WB Raport 2

Konrad Komisarczyk Kacper Grzymkowski Jakub Fołtyn

Analysis of responses

After reading the responses to (Yan et al. 2020) we decided to look for reasons why the model doesn't work and maybe try to rebuild the model to work on external data. We have however found that such a model would likely have to be significantly more complicated and less interpretable.

Genetic differences

Several studies have considered Lactate Dehydrogenase as a possible prognostic tool. In cancer research, a meta-analysis has found a significant genetic difference in LDH expression between Asian and Caucasian ethnicities (Lv et al. 2019). Early COVID studies also have found significantly higher mortality rates than what was observed later in the pandemic. (Wu et al. 2020) This might be one of the reasons for the model performing poorly.

Triage tool

Another reason for the model's poor performance could be it's improper usage as a triage tool. In the Outcomerea dataset, the sample excluded the healthiest patients and the most severely ill patients, which will naturally lower the effectiveness of the model (Dupuis et al. 2021).

Imbalanced tree model

A more technical explanation is the imbalanced decision tree being too trusting of LDH readings. Perhaps balancing the right side of the tree could help with creating a more robust model.

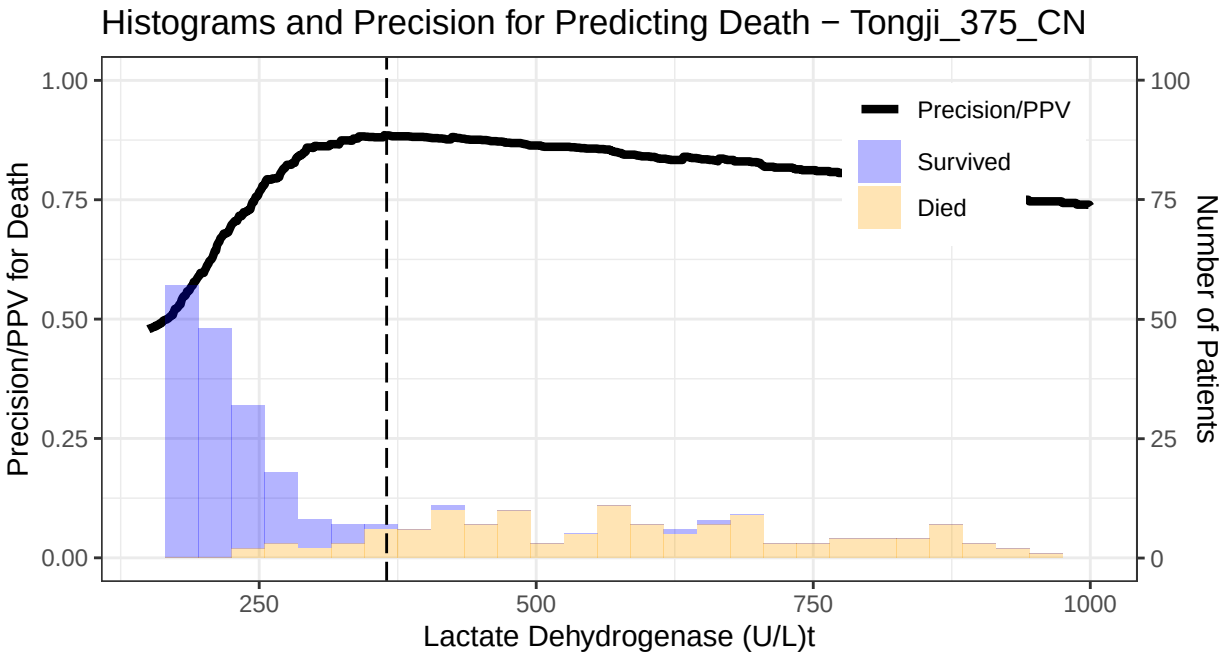
EDA

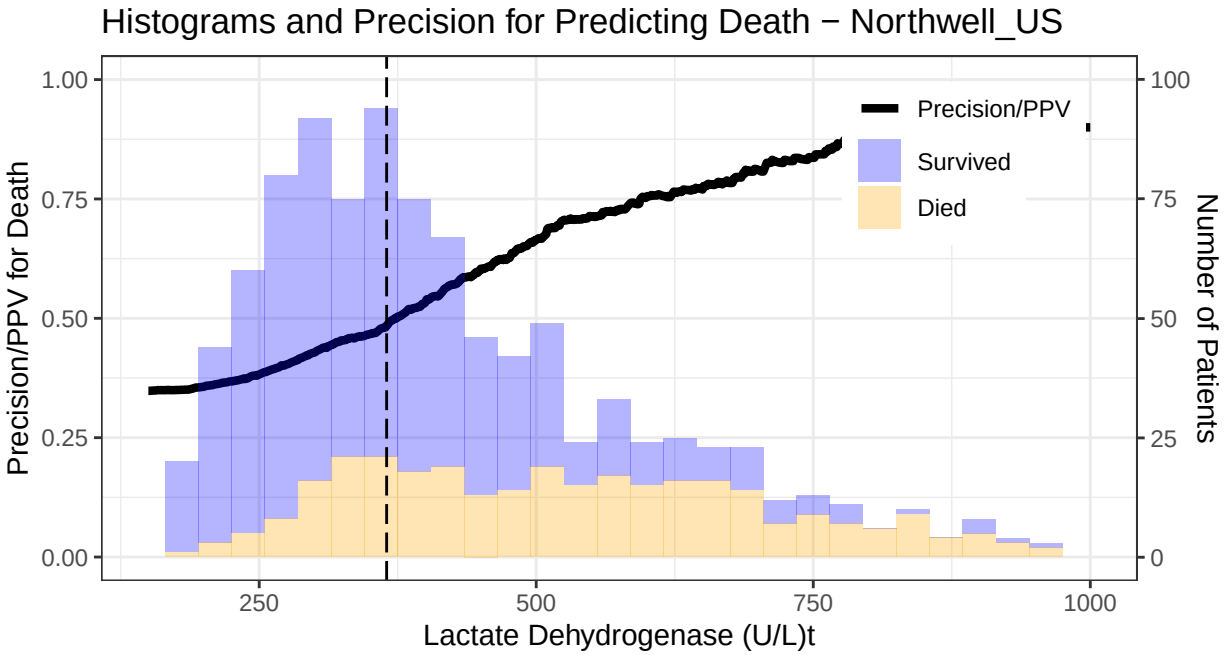
We have decided to take a closer look at patient data from 4 different sources, the original dataset on which the model was trained as well as the dataset from the 3 responses to the article:

- Tongji hospital (Yan et al. 2020)
- Outcomerea database (Dupuis et al. 2021)
- St. Antonius hospital (Quanjel et al. 2021)
- Northwell database (Barish et al. 2021)

