TECHNICAL UNIVERSITY OF MUNICH

TUM Data Innovation Lab

A robust comparison of causal effects from observational data in healthcare

Authors	Thomas Alexander Christoph, Onur Galoglu, Jiaqi Lu, Sarah Lumpp, Daniel Sturm
Mentor(s)	Dr. Narges Ahmidi (Helmholtz Munich Center), Dr. Niki Kilbertus (HelmholtzAI)
Co-Mentor	M.Sc. Konstantin Goebler (Chair of Mathematical Statistics), M.Sc. David Strieder (Chair of Mathematical Statistics)
Project Lead	Dr. Ricardo Acevedo Cabra (Department of Mathematics)
Supervisor	Prof. Dr. Massimo Fornasier (Department of Mathematics)

Jul 2021

Abstract

The implementation of stents to widen blocked coronary arteries can prevent a patient from developing a heart attack and thus save lives. However, like all other medical procedures, this operation is not without risk for the patient. Possible complications include myocardial infarction, cardiac death, and eventual revascularization. Target Lesion Failure (TLF) is a generic term for adverse events and is commonly used as indicator of long-term success for the procedure. In 2016, the German company BIOTRONIK introduced Magmaris, its newest-generation fully bioresorbable stent. Since then, BIOTRONIK is conducting an extensive study with the aim to increase the product's security, and maximize the chance of recovery for every single patient. The study monitors 2,066 patients who received Magmaris from before the procedure until several years after. In this project, two main research questions were considered.

(1) *Given patient data available prior to stent implantation, is it possible to distinguish between the sub-cohorts of patients who develop TLF and those who don't?* If so, this would allow for a more precise benefit-risk estimation and better monitoring for high-risk patients. One approach to this question is to model it as a *Binary Classification* problem. Given a patients data, we try to predict if TLF will happen or not. To solve this, different methods are applied, namely Logistic Regression, Random Forests, Neural Networks, K-Nearest Neighbours and XGboost. Ultimately, none of those achieve desired accuracy. However, a majority of these methods is able to correctly classify the same subset of patients. This suggests that patients can be clustered into two groups: Those for whom TLF can be reliably predicted and those for whom it is difficult. It might be valuable to further investigate this clustering. Another approach is *Survival Analysis*. There are three primary goals of survival analysis: to estimate and interpret hazard functions from the data, to compare survival functions, and to assess the relationship of explanatory variables to survival time. Survival analysis provides a great tool for analyzing the time to an event type of data, which is very common in any clinical trial. The first two tasks are done by Kaplan-Meier estimation and Log-rank test, and we use Cox Proportional Hazards Regression to solve the third task. We select 17 features, of which the survival curve is significantly different among subgroups. Furthermore, we obtain a Cox model that can predict time-to-TLF using 7 features, and achieve an accuracy of 68% in the sense of correct ordering.

(2) *Can causal relationships between procedure parameters and the occurrence of TLF be identified?* If such causal relationships are known, the operating physician can actively intervene on the procedure parameters to achieve a more ideal outcome. In order to quantify such causal relationships, the *Causal Inference* framework is used. To assess the average causal effect of eight treatment parameter ratios, an assumed causal structure of the data generating process based on expert knowledge is defined using a causal graphical model. By applying modified methods from the python framework Ananke-causal, the presumed causal effects are identified and estimated. We consider 95% percentile bootstrap confidence intervals. The estimated effects are in general too close to zero to indicate significance. Therefore, we cannot deduce recommendations for specific treatment ratios that might improve the TLF outcome. However, three of the ratios exhibit a slight shift of the causal effect away from zero. This might indicate some significance of the related treatment decisions for the TLF outcome and could be further investigated.

Nomenclature

ACE	Average Causal Effect
ADMG	Acyclic Directed Mixed Graph
AE	Adverse Event
BMS	Bare Metal Stents
BRS	Bioresorbable Stents
CABG	Coronary Artery Bypass Grafting
CAD	Coronary Artery Disease
CD	Cardiac Death
CK-MB	Creatine Kinase Myocardial Band
CVD	Cardiovascular Disease
DAG	Directed Acyclic Graph
DES	Drug-eluting Stents
GLM	Generalized Linear Model
IPW	Inverse Probability Weighting
ML	Machine Learning
PCI	Percutaneous Coronary Intervention
TLF	Target Lesion Failure
TLR	Target Lesion Revascularization
TV-MI	Target Vessel myocardial infarction

Contents

1 Introduction

Worldwide, cardiovascular diseases amount to up to 32 percent of all deaths. This number is estimated to grow even larger in the following decades [11]. The largest subgroup of CVD are coronary arterial diseases (CAD), where the heart's major arteries become damaged or diseased. In most cases plaque gradually accumulates in those vessels, eventually limiting the blood flow through the heart. This is noticeable through e.g. chest pain or shortness of breath. In the most extreme cases blood begins to clot and the vessel gets completely blocked, causing a heart attack.

The two major approaches in CAD treatment are percutaneous coronary intervention and coronary artery bypass grafting. CABG entails open-heart surgery, and is not the interest of our project.

In PCI a stent, a small tube made of mesh, gets inserted into the affected vessel to widen it and restore the natural blood flow. The whole procedure, depicted in figure 1, can be broken down to the following steps. First, a thin guide wire gets inserted into the artery. Using this wire the physician delivers an inflatable balloon surrounded by the non expanded stent to the damaged vessel region. Under high pressure the balloon is then inflated, causing stent and vessel to expand. The stent now fits tightly against the widened vessel wall and supports the vessel, preventing it from contracting. Subsequently, balloon and guide wire are removed from the artery while the stent remains.

