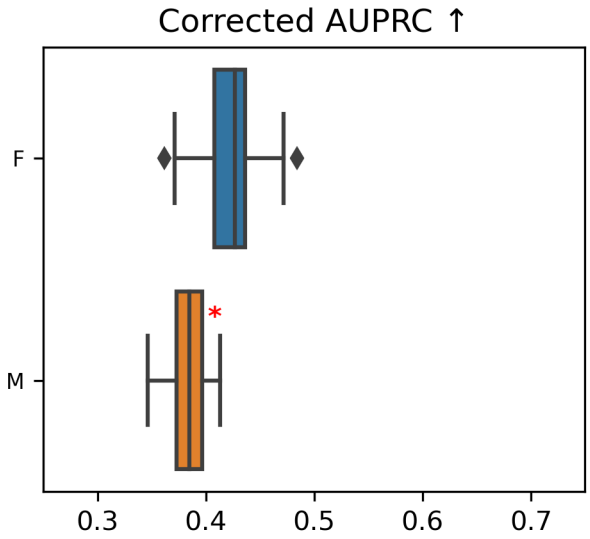
Figure 2.2.1.a

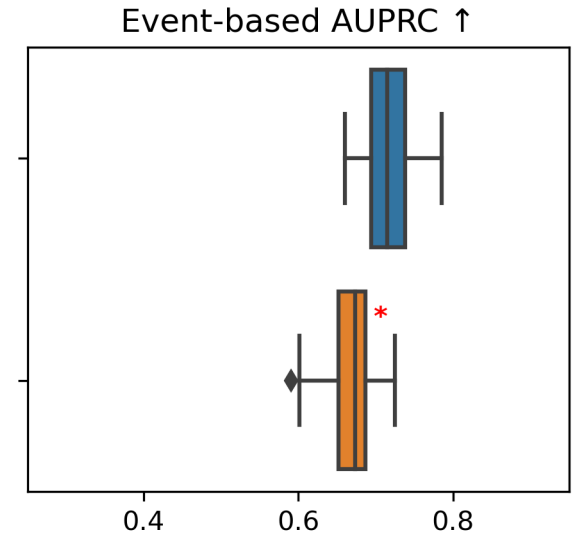



Corrected precision ↑
F
M
*
0.4
0.6
0.8
Corrected NPV ↑
0.95
0.96
0.97
0.98
0.99
1.00
Event-based recall ↑
F
M
*
*
0.4
0.6
0.8
1.0
Calibration error ↓
*
0.00
0.05
0.10
0.15
0.20
0.25
Avg. score on positive class
F
M
0.1
0.2
0.3
0.4
Avg. score on negative class
*
0.00
0.02
0.04
0.06
0.08
0.10









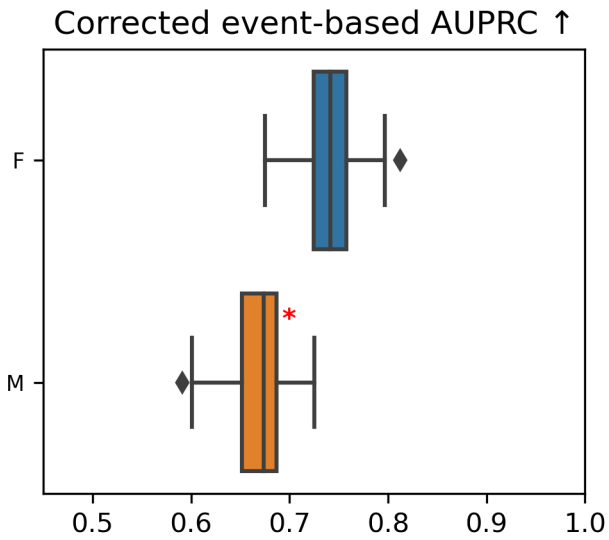


Table 2.2.1.a

Metric	Cohort with the worst metric	P-value	Delta
NPV ↑	M	1.30e-25	0.005
FPR ↓	M	1.89e-12	0.001
Corrected precision ↑	M	1.96e-14	0.039
Event-based recall ↑	M	6.82e-20	0.023
Calibration error ↓	F	5.68e-09	0.008

Avg. score on negative class	M	1.92e-33	0.004
AUROC ↑	M	3.57e-30	0.013
Corrected AUPRC ↑	M	9.56e-26	0.042
Event-based AUPRC ↑	M	1.13e-18	0.041
Corrected event-based AUPRC ↑	M	7.40e-31	0.068

Figure 2.2.1.b

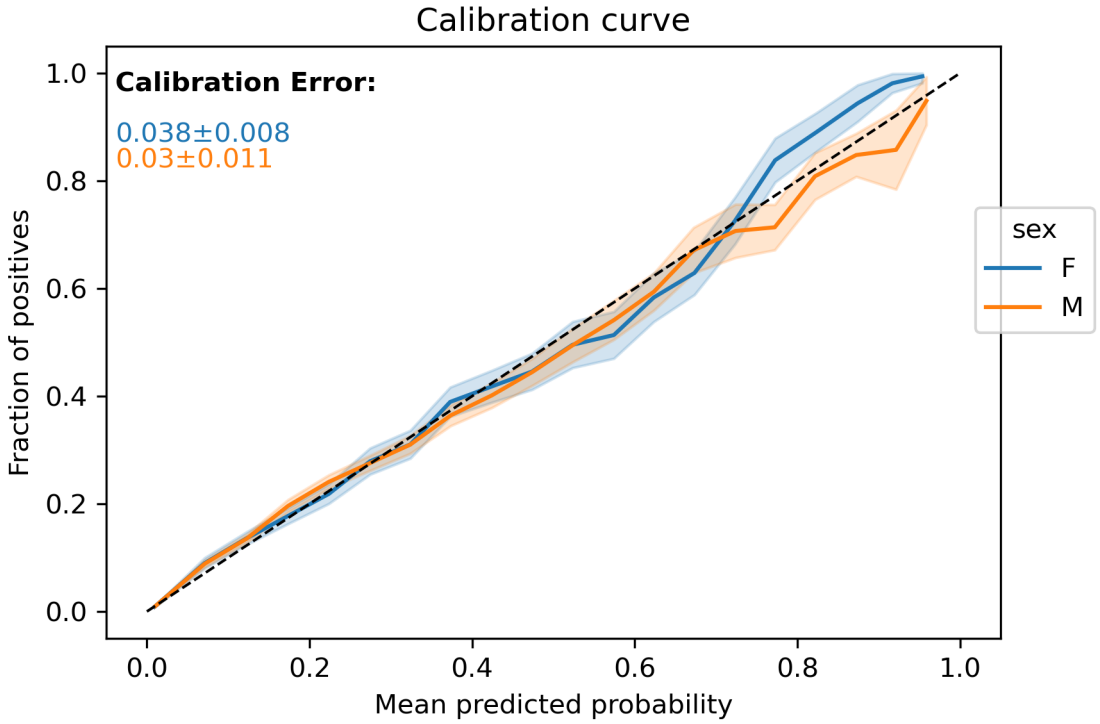


Figure 2.2.1.c

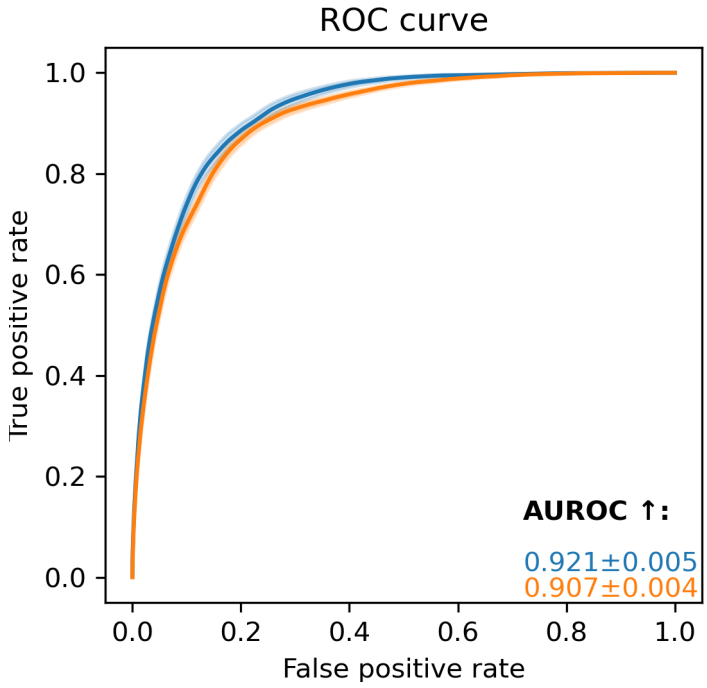
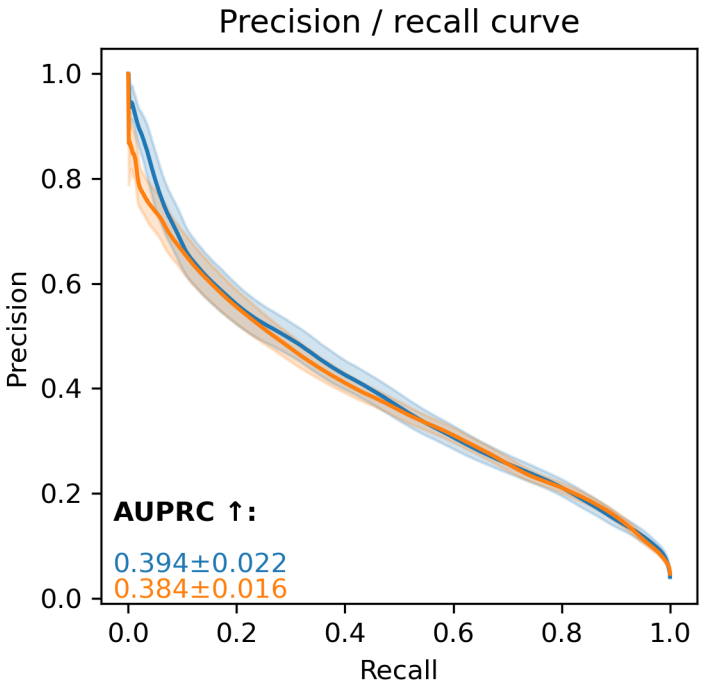
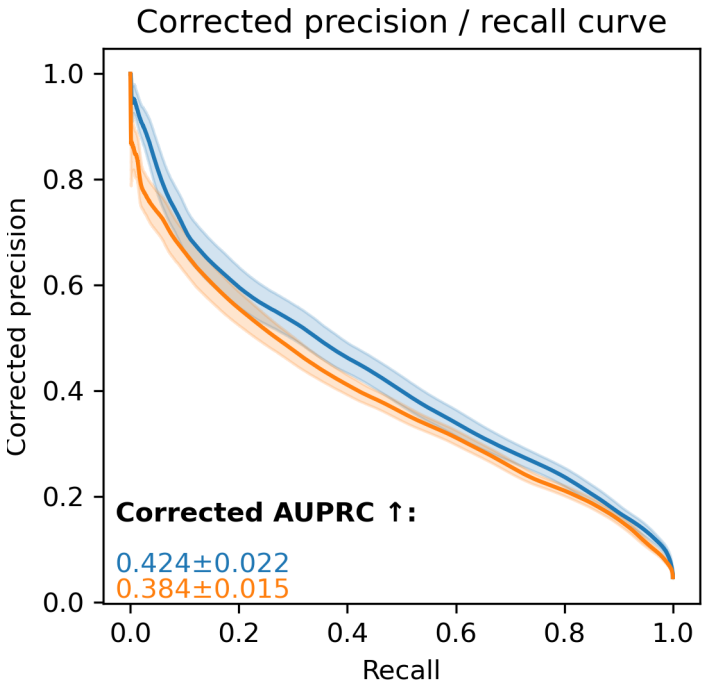
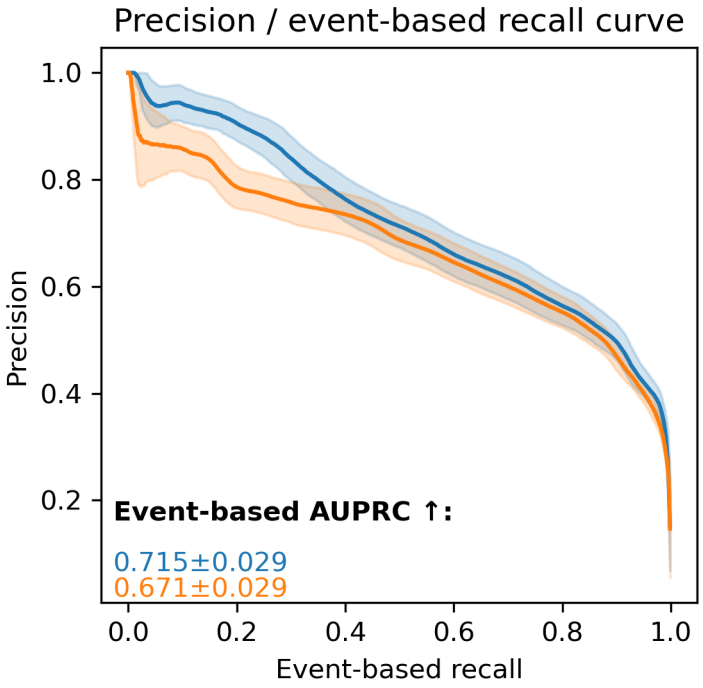
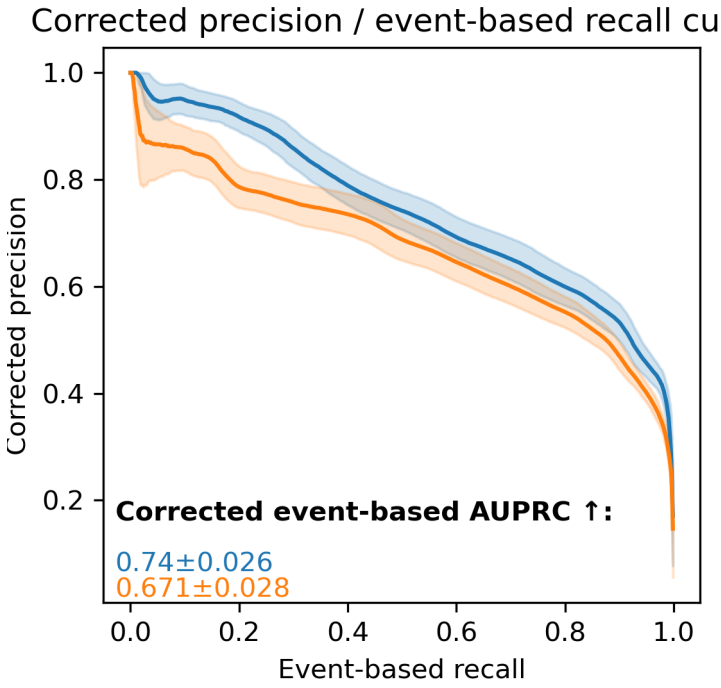


Figure 2.2.1.d


Precision / recall curve
Precision
Recall
AUPRC ↑:
0.394±0.022
0.384±0.016

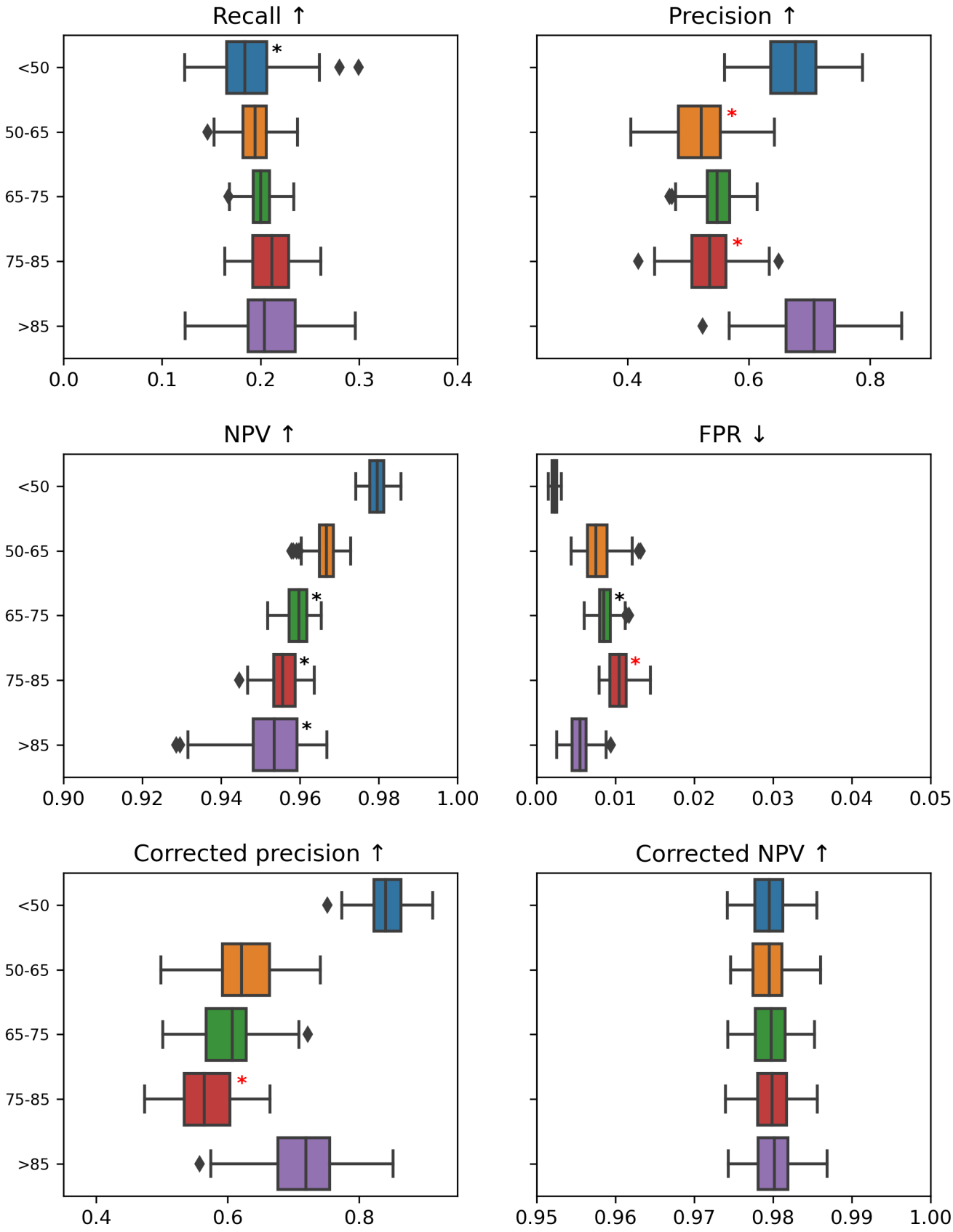

Corrected precision / recall curve
Corrected precision
Recall
Corrected AUPRC ↑:
0.424±0.022
0.384±0.015

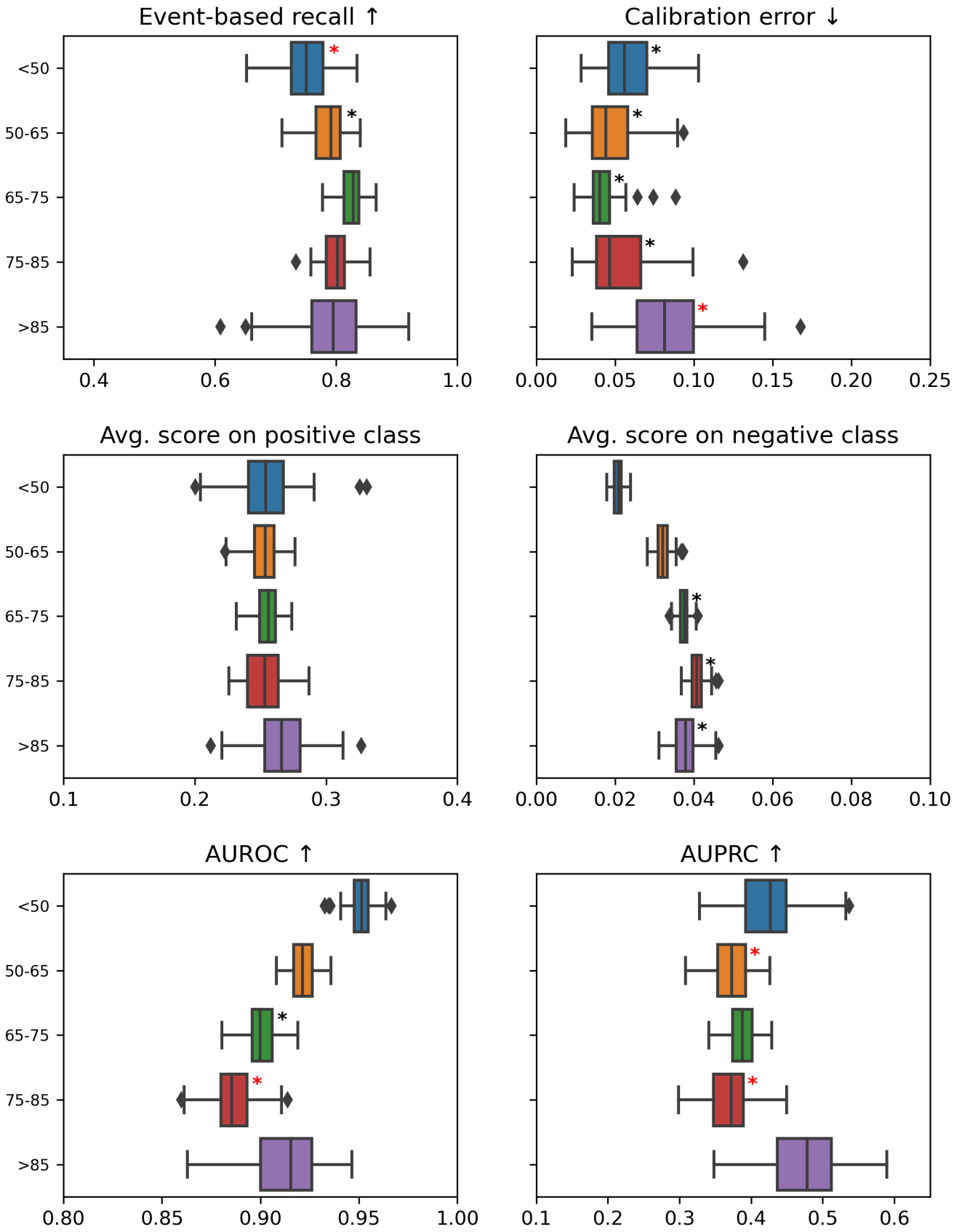

Precision / event-based recall curve
Precision
Event-based recall
Event-based AUPRC ↑:
0.715±0.029
0.671±0.029

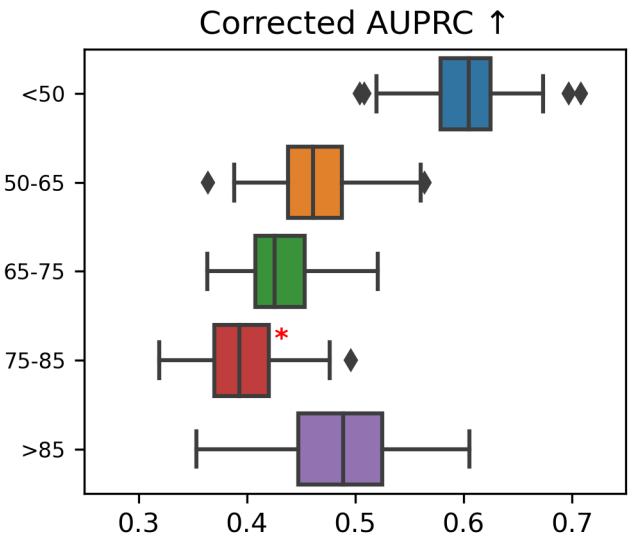

Corrected precision / event-based recall curve
Corrected precision
Event-based recall
Corrected event-based AUPRC ↑:
0.74±0.026
0.671±0.028