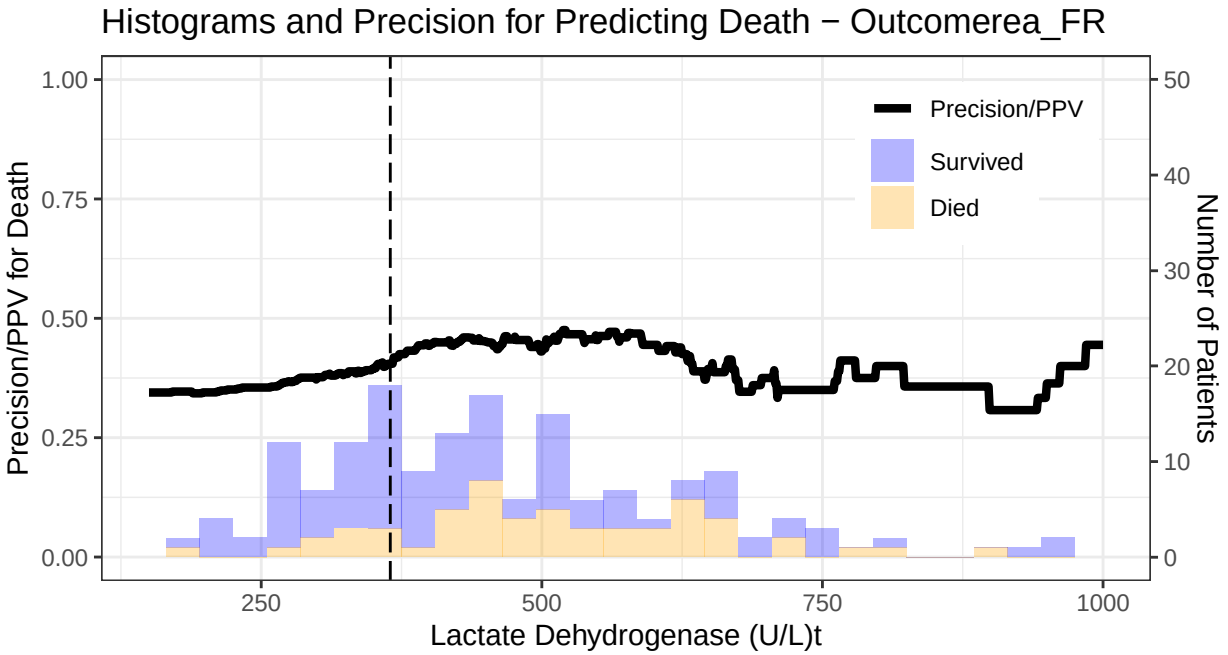
Histograms and PPV

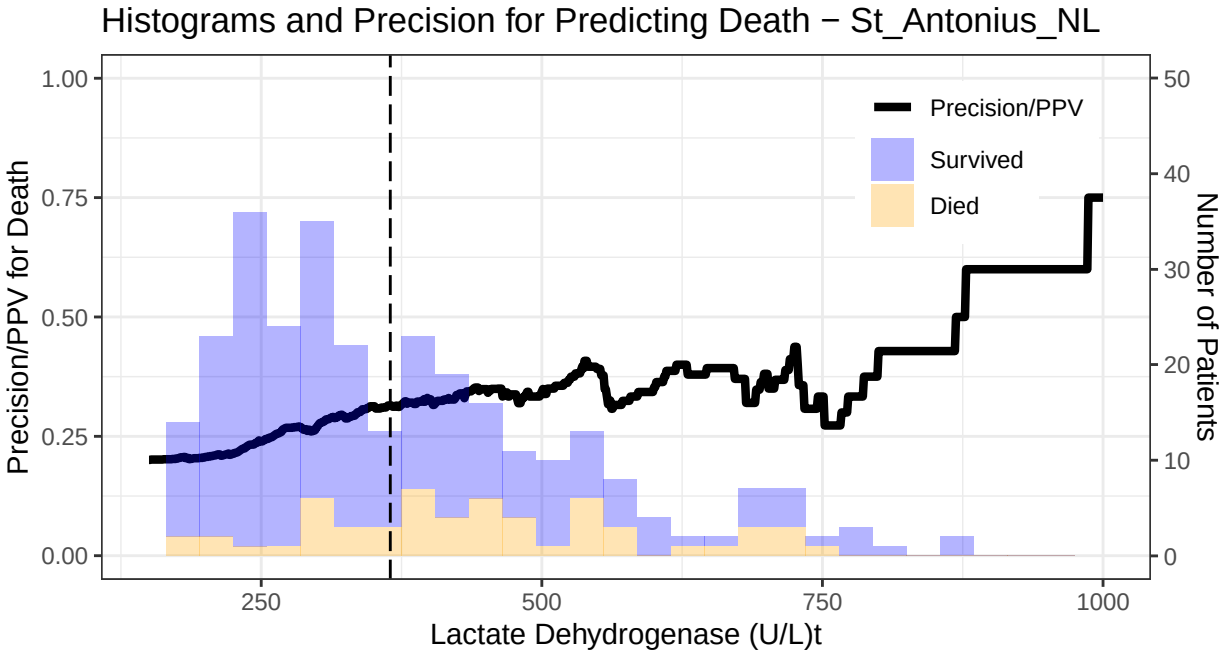
We really liked the visualization from (Barish et al. 2021) as it demonstrated the primary reason of the poor performance of the original model. The result wasn't compared to the Tongji data that Yan trained their models on. We decided to compare the distributions of patients from the two hospitals - Tongji, on which the original model was trained and Northwell, where the model was observed to have poor performance, as well as Outcomerea and St. Antonius, to which we have access.





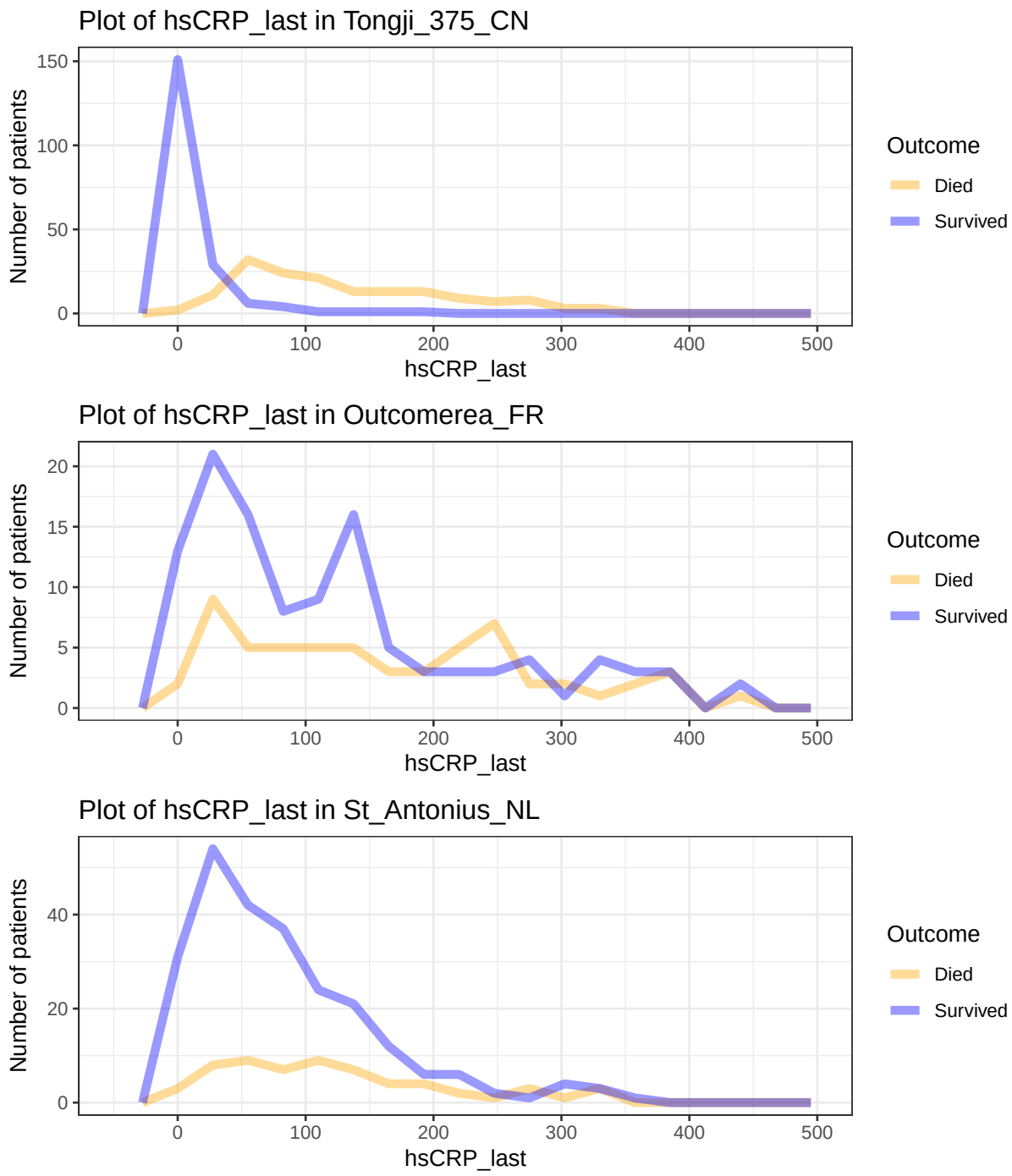
The data we were provided doesn't recreate the figure in the article. We aren't sure why that is. Perhaps we misunderstood the figure presented or we were given a smaller sample of the dataset. Whatever the reason, it's clear that the distributions are different. Particularly interesting is the large number of surviving, low LDH patients from Tongji data. In the Northwell dataset, survivors are more spread out. This is the primary cause of the poor performance of the model and the reported high rate of False Positives.





