# Report v2

Marcelina Kurek 305741, Mateusz Staczek 305757, Jakub Wiśniewski 298850, Hanna Zdulska 298852

April 7, 2021

## Introduction

Our research is based on the article Yan et al. (2020) ”An interpretable mortality prediction model for COVID-19 patients” In the second step of our project, we reproduced the Chinese model and proposed new solutions to the given research problem. We also decided to compare our results with the results described in the reviews Dupuis et al. (2021), Quanjel et al. (2021), and Barish et al. (2021), of the article Yan et al. (2020).

## Life is not binary

Most datasets and articles are about whether a person will die or not. On the other hand, the state of the patient is rarely mentioned. We believe it is important since death is not the only state that the ill person would rather avoid. More categories could be added to each record in data such as whether someone is in ICU or needs special treatment. Similarly, when some unfortunate patient dies, adding a number of days from each blood test to death to each record might provide researchers helpful insights.

## Blood vs age and sex

Quanjel et al. (2021) have compared the model presented in the article by L. Yan et al. with data about Dutch patients with COVID-19. Using their publicly shared data Quanjel et al. (2021) and combining it with Chinese model - XGBoost with certain parameters presented by Yan et al. (2020). it was fairly easy to get the same results as presented in Dutch response article. Overall scores of the model were low and many patients with predicted unfavourable outcomes had actually survived.

After successful recreation of these unsatisfying results three more models were created. Both were XGBoost classifiers with the same parameters as the model presented in the Chinese article with aim to maximize area under PR curve (auprc).

Firstly we created XGBoost model on the Dutch data where information about patient (sex, age) was combined with blood test results. Then we checked what features were the the most important 1 using dalex package Baniecki et al. (2020). The three most important features were Age, LD, and Gender. Then we wanted to check if blood samples are by any means helpful, therefore the first model was trained on age and gender only, while the other was trained using LD, CRP and percentage of lymphocytes. The model with no knowledge about the patient's blood test results scored an average precision of around 0.4 on 3-fold cross validation. The second model theoretically should perform better, as blood provides crucial information about the state of an illness. However it only got the score of 0.23 on the same 3-fold cross validation test. This is somewhat disturbing information - only by briefly looking at person (judging by age and sex) we can make better predictions than after running blood tests.

Next, a similar pair of models was trained on data from China. This time the results were different: model trained on blood samples got an average precision of 0.93 on 3-fold cross validation while model which was using only information about age and gender got precision of 0.72 within the same metric. So for the model, Chinese blood samples are more useful than Dutch.

| | recall | precision | f1 | accuracy | auc |
|---|---|---|---|---|---|
| XGBoostClassifier | 0.545455 | 0.461538 | 0.5 | 0.803279 | 0.792727 |

Table 1: Metrics of Chinese model trained on Dutch data.