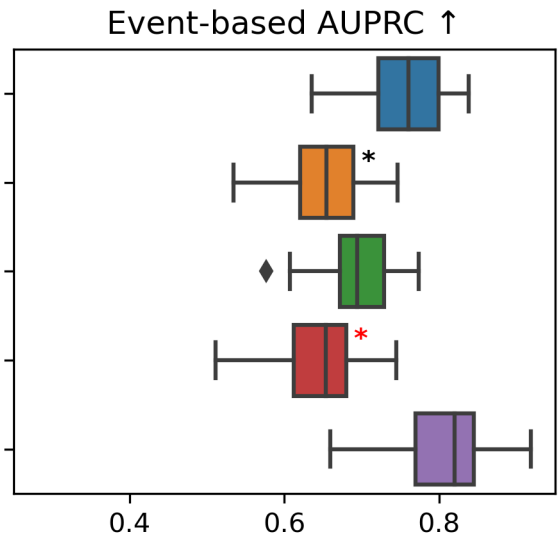
2.2.2. ... age_group

Figure 2.2.2.a




Event-based recall ↑
<50
50-65
65-75
75-85
>85
0.4
0.6
0.8
1.0
Calibration error ↓
0.00
0.05
0.10
0.15
0.20
0.25
Avg. score on positive class
0.1
0.2
0.3
0.4
Avg. score on negative class
0.00
0.02
0.04
0.06
0.08
0.10
AUROC ↑
0.80
0.85
0.90
0.95
1.00
AUPRC ↑
0.1
0.2
0.3
0.4
0.5
0.6





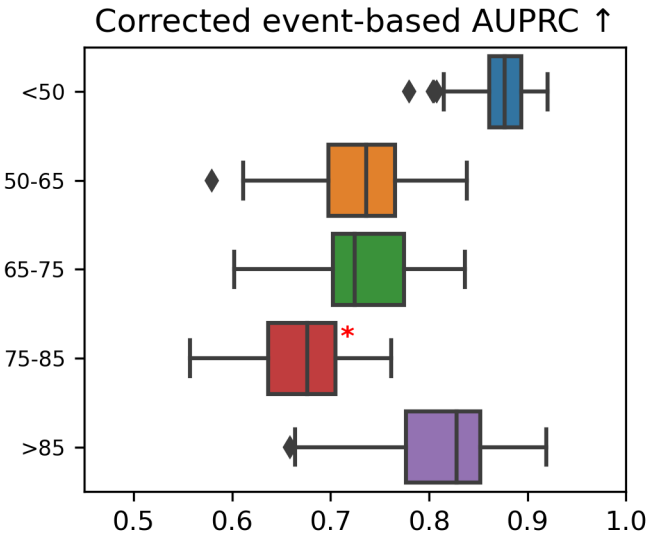


Table 2.2.2.a

Metric	Category	Cohort vs. rest	P-value	Delta
Recall ↑	<50	worse	9.18e-07	0.018
Recall ↑	75-85	better	8.03e-07	0.013
Precision ↑	<50	better	4.50e-34	0.13
Precision ↑	50-65	worse	7.48e-14	0.048
Precision ↑	75-85	worse	6.77e-07	0.03
Precision ↑	>85	better	6.61e-33	0.162
NPV ↑	<50	better	1.28e-34	0.019
NPV ↑	50-65	better	3.88e-19	0.004
NPV ↑	65-75	worse	1.93e-28	0.006
NPV ↑	75-85	worse	2.48e-34	0.011
NPV ↑	>85	worse	1.70e-27	0.011
FPR ↓	<50	better	1.28e-34	0.006
FPR ↓	65-75	worse	2.98e-19	0.001
FPR ↓	75-85	worse	4.50e-34	0.004
FPR ↓	>85	better	2.66e-25	0.002
Corrected precision ↑	<50	better	1.28e-34	0.294
Corrected precision ↑	50-65	better	5.62e-11	0.04

Corrected precision ↑	75-85	worse	2.27e-09	0.04
Corrected precision ↑	>85	better	5.35e-32	0.146
Corrected NPV ↑	<50	better	1.28e-34	0.013
Corrected NPV ↑	50-65	better	1.28e-34	0.013
Corrected NPV ↑	65-75	better	1.28e-34	0.013
Corrected NPV ↑	75-85	better	1.28e-34	0.014
Corrected NPV ↑	>85	better	1.28e-34	0.014
Event-based recall ↑	<50	worse	6.44e-24	0.056
Event-based recall ↑	50-65	worse	2.22e-06	0.013
Event-based recall ↑	65-75	better	6.23e-27	0.041
Calibration error ↓	<50	worse	2.23e-29	0.029
Calibration error ↓	50-65	worse	1.03e-25	0.021
Calibration error ↓	65-75	worse	1.01e-20	0.014
Calibration error ↓	75-85	worse	7.78e-26	0.019
Calibration error ↓	>85	worse	8.92e-34	0.056
Avg. score on positive class	>85	better	3.18e-08	0.012
Avg. score on negative class	<50	better	1.28e-34	0.016
Avg. score on negative class	50-65	better	1.82e-20	0.002
Avg. score on negative class	65-75	worse	2.13e-34	0.005
Avg. score on negative class	75-85	worse	1.28e-34	0.009
Avg. score on negative class	>85	worse	1.51e-21	0.004
AUROC ↑	<50	better	1.28e-34	0.048
AUROC ↑	50-65	better	5.09e-29	0.013
AUROC ↑	65-75	worse	1.59e-30	0.017
AUROC ↑	75-85	worse	1.49e-34	0.034
AUPRC ↑	<50	better	6.73e-13	0.044
AUPRC ↑	50-65	worse	2.03e-07	0.016
AUPRC ↑	75-85	worse	1.33e-09	0.023
AUPRC ↑	>85	better	1.44e-29	0.097
Corrected AUPRC ↑	<50	better	1.28e-34	0.222
Corrected AUPRC ↑	50-65	better	4.89e-26	0.061
Corrected AUPRC ↑	75-85	worse	7.34e-14	0.035
Corrected AUPRC ↑	>85	better	3.98e-27	0.086
Event-based AUPRC ↑	<50	better	7.33e-25	0.083
Event-based AUPRC ↑	50-65	worse	8.e-11	0.041
Event-based AUPRC ↑	75-85	worse	4.99e-15	0.046
Event-based AUPRC ↑	>85	better	1.13e-30	0.145
Corrected event-based AUPRC ↑	<50	better	1.28e-34	0.199
Corrected event-based AUPRC ↑	50-65	better	2.24e-06	0.031
Corrected event-based AUPRC ↑	75-85	worse	1.37e-18	0.053
Corrected event-based AUPRC ↑	>85	better	5.53e-29	0.132

Figure 2.2.2.b

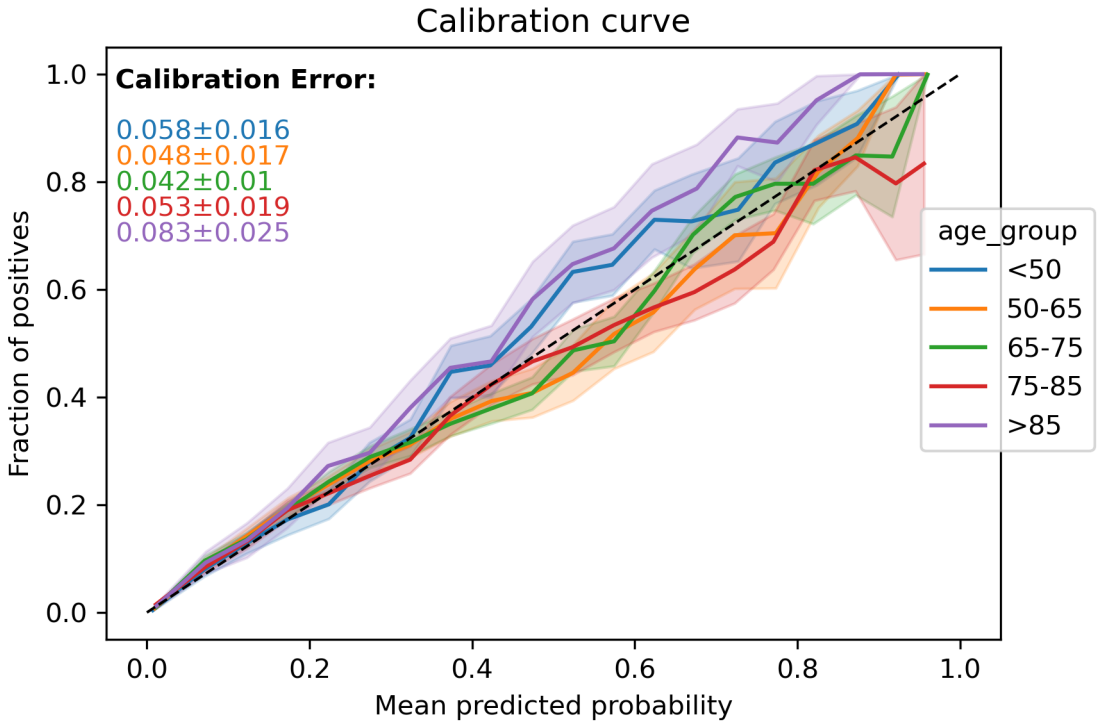


Figure 2.2.2.c

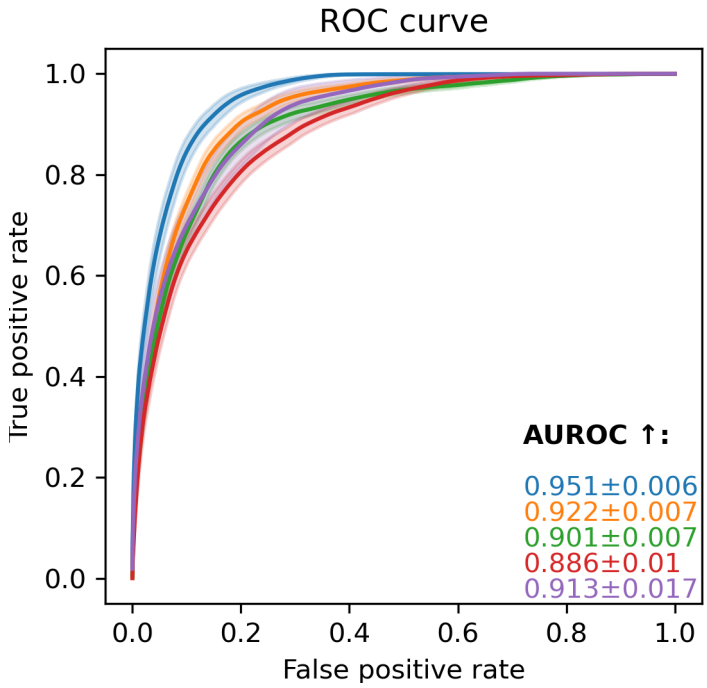
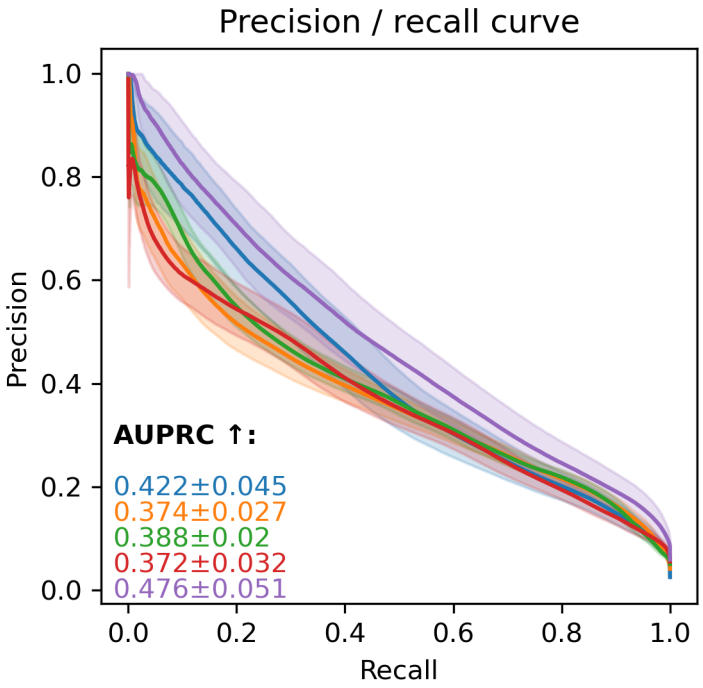
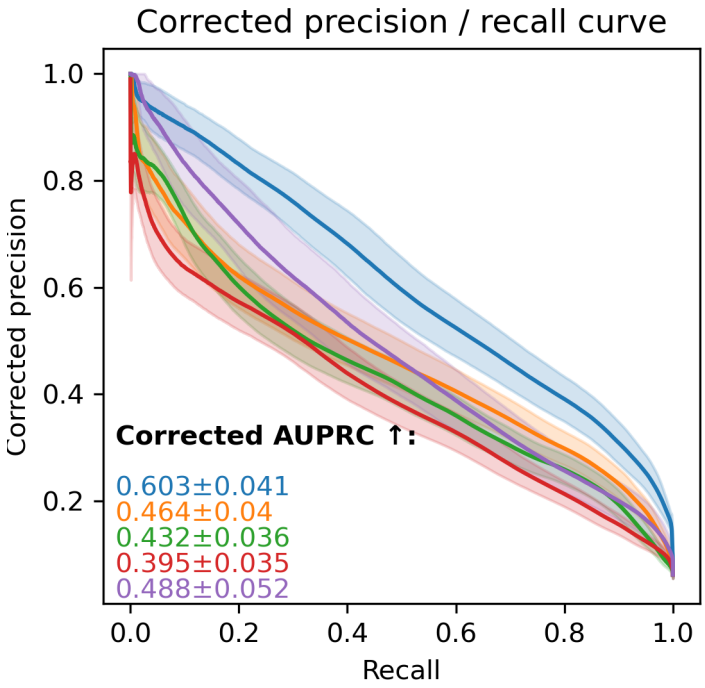
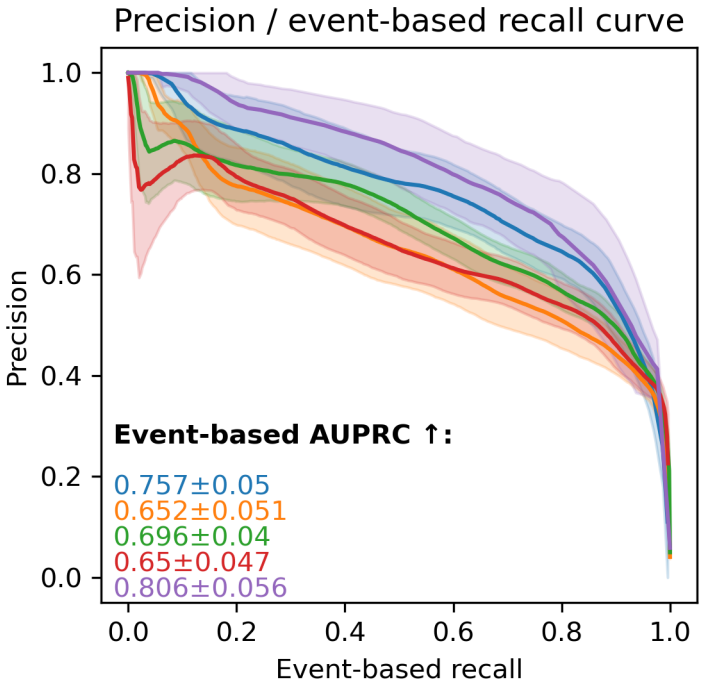
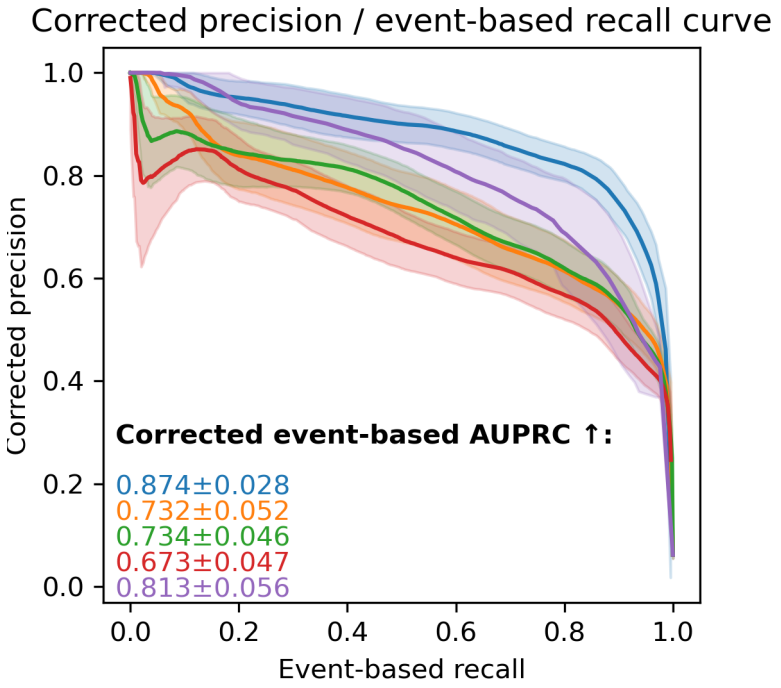


Figure 2.2.2.d


Precision / recall curve
Precision
Recall
AUPRC ↑:
0.422±0.045
0.374±0.027
0.388±0.02
0.372±0.032
0.476±0.051

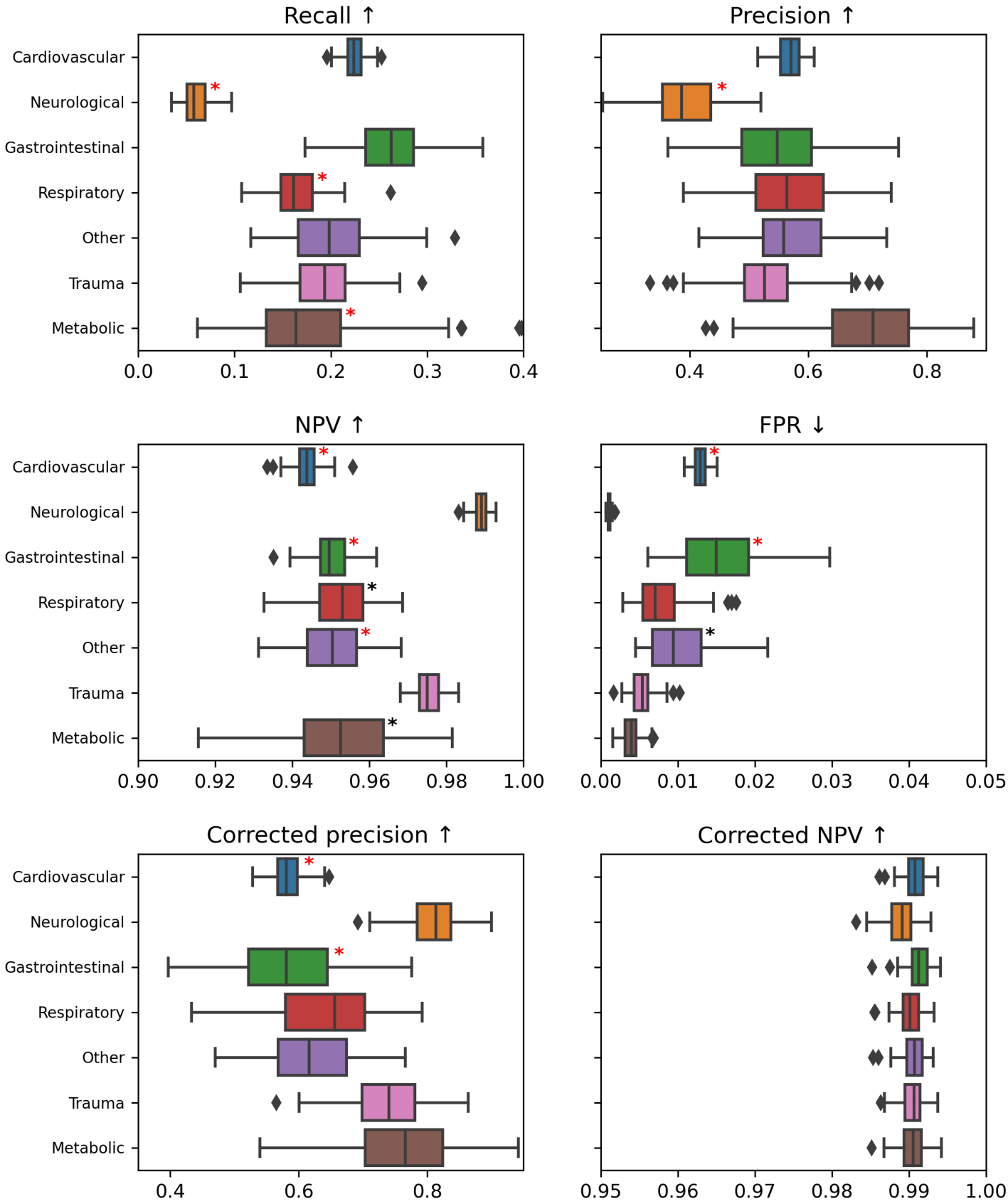

Corrected precision / recall curve
Corrected precision
Recall
Corrected AUPRC ↑:
0.603±0.041
0.464±0.04
0.432±0.036
0.395±0.035
0.488±0.052

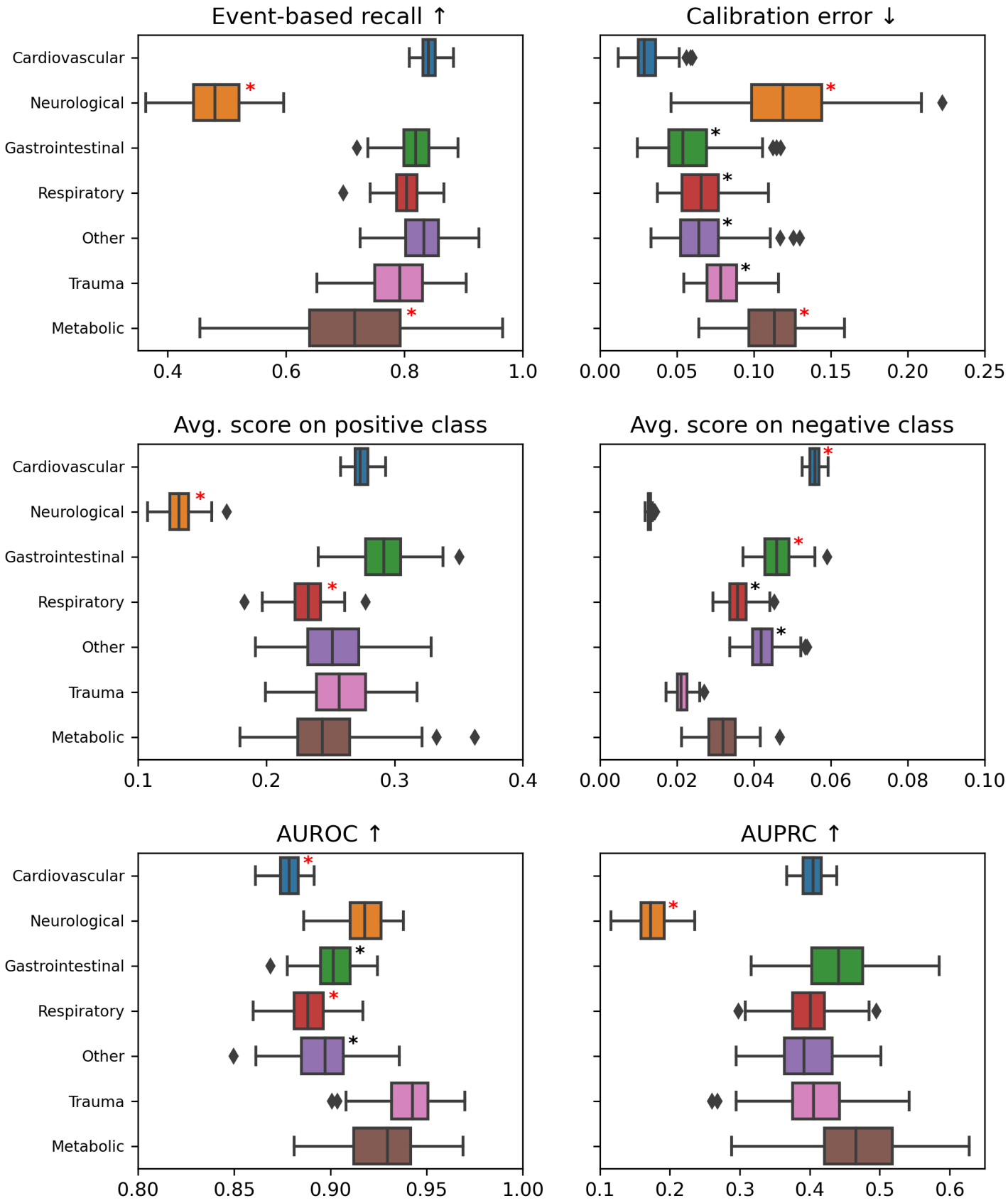

Precision / event-based recall curve
Precision
Event-based recall
Event-based AUPRC ↑:
0.757±0.05
0.652±0.051
0.696±0.04
0.65±0.047
0.806±0.056