We also decided to look at the Outcomerea and St. Antonius data through the same lens. The small size of these datasets makes these plots difficult to interpret. Even so, these plots highlight the differences in distributions and the limited predictive ability of lactate dehydrogenase biomarker. One likely cause is the exclusion of worst outcome patients and best outcome patients.

Frequency polygons

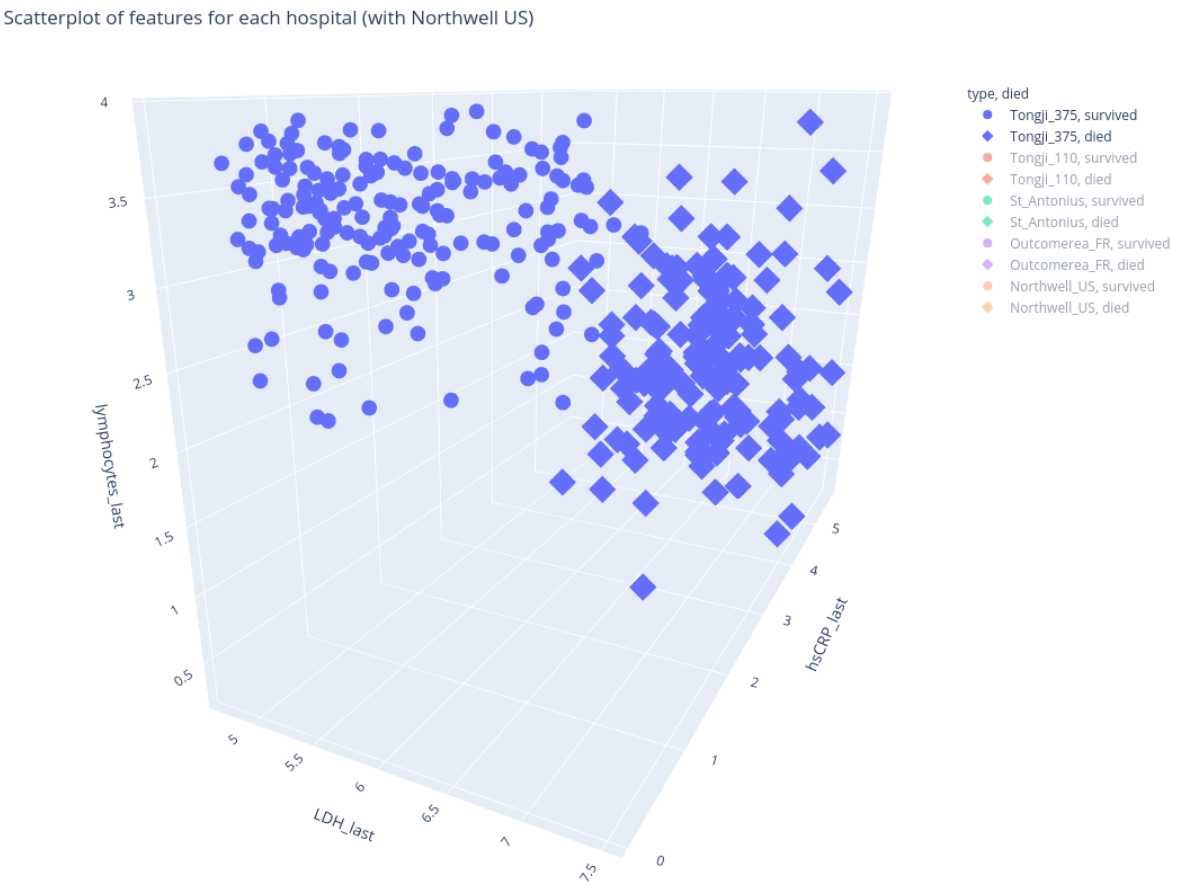


We also created density plots for other features, however the interpretation is similar to the LDH histograms. Tongji data when split by negative outcome has a "fat tail" that starts above a certain threshold. Data from other sources is more evenly distributed. It may be possible to split the distributions using a simple decision rule, such a split might prove ineffective.

3-dimensional scatter plots

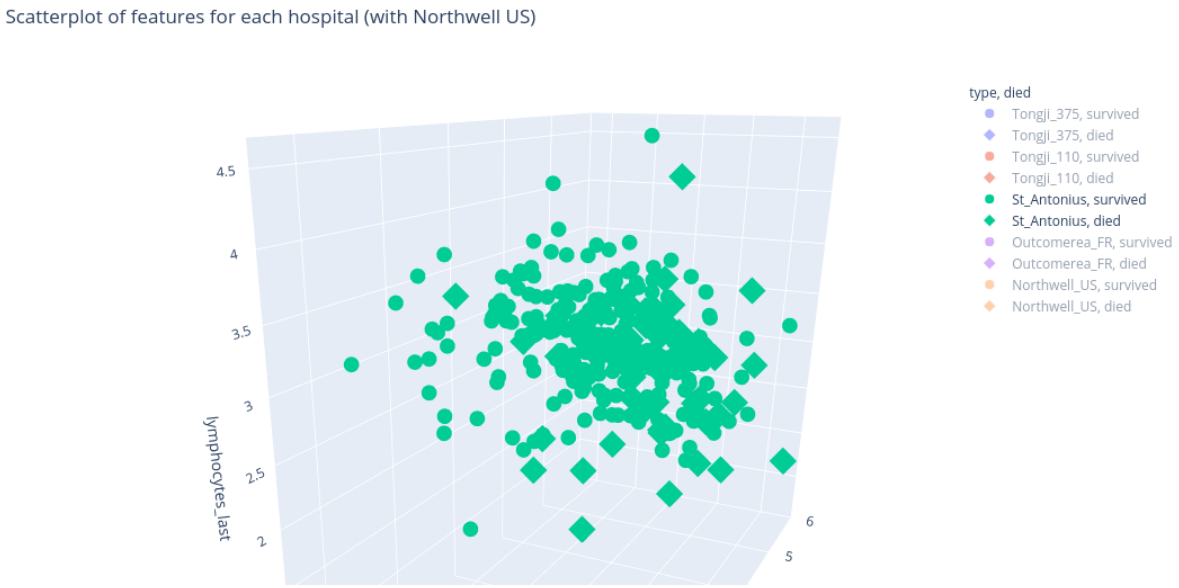
We have used plotly as a tool to create easily explorable 3D scatter plots of the 3 common biomarkers present in all 3 datasets. They are available in the code repository as .html files (plot with Northwell data has to be generated due to data availability issues).

Examination of Tongji data led us to belief that it exhibits clustering behavior. Applying to all the features a $\log(x) + 1$ transformation made the clusters even more visible.

