Methodological considerations

*The statistical methods suggested for the paper are not trivial. I will try to explain why I suggest using those methods. This is based on my initial proposal and with some additional modifications based on a literature review after Szilard’s suggestions.*

# Interpretable model

An interpretable model would be ideal. This would favor a model with as few predictors as possible, combined in a straightforward manner. A more advanced black-box method might yield better predictions however [1]. To motivate the use of such method in clinical practice would however be difficult.

A multivariable logistic regression model might seem as easily interpretable, but it is not. Interdependencies among covariates yields non-trivial relations between predictors and outcome (not to be confused by a linear model) [2]. This is pronounced further if some predictors are inherently dependent.

The opposite attempt could be for example a model with complex variable transformations of principal components. This might be a good prediction model with nice theoretical properties, but it will confiscate the original variables and their scales. The model would be impossible to interpret.

There might be no such thing as a “semi-interpretable method”. Moreover, if there is, it might be even worse than a non-interpretable method, since it might further invite to misleading interpretations. Nevertheless, it seems like some methods are described as such in the applied literature. A least absolute shrinkage and selection operator (Lasso) has been used at least twice recently to increase “transparency” for prediction of 30-day mortality after total hip and knee arthroplasties [3,4]. Also, a hierarchical Tree-Lasso logistic regression has been explicitly presented as an “interpretable predictive model” in the field of pediatrics [5].

# Lasso regression

We must note that logistic Lasso regression has the same flaws as ordinary logistic regression considering interpretability (above). The difference lies in the use of a regularizing penalty term when estimating the regression coefficients. An -norm, multiplied by an arbitrary constant λ (usually found by some cross-validated optimization), is used to force some coefficients to zero, keeping only a subset of the initial predictors.

This is most relevant for settings with a vast number of candidate predictors. To find a minimal subset of predictors is however one of the explicit goals of the project, wherefore this type of variable selection might still be relevant.

Hence, we might argue that the lasso regression model might not be more interpretable than any other opaque prediction method. It seems however that it is considered as such.

# Ensemble Lasso

Variable selection by Lasso is known to include potential predictors that are false positives (that should not be included in the final model). The method is also associated with some bias. A robust alternative is ensembles of multiple Lassos performed to different resamples of the original data set. The Bolasso method, suggested in 2008 combines bootstrap replicates and Lasso to a consistent estimator [6]. Consistency in variable selection was also proven for “stability selection” more broadly in 2010 [7]. Then in 2015, some additional improvements were made to the method, suggesting the “bootstrap ranking procedure”, applying Lasso to bootstrap samples using variable importance for variable ranking and variable selection [8].

# Class imbalance

We have a highly imbalanced data set with very few deaths compared to survivors. This makes ordinary methods such as logistic regression (and therefore Lasso) unfeasible.

There are some strategies to this problem:

1. **Ignore** it and use standard methods anyway. This is easy and not necessary as bad as it sounds, since no method is perfect.
2. **Cost-sensitive learning**, assuming there is a known cost imbalance difference between misclassification of either deaths as survivors and survivors as deaths. Those costs are rarely known in practice however. We will therefore not consider this method further.
3. **Data preprocessing re-sampling approaches**. Bypassing the imbalance by using an artificial resample of the original data. This strategy seems to be the most common.

## Re-sampling approaches

Re-sampling is made by up sampling (of deaths), down sampling (of survivors), a combination of the two, or some more advanced sampling technique [9].

Up sampling does have its merits but computational requirements would in our case be infeasible due to the combination of bootstrap resampling. A review of different methods have shown that down sampling is in fact preferred due to simplicity, low risk of over-fitting and superior result when evaluating predictive power on an independent test data set [10].

Down sampling with ensemble Lasso (stability selection) has been used to discover adverse drug reactions by sampling four times as many cases than controls [11].

A modified down sampling procedure with pre-clustered data and balanced sampling from each cluster has also been used to predict credit scores [12]. The method slightly outperformed random forest and some other traditional methods if including well-designed prediction features as independent variables. The result was however inferior before variable transformation and binning. We will therefore not consider this method further.