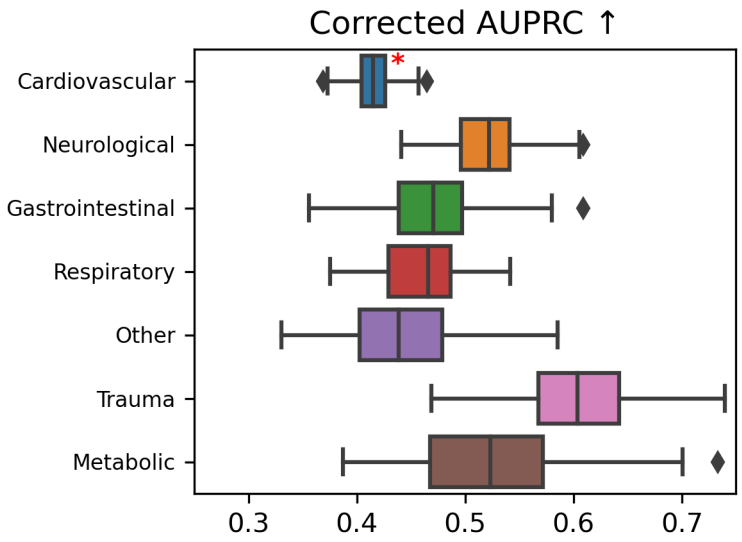

Corrected precision / event-based recall curve
Corrected precision
Event-based recall
Corrected event-based AUPRC ↑:
0.874±0.028
0.732±0.052
0.734±0.046
0.673±0.047
0.813±0.056

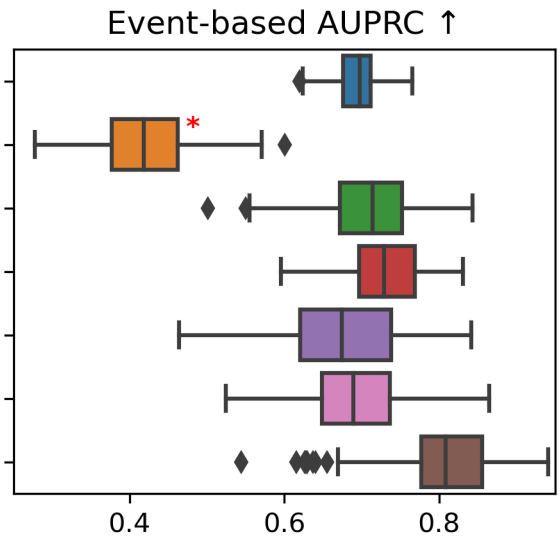
2.2.3. ... APACHE_group

Figure 2.2.3.a




Event-based recall ↑
Calibration error ↓
Avg. score on positive class
Avg. score on negative class
AUROC ↑
AUPRC ↑
Cardiovascular
Neurological
Gastrointestinal
Respiratory
Other
Trauma
Metabolic
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
0.20
0.25
0.1
0.2
0.3
0.4
0.00
0.02
0.04
0.06
0.08
0.10
0.80
0.85
0.90
0.95
1.00
0.1
0.2
0.3
0.4
0.5
0.6





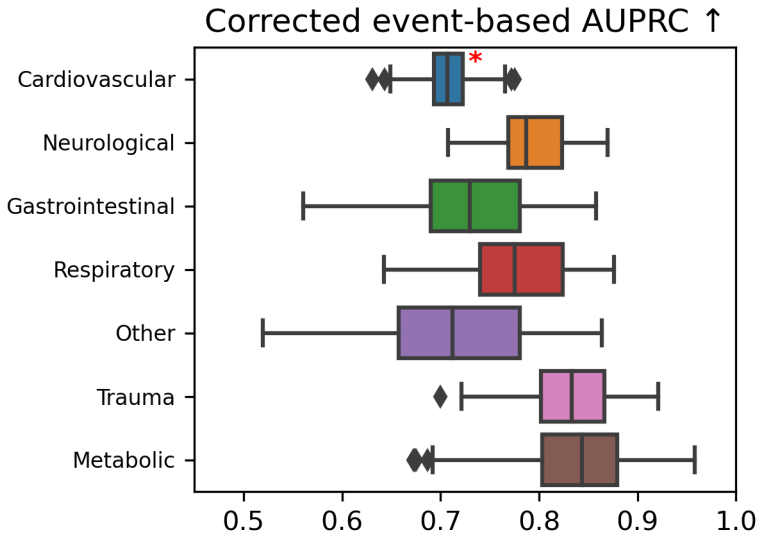


Table 2.2.3.a

Metric	Category	Cohort vs. rest	P-value	Delta
Recall ↑	Cardiovascular	better	1.38e-31	0.044
Recall ↑	Neurological	worse	1.28e-34	0.156
Recall ↑	Gastrointestinal	better	7.78e-32	0.071
Recall ↑	Respiratory	worse	7.05e-28	0.047
Recall ↑	Metabolic	worse	1.05e-07	0.038
Precision ↑	Cardiovascular	better	2.35e-08	0.027
Precision ↑	Neurological	worse	1.44e-34	0.176
Precision ↑	Metabolic	better	1.64e-22	0.155
NPV ↑	Cardiovascular	worse	1.28e-34	0.028
NPV ↑	Neurological	better	1.28e-34	0.038
NPV ↑	Gastrointestinal	worse	1.63e-34	0.016
NPV ↑	Respiratory	worse	2.16e-30	0.012
NPV ↑	Other	worse	5.91e-30	0.014
NPV ↑	Trauma	better	1.28e-34	0.012
NPV ↑	Metabolic	worse	4.35e-10	0.011
FPR ↓	Cardiovascular	worse	1.28e-34	0.007
FPR ↓	Neurological	better	1.28e-34	0.01

FPR ↓	Gastrointestinal	worse	4.52e-33	0.008
FPR ↓	Other	worse	2.04e-06	0.002
FPR ↓	Trauma	better	9.45e-22	0.002
FPR ↓	Metabolic	better	1.13e-33	0.004
Corrected precision ↑	Cardiovascular	worse	5.07e-34	0.104
Corrected precision ↑	Neurological	better	1.28e-34	0.252
Corrected precision ↑	Gastrointestinal	worse	2.1e-10	0.067
Corrected precision ↑	Trauma	better	1.84e-27	0.114
Corrected precision ↑	Metabolic	better	1.03e-25	0.134
Corrected NPV ↑	Cardiovascular	better	1.28e-34	0.019
Corrected NPV ↑	Neurological	better	1.28e-34	0.016
Corrected NPV ↑	Gastrointestinal	better	1.28e-34	0.019
Corrected NPV ↑	Respiratory	better	1.28e-34	0.018
Corrected NPV ↑	Other	better	1.28e-34	0.018
Corrected NPV ↑	Trauma	better	1.28e-34	0.018
Corrected NPV ↑	Metabolic	better	1.28e-34	0.018
Event-based recall ↑	Cardiovascular	better	1.32e-34	0.081
Event-based recall ↑	Neurological	worse	1.28e-34	0.346
Event-based recall ↑	Gastrointestinal	better	5.90e-10	0.02
Event-based recall ↑	Other	better	3.13e-13	0.035
Event-based recall ↑	Metabolic	worse	1.55e-13	0.088
Calibration error ↓	Cardiovascular	better	2.11e-11	0.01
Calibration error ↓	Neurological	worse	1.28e-34	0.096
Calibration error ↓	Gastrointestinal	worse	3.38e-30	0.032
Calibration error ↓	Respiratory	worse	5.88e-33	0.041
Calibration error ↓	Other	worse	4.93e-33	0.04
Calibration error ↓	Trauma	worse	1.28e-34	0.056
Calibration error ↓	Metabolic	worse	1.28e-34	0.09
Avg. score on positive class	Cardiovascular	better	1.58e-34	0.035
Avg. score on positive class	Neurological	worse	1.28e-34	0.133
Avg. score on positive class	Gastrointestinal	better	1.94e-32	0.043
Avg. score on positive class	Respiratory	worse	1.57e-27	0.026
Avg. score on negative class	Cardiovascular	worse	1.28e-34	0.031
Avg. score on negative class	Neurological	better	1.28e-34	0.032
Avg. score on negative class	Gastrointestinal	worse	1.28e-34	0.013
Avg. score on negative class	Respiratory	worse	1.71e-10	0.002
Avg. score on negative class	Other	worse	3.15e-34	0.009
Avg. score on negative class	Trauma	better	1.28e-34	0.014
AUROC ↑	Cardiovascular	worse	1.28e-34	0.045
AUROC ↑	Neurological	better	2.e-29	0.023
AUROC ↑	Gastrointestinal	worse	1.18e-12	0.011
AUROC ↑	Respiratory	worse	2.99e-33	0.027
AUROC ↑	Other	worse	1.03e-18	0.016

AUROC ↑	Trauma	better	1.34e-30	0.033
AUROC ↑	Metabolic	better	1.05e-10	0.019
AUPRC ↑	Cardiovascular	better	8.15e-19	0.028
AUPRC ↑	Neurological	worse	1.28e-34	0.233
AUPRC ↑	Gastrointestinal	better	1.40e-18	0.062
AUPRC ↑	Metabolic	better	4.27e-17	0.081
Corrected AUPRC ↑	Cardiovascular	worse	1.53e-34	0.084
Corrected AUPRC ↑	Neurological	better	1.28e-34	0.117
Corrected AUPRC ↑	Trauma	better	1.63e-34	0.158
Corrected AUPRC ↑	Metabolic	better	7.75e-17	0.072
Event-based AUPRC ↑	Neurological	worse	1.28e-34	0.281
Event-based AUPRC ↑	Respiratory	better	4.99e-15	0.047
Event-based AUPRC ↑	Metabolic	better	1.99e-23	0.126
Corrected event-based AUPRC ↑	Cardiovascular	worse	8.48e-32	0.079
Corrected event-based AUPRC ↑	Neurological	better	8.35e-33	0.088
Corrected event-based AUPRC ↑	Respiratory	better	5.11e-08	0.03
Corrected event-based AUPRC ↑	Trauma	better	3.23e-26	0.095
Corrected event-based AUPRC ↑	Metabolic	better	8.35e-23	0.101

Figure 2.2.3.b

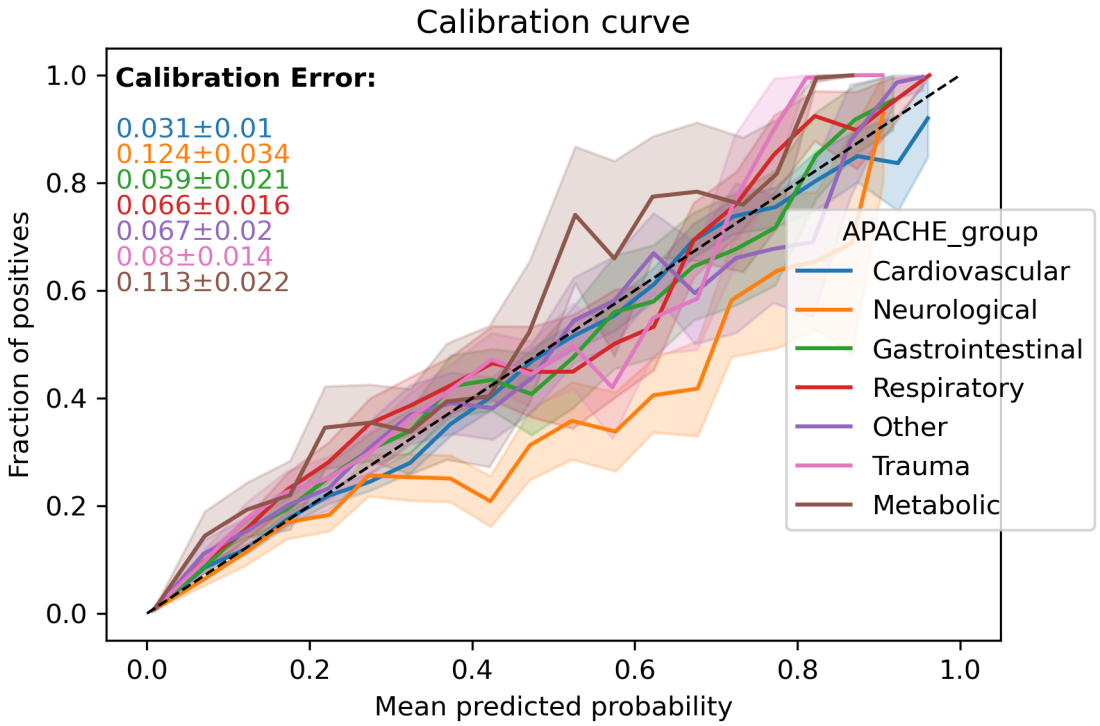


Figure 2.2.3.c

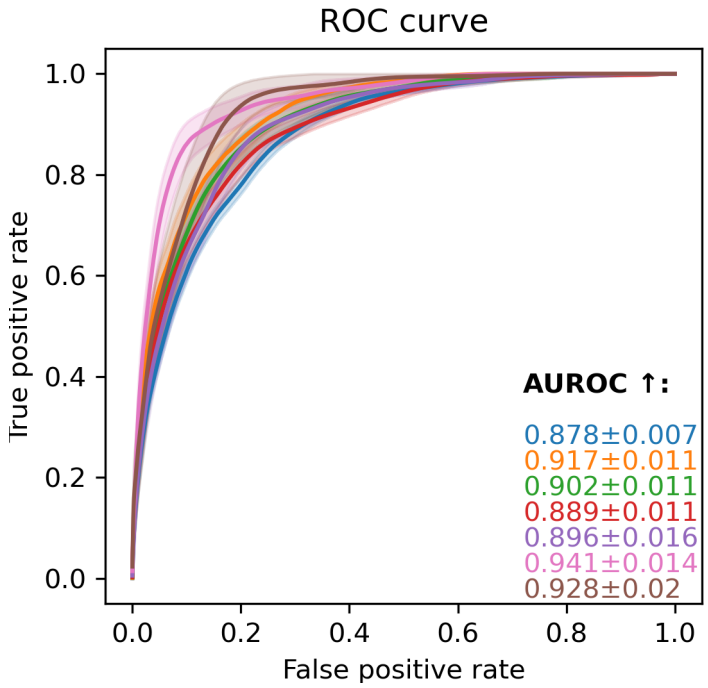
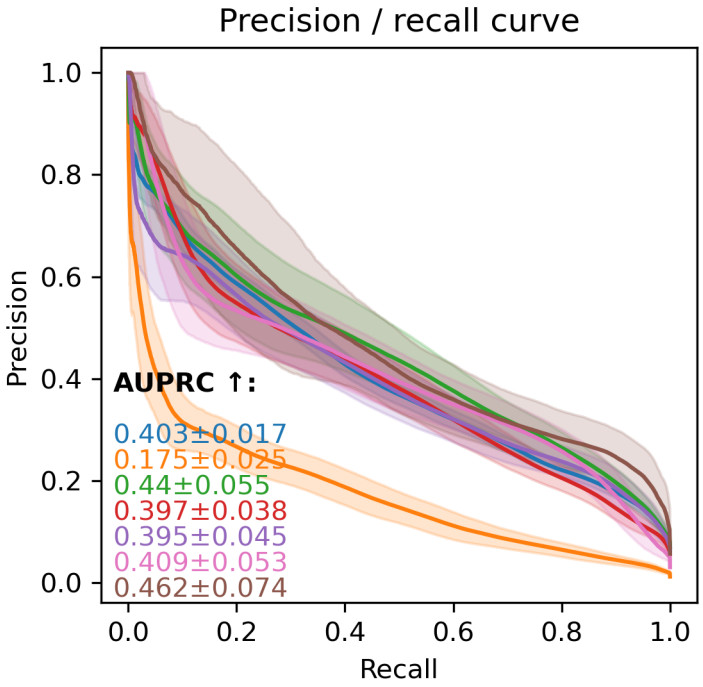
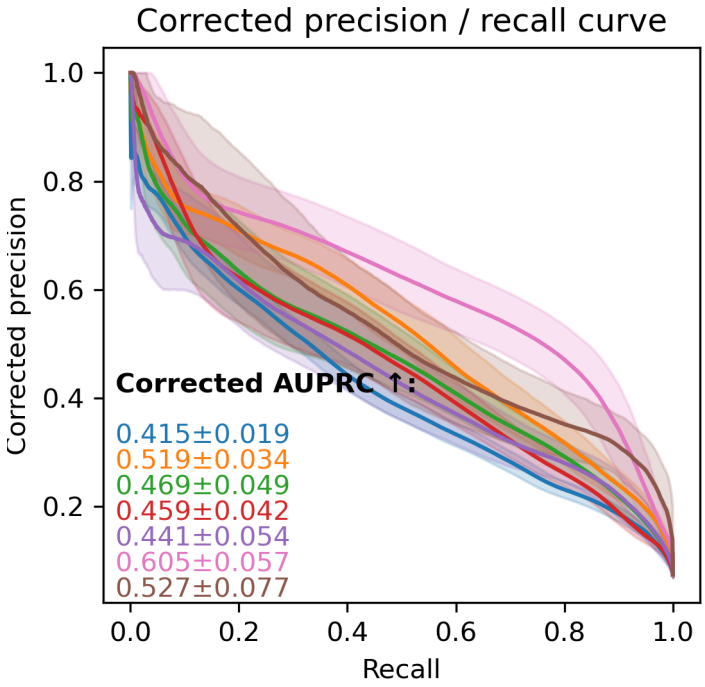
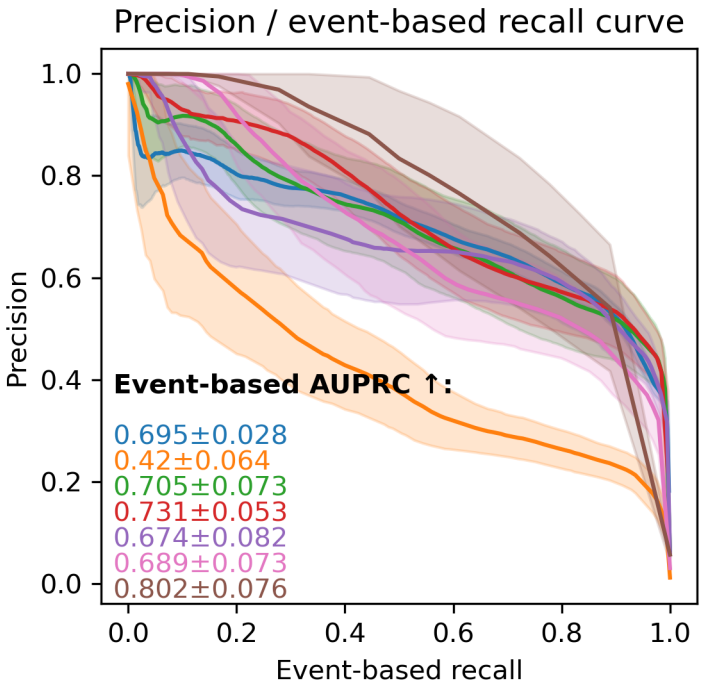
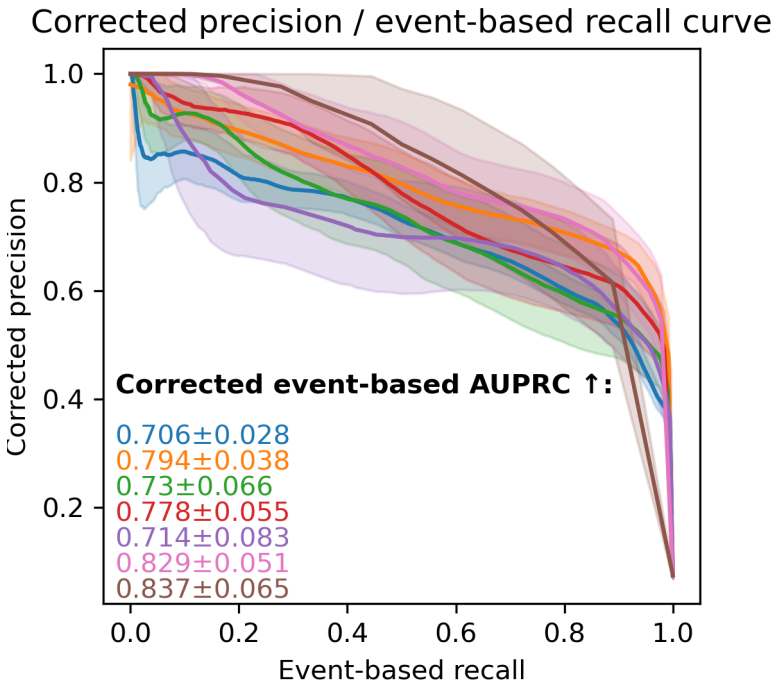

ROC curve
True positive rate
False positive rate
AUROC ↑:
0.878±0.007
0.917±0.011
0.902±0.011
0.889±0.011
0.896±0.016
0.941±0.014
0.928±0.02

Figure 2.2.3.d


Precision / recall curve
Precision
Recall
AUPRC ↑:
0.403±0.017
0.175±0.025
0.44±0.055
0.397±0.038
0.395±0.045
0.409±0.053
0.462±0.074

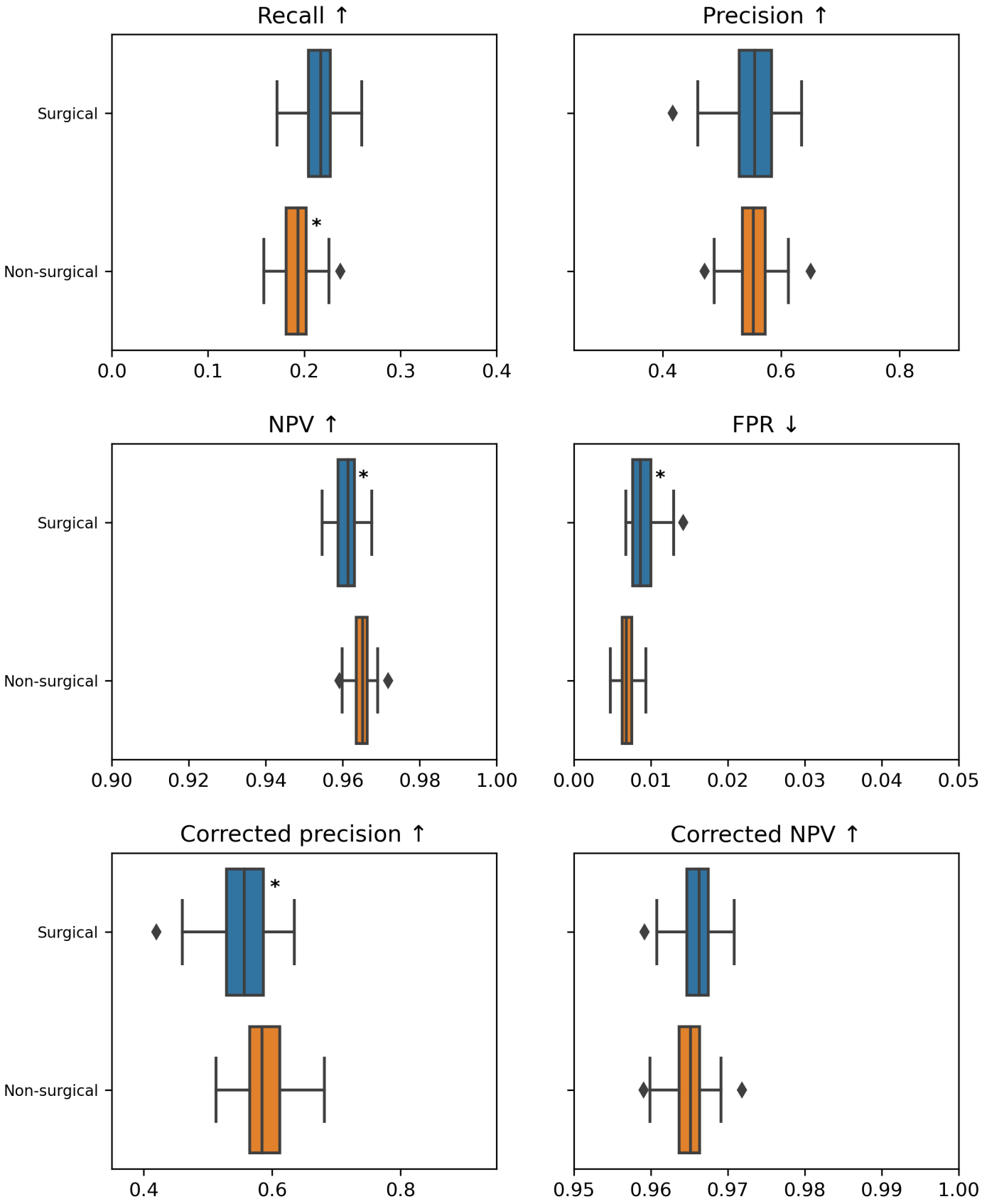

Corrected precision / recall curve
Corrected precision
Recall
Corrected AUPRC ↑:
0.415±0.019
0.519±0.034
0.469±0.049
0.459±0.042
0.441±0.054
0.605±0.057
0.527±0.077