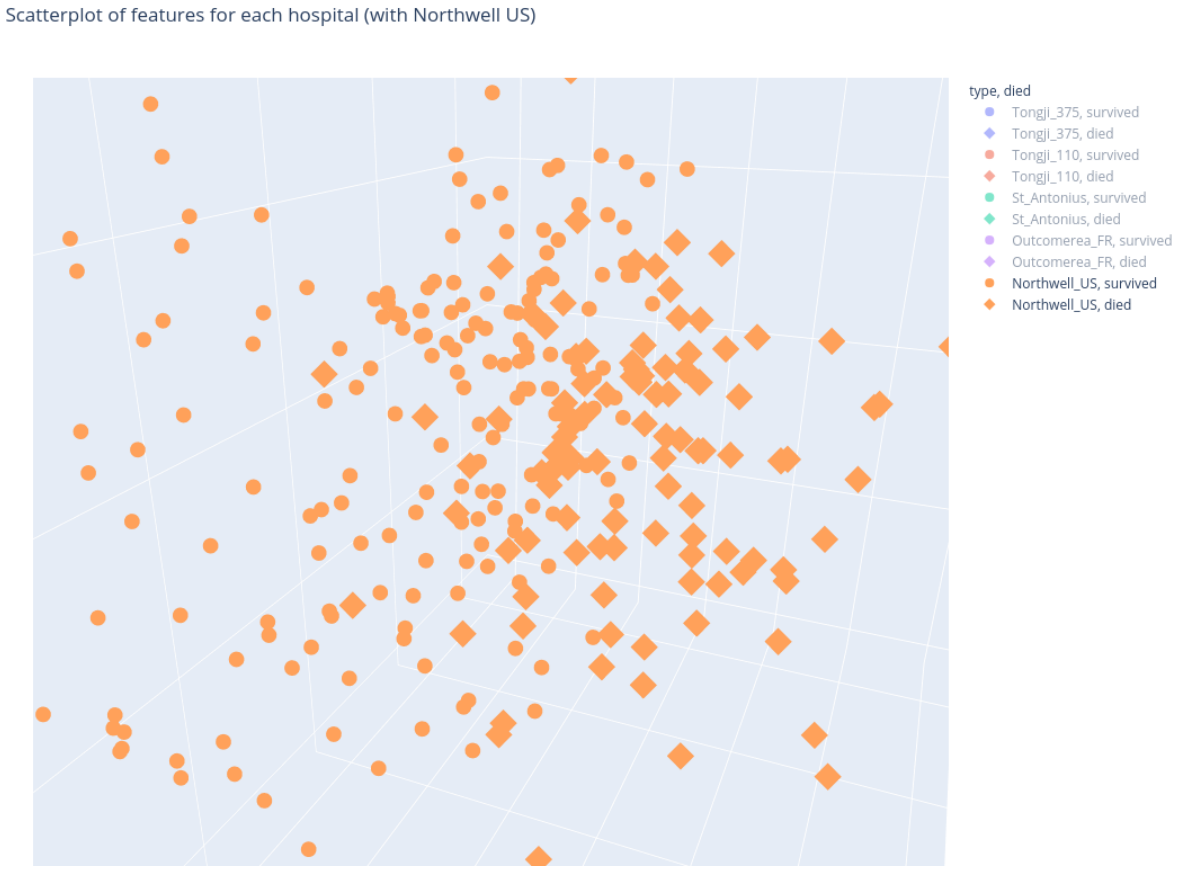


In further part of our work we test this hypothesis.

This behavior is not present to such a degree in any other dataset.

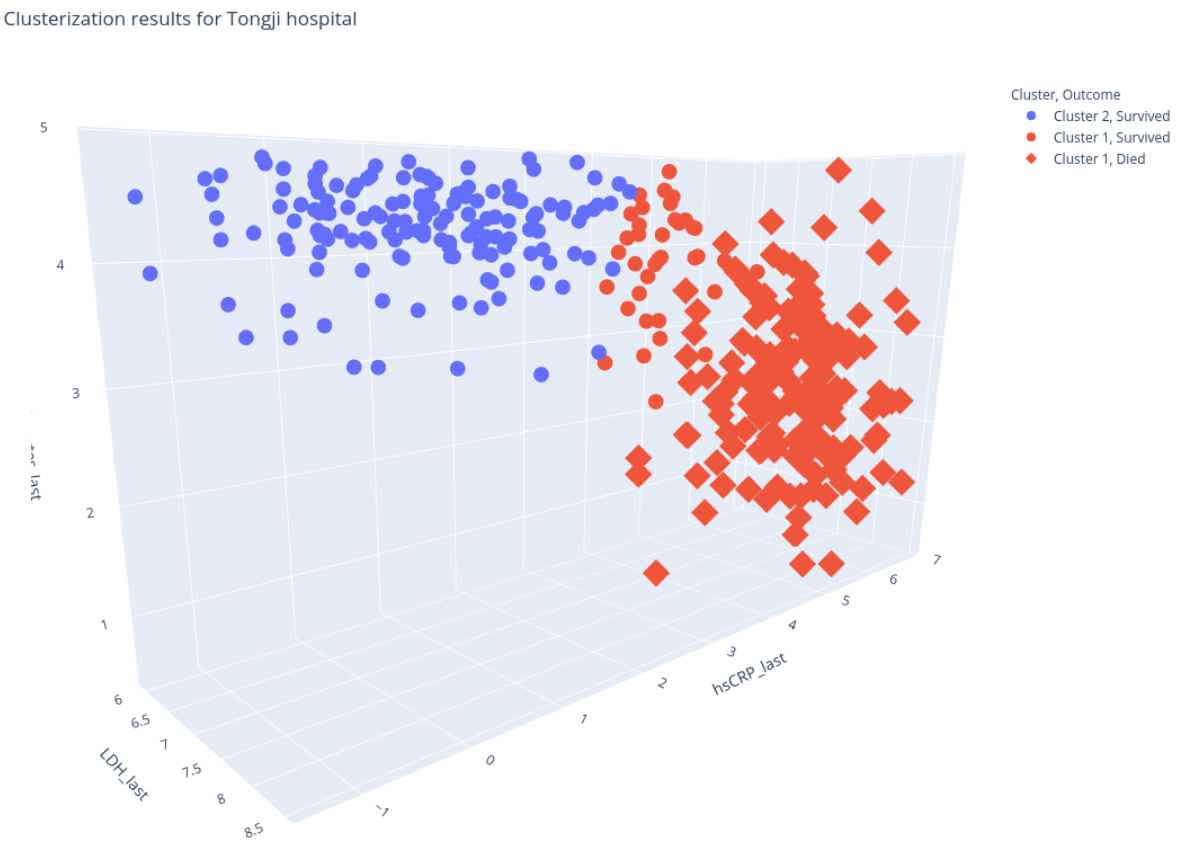


St. Antonius dataset was particularly unclustered. Creating a decision boundary wouldn't be very simple.



Northwell dataset shows some clustering behavior. Unlike the Tongji dataset, the boundary would have to go through the densest part of the plot.

Clustering Tongji data



We have ran a k-means algorithm on the original Tongji data. The idea behind this experiment is to show that even without labels the model can notice and create a decision boundary similar to the one in the article. This seems to be the case here, as there are no patients belonging to Cluster 2 who have also died and the majority of surviving patients belong to Cluster 2. This means that it shouldn't be surprising that models trained on this data perform well internally, but so poorly when validated with an external dataset.

We have ran a k-means algorithm also on all the other datasets but only for Tongji classes were visible as clusters. Removing outliers and transforming data did not improve the result.