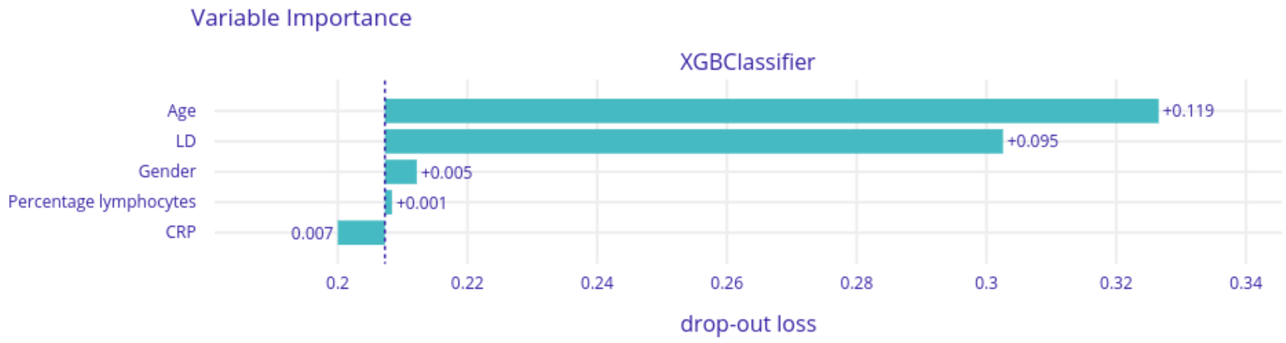

Figure 1: Variable importance for XGBoost Classifier trained on Dutch data shows importance of age and gender compared to blood sample. Parameters of this model have been described by L. Yan et al.Yan et al. (2020)

This suggests that the information about Dutch blood is less useful than Chinese and so are the results. Although, it is surprising to see the difference. It may be argued whether such comparisons should be made when using only small datasets (Chinese had about 450 records and Dutch only 300) but they show the point that a model precision may vary across countries.

## New York data

The New York dataset is not attached to the article (Barish et al., 2021) as it contains confidential data. However, due to the civility of the authors, we were provided with the dataset. The New York dataset contains 1000 observations, which is the highest number of records in comparison to the Chinese, Netherlands and French datasets. The precision score of the model proposed in the article (Yan et al., 2020) fitted to the New York dataset depends on the time when the blood sample was taken. The precision of the model trained on first blood samples was 0.47 on the 3-fold cross validation test, but on last blood samples it reached 0.76. It confirms that the prediction tends to be more accurate when the death or discharge date is near. 3

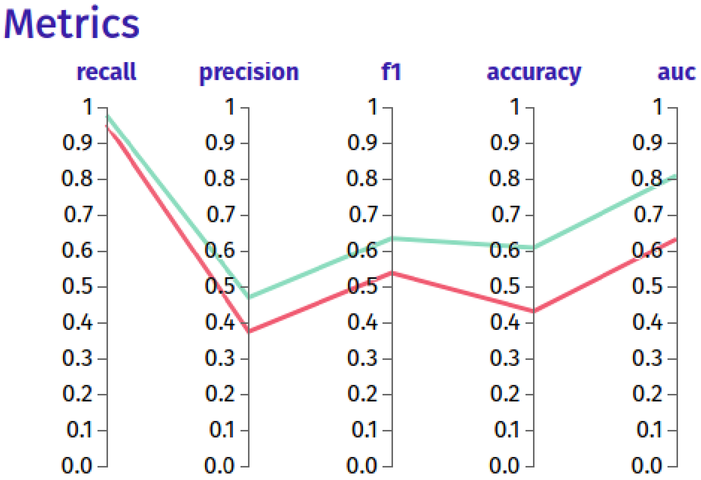

Figure 2: Teal line represents scores for last measurement red for first measurement

|  | recall | precision | f1 | accuracy | auc |
|---|---|---|---|---|---|
| XGBClassifier | 0.952778 | 0.374046 | 0.537197 | 0.430636 | 0.632223 |

Table 2: Performance of model trained on New York data (first measurements).

|  | recall | precision | f1 | accuracy | auc |
|---|---|---|---|---|---|
| XGBClassifier | 0.977778 | 0.469333 | 0.634234 | 0.608863 | 0.808837 |

Table 3: Performance of model trained on New York data (last measurements).