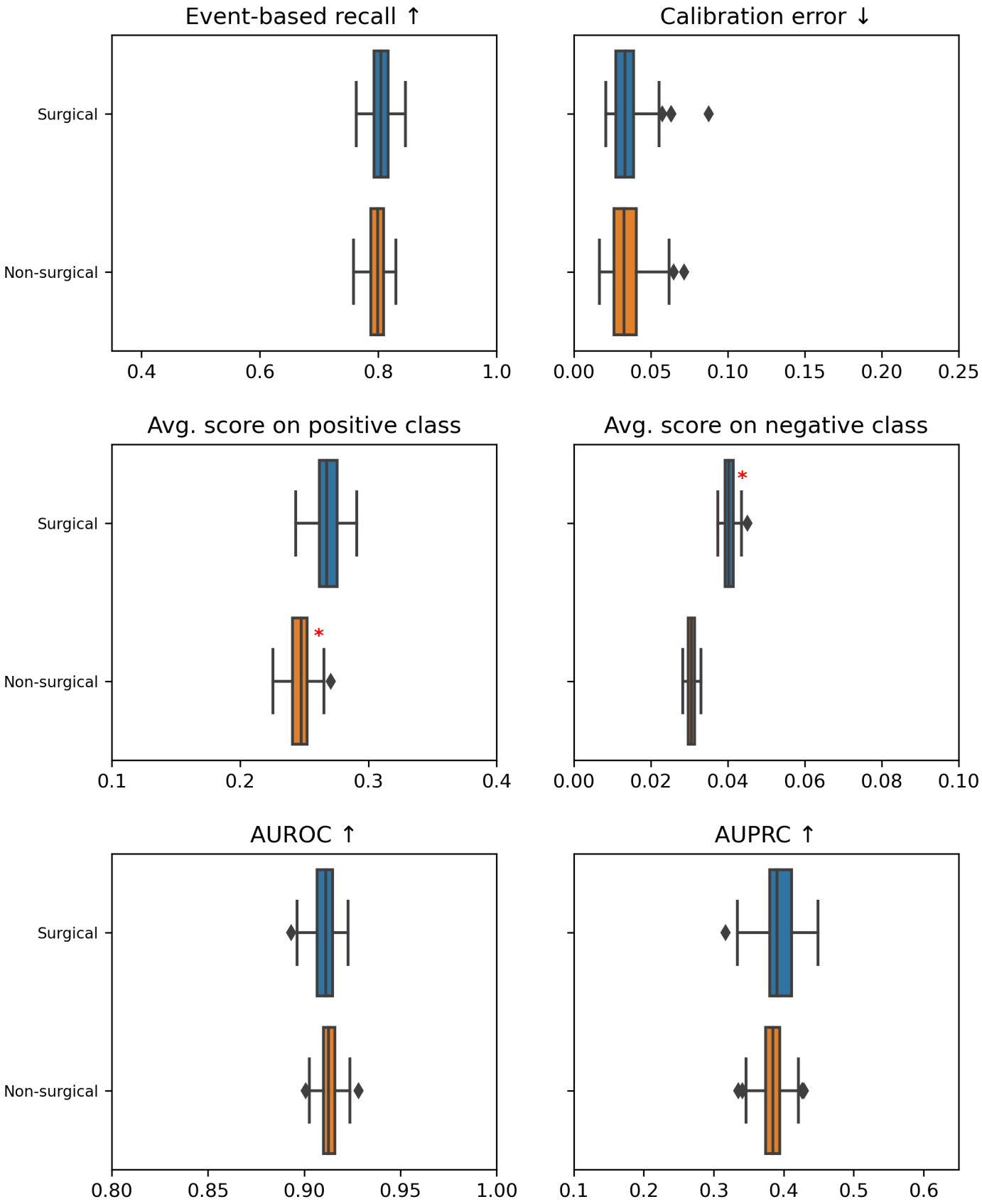

Precision / event-based recall curve
Precision
Event-based recall
Event-based AUPRC ↑:
0.695±0.028
0.42±0.064
0.705±0.073
0.731±0.053
0.674±0.082
0.689±0.073
0.802±0.076

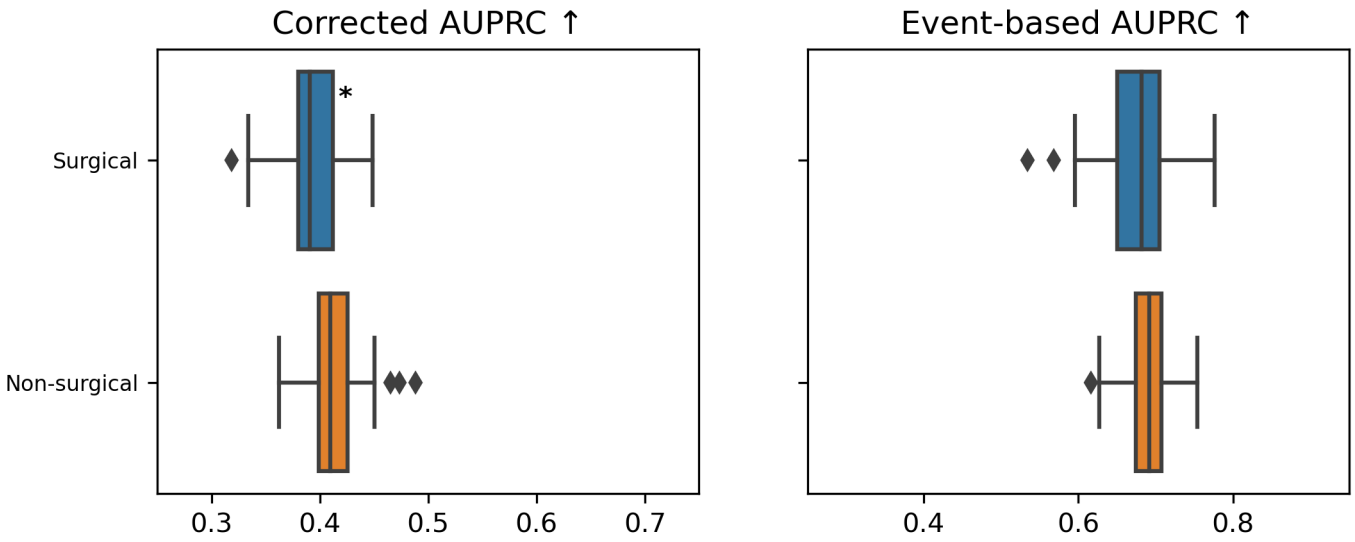

Corrected precision / event-based recall curve
Corrected precision
Event-based recall
Corrected event-based AUPRC ↑:
0.706±0.028
0.794±0.038
0.73±0.066
0.778±0.055
0.714±0.083
0.829±0.051
0.837±0.065

2.2.4. ... surgical_status

Figure 2.2.4.a




Event-based recall ↑
Surgical
Non-surgical
0.4
0.6
0.8
1.0
Calibration error ↓
0.00
0.05
0.10
0.15
0.20
0.25
Avg. score on positive class
Surgical
Non-surgical
*
0.1
0.2
0.3
0.4
Avg. score on negative class
*
0.00
0.02
0.04
0.06
0.08
0.10
AUROC ↑
Surgical
Non-surgical
0.80
0.85
0.90
0.95
1.00
AUPRC ↑
0.1
0.2
0.3
0.4
0.5
0.6



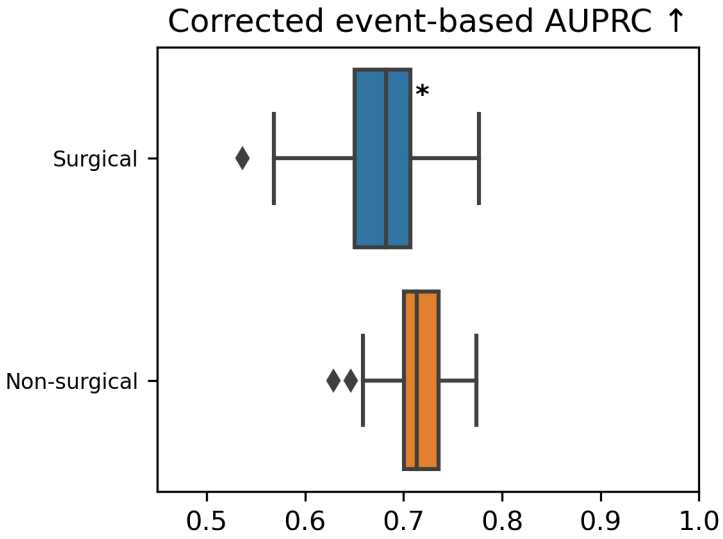


Table 2.2.4.a

Metric	Cohort with the worst metric	P-value	Delta
Recall ↑	Non-surgical	2.1e-19	0.024
NPV ↑	Surgical	8.23e-18	0.004
FPR ↓	Surgical	1.62e-21	0.002
Corrected precision ↑	Surgical	1.57e-08	0.027
Avg. score on positive class	Non-surgical	3.41e-28	0.02
Avg. score on negative class	Surgical	1.28e-34	0.01
Corrected AUPRC ↑	Surgical	6.08e-08	0.019
Corrected event-based AUPRC ↑	Surgical	1.08e-09	0.031

Figure 2.2.4.b

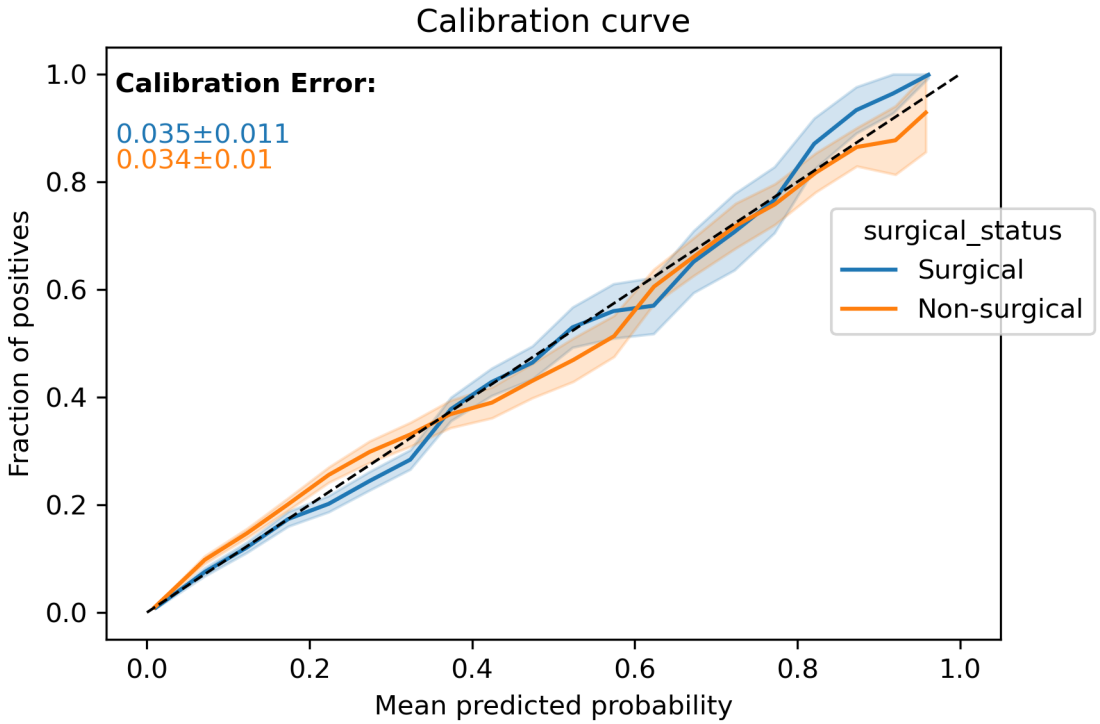


Figure 2.2.4.c

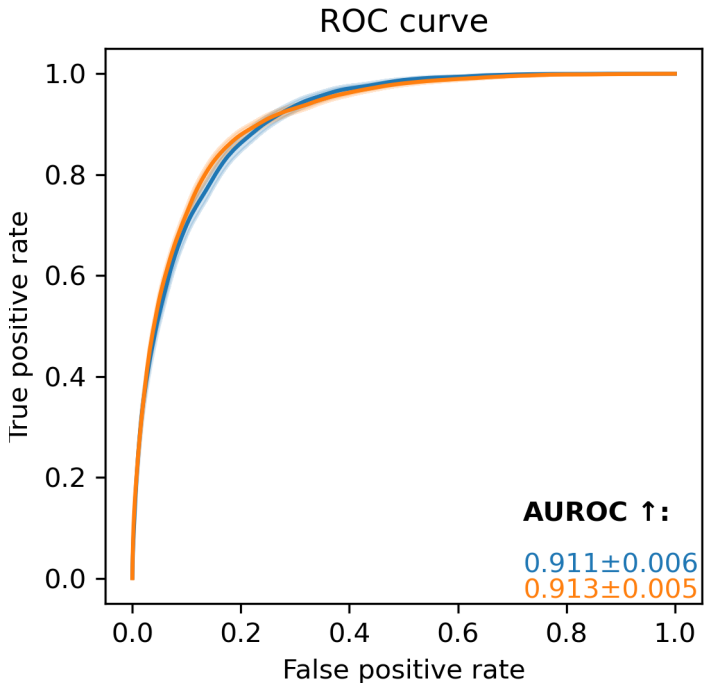
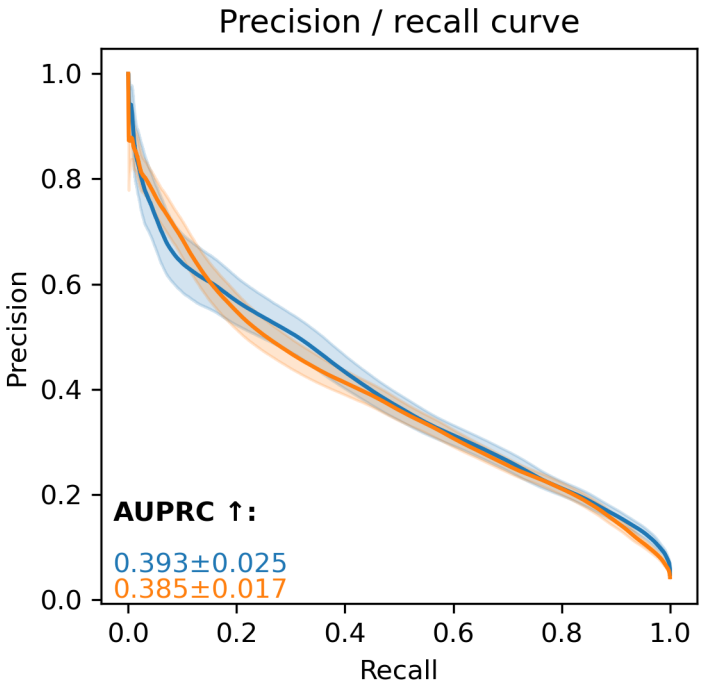
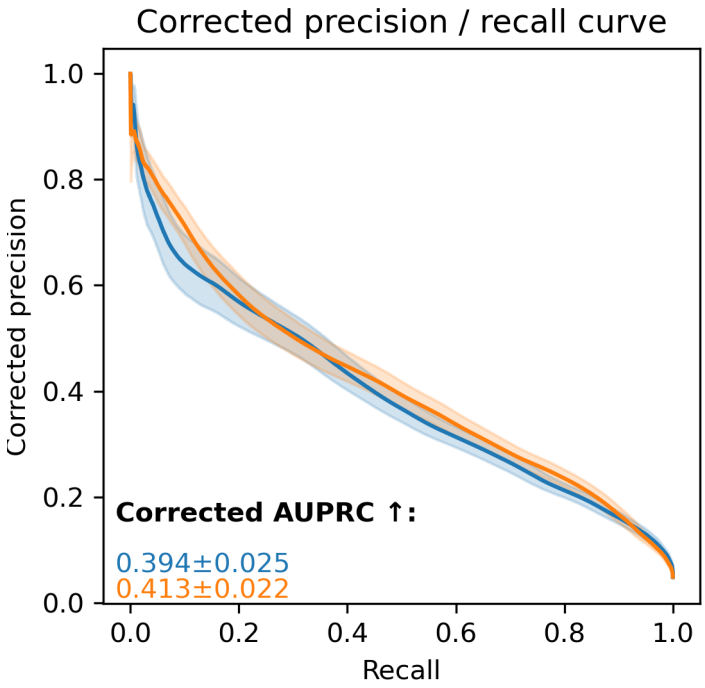
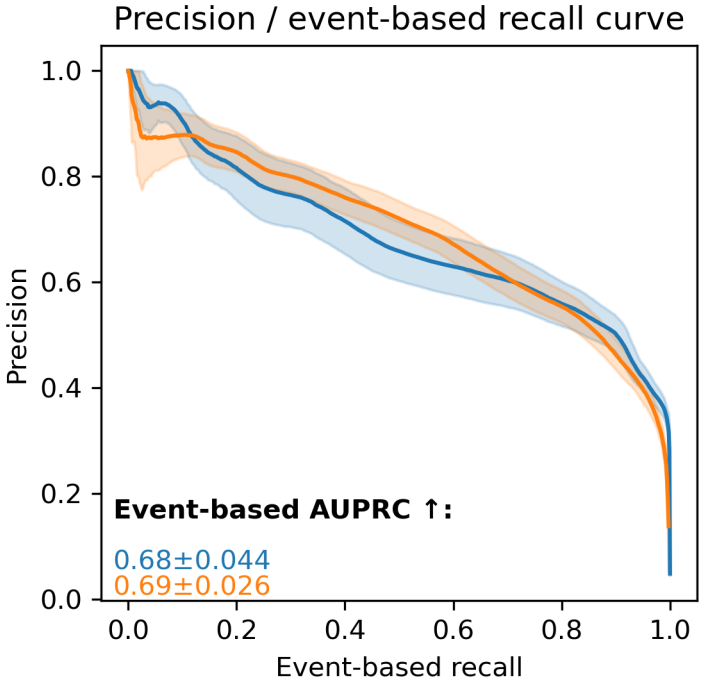
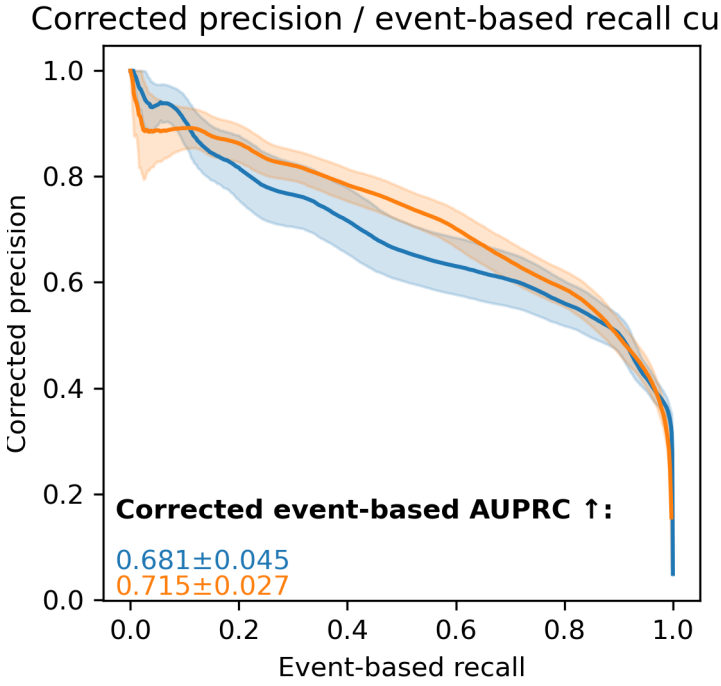


Figure 2.2.4.d


Precision / recall curve
Precision
Recall
AUPRC ↑:
0.393±0.025
0.385±0.017


Corrected precision / recall curve
Corrected precision
Recall
Corrected AUPRC ↑:
0.394±0.025
0.413±0.022


Precision / event-based recall curve
Precision
Event-based recall
Event-based AUPRC ↑:
0.68±0.044
0.69±0.026


Corrected precision / event-based recall curve
Corrected precision
Event-based recall
Corrected event-based AUPRC ↑:
0.681±0.045
0.715±0.027

3. Time Gap Analysis

Goal: Checking whether the time gap between the first correct alarm and the start of the corresponding event are similar across cohorts of patients

3.1. Aggregated views

3.1.1. Summary statistics of median time gap per grouping

For event starting in the window 0-3h, the overall macro-averaged median time gap is 48.4 (in minutes).
For event starting in the window 3-6h, the overall macro-averaged median time gap is 218.2 (in minutes).
For event starting in the window 6-12h, the overall macro-averaged median time gap is 394.4 (in minutes).
For event starting in the window >12h, the overall macro-averaged median time gap is 66.6 (in minutes).

Grouping by sex

Table 3.1.1.a

Start event	Macro-average (in minutes)	Minimum (category)	For minority category
0-3h	47.5	45.0 (M)	50.0
3-6h	217.5	215.0 (F)	215.0
6-12h	405.0	395.0 (M)	415.0
>12h	30.625	26.25 (F)	26.25

Grouping by age_group

Table 3.1.1.b

Start event	Macro-average (in minutes)	Minimum (category)	For minority category
0-3h	43.0	25.0 (>85)	25.0
3-6h	216.25	210.0 (50-65)	215.0
6-12h	403.0	375.0 (<50)	432.5
>12h	38.5	20.0 (75-85)	55.0

Grouping by APACHE_group

Table 3.1.1.c

Start event	Macro-average (in minutes)	Minimum (category)	For minority category
0-3h	55.357	35.0 (Neurological)	90.0
3-6h	220.893	190.0 (Respiratory)	255.0
6-12h	367.143	157.5 (Neurological)	370.0
>12h	170.357	20.0 (Cardiovascular)	140.0

Grouping by surgical_status

Table 3.1.1.d

Start event	Macro-average (in minutes)	Minimum (category)	For minority category
0-3h	47.5	45.0 (Surgical)	45.0
3-6h	217.5	215.0 (Surgical)	215.0
6-12h	405.0	400.0 (Surgical)	400.0
>12h	32.5	30.0 (Non-surgical)	35.0

3.1.2. Top 3 cohorts with the biggest time gap discrepancies

In the following table, we show for each start of the event window the 3 cohorts with the biggest delta that are significantly worse off than the rest of the patients. If some cells are empty, this means that there are fewer than 3 cohorts, possibly none, that are significantly worse than the rest of the patients for this particular start of the event window.