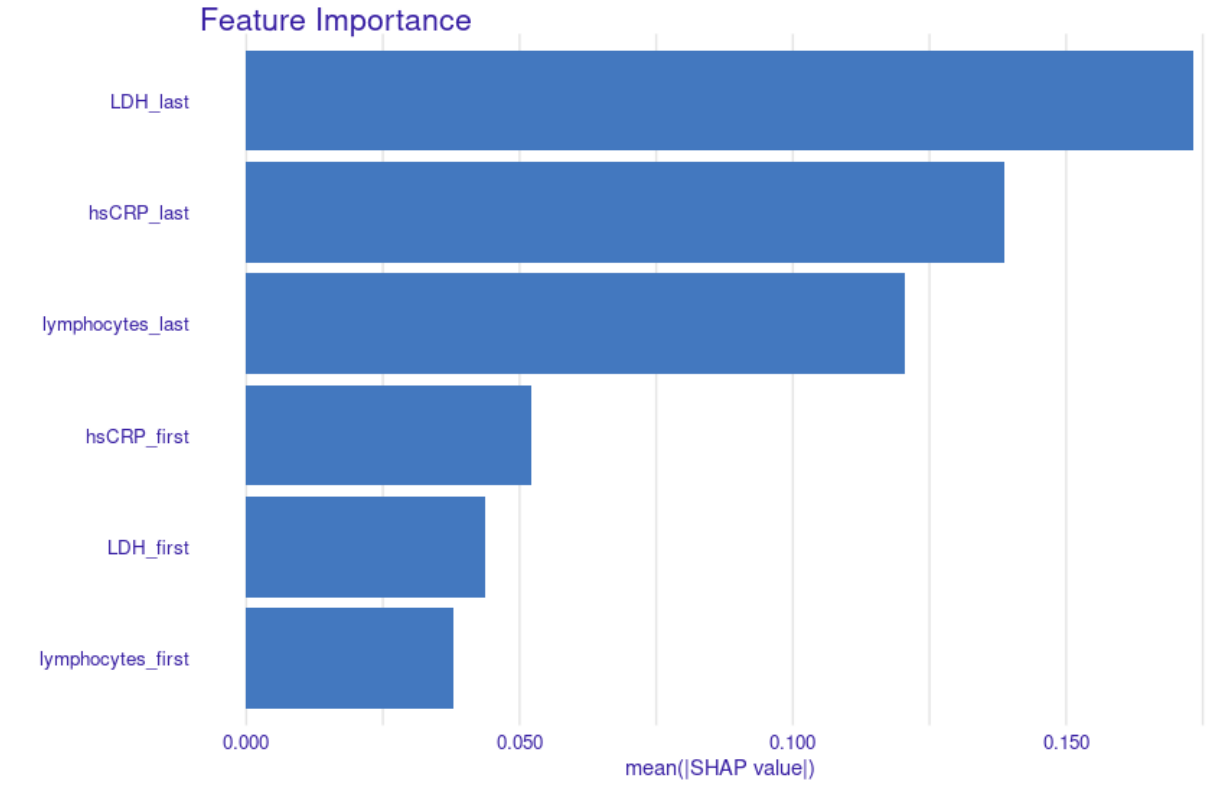
XGBoost and decision tree

Part of the criticism directed against (Yan et al. 2020) refered to the data used by authors. Main concerns were that original model scored badly on outside data. Responses to the article delivered other datasets that can be used to train model addressing the same problem. We decided to combine data from Tongji, St. Antonius and Outcomerea and Northwell to build a decision tree repeating steps from the original article. Our aim was to achieve acceptable performance on all the used sets.

Feature selection

Only common features of all combined sets were LDH, hsCRP and lymphocytes assays. Although, every set contained at least two measurements of each marker - first and last.

We trained an XGBoost model and calculated feature importances to compare impact of first and last measurements on the model. Metrics of feature importance we used were mean of absolute values of SHAP values (Lundberg and Lee (2017)) and Gain. SHAP based feature importance visualized below shows that for every marker last measurement had significantly greater impact on the model's prediction. Gain metric shows similar results.



We decided to keep only 3 most important features.

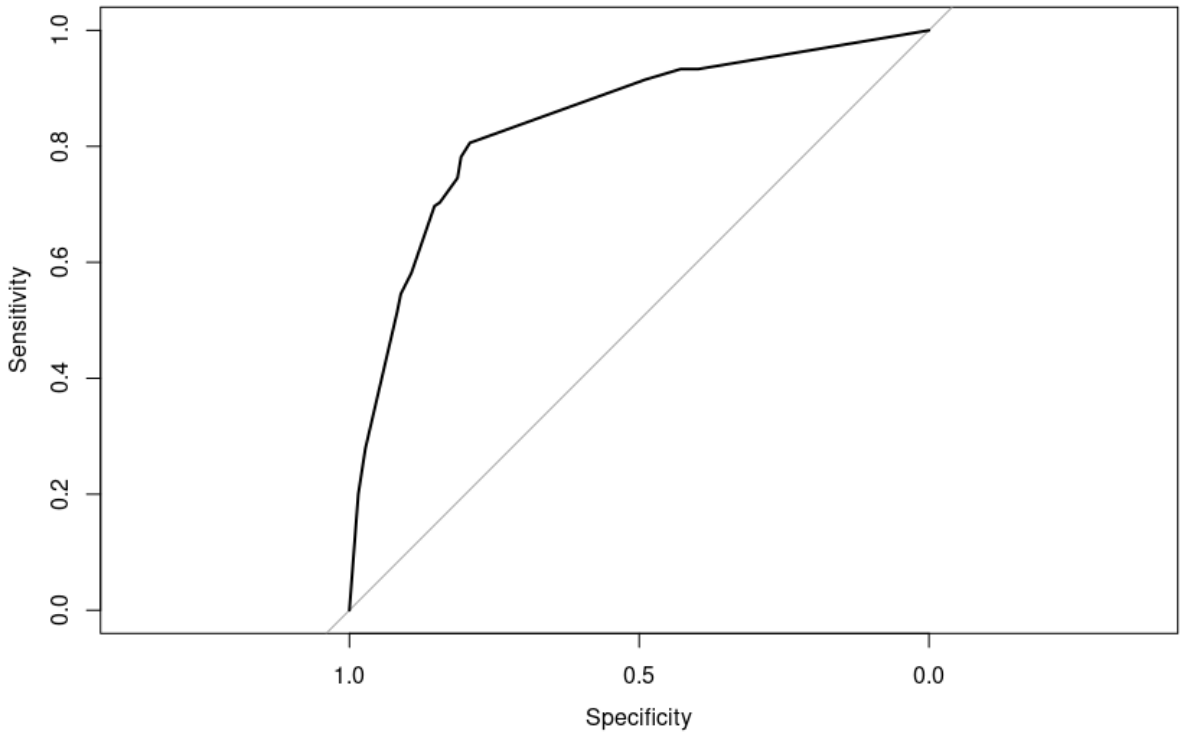
Decision tree

Using selected features we trained XGBoost models with different hyperparameters. Evaluation of their AUC scores led us to conclusion that a single decision tree scores comparably to larger models. Maximum AUC achieved was around 0.875 what is a good result. A single tree of depth 4 scored 0.84 and a tree of depth 3 scored 0.81. Both of these are satisfactory scores, but we decided the difference was great enough to choose the deeper one.

