Supplementary

The supplementary is divided into three chapters.

Accuracy and standard deviation

In this chapter we report about the mean and standard deviation of the different experiments. The standard derivation is reported in the brackets.
This experiment is the classification over the full range of LOS.

Method	Accuracy	F1	balanced Accuracy	MCC	AUC
LR	0.784 (0.003)	0.435 (0.0010)	0.608 (0.0033)	0.303 (0.010)	0.735 (0.0069)
SVM	0.792 (0.003)	0.438 (0.0010)	0.610 (0.0033)	0.326 (0.0076)	0.800 (0.0045)
RF	0.810 (0.003)	0.442 (0.0009)	0.653 (0.0024)	0.406 (0.0050)	0.740 (0.0041)
XG	0.802 (0.004)	0.438 (0.0015)	0.669 (0.0025)	0.398 (0.0072)	0.777 (0.0035)

This experiment is the regression over the full LOS from 1 to 21.

Method	RMSD	MAE	MAPE	R^2
LR	2.893 (0.0454)	1.916 (0.0181)	70.655 (0.5755)	0.439 (0.0097)
SVM	2.908 (0.0437)	1.684 (0.0178)	48.369 (0.3020)	0.483 (0.0100)
RF	2.807 (0.0347)	1.868 (0.0136)	69.561 (0.5615)	0.493 (0.0078)
XG	3.001 (0.0301)	1.990 (0.0042)	72.075 (0.8323)	0.434 (0.0062)

This table is the experiment where we only predict the exact LOS but only in the range from 1-4 days

Method	RMSD	MAE	MAPE	R^2
LR	0.772 (0.0059)	0.636 (0.0053)	35.169 (0.3375)	0.283 (0.0102)
SVM	0.781 (0.0080)	0.612 (0.0064)	31.269 (0.3290)	0.326 (0.0100)
RF	0.755 (0.0062)	0.621 (0.0056)	34.542 (0.3456)	0.350 (0.0094)
XG	0.825 (0.0050)	0.664 (0.0042)	36.352 (0.2385)	0.252 (0.0075)

This is the results of the experiment where we first classify the stays into short and long and use the predicted short stay for a regression task. The used model is the random forest.

Method	RMSD	MAE	MAPE	R^2
LR	1.073 (0.0368)	0.743 (0.0077)	38.869 (0.2340)	0.199 (0.0135)
SVM	1.096 (0.0374)	0.711 (0.0084)	33.107 (0.2407)	0.228 (0.0121)
RF	1.068 (0.0381)	0.741 (0.0081)	39.294 (0.2409)	0.230 (0.0171)
XG	1.166 (0.0348)	0.805 (0.0052)	41.682 (0.4674)	0.160 (0.0114)

This is the results of the experiment where we first classify the stays into short and long and use the predicted short stay for a regression task. The used model is the XGBoost.

Method	RMSD	MAE	MAPE	R^2
LR	1.138 (0.0351)	0.763 (0.0063)	39.529 (0.4295)	0.217 (0.0114)
SVM	1.156 (0.0362)	0.726 (0.0077)	33.491 (0.2444)	0.244 (0.0116)
RF	1.131 (0.0364)	0.761 (0.0067)	39.938 (0.4579)	0.251 (0.0130)
XG	1.227 (0.0288)	0.828 (0.0063)	42.317 (0.5475)	0.176 (0.0083)

Distribution train and test data

Run	% Train data short stays	% Test data short stays
1	75.70	76.13
2	75.60	76.54
3	75.64	76.36
4	75.89	75.40
5	75.70	76.11
6	75.80	75.75
7	77.05	75.71
8	76.12	75.79
9	75.77	75.82

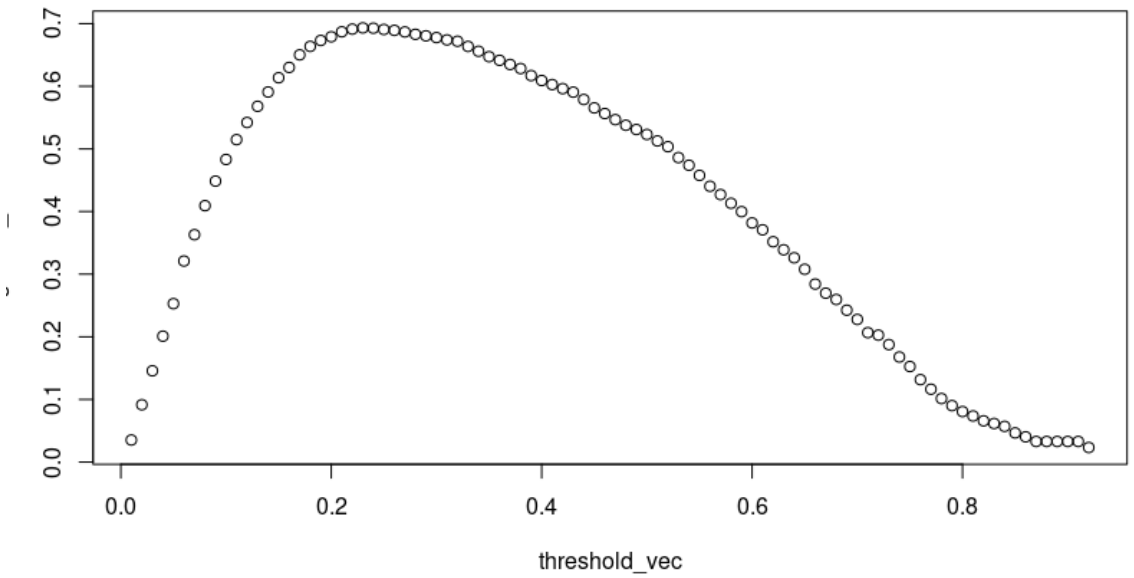
10	75.79	75.79

Gmean and confusion matrix

For the selection of the threshold in the classification task experiment with the stepwise regression where there is a classification and then a regression. For the selection of the short stays we used two models: random forest and XGBoost. For the calculation of the best threshold we use the Gmean (geometric mean) and therefore present the results of the calculation.