Table 3.1.2.a

Start event	Cohort 1 (Δ in minutes)	Cohort 2 (Δ in minutes)	Cohort 3 (Δ in minutes)
0-3h	>85 (25.0)	Neurological (15.0)	Cardiovascular (5.0)
3-6h	Respiratory (30.0)	Other (15.0)	50-65 (10.0)
6-12h	Neurological (247.5)	Trauma (40.0)	<50 (30.0)
>12h	Cardiovascular (70.0)	75-85 (17.5)	Other (12.5)

3.2. Grouping by

For each grouping, we display box plots that show the median time gap between alarm and event for the different categories of patients depending on the period of the stay when the event began. For each start of event window, we emphasize with a black star the cohorts that are significantly worse off compared to the rest of the patients and with a red star the cohorts that appear in the table **Top 3 cohorts with the biggest time gap discrepancies**.
For each grouping, we propose a table that presents the results of the statistical analysis: comparing the time gap from alarm to event for one cohort against the rest of the patients. P-values are obtained by running the Mann-Whitney U test with Bonferroni correction. We display only start of event windows and cohorts with a significant p-value (smaller than 0.001/number of comparisons) and whose delta is bigger than 0. For binary grouping, we display the category with the worst time gap distribution for each start of event window. While for multicategorical grouping we display whether the distribution for the category is better or worse than for the rest of patients

3.2.1. ... sex

Figure 3.2.1.a

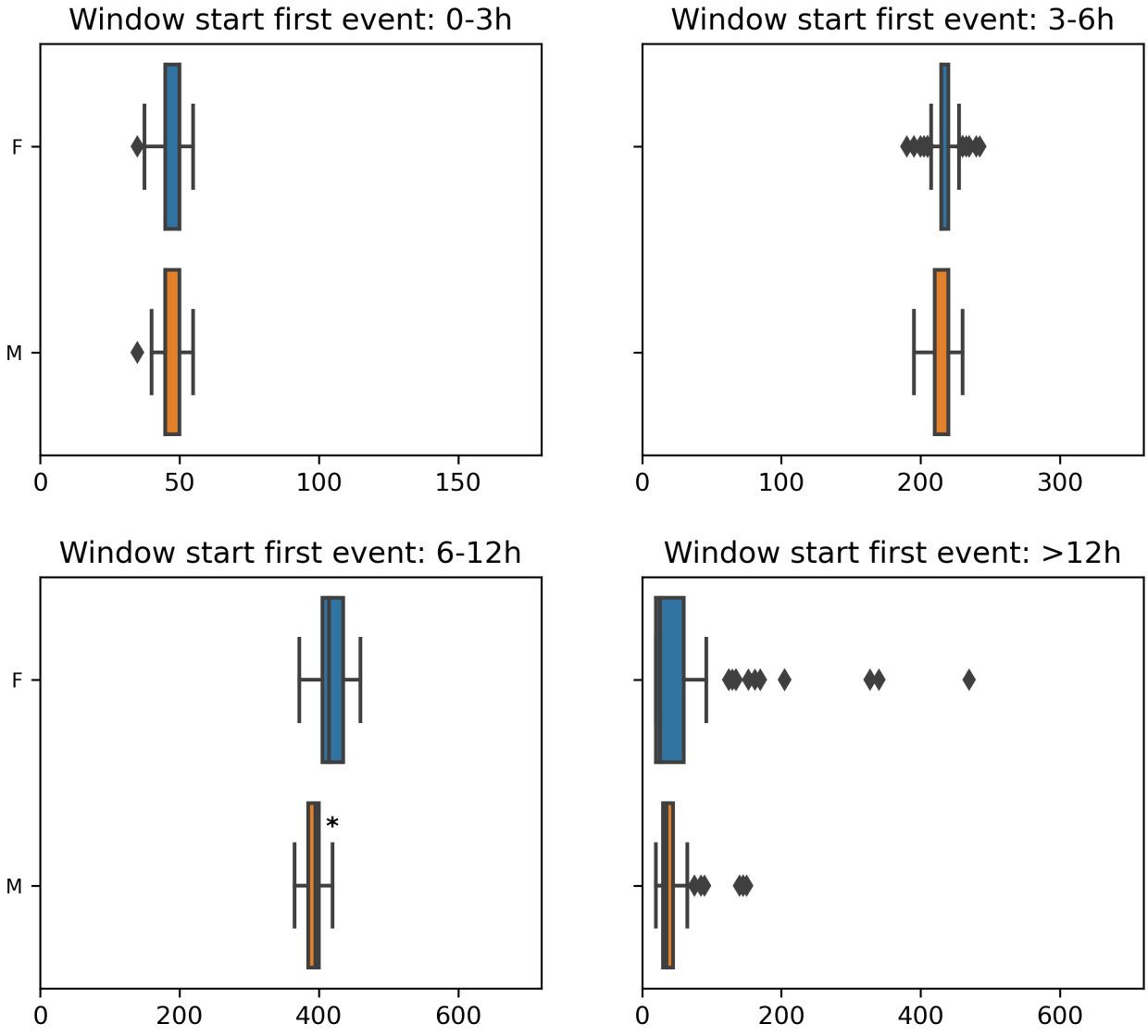


Table 3.2.1.a

Start event	Cohort with the worst time gap	P-value	Delta (in minutes)
6-12h	M	2.58e-20	20.0

3.2.2. ... age_group

Figure 3.2.2.a

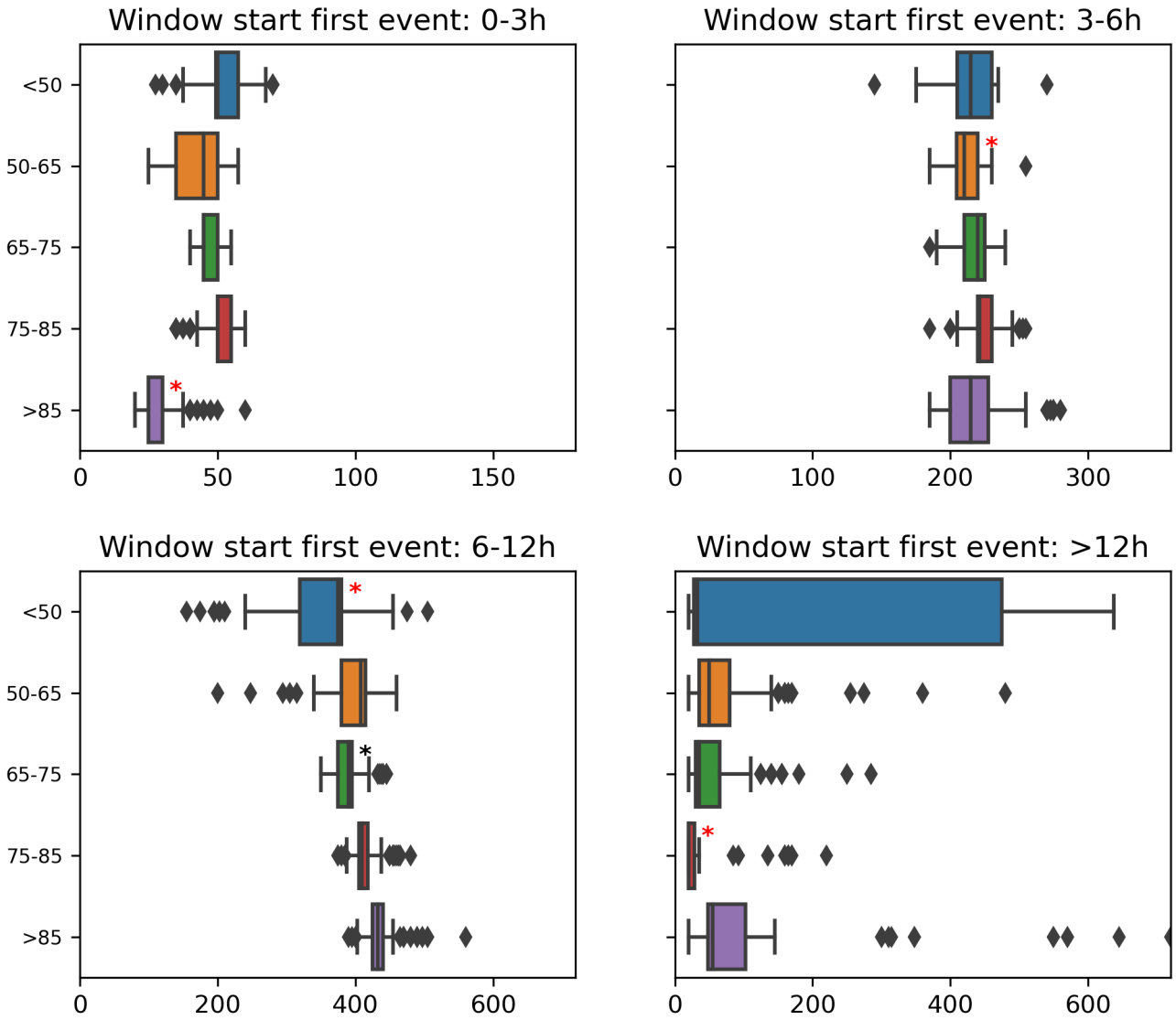


Table 3.2.2.a

Start event	Category	Cohort vs. rest	P-value	Delta (in minutes)
0-3h	<50	better	1.48e-12	5.0
0-3h	75-85	better	6.25e-11	5.0
0-3h	>85	worse	3.53e-32	25.0
3-6h	50-65	worse	6.9e-09	10.0
3-6h	75-85	better	3.18e-11	6.25
6-12h	<50	worse	2.1e-20	30.0
6-12h	65-75	worse	1.1e-16	20.0
6-12h	75-85	better	1.05e-09	15.0
6-12h	>85	better	2.75e-29	37.5
>12h	50-65	better	3.62e-13	20.0
>12h	75-85	worse	4.78e-18	17.5
>12h	>85	better	7.04e-16	25.0

3.2.3. ... APACHE_group

Figure 3.2.3.a

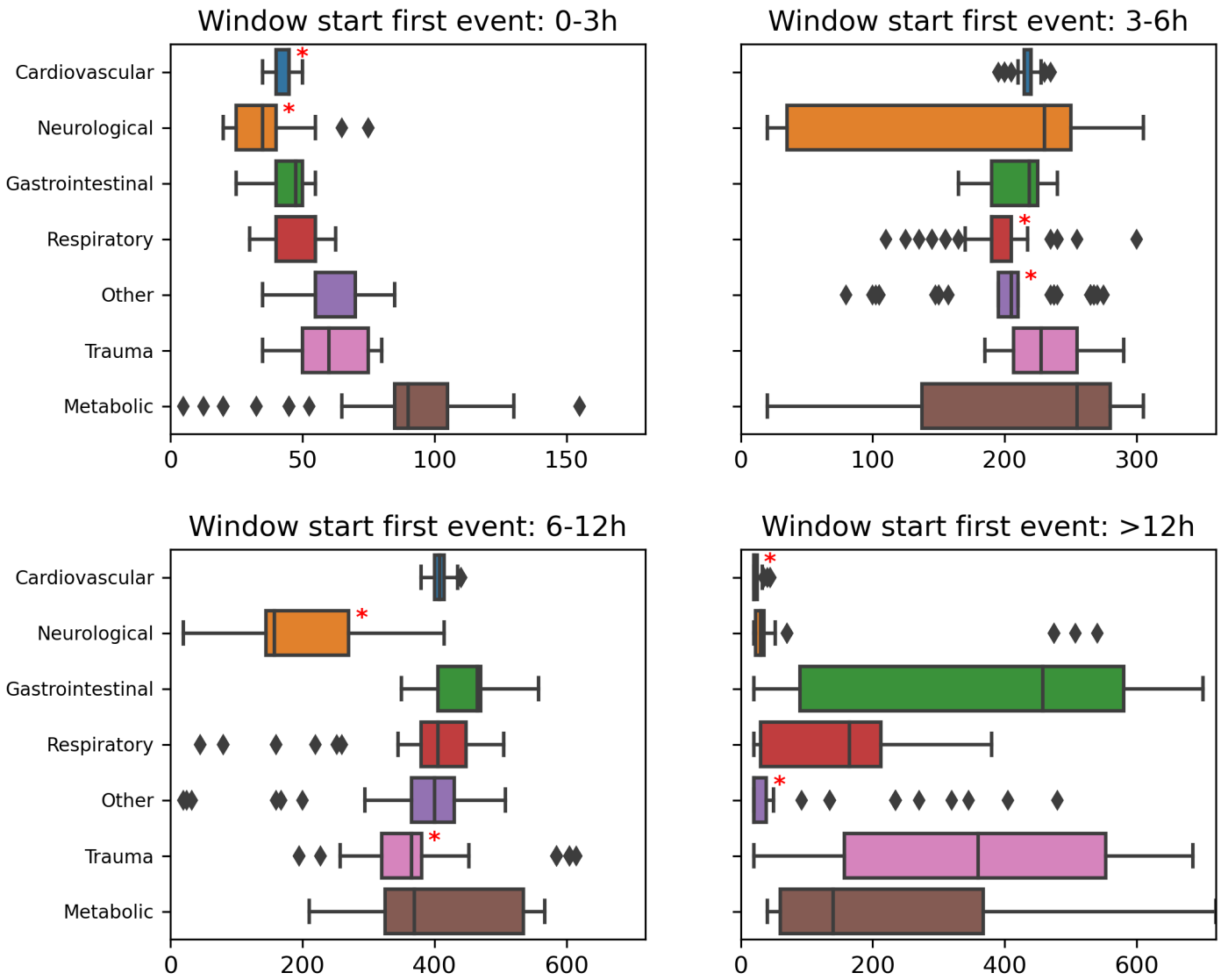


Table 3.2.3.a

Start event	Category	Cohort vs. rest	P-value	Delta (in minutes)
0-3h	Cardiovascular	worse	1.71e-14	5.0
0-3h	Neurological	worse	4.61e-13	15.0
0-3h	Respiratory	better	7.03e-06	10.0
0-3h	Other	better	3.51e-18	10.0
0-3h	Trauma	better	1.58e-14	15.0
0-3h	Metabolic	better	1.69e-21	45.0
3-6h	Cardiovascular	better	1.99e-06	10.0
3-6h	Respiratory	worse	5.34e-24	30.0
3-6h	Other	worse	3.59e-11	15.0
6-12h	Cardiovascular	better	1.19e-09	16.25
6-12h	Neurological	worse	1.28e-25	247.5
6-12h	Gastrointestinal	better	1.68e-12	65.0
6-12h	Trauma	worse	1.23e-18	40.0
>12h	Cardiovascular	worse	1.72e-33	70.0
>12h	Gastrointestinal	better	1.83e-26	427.5
>12h	Respiratory	better	1.4e-08	135.0

>12h	Other	worse	1.19e-08	12.5
>12h	Trauma	better	6.72e-31	330.0
>12h	Metabolic	better	9.84e-30	110.0

3.2.4. ... surgical_status

Figure 3.2.4.a

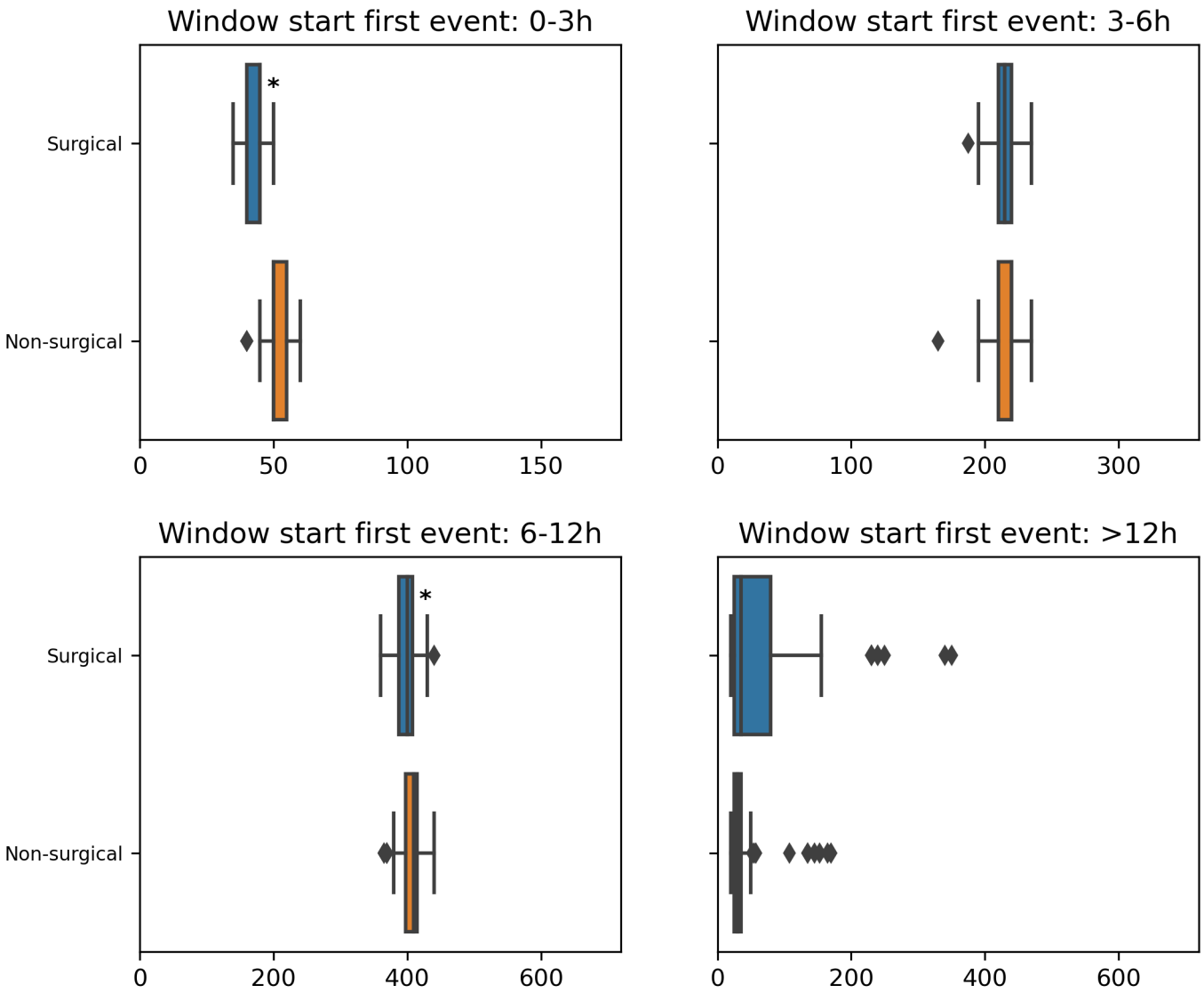


Table 3.2.4.a

Start event	Cohort with the worst time gap	P-value	Delta (in minutes)
0-3h	Surgical	5.81e-23	5.0
6-12h	Surgical	5.46e-05	10.0

4. Medical Variable Analysis

Goal: Comparing the median value of relevant medical variables across cohorts

We check the following variables: a_Lac, ABPm

4.1. Aggregated views

4.1.1. Top 3 cohorts with the biggest differences in the medical variables distributions

In the following table, for each of the selected medical variables and median computation condition, we show the 3 cohorts with the biggest delta that are significantly different than the rest of the patients. If some cells are empty, that means that there are less than 3 cohorts (possibly none) that are significantly different than the rest of the patients for this particular medical variable and median computation condition.

Table 4.1.1.a

Medical Variable	Cohort 1 (Δ)	Cohort 2 (Δ)	Cohort 3 (Δ)
a_Lac (mmol/l)	Gastrointestinal (0.25)	Cardiovascular (0.25)	Neurological (0.25)
a_Lac - Not in event (mmol/l)	Gastrointestinal (0.25)	Cardiovascular (0.25)	Neurological (0.25)
a_Lac - Never in event (mmol/l)	surgical_status (0.2)	Cardiovascular (0.15)	<50 (0.1)
ABPm (mmHg)	Neurological (14.0)	Cardiovascular (10.0)	<50 (5.0)
ABPm - Not in event (mmHg)	Neurological (14.0)	Cardiovascular (10.0)	<50 (5.0)
ABPm - Never in event (mmHg)	Neurological (12.0)	Cardiovascular (10.0)	<50 (4.0)

4.2. Grouping by

For each grouping, we display box plots that show the median value of the selected medical variables for three conditions: all time points during the entire stay, time points while not in an event, and time points from patients not experiencing any event. For each variable and condition, we emphasize with a black star the cohorts that are significantly different compared to the rest of the patients and with a red star the cohorts that appear in the table **Top 3 cohorts with the biggest differences in the medical variables values**.
For each grouping, we propose a table that presents the results of the statistical analysis: comparing the medical variables' median value for one cohort against the rest of the patients. P-values are obtained by running the Mann-Whitney U test with Bonferroni correction. We display only medical variables and cohorts with a significant p-value (smaller than 0.001/number of comparisons) and whose delta is bigger than 0. For binary grouping, we display the category with the greatest median value for each of the selected medical variables and median computation condition. While for multicategorical grouping we display whether the median value for the category is greater or less than for the rest of patients

4.2.1. ... sex

Figure 4.2.1.a

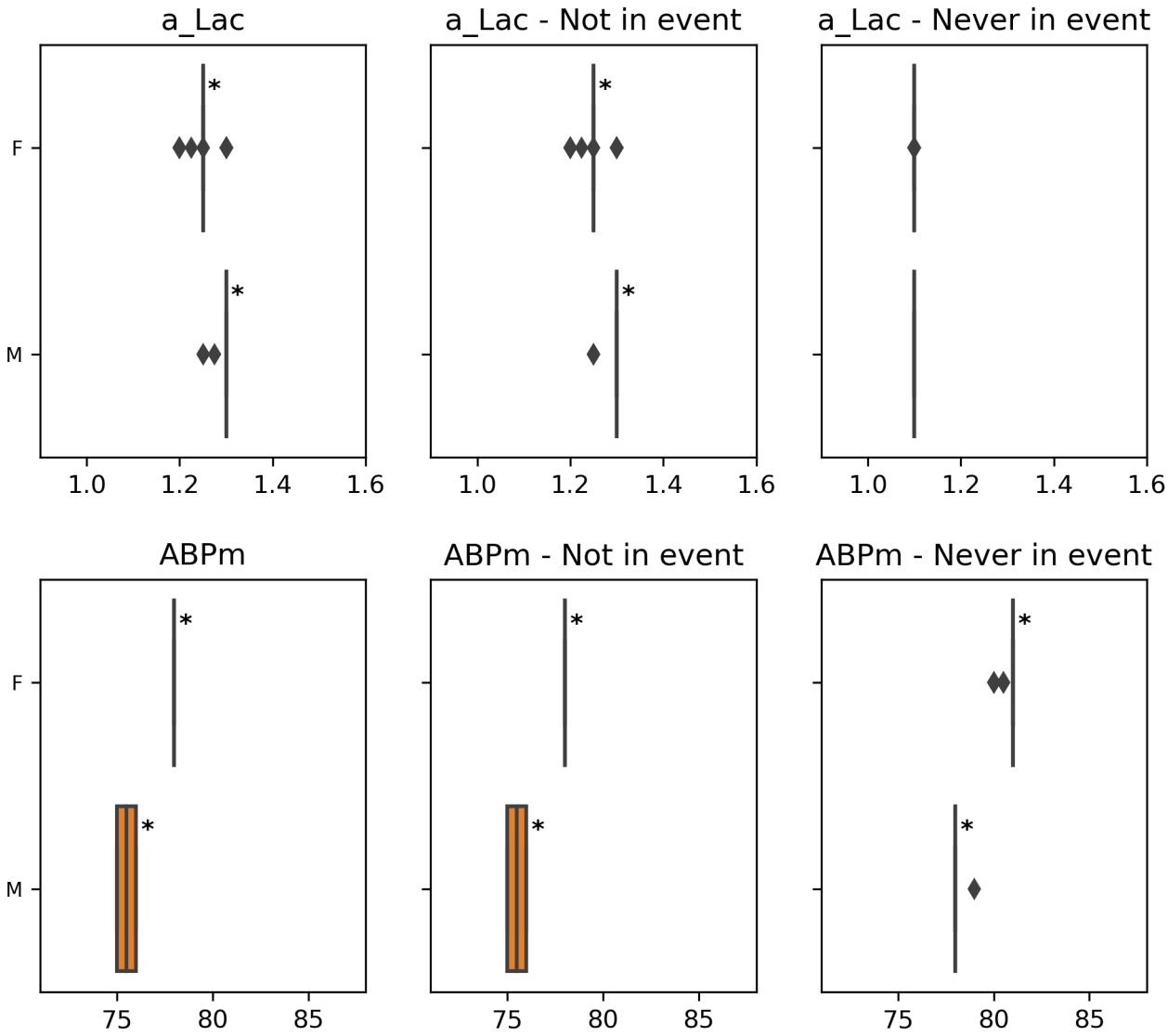


Table 4.2.1.a

Medical Variable	Cohort with greater median value	P-value	Delta
a_Lac	M	4.79e-30	0.05
a_Lac - Not in event	M	4.95e-30	0.05
ABPm	F	3.98e-40	2.5
ABPm - Not in event	F	3.77e-40	2.5