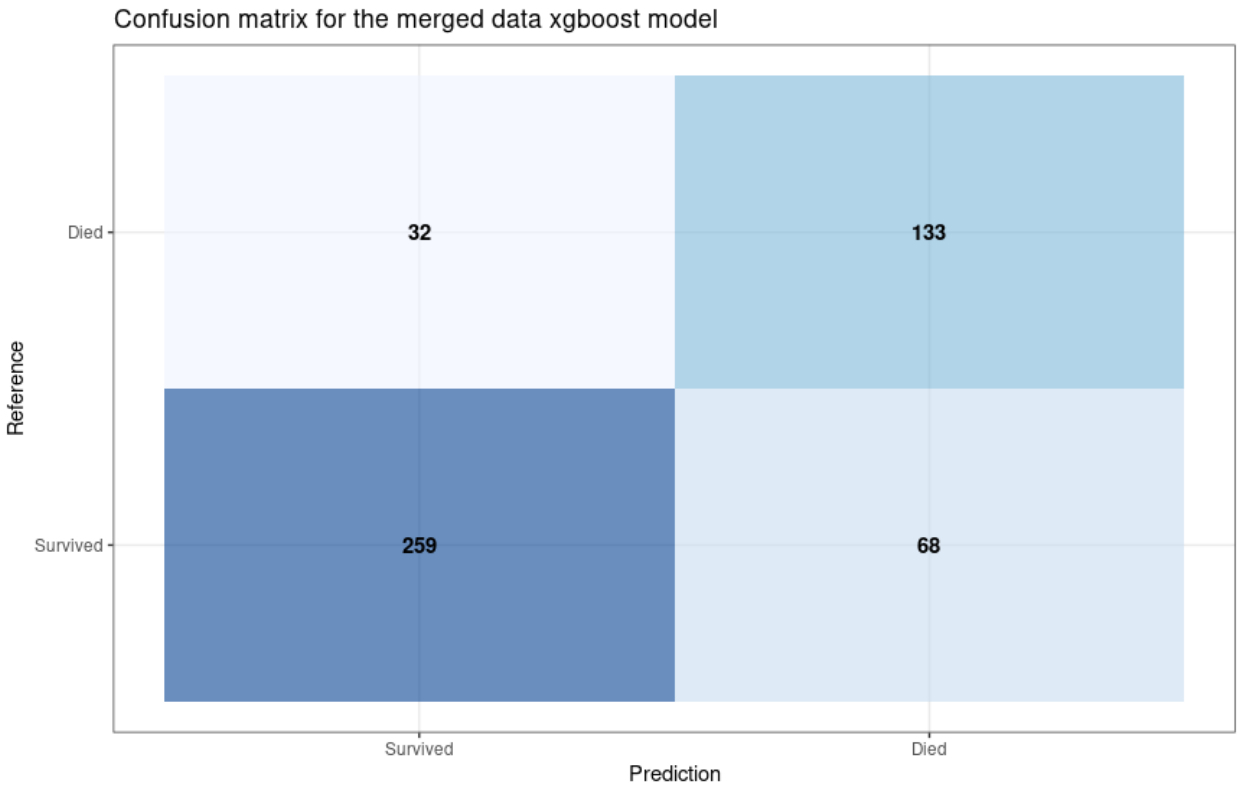
We checked the tree structure and the tree is balanced. Result does not rely on only one feature in any case.

Results at the combined set

Below we show ROC curve for the decision tree tested on combined sets:



We decided to establish the threshold at 0.42 it was the point maximalizing Sensitivity and Specificity sum. Now the confusion matrix looks like this:

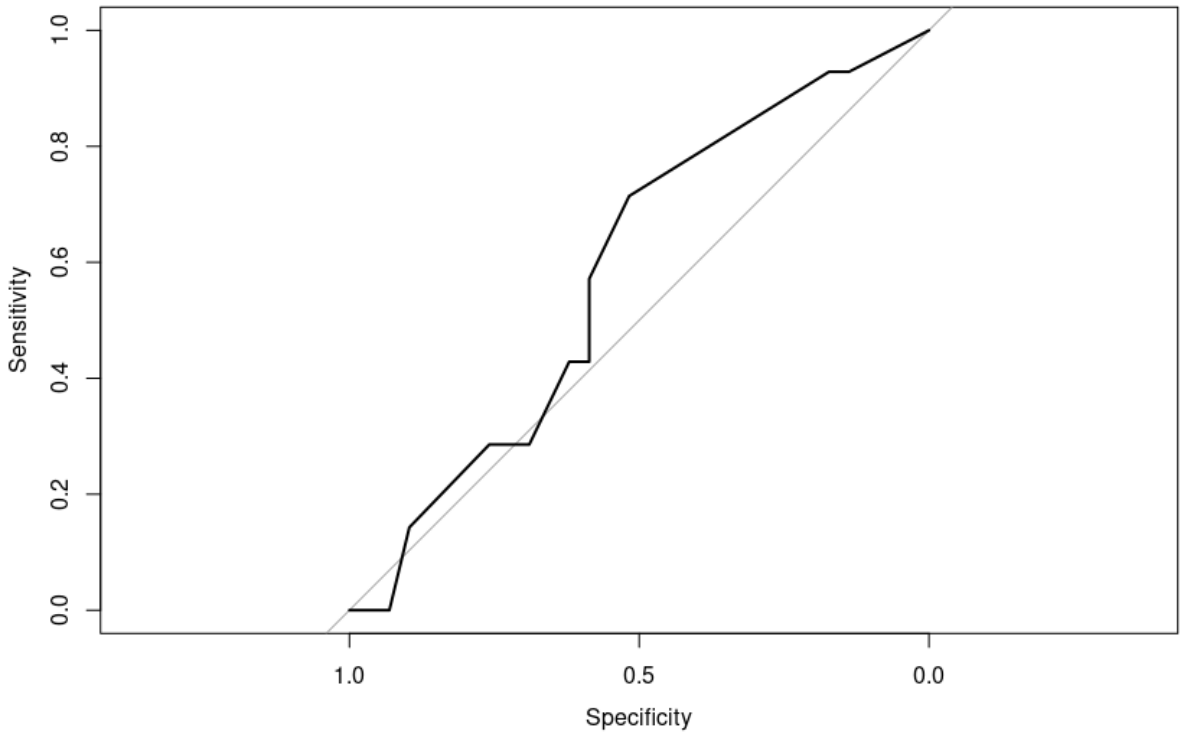


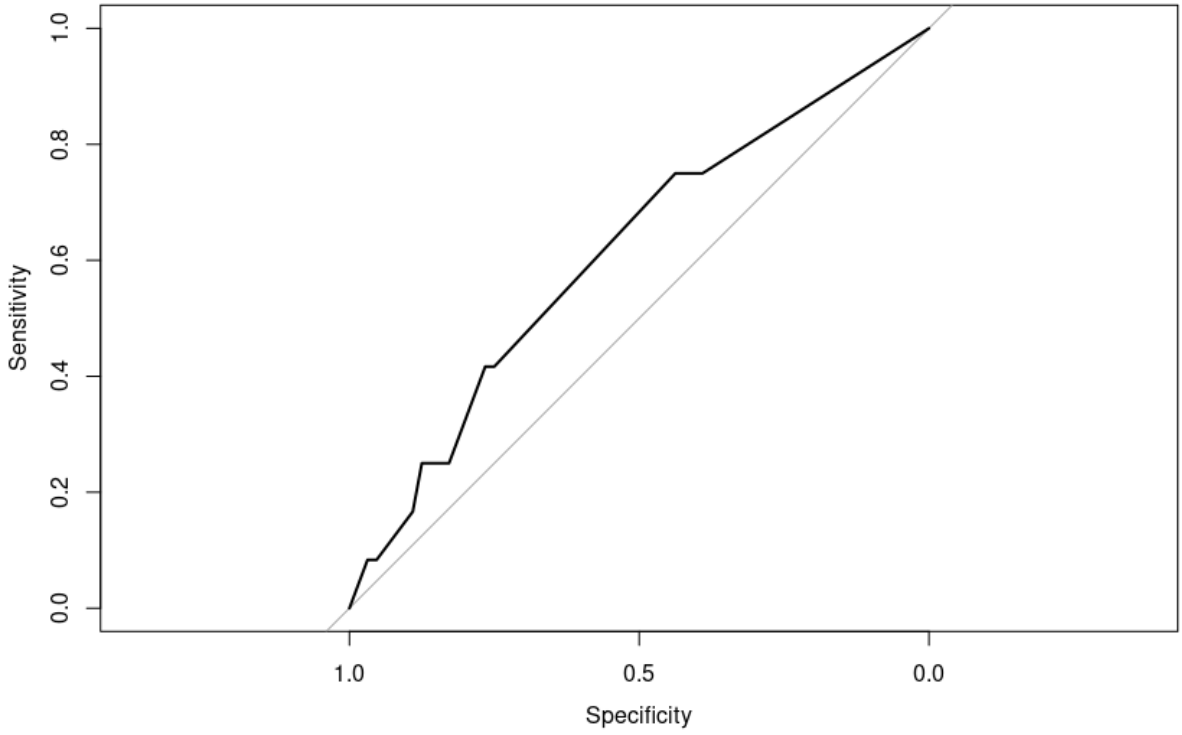
Model achieved following metrics:

```
Accuracy : 0.8
Sensitivity : 0.79
Specificity : 0.81
```

Results at the single sets

We examined performance of the model at the data from each source individually. On the data from Tongji and Northwell model scores were comparable to the scores on combined data. On the other side, performance on Outcomera and St. Antonius was close to random. Below we attach ROC curves for Outcomera and St.Antonius test sets respectively:


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0
Specificity


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0
Specificity

Increasing weights of observations

To achieve satisfying scores on all the sets we increased weight of observations from Outcomera and St. Antonius in XGBoost training set. Trying different weights (increased $1.5, 2, 2.5, 3, 4$ times for chosen sets) and different tree depths (3 - 5) we failed to develop a decision tree that would satisfy our assumed requirements of Accuracy > 0.7, Sensitivity > 0.65 and Specificity > 0.65 for both Outcomera and St. Antonius. Even training decision tree using only these two sets produced no acceptable results.