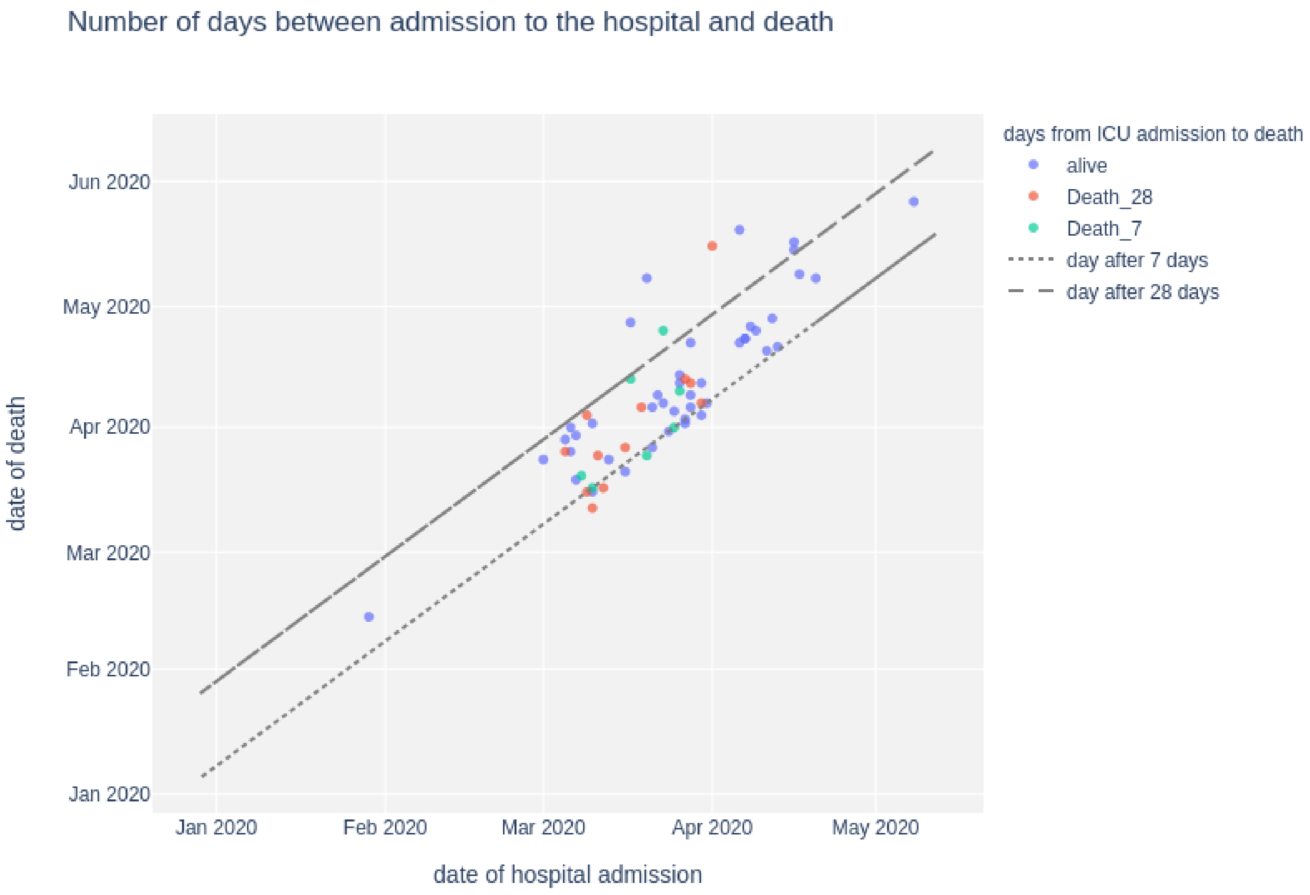

Figure 3: Incoherence of the French data

## French data

French dataset provided in supplementary information in the article by (Dupuis et al., 2021) assumed ICU discharge as target value - this does not mean death or survival. Patient is discharged from Intensive Care Unit when he or she is no longer in need of life support or no longer requires monitoring or treatment. This means that the model predicts whether a patient will be in need of intensive medical care or not. Dataset additionally had few interesting features - Death_D7, Death_D28, Death_date, which could be used to predict more than just "Survival/Death". Unfortunately this data contradicts itself - from 178 records, at least 43 are incorrect - some have Death_D7 marked as 1 and Death_D28 marked as 0, others have died according to death_date in less than 28 days from hospital_admission _date, yet have both Death_D7 and Death_D28 marked as 0. Moreover 42 have at least one Death column marked as 1 without death_date. Authors of article Dupuis et al. (2021) did not respond to our email inquires before the day of writing this report.