ABPm - Never in event	F	2.97e-43	3.0

4.2.2. ... age_group

Figure 4.2.2.a

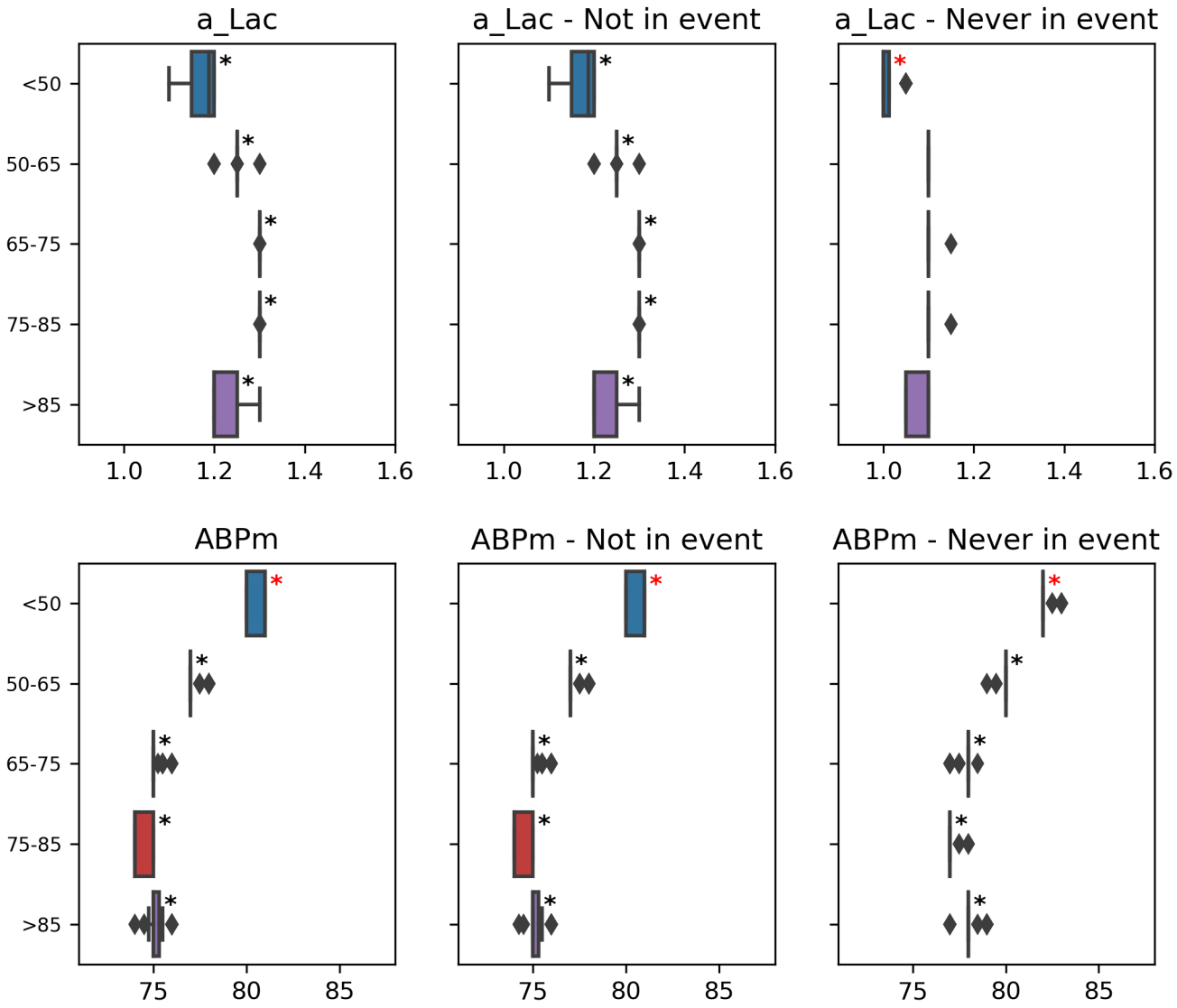


Table 4.2.2.a

Medical Variable	Category	Cohort vs. rest	P-value	Delta
a_Lac	<50	less	1.35e-39	0.112
a_Lac	50-65	less	6.63e-36	0.05
a_Lac	65-75	greater	8.44e-38	0.05
a_Lac	75-85	greater	4.86e-32	0.05
a_Lac	>85	less	8.42e-33	0.1
a_Lac - Not in event	<50	less	1.41e-39	0.112
a_Lac - Not in event	50-65	less	6.63e-36	0.05
a_Lac - Not in event	65-75	greater	4.76e-39	0.05
a_Lac - Not in event	75-85	greater	1.09e-32	0.05
a_Lac - Not in event	>85	less	5.44e-33	0.1
a_Lac - Never in event	<50	less	1.35e-41	0.1
ABPm	<50	greater	1.11e-37	5.0
ABPm	50-65	greater	9.06e-44	1.0
ABPm	65-75	less	2.07e-43	2.0
ABPm	75-85	less	2.23e-40	3.0
ABPm	>85	less	8.96e-29	1.0

ABPm - Not in event	<50	greater	1.11e-37	5.0
ABPm - Not in event	50-65	greater	9.06e-44	1.0
ABPm - Not in event	65-75	less	2.97e-43	2.0
ABPm - Not in event	75-85	less	2.23e-40	3.0
ABPm - Not in event	>85	less	6.84e-29	1.0
ABPm - Never in event	<50	greater	5.93e-41	4.0
ABPm - Never in event	50-65	greater	9.12e-43	1.0
ABPm - Never in event	65-75	less	2.21e-40	1.0
ABPm - Never in event	75-85	less	3.27e-40	3.0
ABPm - Never in event	>85	less	1.14e-29	1.0

4.2.3. ... APACHE_group

Figure 4.2.3.a

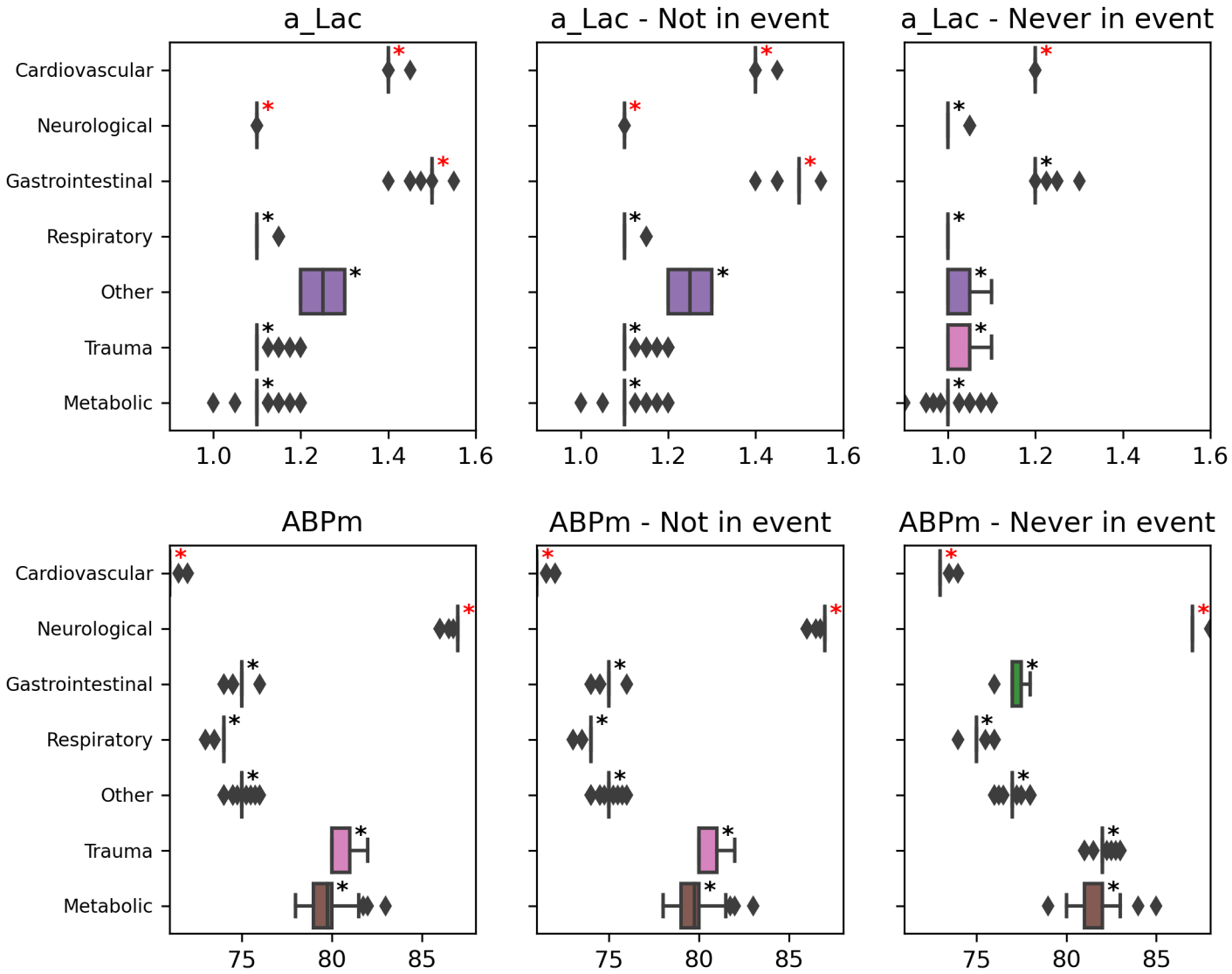


Table 4.2.3.a

Medical Variable	Category	Cohort vs. rest	P-value	Delta
a_Lac	Cardiovascular	greater	5.77e-40	0.25
a_Lac	Neurological	less	8.43e-41	0.25
a_Lac	Gastrointestinal	greater	2.23e-43	0.25
a_Lac	Respiratory	less	4.14e-44	0.2
a_Lac	Other	less	4.51e-21	0.05
a_Lac	Trauma	less	1.37e-41	0.2
a_Lac	Metabolic	less	4.28e-42	0.2
a_Lac - Not in event	Cardiovascular	greater	5.77e-40	0.25
a_Lac - Not in event	Neurological	less	9.91e-41	0.25
a_Lac - Not in event	Gastrointestinal	greater	3.22e-43	0.25
a_Lac - Not in event	Respiratory	less	4.14e-44	0.2
a_Lac - Not in event	Other	less	4.51e-21	0.05
a_Lac - Not in event	Trauma	less	1.37e-41	0.2
a_Lac - Not in event	Metabolic	less	4.28e-42	0.2
a_Lac - Never in event	Cardiovascular	greater	1.90e-42	0.15
a_Lac - Never in event	Neurological	less	3.20e-41	0.1

a_Lac - Never in event	Gastrointestinal	greater	1.32e-43	0.1
a_Lac - Never in event	Respiratory	less	1.76e-45	0.1
a_Lac - Never in event	Other	less	8.31e-37	0.1
a_Lac - Never in event	Trauma	less	2.33e-35	0.05
a_Lac - Never in event	Metabolic	less	7.89e-39	0.1
ABPm	Cardiovascular	less	6.14e-44	10.0
ABPm	Neurological	greater	1.76e-42	14.0
ABPm	Gastrointestinal	less	2.10e-37	2.0
ABPm	Respiratory	less	2.95e-41	3.0
ABPm	Other	less	5.26e-36	1.0
ABPm	Trauma	greater	4.33e-41	4.0
ABPm	Metabolic	greater	6.85e-39	3.75
ABPm - Not in event	Cardiovascular	less	9.05e-44	10.0
ABPm - Not in event	Neurological	greater	1.76e-42	14.0
ABPm - Not in event	Gastrointestinal	less	2.10e-37	2.0
ABPm - Not in event	Respiratory	less	2.95e-41	3.0
ABPm - Not in event	Other	less	5.26e-36	1.0
ABPm - Not in event	Trauma	greater	4.33e-41	4.0
ABPm - Not in event	Metabolic	greater	6.85e-39	3.75
ABPm - Never in event	Cardiovascular	less	4.65e-45	10.0
ABPm - Never in event	Neurological	greater	7.39e-45	12.0
ABPm - Never in event	Gastrointestinal	less	2.35e-41	2.0
ABPm - Never in event	Respiratory	less	2.8e-38	4.0
ABPm - Never in event	Other	less	6.96e-42	2.0
ABPm - Never in event	Trauma	greater	2.6e-41	3.0
ABPm - Never in event	Metabolic	greater	3.47e-39	3.0

4.2.4. ... surgical_status

Figure 4.2.4.a

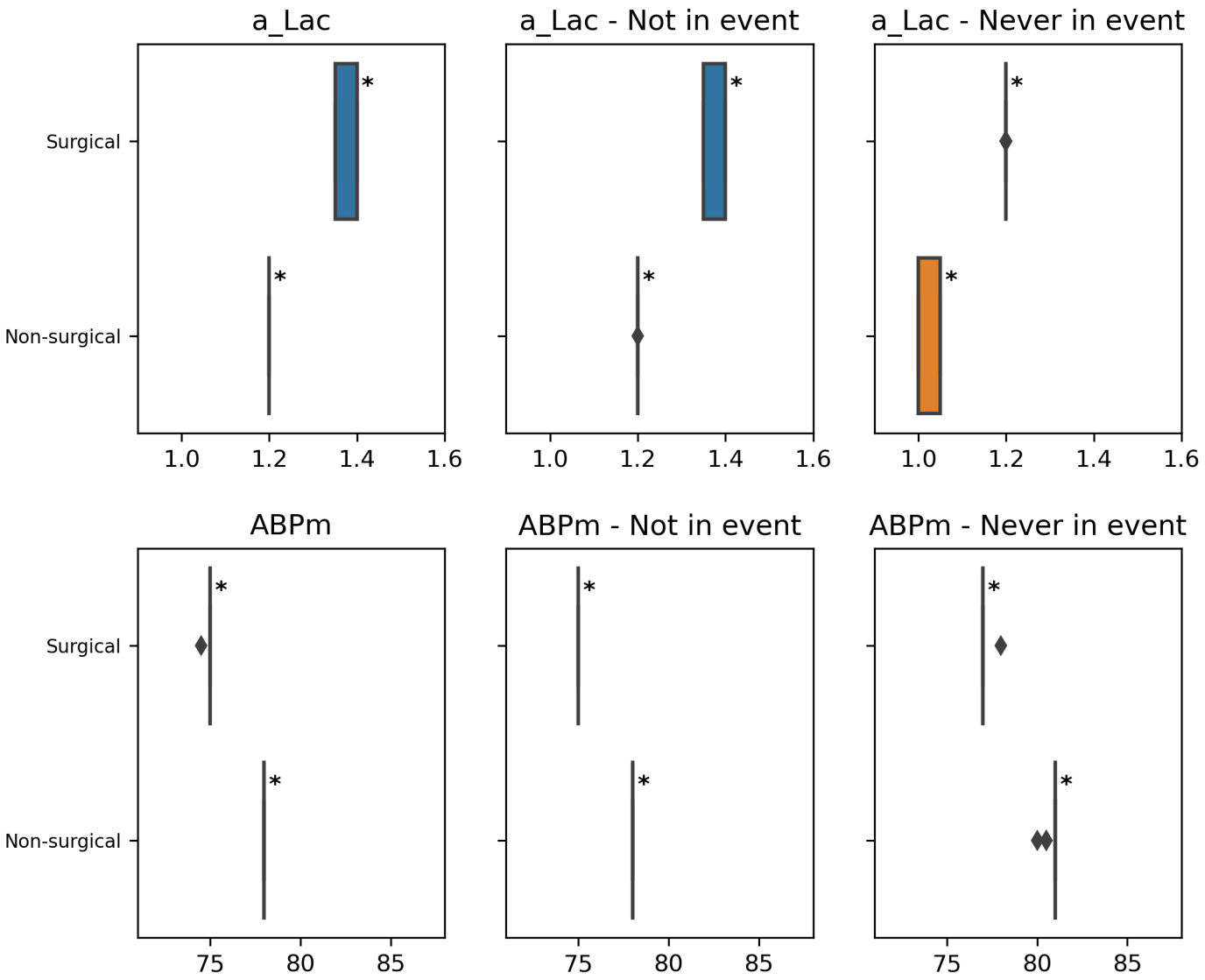


Table 4.2.4.a

Medical Variable	Cohort with greater median value	P-value	Delta
a_Lac	Surgical	4.63e-41	0.2
a_Lac - Not in event	Surgical	1.34e-40	0.2
a_Lac - Never in event	Surgical	5.45e-38	0.2
ABPm	Non-surgical	2.88e-45	3.0
ABPm - Not in event	Non-surgical	1.76e-45	3.0
ABPm - Never in event	Non-surgical	9.73e-44	4.0

5. Feature importance Analysis

Goal: Comparing the top 15 most important features across cohorts

5.1. Aggregated views

5.1.1 Similarity of feature ranking per groupig

The following table displays the RBO (similarity measure) between the feature ranking for a patients' cohort and the general feature ranking. We consider the feature ranking for a specific cohort to be significantly different when its RBO is smaller than 0.627 (colored in red in the table).

Table 5.1.1.a

Grouping	Category	RBO
sex	F	0.635
sex	M	0.627
age_group	<50	0.574
age_group	50-65	0.618
age_group	65-75	0.617
age_group	75-85	0.633
age_group	>85	0.627
APACHE_group	Cardiovascular	0.611
APACHE_group	Neurological	0.618
APACHE_group	Gastrointestinal	0.622
APACHE_group	Respiratory	0.615
APACHE_group	Other	0.619
APACHE_group	Trauma	0.607
APACHE_group	Metabolic	0.597
surgical_status	Surgical	0.627
surgical_status	Non-surgical	0.635

5.2. Grouping by

We will now display for each grouping, the top 15 most important features. When the feature's rank changes compared to the general ranking, we put the rank difference in parentheses.
We color in red the features that aren't in the general top 15 features and in blue the ones that change place within the top 15, when their delta of inverse rank is significantly large.

5.2.1. ... sex

Table 5.2.1.a

Top 15	Top 15 F	Top 15 M
a_Lac	a_Lac	a_Lac
datetime	datetime	datetime
ABPm	ABPm	ABPm
ABPs	ABPs	ABPs
HR	HR	HR
ABPd	ABPd	Spitzendruck (↑ 1)
Spitzendruck	Spitzendruck	ABPd (↓ 1)
RASS	RASS	RASS
age	age	age
a-BE	a-BE	a-BE
creatinine	creatinine	creatinine
norepinephrine	norepinephrine	norepinephrine
ETCO2	ETCO2	ETCO2
glucose	glucose	glucose
INR	INR	INR

5.2.2. ... age_group

Table 5.2.2.a

Top 15	Top 15 <50	Top 15 50-65	Top 15 65-75
a_Lac	a_Lac	a_Lac	a_Lac
datetime	datetime	datetime	datetime
ABPm	age (↑ 6)	ABPm	ABPm
ABPs	ABPm (↓ 1)	ABPs	ABPs
HR	HR	HR	HR
ABPd	ABPs (↓ 2)	ABPd	Spitzendruck (↑ 1)
Spitzendruck	ABPd (↓ 1)	Spitzendruck	ABPd (↓ 1)
RASS	Spitzendruck (↓ 1)	RASS	RASS
age	RASS (↓ 1)	a-BE (↑ 1)	a-BE (↑ 1)
a-BE	a-BE	creatinine (↑ 1)	creatinine (↑ 1)
creatinine	creatinine	norepinephrine (↑ 1)	norepinephrine (↑ 1)
norepinephrine	norepinephrine	ETCO2 (↑ 1)	age (↓ 3)
ETCO2	ETCO2	glucose (↑ 1)	ETCO2
glucose	glucose	INR (↑ 1)	glucose
INR	INR	age (↓ 6)	INR

Top 15 75-85	Top 15 >85
a_Lac	a_Lac
datetime	datetime

ABPm	ABPm
ABPs	ABPs
HR	HR
ABPd	ABPd
Spitzendruck	Spitzendruck
RASS	age (↑ 1)
age	RASS (↓ 1)
a-BE	a-BE
creatinine	creatinine
ETCO2 (↑ 1)	ETCO2 (↑ 1)
norepinephrine (↓ 1)	norepinephrine (↓ 1)
glucose	glucose
INR	INR

5.2.3. ... APACHE_group

Table 5.2.3.a

Top 15	Top 15 Cardiovascular	Top 15 Neurological	Top 15 Gastrointestinal
a_Lac	a_Lac	a_Lac	a_Lac
datetime	datetime	datetime	datetime
ABPm	ABPm	ABPm	ABPm
ABPs	ABPs	ABPs	ABPs
HR	Spitzendruck (↑ 2)	HR	HR
ABPd	HR (↓ 1)	ABPd	ABPd
Spitzendruck	ABPd (↓ 1)	RASS (↑ 1)	Spitzendruck
RASS	RASS	Spitzendruck (↓ 1)	a-BE (↑ 2)
age	age	age	age
a-BE	a-BE	norepinephrine (↑ 2)	RASS (↓ 2)
creatinine	creatinine	creatinine	creatinine
norepinephrine	ETCO2 (↑ 1)	a-BE (↓ 2)	norepinephrine
ETCO2	INR (↑ 2)	glucose (↑ 1)	INR (↑ 2)
glucose	norepinephrine (↓ 2)	ETCO2 (↓ 1)	ETCO2 (↓ 1)
INR	glucose (↓ 1)	NIBPm (↑ 1)	glucose (↓ 1)
Top 15 Respiratory	**Top 15 Other**	**Top 15 Trauma**	**Top 15 Metabolic**
a_Lac	a_Lac	a_Lac	a_Lac
datetime	datetime	datetime	datetime
ABPm	ABPm	ABPm	ABPm
HR (↑ 1)	ABPs	HR (↑ 1)	ABPs
ABPs (↓ 1)	HR	ABPs (↓ 1)	HR
ABPd	ABPd	Spitzendruck (↑ 1)	ABPd
Spitzendruck	Spitzendruck	ABPd (↓ 1)	a-BE (↑ 3)
RASS	a-BE (↑ 2)	age (↑ 1)	age (↑ 1)
a-BE (↑ 1)	age	RASS (↓ 1)	Spitzendruck (↓ 2)
age (↓ 1)	RASS (↓ 2)	a-BE	glucose (↑ 4)

creatinine	norepinephrine (↑ 1)	creatinine	RASS (↓ 3)
ETCO2 (↑ 1)	creatinine (↓ 1)	norepinephrine	creatinine (↓ 1)
norepinephrine (↓ 1)	INR (↑ 2)	ETCO2	NIBPm (↑ 3)
glucose	ETCO2 (↓ 1)	glucose	norepinephrine (↓ 2)
INR	glucose (↓ 1)	GCS Motorik (↑ 2)	ETCO2 (↓ 2)

5.2.4. ... surgical_status

Table 5.2.4.a

Top 15	Top 15 Surgical	Top 15 Non-surgical
a_Lac	a_Lac	a_Lac
datetime	datetime	datetime
ABPm	ABPm	ABPm
ABPs	ABPs	ABPs
HR	HR	HR
ABPd	Spitzendruck (↑ 1)	ABPd
Spitzendruck	ABPd (↓ 1)	Spitzendruck
RASS	RASS	RASS
age	age	age
a-BE	a-BE	a-BE
creatinine	creatinine	creatinine
norepinephrine	norepinephrine	norepinephrine
ETCO2	ETCO2	ETCO2
glucose	glucose	glucose
INR	INR	INR

6. Missingness Analysis

Goal: Comparing the intensity of measurements across cohorts of patients and its impact of performance

Binary metrics computed with a threshold on score of 0.445.

6.1. Aggregated views

6.1.1. a_Lac

Groupings that are statistically dependent on the intensity of measurements:

Table 6.1.1.a

Group name	Category with the biggest rate of no_msrt	Category with the biggest rate of insufficient
sex	F	M
age_group	<50	75-85
APACHE_group	Neurological	Cardiovascular
surgical_status	Surgical	Surgical

Summary of the impact of missingness on performance.

missing_msrt: 36.4% of metrics are worse than for with measurement time points, with the biggest delta 0.115 for metric Recall.

6.1.2. Spitzendruck

Groupings that are statistically dependent on the intensity of measurements:

Table 6.1.2.a

Group name	Category with the biggest rate of no_msrt	Category with the biggest rate of insufficient
sex	F	M
age_group	<50	75-85
APACHE_group	Metabolic	Cardiovascular
surgical_status	Non-surgical	Surgical

Summary of the impact of missingness on performance.

no_msrt: 45.5% of metrics are worse than for with measurement time points, with the biggest delta 0.175 for metric AUPRC.
missing_msrt: 45.5% of metrics are worse than for with measurement time points, with the biggest delta 0.155 for metric AUPRC.