Random forest

Here is the random forest gmean results which is shown in the graph below. The graph has a optimum at 0.23 with a gmean of 0.693. The thresholds ranges from 0.01 to 0.99.



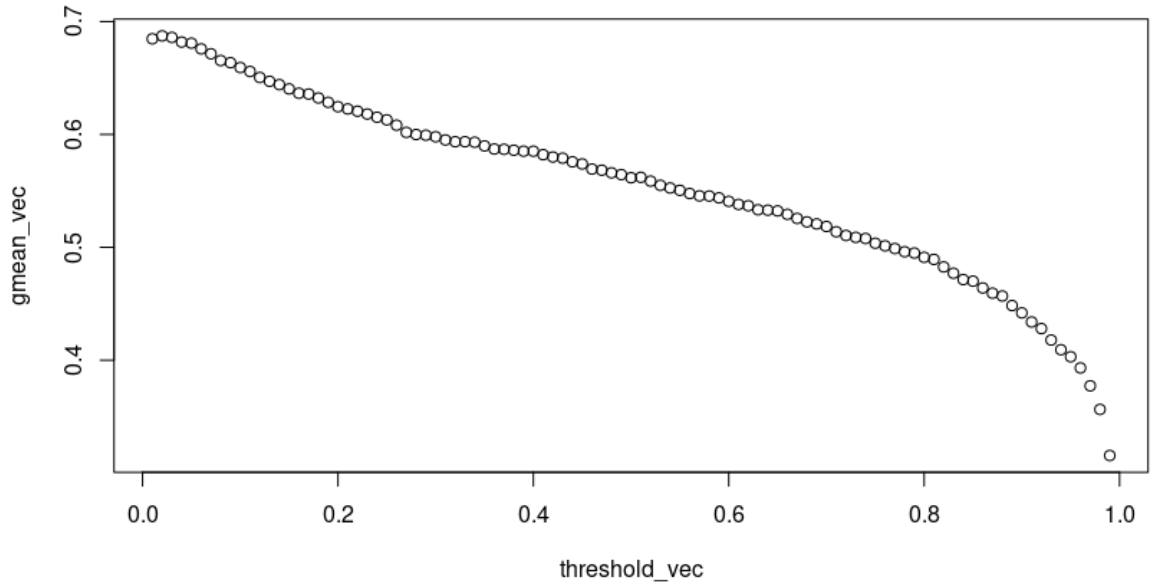
For the highest gmean we also show the confusion matrix. Note that the metrics of this matrix is much better than the reported because the gmean was optimised which is not the case for the calculation in the paper.

	prediction long stay	prediction short stay
actual long stay	9108	516
actual short stay	2469	29380

XGBoost

For the XGBoost we show the same results. The XGBoost has a different shape of the “curve” as the random forest with its optimum at 0.02 with and gmean of 0.687. We explain

this behavior from the extreme gradient boosting inside of the XGBoost algorithm which pushes the probabilities through the limit and therefore the threshold we will be heavily on one side.



For the highest gmean we also show the confusion matrix. Note that the metrics of this matrix is much better than the reported because the gmean was optimised which is not the case for the calculation in the paper.

	prediction long stay	prediction short stay
actual long stay	9007	617
actual short stay	1873	29976

Diagnosis MIMIC IV and ICD transformation

The ICD transformation is briefly described in the manuscript but here we would like to make a practical example of how it works.

Diagnoses in ICD 9 and 10 available;
total diagnoses ICD 9 : ~55%
total diagnoses ICD 10: ~45%

Converting 9 to 10 with the Touch package -> 4 cases

1 case: icd 9 = 5723 to icd 10=K766 (73.89%)
2 case: icd 9 = 5715 to icd 10= K740,K7460, K7469 (23.26 %)
3 case: icd 9 = 8603 to icd 10 = S271XXA+S21309A (0.65 %)
4 case: icd 9 = E9331 to icd 10 = (2.00%)

example:
For the example we choose two ICD 9 codes from the MIMIC IV database. This represent a usecase is an example.
This comes out of the conversion from ICD version 9 to 10.

icd_code	icd_version	icd_code_10
5715	9	K740,K7460,K7469
30981	9	F4310,F4312

Now our goal was to have a single ICD code for the ‘icd_code_10’ column. This allows us the proceed further like all ICD Codes would be in ICD 10.
First we shorten the ‘icd_code_10 shorten’ to the smallest number of digits from the ICD codes available (in this case 4) -> K740, K746, K746

Then we start a loop where we reduce the number of symbols in each string by 1 till the a ICD 10 code is found or there is no solution.

find overlapping between icd_cod_10

K746 = 2
K740 = 1

reducing length string by 1 : K74, K74, K74

K74 = 3 -> only one ICD 10 code present in the data

If there is no ICD 10 code found then we declare the conversion as failed (NA).

The result:

icd_code	icd_version	icd_code_10
5715	9	K74
30981	9	F431