Conclusions and discussion

We failed in our attempts to develop an interpretable decision tree model that would correctly classify patients from St. Antonius and Outcomera sets. It is possible to build such a tree that will produce acceptable results when tested on the combination of all selected datasets.

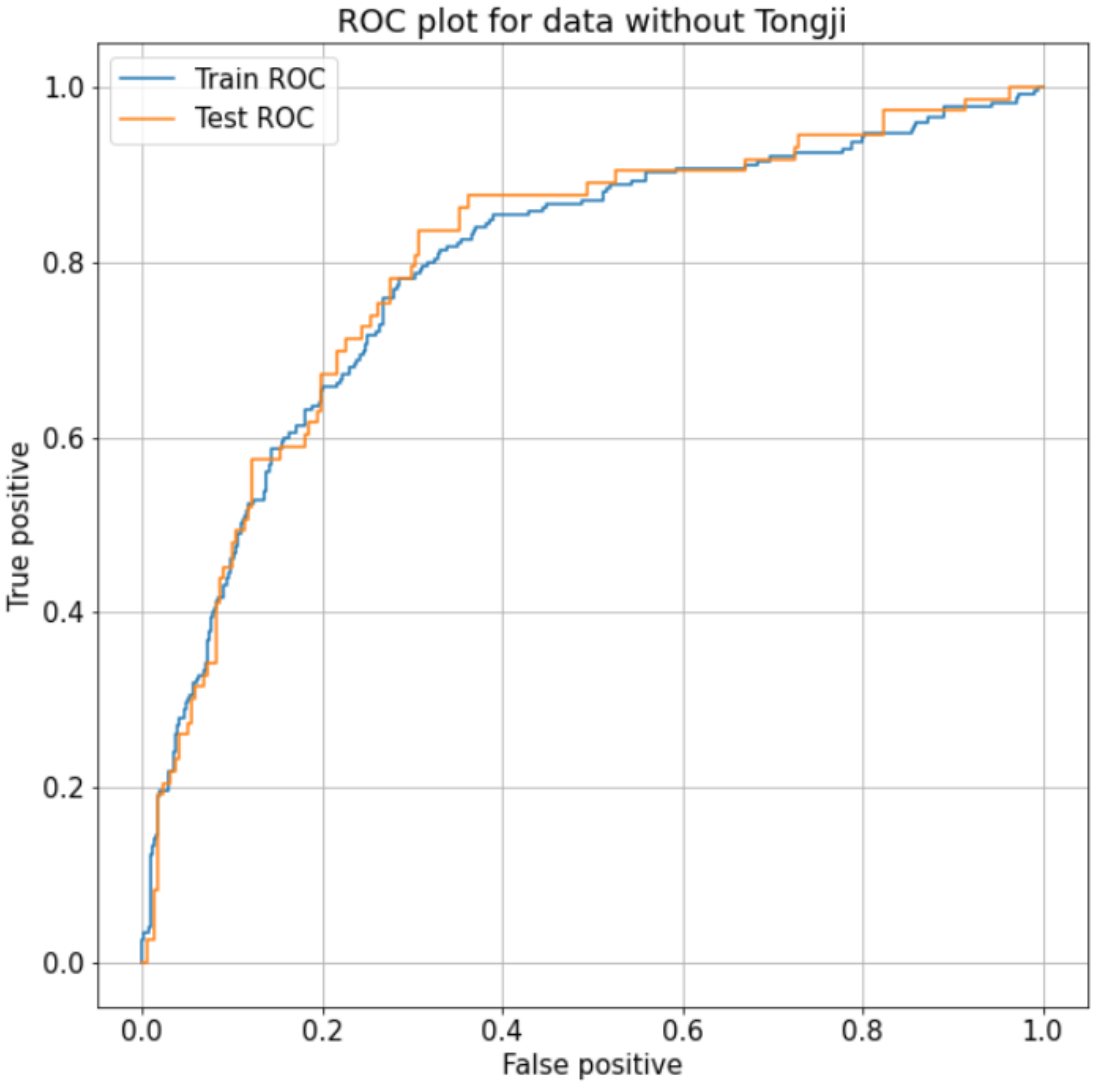
If the aim is to create a dataset representing general population, our method of simply combining available data may not be optimal. Balancing data by choosing some subsets of points from the sets is an option, but how to subset individual datasets has yet to be considered.

What is more, due to the differences between populations of different regions, such as genetic differences linked to the ethnicity, independently fitted models may be needed.