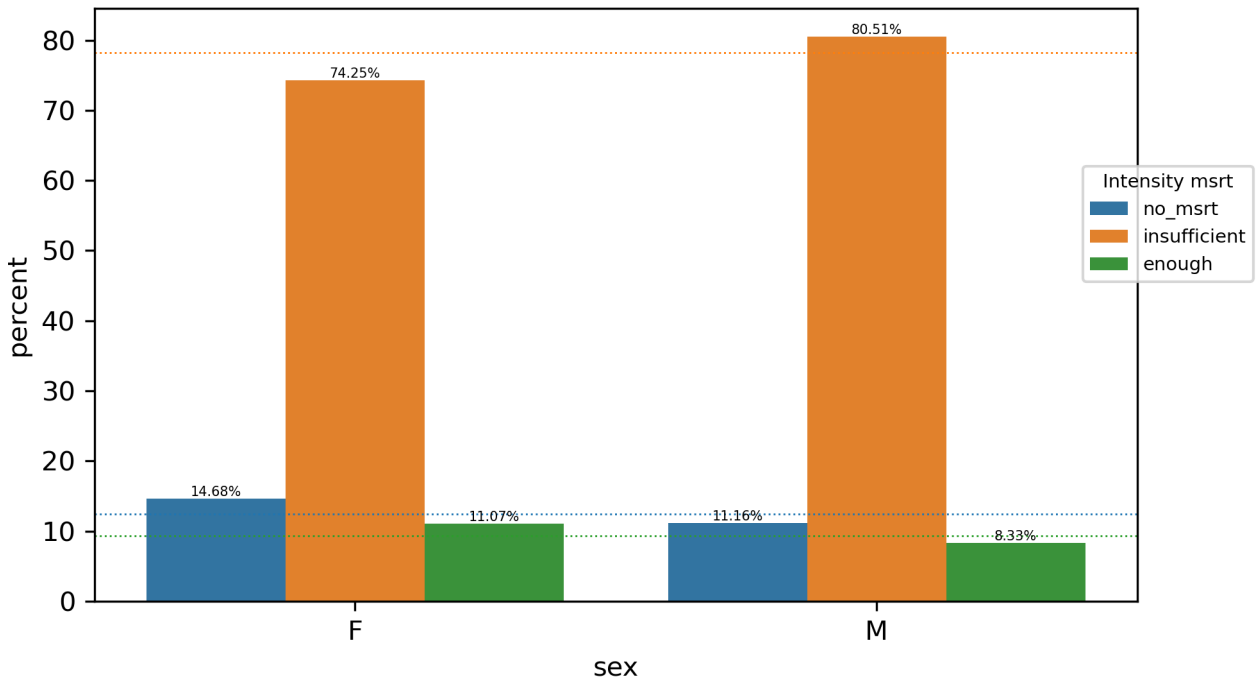
For each grouping, we display a bar plot that shows the percentage of each intensity of measurement category within a cohort of patients. The dashed lines represent the percentage of each intensity of measurement category with respect to the entire patient population. We run the Chi-squared independence test (with significance level 0.001) to assess the depence between the intensity of measurement and the grouping.
In the impact on performance subsection, we present box plots that show the metrics' distribution for each of the missingness categories. For each metric, we mark with a black star the missingness categories that are significantly worse compared to metrics computed on data points with present measurement.
We also propose tables presenting the results of the impact on performance statistical analysis, we display only metrics and missingness categories with a significant p-value (smaller than 0.001/number of comparisons) and whose delta is bigger than 0. We compare the metrics for missingness categories *missing_msrt* and *no_msrt* (when relevant) against the *with_msrt* category. P-values are obtained by running the Mann-Whitney U test with Bonferroni correction.

6.2. Study of the variable a_Lac

6.2.1. Intensity of measurement per grouping

Grouping by sex

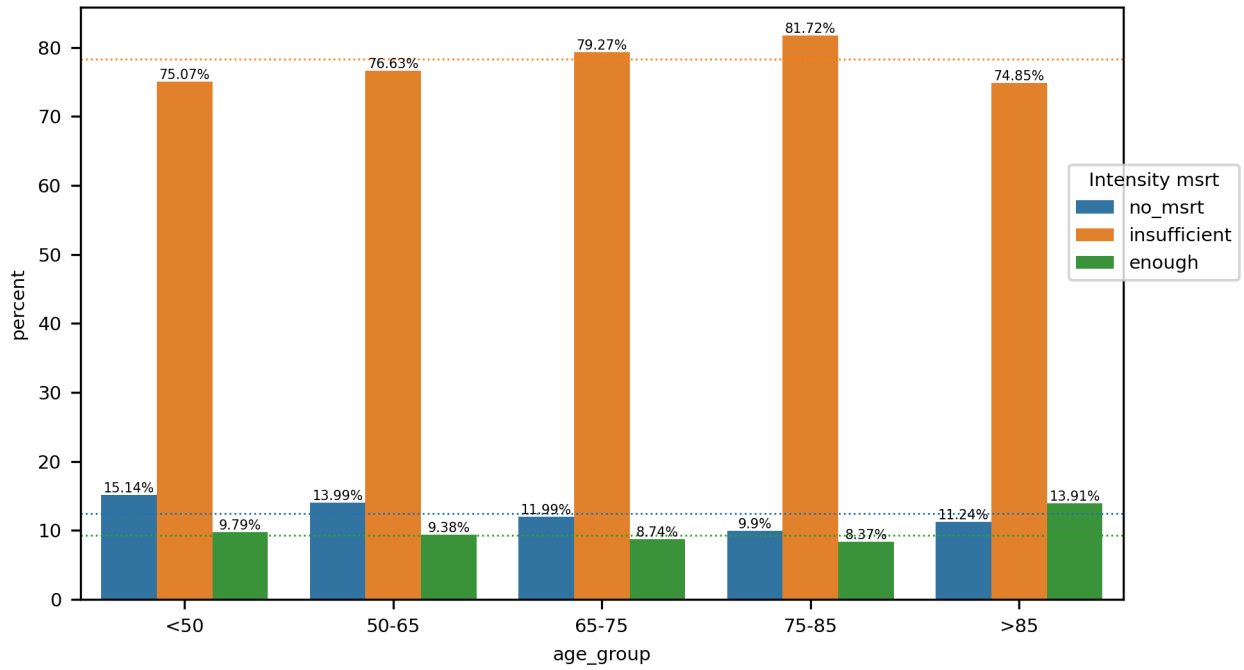
Figure 6.2.1.a



The intensity of measurements of a_Lac and sex attributes are dependent.

Grouping by age_group

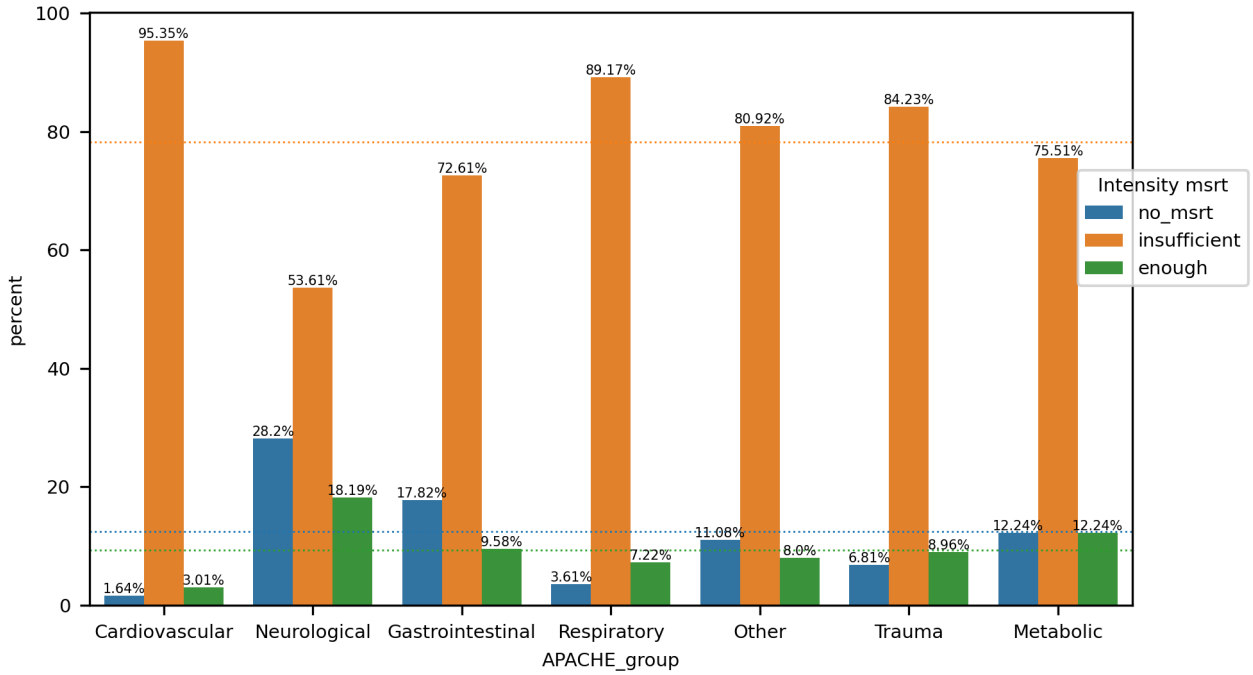
Figure 6.2.1.b



The intensity of measurements of a_Lac and age_group attributes are dependent.

Grouping by APACHE_group

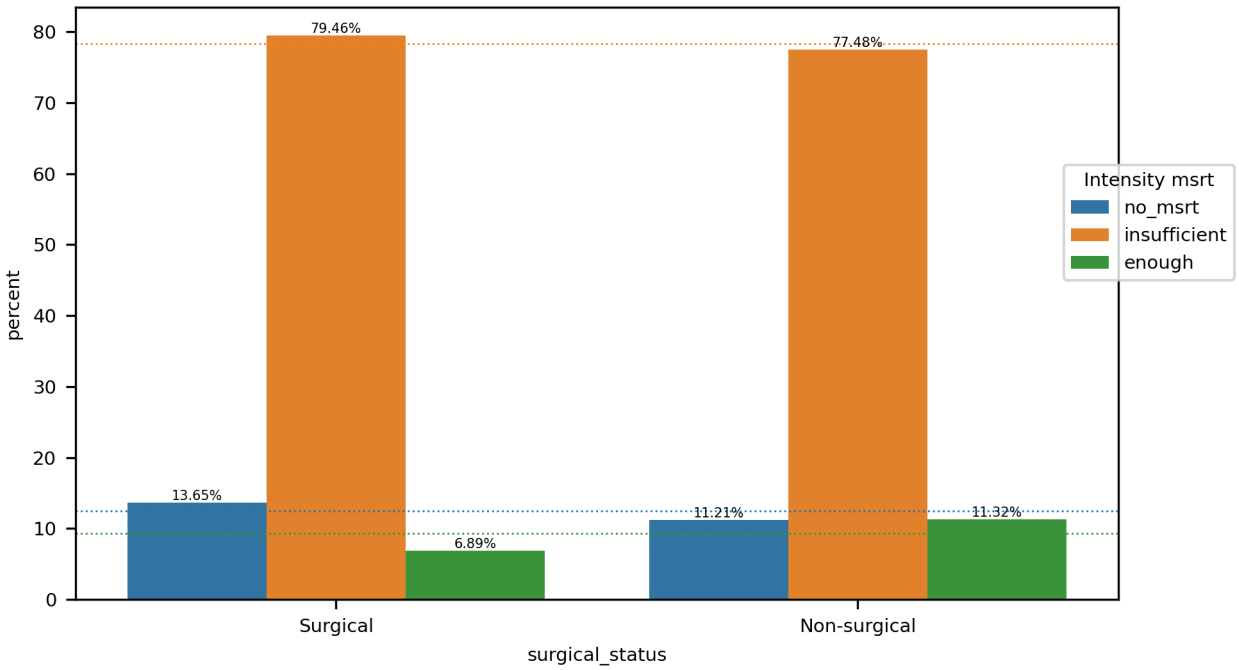
Figure 6.2.1.c



The intensity of measurements of a_Lac and APACHE_group attributes are dependent.

Grouping by surgical_status

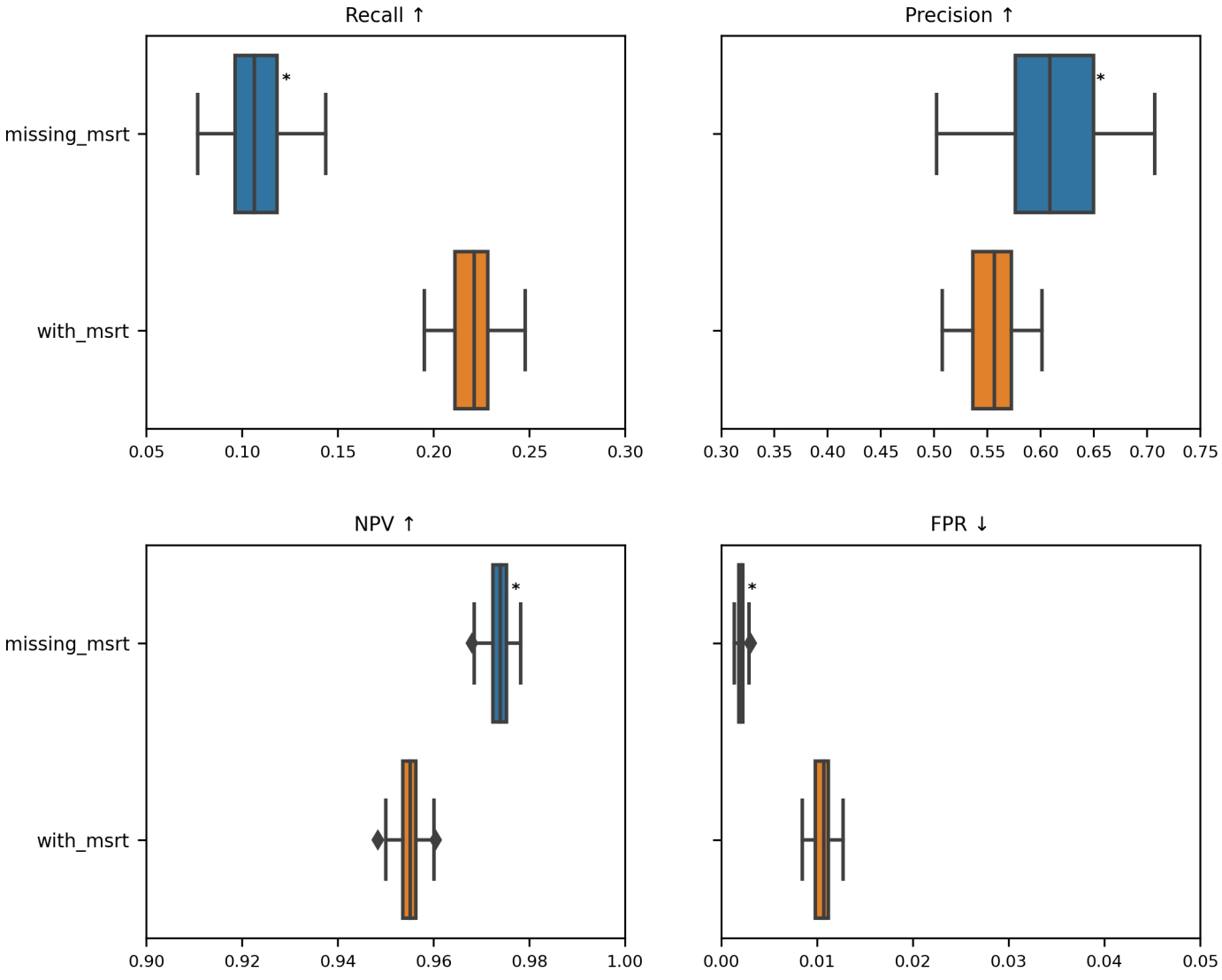
Figure 6.2.1.d

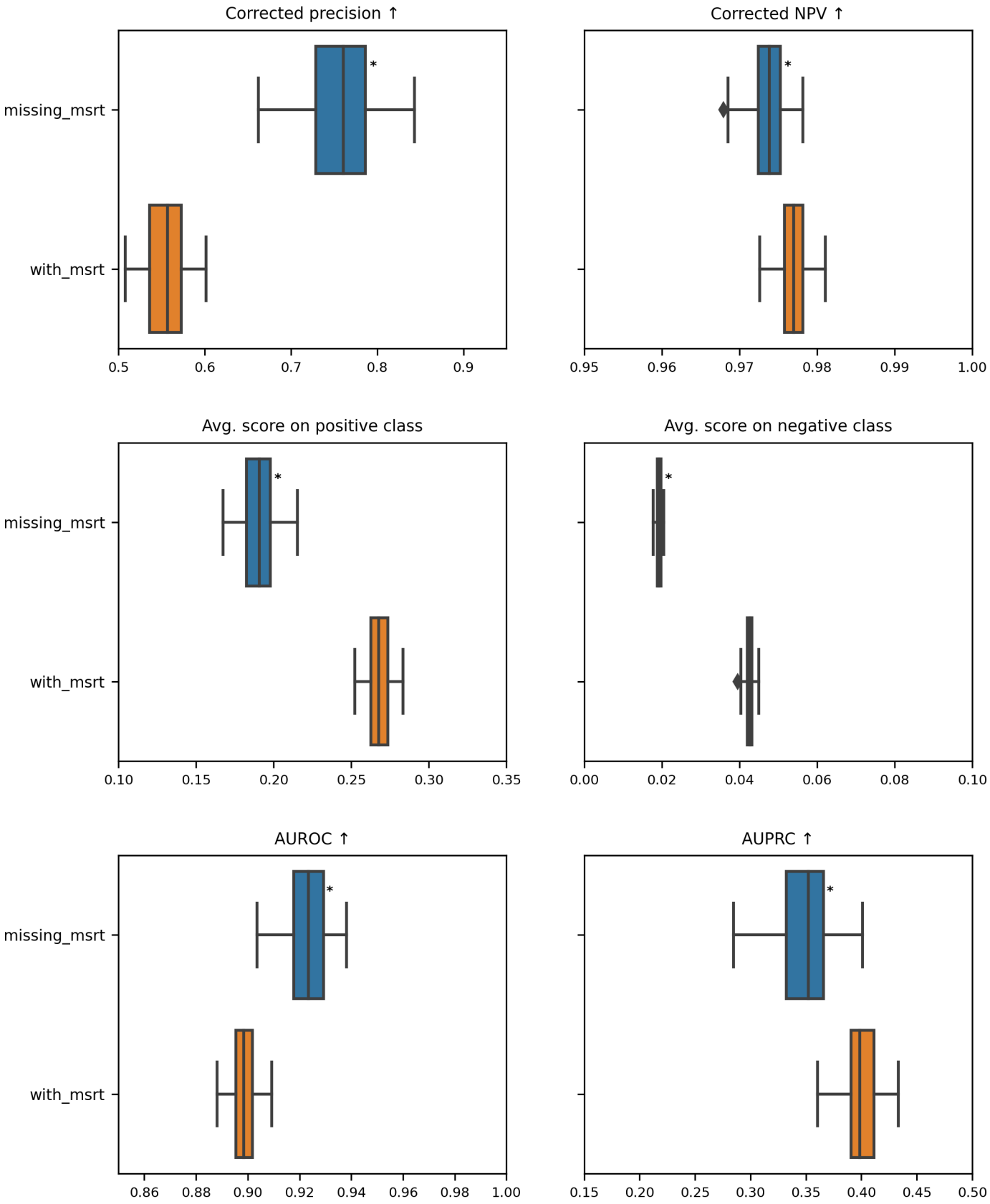


The intensity of measurements of a_Lac and surgical_status attributes are dependent.

6.2.2. Impact on performance

Figure 6.2.2.a




Corrected precision ↑
missing_msrt
with_msrt
*
0.5
0.6
0.7
0.8
0.9
Corrected NPV ↑
*
0.95
0.96
0.97
0.98
0.99
1.00
Avg. score on positive class
missing_msrt
with_msrt
*
0.10
0.15
0.20
0.25
0.30
0.35
Avg. score on negative class
*
0.00
0.02
0.04
0.06
0.08
0.10
AUROC ↑
missing_msrt
with_msrt
*
0.86
0.88
0.90
0.92
0.94
0.96
0.98
1.00
AUPRC ↑
*
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50

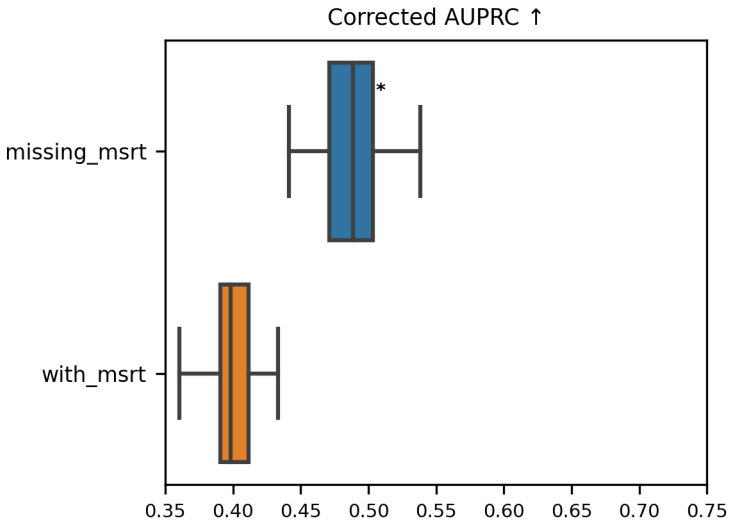


Table 6.2.2.a

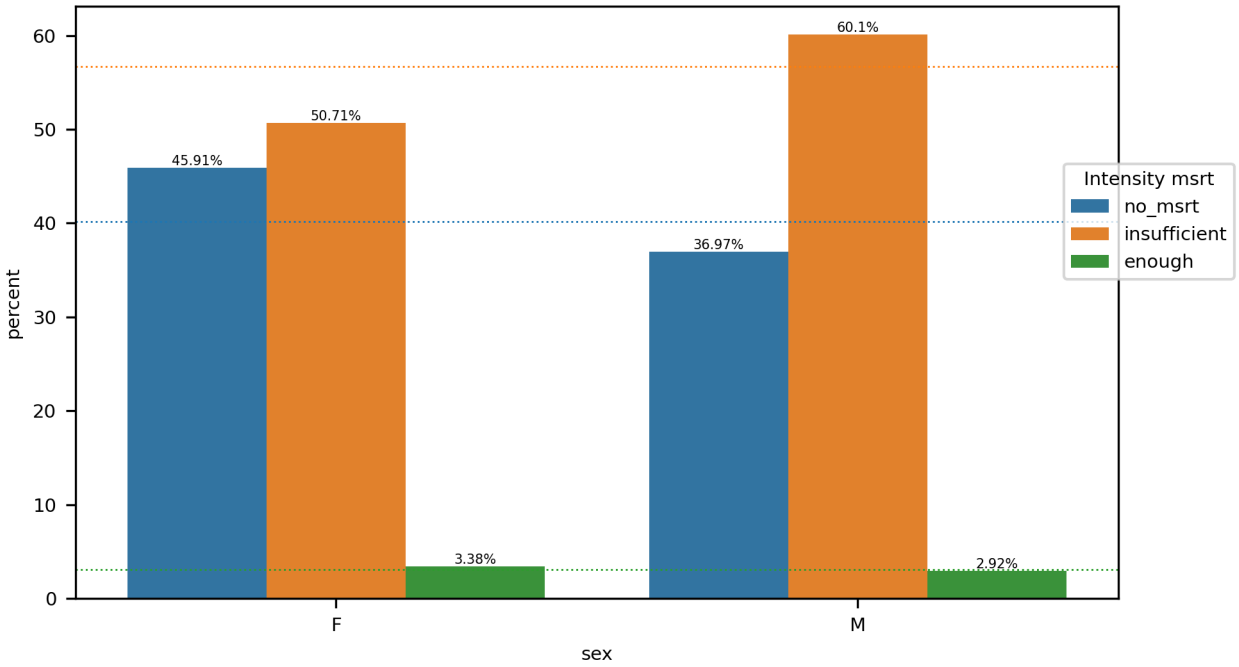
Metric	Missingness category	Category vs with msrt	P-value	Delta
Recall ↑	missing_msrt	worse	1.28e-34	0.115
Precision ↑	missing_msrt	better	8.76e-17	0.052
NPV ↑	missing_msrt	better	1.28e-34	0.019
FPR ↓	missing_msrt	better	1.28e-34	0.009
Corrected precision ↑	missing_msrt	better	1.28e-34	0.204
Corrected NPV ↑	missing_msrt	worse	2.29e-20	0.003
Avg. score on positive class	missing_msrt	worse	1.28e-34	0.077
Avg. score on negative class	missing_msrt	better	1.28e-34	0.023
AUROC ↑	missing_msrt	better	2.88e-34	0.025
AUPRC ↑	missing_msrt	worse	4.32e-31	0.046
Corrected AUPRC ↑	missing_msrt	better	1.28e-34	0.09

6.3. Study of the variable Spitzendruck

6.3.1. Intensity of measurement per grouping

Grouping by sex

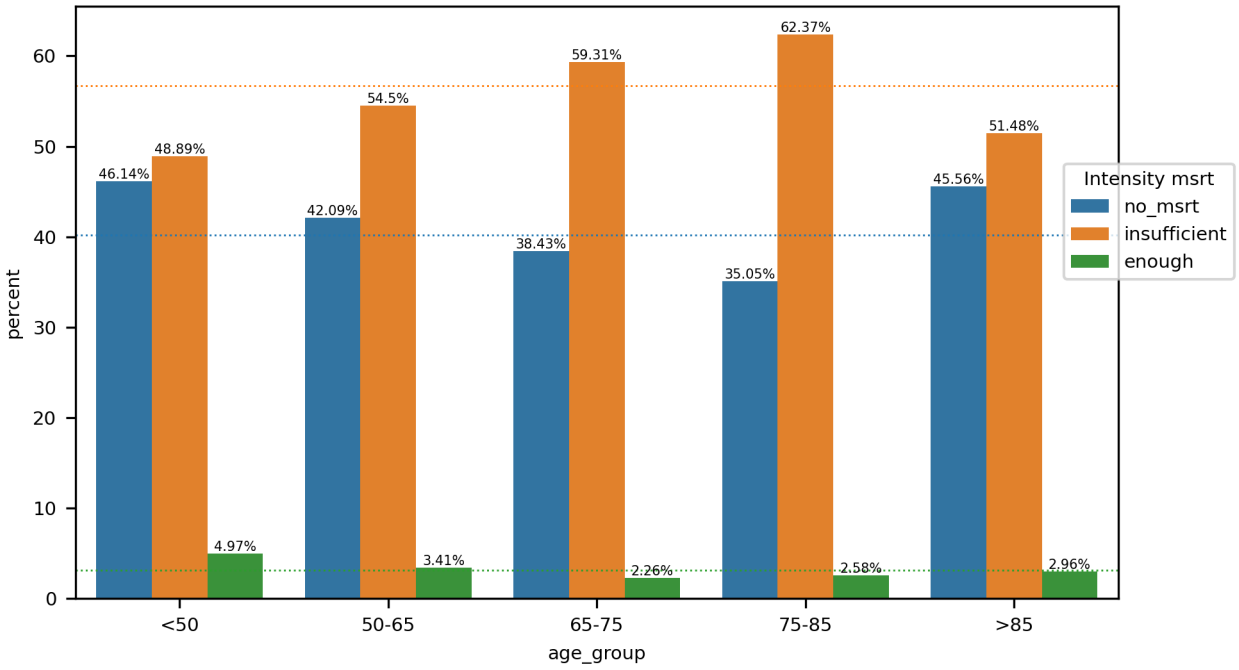
Figure 6.3.1.a



The intensity of measurements of Spitzendruck and sex attributes are dependent.