## Biased probability estimates

Re-sampling methods are commonly used in classification (outcome labelling as either dead or alive). We aim for regression however. We model probabilities of death, not simply a statement of death or not.

Re-sampling changes the underlying distribution of the data. Thus, the scale of estimated probabilities will be different. This might go unnoticed if models are chosen simply based on predictive performance. A high AUC value indicates that for one survivor and one dead chosen randomly, the dead had a higher predicted probability to die. This relation is maintained through re-sampling since the rank order of estimated probabilities goes unchanged. The estimated class probabilities on the other hand is not unchanged.

We therefore need to apply a simple bias-correction for the estimated probabilities based on the re-sampled data. Such correction was suggested by Pozollo et al. in [13]:

where and are the estimated prior and posterior probabilities after adjustment, and where β is the proportion of positive to negative outcomes in the original data set.

# Bibliography

[1] Breiman L. Statistical modeling: the two cultures. Stat Sci 2001;16:199–215. doi:10.1214/ss/1009213726.

[2] Westreich D, Greenland S. The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. Am J Epidemiol 2013;177:292–8. doi:10.1093/aje/kws412.

[3] Harris AH, Kuo AC, Bowe T, Gupta S, Nordin D, Giori NJ. Prediction Models for 30-Day Mortality and Complications After Total Knee and Hip Arthroplasties for Veteran Health Administration Patients With Osteoarthritis. J Arthroplasty 2018;33:1539–45. doi:10.1016/J.ARTH.2017.12.003.

[4] Harris AHS, Kuo AC, Weng Y, Trickey AW, Bowe T, Giori NJ. Can Machine Learning Methods Produce Accurate and Easy-to-use Prediction Models of 30-day Complications and Mortality After Knee or Hip Arthroplasty? Clin Orthop Relat Res 2019;477:452–60. doi:10.1097/CORR.0000000000000601.

[5] Jovanovic M, Radovanovic S, Vukicevic M, Van Poucke S, Delibasic B. Building interpretable predictive models for pediatric hospital readmission using Tree-Lasso logistic regression. Artif Intell Med 2016;72:12–21. doi:10.1016/J.ARTMED.2016.07.003.

[6] Bach FR, R. F. Bolasso. Proc. 25th Int. Conf. Mach. Learn. - ICML ’08, New York, New York, USA: ACM Press; 2008, p. 33–40. doi:10.1145/1390156.1390161.

[7] Meinshausen N, Bühlmann P. Stability selection. J R Stat Soc Ser B (Statistical Methodol 2010;72:417–73. doi:10.1111/J.1467-9868.2010.00740.X@10.1111/(ISSN)1467-9868.TOP\_SERIES\_B\_RESEARCH.

[8] Guo P, Zeng F, Hu X, Zhang D, Zhu S, Deng Y, et al. Improved Variable Selection Algorithm Using a LASSO-Type Penalty, with an Application to Assessing Hepatitis B Infection Relevant Factors in Community Residents. PLoS One 2015;10:e0134151. doi:10.1371/journal.pone.0134151.

[9] Lunardon N, Menardi G, Torelli N. ROSE: A Package for Binary Imbalanced Learning. R J 2014;6.

[10] Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Trans Syst Man, Cybern Part C (Applications Rev 2012;42:463–84. doi:10.1109/TSMCC.2011.2161285.

[11] Ahmed I, Pariente A, Tubert-Bitter P. Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions. Stat Methods Med Res 2018;27:785–97. doi:10.1177/0962280216643116.

[12] Wang H, Xu Q, Zhou L. Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble. PLoS One 2015;10:e0117844. doi:10.1371/journal.pone.0117844.

[13] Dal Pozzolo A, Caelen O, Bontempi G, Johnson RA. Calibrating Probability with Undersampling for Unbalanced Classification Fraud detection View project Volatility forecasting View project Calibrating Probability with Undersampling for Unbalanced Classification n.d. doi:10.1109/SSCI.2015.33.