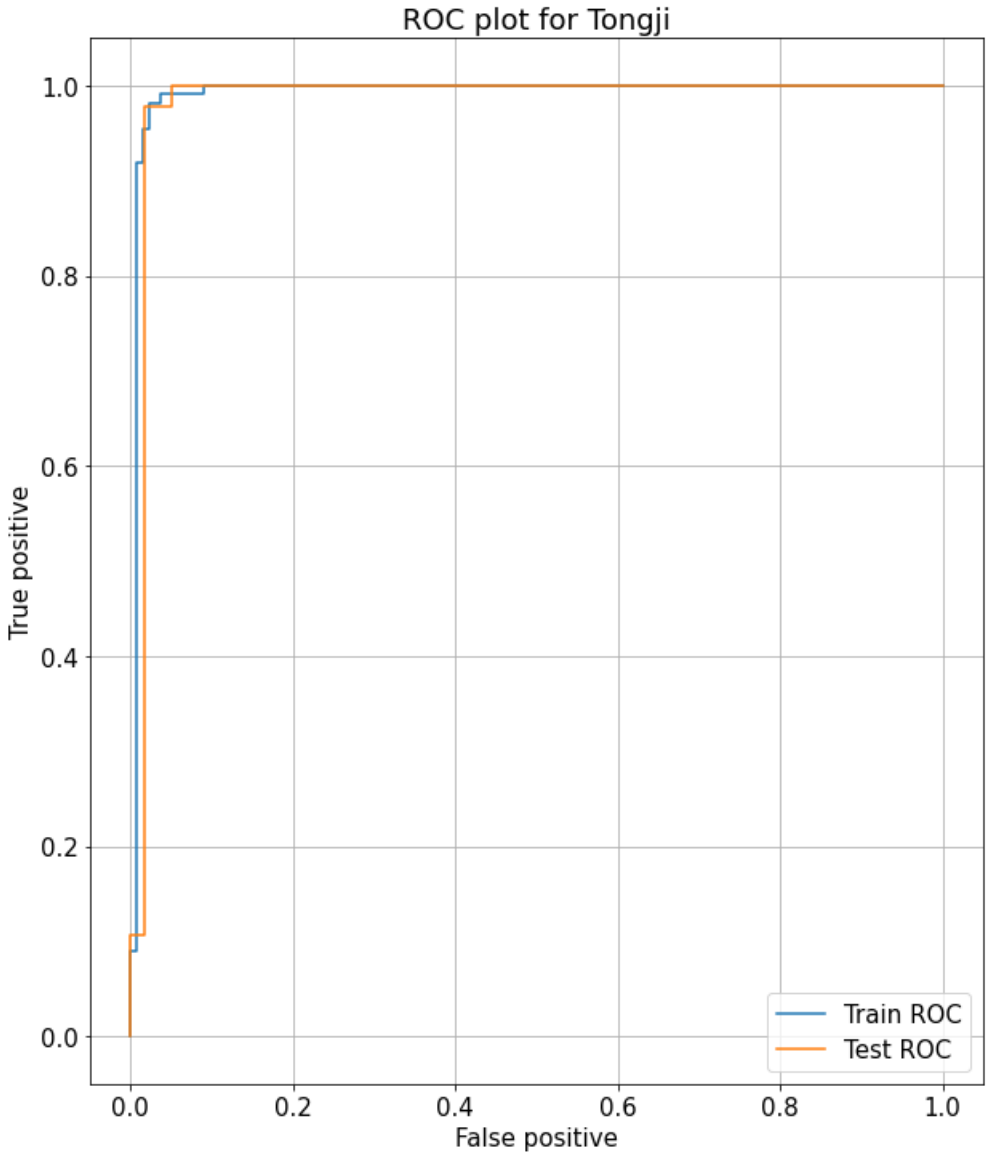
SVM testing

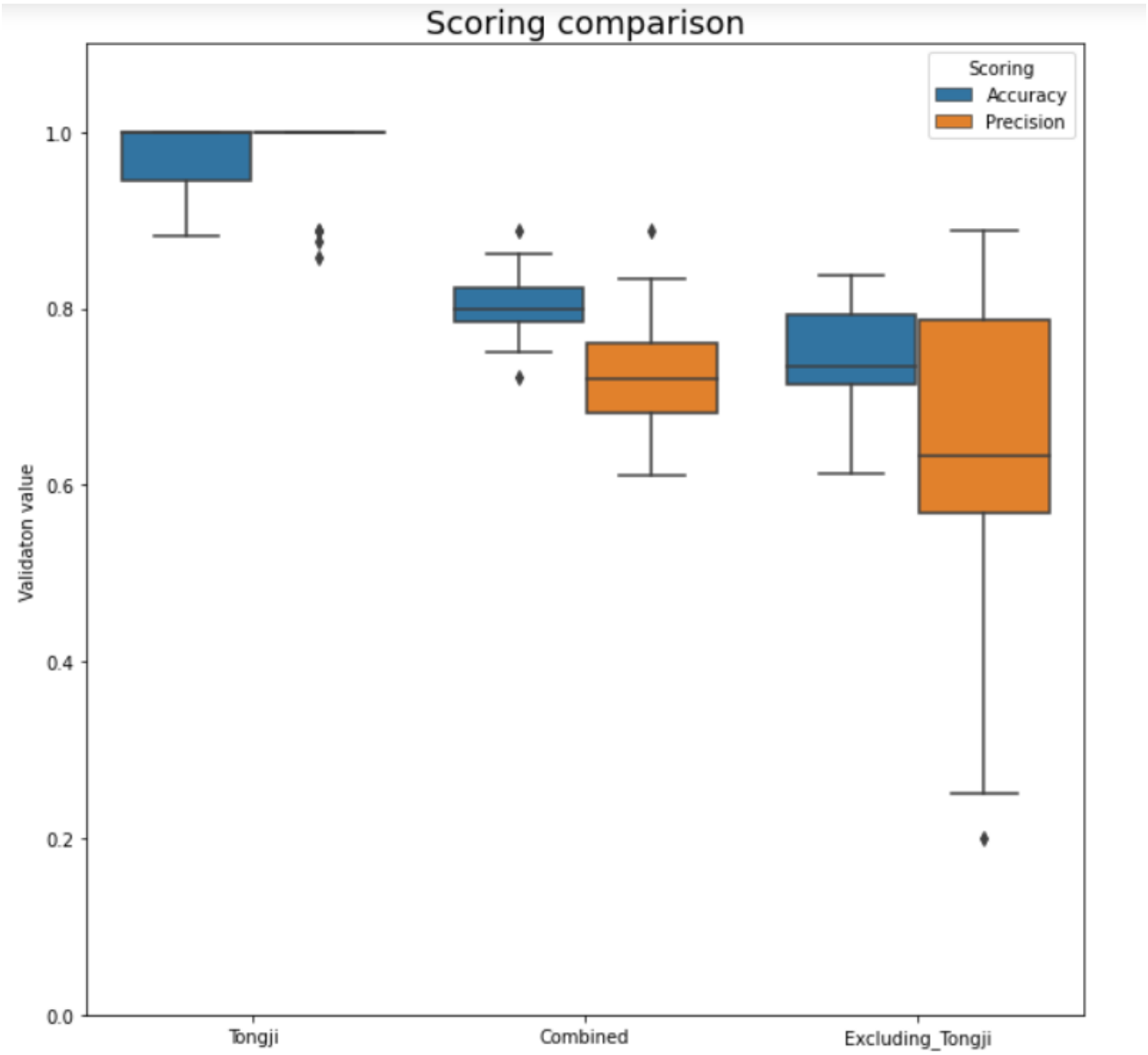
After examining 3D scatterplots for Tongji hospital, we decided that SVM might be a suitable model for that data. As expected, SVM trained on the original Tongji hospital data has produced particullary high scoring for both accuracy and precision. To compare, we also trained SVM on combined dataset, as well as on a dataset which excluded the Tongji hospital. The difference in shapes is best shown by ROC curves for data from Tongji hospital alone and for data excluding Tongji hospital.

```
AUC for train ROC is 0.7911014282256112
AUC for test ROC is 0.7991074195747846
```



```
AUC for train ROC is 0.9914742212674543
AUC for test ROC is 0.9841327082582041
```




Scoring comparison
Scoring
Accuracy
Precision
Validaton value
1.0
0.8
0.6
0.4
0.2
0.0
Tongji
Combined
Excluding_Tongji

References

Barish, Matthew, Siavash Bolourani, Lawrence F. Lau, Sareen Shah, and Theodoros P. Zanos. 2021. "External Validation Demonstrates Limited Clinical Utility of the Interpretable Mortality Prediction Model for Patients with COVID-19." *Nature Machine Intelligence* 3 (1): 25–27. https://doi.org/10.1038/s42256-020-00254-2.

Dupuis, C., E. De Montmollin, M. Neuville, B. Mourvillier, S. Ruckly, and J. F. Timsit. 2021. "Limited Applicability of a COVID-19 Specific Mortality Prediction Rule to the Intensive Care Setting." *Nature Machine Intelligence* 3 (1): 20–22. https://doi.org/10.1038/s42256-020-00252-4.

Lundberg, Scott, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions," December.

Lv, Jiancheng, Zijian Zhou, Jingzi Wang, Hao Yu, Hongcheng Lu, Baorui Yuan, Jie Han, et al. 2019. "Prognostic Value of Lactate Dehydrogenase Expression in Different Cancers: A Meta-Analysis." *The American Journal of the Medical Sciences* 358 (6): 412–21. https://doi.org/10.1016/j.amjms.2019.09.012.

Quanjel, Marian J. R., Thijs C. van Holten, Pieternel C. Gunst-van der Vliet, Jette Wielaard, Bekir Karakaya, Maaike Söhne, Hazra S. Moeniralam, and Jan C. Grutters. 2021. "Replication of a Mortality Prediction Model in Dutch Patients with COVID-19." *Nature Machine Intelligence* 3 (1): 23–24. https://doi.org/10.1038/s42256-020-00253-3.

Wu, Mei-ying, Lin Yao, Yi Wang, Xin-yun Zhu, Xia-fang Wang, Pei-jun Tang, and Cheng Chen. 2020. "Clinical Evaluation of Potential Usefulness of Serum Lactate Dehydrogenase (LDH) in 2019 Novel Coronavirus (COVID-19) Pneumonia." *Respiratory Research* 21 (1): 171. https://doi.org/10.1186/s12931-020-01427-8.

Yan, Li, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, et al. 2020. "An Interpretable Mortality Prediction Model for COVID-19 Patients." *Nature Machine Intelligence* 2 (5): 283–88. https://doi.org/10.1038/s42256-020-0180-7.