Grouping by age_group

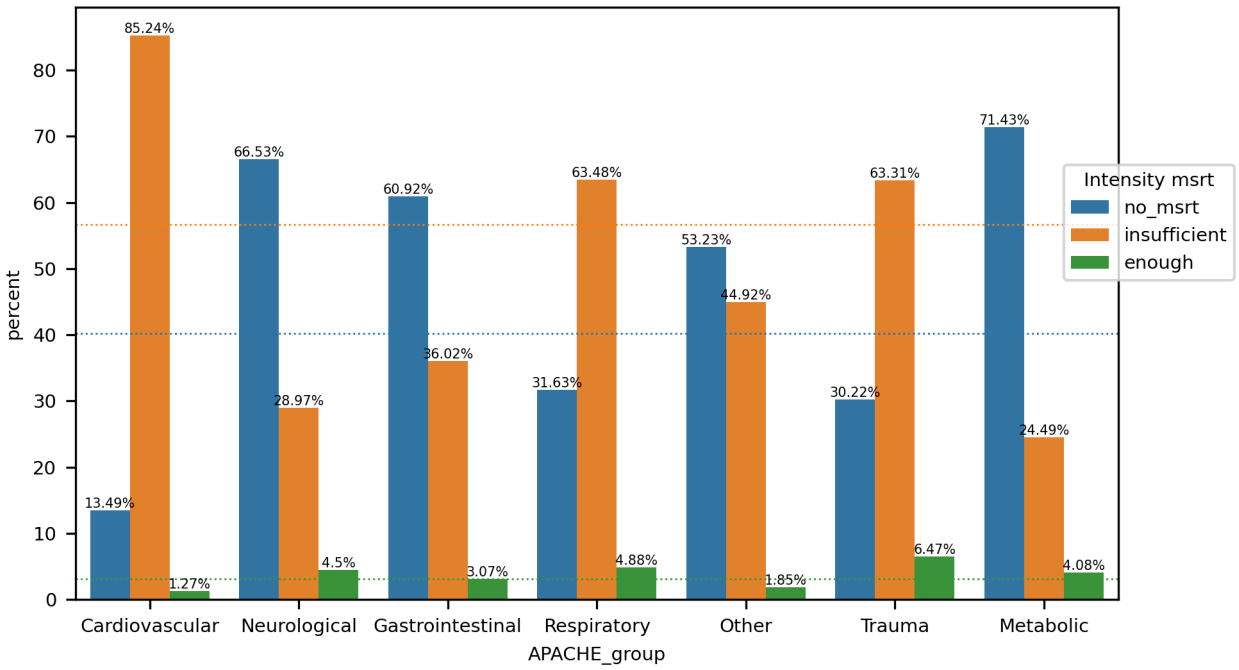
Figure 6.3.1.b



The intensity of measurements of Spitzendruck and age_group attributes are dependent.

Grouping by APACHE_group

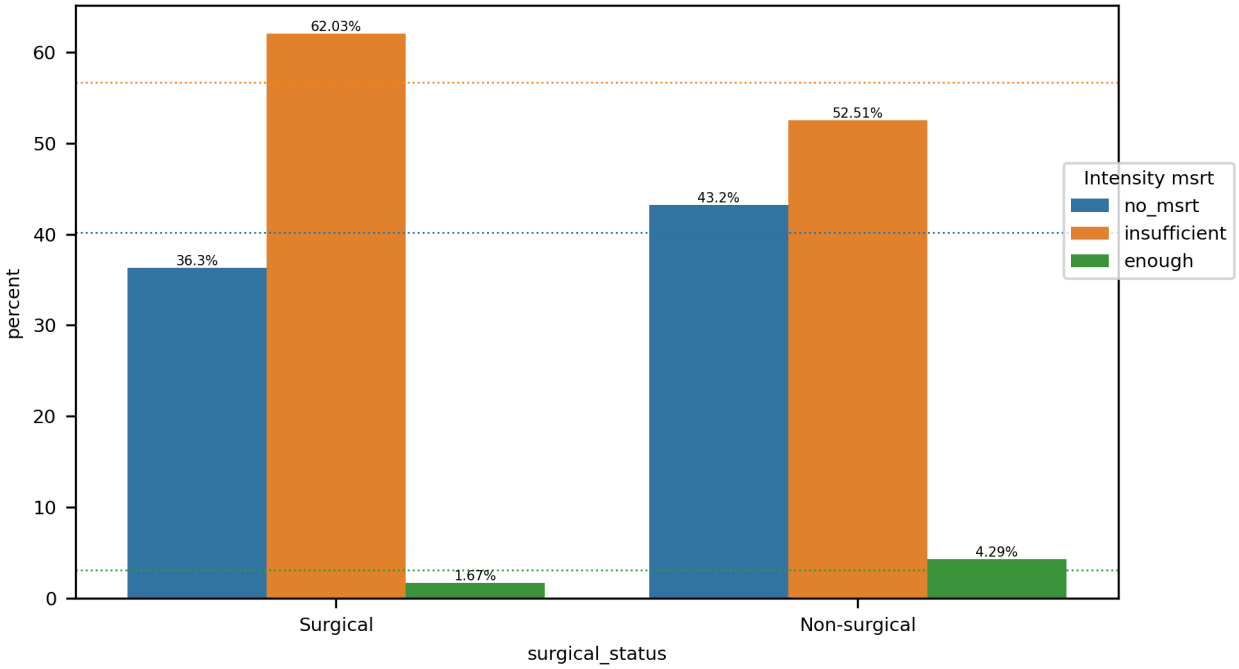
Figure 6.3.1.c



The intensity of measurements of Spitzendruck and APACHE_group attributes are dependent.

Grouping by surgical_status

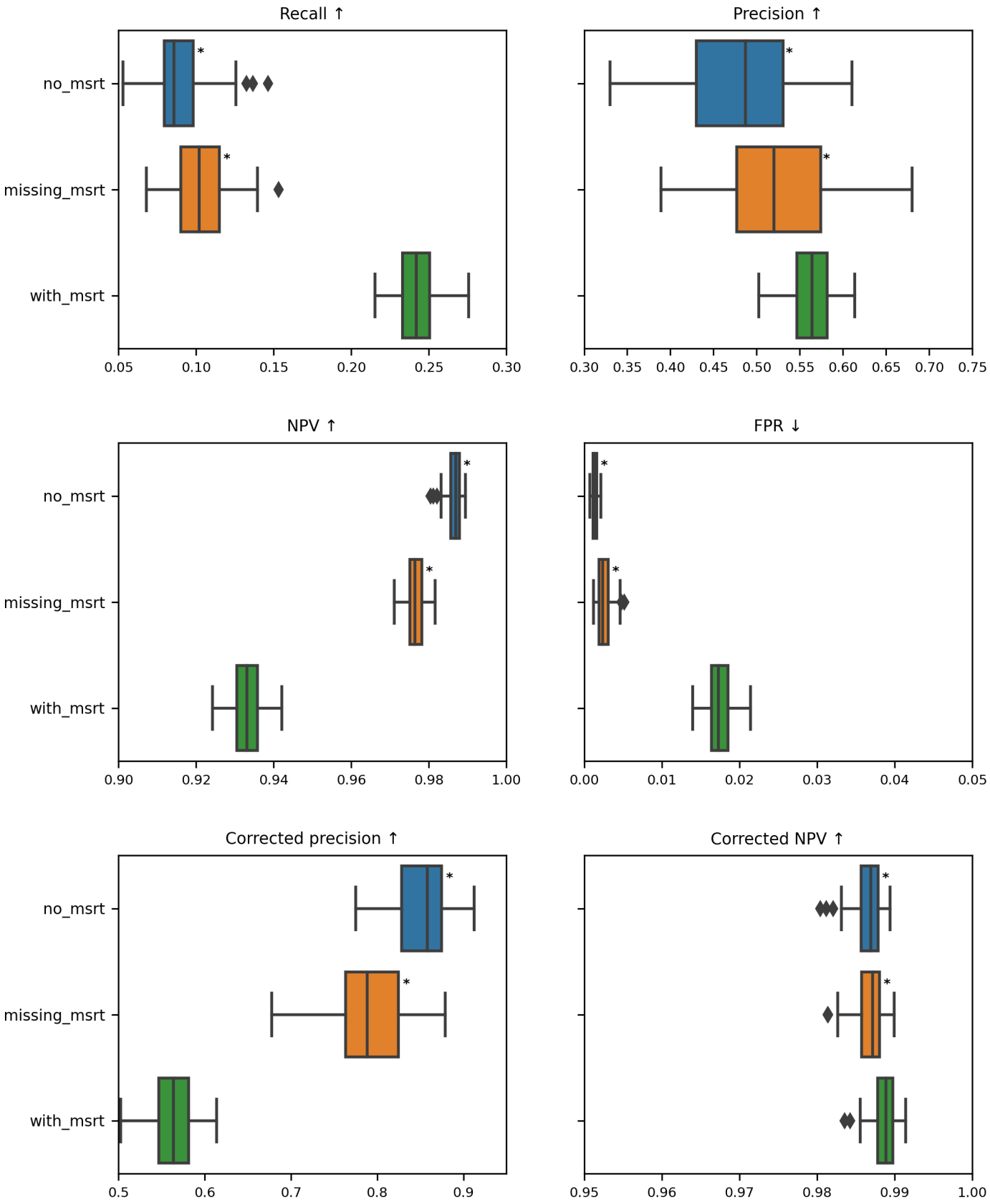
Figure 6.3.1.d

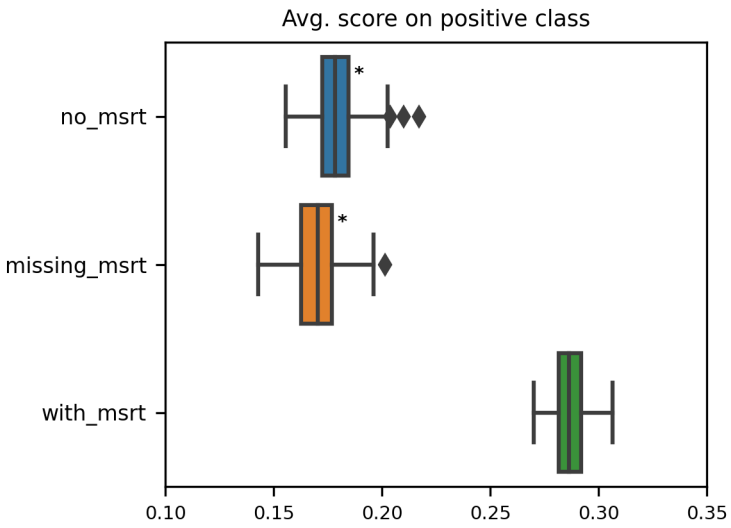


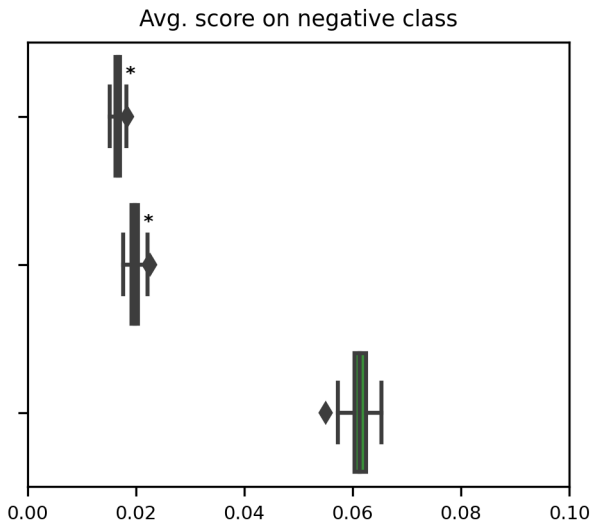
The intensity of measurements of Spitzendruck and surgical_status attributes are dependent.

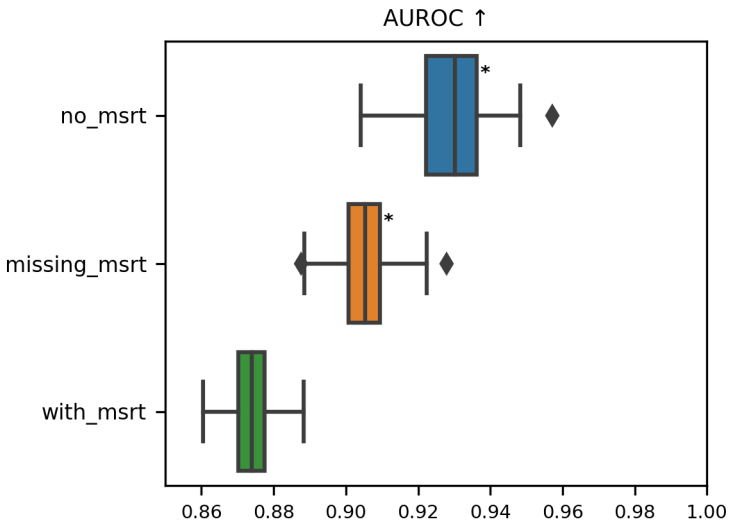
6.3.2. Impact on performance

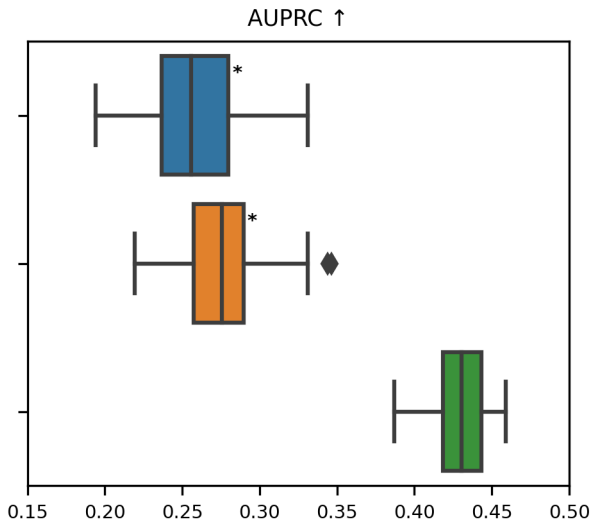
Figure 6.3.2.a


Recall ↑
no_msrt
missing_msrt
with_msrt
0.05
0.10
0.15
0.20
0.25
0.30
Precision ↑
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
NPV ↑
0.90
0.92
0.94
0.96
0.98
1.00
FPR ↓
0.00
0.01
0.02
0.03
0.04
0.05
Corrected precision ↑
0.5
0.6
0.7
0.8
0.9
Corrected NPV ↑
0.95
0.96
0.97
0.98
0.99
1.00









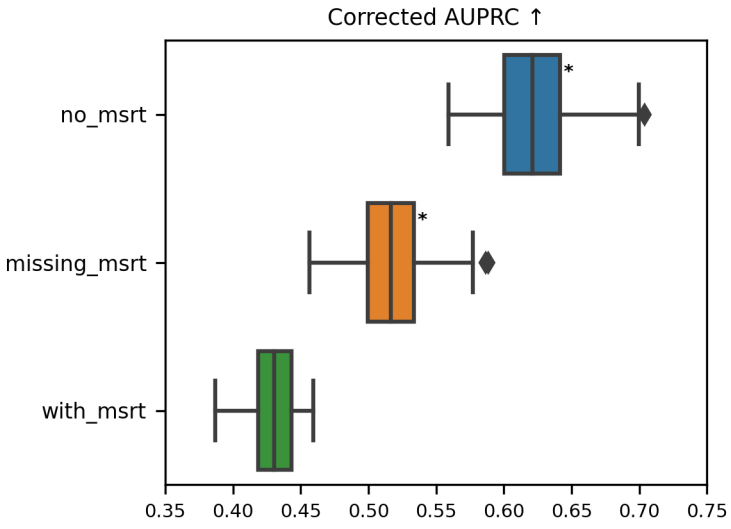


Table 6.3.2.a

Metric	Missingness category	Category vs with msrt	P-value	Delta
Recall ↑	no_msrt	worse	1.28e-34	0.156
Recall ↑	missing_msrt	worse	1.28e-34	0.14
Precision ↑	no_msrt	worse	2.56e-19	0.077
Precision ↑	missing_msrt	worse	6.21e-07	0.044
NPV ↑	no_msrt	better	1.28e-34	0.054

NPV ↑	missing_msrt	better	1.28e-34	0.043
FPR ↓	no_msrt	better	1.28e-34	0.016
FPR ↓	missing_msrt	better	1.28e-34	0.015
Corrected precision ↑	no_msrt	better	1.28e-34	0.294
Corrected precision ↑	missing_msrt	better	1.28e-34	0.225
Corrected NPV ↑	no_msrt	worse	1.79e-15	0.002
Corrected NPV ↑	missing_msrt	worse	1.16e-13	0.002
Avg. score on positive class	no_msrt	worse	1.28e-34	0.108
Avg. score on positive class	missing_msrt	worse	1.28e-34	0.116
Avg. score on negative class	no_msrt	better	1.28e-34	0.045
Avg. score on negative class	missing_msrt	better	1.28e-34	0.042
AUROC ↑	no_msrt	better	1.28e-34	0.056
AUROC ↑	missing_msrt	better	1.32e-34	0.031
AUPRC ↑	no_msrt	worse	1.28e-34	0.175
AUPRC ↑	missing_msrt	worse	1.28e-34	0.155
Corrected AUPRC ↑	no_msrt	better	1.28e-34	0.191
Corrected AUPRC ↑	missing_msrt	better	1.44e-34	0.086

7. Glossary

7.1. General concepts

Event: Failure or more generally health condition that the model aims to predict. We assume that it has some duration.
Grouping / Group name: This refers to an attribute used to form the cohorts of patients.
Category: (abbreviation: **Cat.**) This refers to the value taken by the grouping attribute, it characterizes a specific cohort. It can also be used to directly designate a cohort.
Cohort: This is used to designate a particular category of patients (i.e. a set of patients that share a common grouping attribute value).
Macro-average: Consider a grouping with *n* categories, and each category *i* has a metric value *m_i*, then the macro-average is *(m_1 + m_2 + ... + m_n)/n*.
Delta: (abbreviation: Δ) Each stage is associated with certain metrics, the delta for a metric and a cohort corresponds to the absolute difference in median metric between patients of this cohort and the rest of the patients.
Threshold on score: Binary classifier outputs probability between 0 and 1, to obtain a binary output the user has to decide on a threshold value below which the output class will be 0 and above which it will be 1.

7.2. Model Performance Analysis concepts

Metrics Definitions:
P number of positive labels, **N** number of negative labels, **TP** number of correctly predicted positive labels, **TN** number of correctly predicted negative labels, **FP** number of wrongly predicted negative labels, **FN** number of wrongly predicted positive labels.
$\uparrow$: Means that the larger the metric value, the better it is.
$\downarrow$: Means that the lower the metric value, the better it is.
Recall: *TP/P*
Precision: *TP/(TP+FP)*
NPV: Negative predictive value, *TN/(TN+FN)*
FPR: False positive rate, *FP/(FP+TN)*
Corrected precision: Precision corrected for the cohort prevalence of positive labels, *TP/(TP+s*FP)* with *s* the correcting factor that depends on the cohort prevalence and the maximum prevalence for the grouping.
Corrected NPV: NPV corrected for the cohort prevalence of positive labels, *TN/(TN+s*FN)* with *s* the correcting factor that depends on the cohort prevalence and the minimum prevalence for the grouping.
Event-based recall: Number of detected events over the total number of events.
Calibration curve: Illustrates how well the probabilistic predictions of the model are calibrated (whether they can be interpreted as true probabilities), x-axis mean predicted probabilities, y-axis frequency of positive labels. The perfect calibration line (dashed line in the figures) acts as a reference.
Calibration error: Area between the calibration curve and the perfect calibration line.
Avg. score on positive class: for all positive labels, average of the output scores.
Avg. score on negative class: for all negative labels, average of the output scores.
ROC curve: Receiver operating characteristic curve, x-axis FPR, y-axis TPR.
AUROC: Area under the ROC curve.
PR curve: Precision-recall curve, x-axis recall, y-axis precision. It can be drawn also for event-based recall and corrected precision.
AUPRC: Area under the PR curve. It can be computed for the PR curve drawn with event-based recall and/or corrected precision.
Ratio of significantly worse metrics: For a specific category of patients, it refers to the number of metrics for which the category is significantly worse off compared to the rest of the population divided by the total number of metrics.
Worst ratio: Refers to the largest **ratio of significantly worse metrics** (for a grouping or for the overall analysis).
Worst delta: Refers to the largest **delta** in performance metrics (for a grouping or for the overall analysis).

7.3. Time Gap Analysis concepts

Time gap: Amount of time between the trigger of the first correct alarm and the event occurence.
Start event: Considered split of the alarm horizon. We split the alarm horizon into different windows (chosen by the user) based on how much time in advance the alarm can be triggered. The available

prediction horizon can not be longer than the time between the start of the considered event and the start of the stay or between the start of the considered event and the time when the previous event finished.

7.4. Medical Variable Analysis concepts

Not in event: Refers to the median value computed on time points when patients aren't undergoing an event.
Never in event: Refers to the median value computed for patients without any event during their stay.

7.5. Feature Importance Analysis concepts

Feature importance: Approximates how useful is a feature for the prediction task. We use SHAP values to estimate it.
RBO (Rank-biased overlap): Similarity measure between two lists that focuses more on the head of the list (i.e it penalizes more mismatches that occur at the beginning). We use this measure to compare two feature rankings.
General feature ranking: Refers to the ranking of features based on their importance (from the most important to the least important), obtained on the entire set of patients. In contrast to cohort-based rankings, that are obtained on a specific cohort of patients.
Delta of inverse rank: For a feature that has rank *rk_0* in the cohort-based ranking and *rk_all* in the general ranking, it is defined as |*1/rk_0 - 1/rk_all*|. If it is big enough, we consider the change in rank of the feature from the general to the cohort-based ranking to be significant.
Top 15 (cohort): refers to the first 15 features of the general (or cohort-based) ranking.

7.6. Missingness Analysis concepts

Performance metrics definitions:
All metrics have already been defined in the **Model Performance Analysis concepts**.
Intensity of measurement categories
no_msrt: Refers to patients without any measurement for a variable.
insufficient: Refers to patients with between 0% (not included) and 90% of valid measurements (over the number of expected measurements).
The number of expected measurements is computed from the medical variable's expected sampling interval *t_e* (input from the user) and the patient's length of stay *los* as *los* / *t_e*.
enough: Refers to patients with between 90% (not included) and 100% of valid measurements (over the number of expected measurements).
The number of expected measurements is computed from the medical variable's expected sampling interval *t_e* (input from the user) and the patient's length of stay *los* as *los* / *t_e*.
Missingness categories:
no_msrt: Refers to patients without any measurement for a variable (before full data imputation).
missing_msrt: Refers to data points without valid measurement for a variable (before full data imputation but after forward propagation of measurements based on the variable's expected sampling interval).
with_msrt: Refers to data points with valid measurements for a variable (before full data imputation but after forward propagation of measurements based on the variable's expected sampling interval).
Dependent/Independent: Refers to the result of the Chi-squared independence test.