## French model

Despite the questionable reliability of the French dataset, we decided to test the Chinese model on the provided data. Theoretically, the dataset contained maximum of 5 measurements of the LD, CRP and leukocyte percentage for each patient, but there were numerous instances of missing observations. The missing data was therefore imputed with the column means, minimums and maximums. Then the model was trained using each imputation strategy. In the cases mentioned above, the precision of Chinese model was 0. There was no utility of the XGBoost model performed on this dataset, probably due to the size of the French dataset (only 173 records). We have also compared the performance of model trained on blood samples data with a model containing only age and gender of patients. The precision was 0.27 for the age & gender model and 0.38 for the blood model.

## Is making a model on joint data safe?

The significant asset of developed models is an ability to explain its behaviour. To ensure that our models work properly, we checked its fairness.

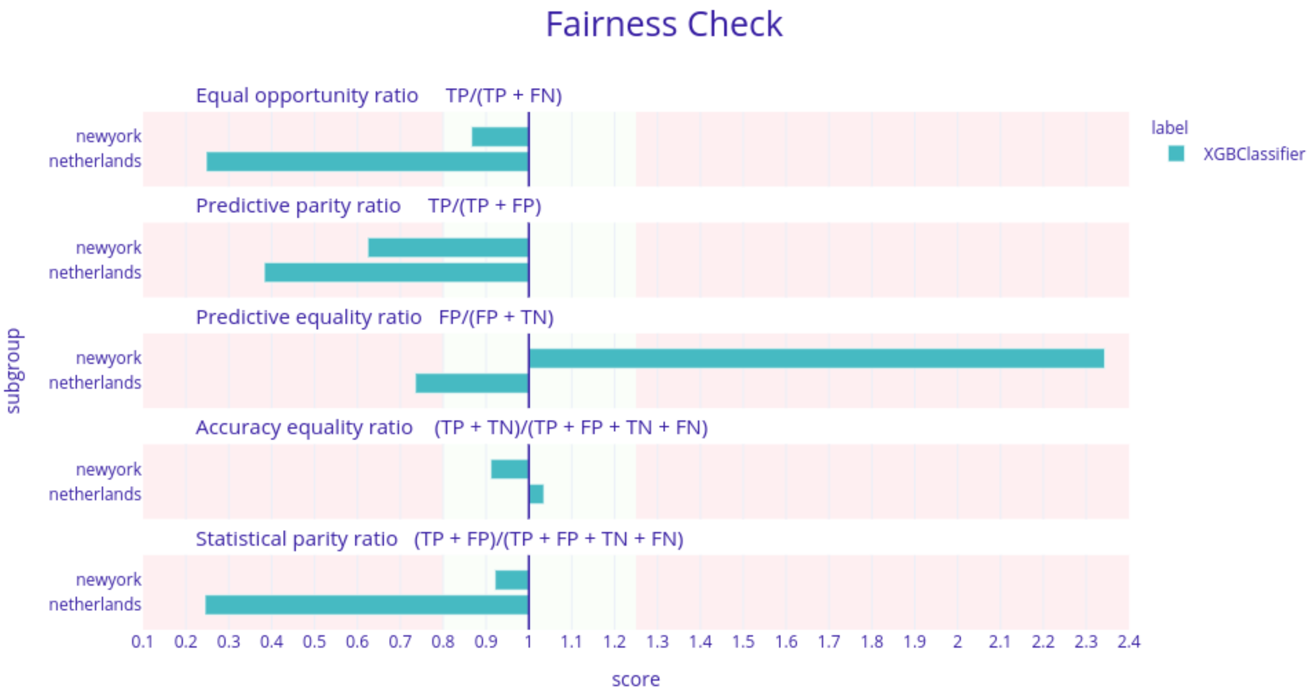

Figure 4: Fairness check plot where the privileged group is 'china'. The only metric where there is no issues with fairness is accuracy.

When exploring datasets and models we had a hypothesis that the models created on merged data from the three said sources will have a problem with discrimination. Briefly let's remind ourselves what a fair model is. The model is fair when it gives similar predictions and mispredictions to all groups of people. Formally the protected attribute (in our case the the origin of data) is independent of the prediction. It can be checked using a few fairness metrics. We used the model from the Chinese article that was previously mentioned. It turned out that our hypothesis was true. The model had severe issues in terms of fairness. Not only the Predictive parity ratio (Precision, PPV) was biased in favour of people from China but also the Equal opportunity ratio (TPR). Predictive equality ratio (FPR) was almost 2.5 times higher for the data from New York than from China. Statistical parity ratio (positive predictions) was really low in the Netherlands. It means that the model predicted far less deaths among the people from The Netherlands than from New York. Many factors might have contributed to this outcome. For example, biological differences between people, efficiency of the medical system, better and more experienced staff, or measurement devices. This analysis shows the dangers of merging the data or using models trained on data from many countries.

# Conclusion

In conclusion, our analysis supports hypothesis that Chinese model cannot be applied to people from rest of the world. However, the comparison of scores for records from different countries shows the vastly different behaviour of models. We do not have enough data to confirm if it is due to the nature of small datasets or differences in blood between people from different parts of the world. Additionally, our research shows that in some cases age and sex may contribute more to predictions than blood samples.

# References

Li Yan, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jing, Mingyang Zhang, Xiang Huang, Ying Xiao, Haosen Cao, Yanyan Chen, Tongxin Ren, Fang Wang, Yaru Xiao, Sufang Huang, Xi Tan, Niannian Huang, Bo Jiao, Cheng Cheng, Yong Zhang, Ailin Luo, Laurent Mombaerts, Junyang Jin, Zhiguo Cao, Shusheng Li, Hui Xu, and Ye Yuan. An interpretable mortality prediction model for covid-19 patients. *Nature Machine Intelligence*, 2(5):283–288, May 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-0180-7. URL https://doi.org/10.1038/s42256-020-0180-7.

C. Dupuis, E. De Montmollin, M. Neuville, B. Mourvillier, S. Ruckly, and J. F. Timsit. Limited applicability of a covid-19 specific mortality prediction rule to the intensive care setting. *Nature Machine Intelligence*, 3 (1):20–22, Jan 2021. ISSN 2522-5839. doi: 10.1038/s42256-020-00252-4. URL https://doi.org/10.1038/s42256-020-00252-4.

Marian J. R. Quanjel, Thijs C. van Holten, Pieternel C. Gunst-van der Vliet, Jette Wielaard, Bekir Karakaya, Maaike Söhne, Hazra S. Moeniralam, and Jan C. Grutters. Replication of a mortality prediction model in dutch patients with covid-19. *Nature Machine Intelligence*, 3(1):23–24, Jan 2021. ISSN 2522-5839. doi: 10.1038/s42256-020-00253-3. URL https://doi.org/10.1038/s42256-020-00253-3.

Matthew Barish, Siavash Bolourani, Lawrence F. Lau, Sareen Shah, and Theodoros P. Zanos. External validation demonstrates limited clinical utility of the interpretable mortality prediction model for patients with covid-19. *Nature Machine Intelligence*, 3(1):25–27, Jan 2021. ISSN 2522-5839. doi: 10.1038/s42256-020-00254-2. URL https://doi.org/10.1038/s42256-020-00254-2.

Hubert Baniecki, Wojciech Kretowicz, Piotr Piatyszek, Jakub Wisniewski, and Przemyslaw Biecek. dalex: Responsible machine learning with interactive explainability and fairness in python. 2020. URL https://arxiv.org/abs/2012.14406.