Different types of coronary stents are available. First-generation, bare-metal stents (BMS) only consist of metal mesh. Drug-eluting stents (DES) are additionally coated with drugs, inhibiting tissue growth and reducing the risk of restenosis. Although DES were a huge improvement over BMS, they still pose a high risk for late stent thrombosis. This led to the development of fully bioresorbable stents (BRS). After temporarily supporting the vessel wall, BRS will eventually be resorbed when no longer needed. This leaves no triggers for late stent thrombosis while additionally not limiting future treatment options [10].

Magmaris is a sirolimus-eluting bioresorbable magnesium scaffold developed by the German company Biotronik. In two premarket studies BIOSOLVE-II and -III with 184 enrolled patients, Magmaris showed good behavior and was CE-certified in June 2016.[19]

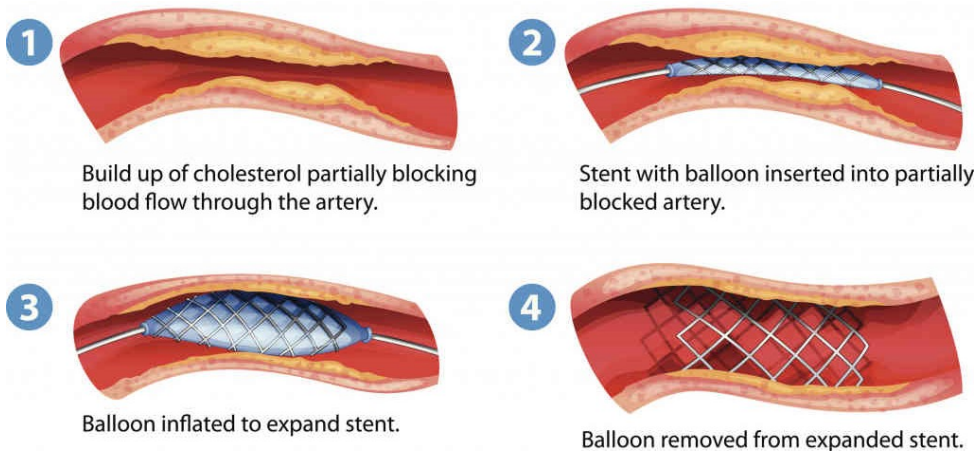

Figure 1: Stent implantation procedure (PCI) [12]

After commercial launch, to "assess the safety and performance of the Magmaris sirolimus-eluting bioresorbable magnesium scaffold in a large patient population"[19], Biotronik initiated the BIOSOLVE-IV study. The study accompanies patients who receive Magmaris implants starting before the procedure until five years later. In total 2,066 patients from 23 countries were enrolled. At time of entry patients need to fulfill strict inclusion and exclusion criteria. Among others, a maximum of two single de novo lesions in two different large epicardial vessels is required.

The study's main objective is to track target lesion failure (TLF), which consists of "cardiac death, target-vessel myocardial infarction, coronary artery bypass grafting, and clinically driven target lesion revascularization (TLR)"[19].

The goals of this project were

1. to identify patients likely to experience TLF after receiving Magmaris. This is done using data from BIOSOLVE-IV up until procedure and applying classic machine learning approaches.The incentive is to allow for more precise, patient tailored, benefit risk estimations prior to implanting Magmaris;

2. to perform time-to-event analyses in order to estimate the risk of getting TLF at different time points within the 5 years of study, based on information from screening and procedure;

3. to measure the impact of ratios of procedure parameters like lesion length to stent length on TLF, by using methods from causal inference, as the available BIOSOLVE-IV data is non-randomized.

In general, our main motivation is to aid in increasing the performance of Magmaris by efficiently choosing procedure parameters.

2 Data

2.1 Study Structure

Data collection for BIOSOLVE-IV patients starts with study enrolment and continues up to five years after Magmaris implantation. Figure 2 depicts the principal steps of the study and essential features recorded.

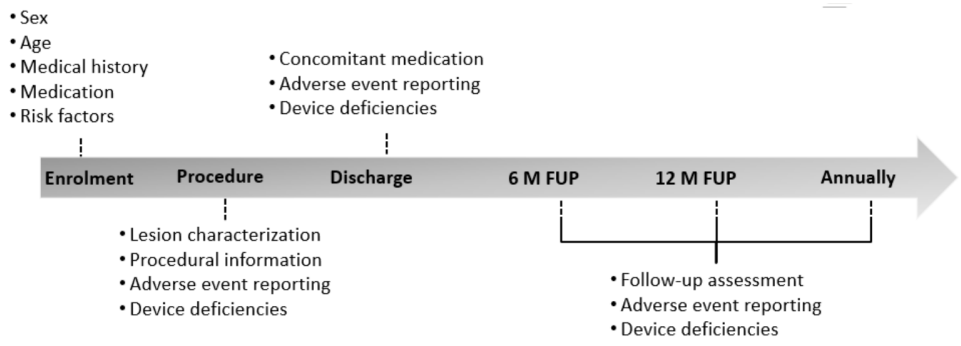

Figure 2: Study design of BIOSOLVE-IV [12]

After enrolment, baseline clinical data is collected for each patient. This includes demographic features, inclusion/exclusion criteria and medical history as well as known risk factors. Additionally, the patients answer the EQ5d questionnaire for assessing their quality of life. Prior to PCI a 12-lead ECG is performed, the patient's ischemic status is assessed, and cardiac enzymes are measured.

Prior to PCI, the physician examines the lesion. Here, location of the damaged vessel and main lesion characteristics like length, angulation or calcification are recorded. Additionally, the stenosis pre-procedure is measured in percent.

During PCI, all device related features are documented. For balloons this includes for example the type of balloon as well as length, diameter pre- and post-dilatation, or the maximal applied pressure. For stents, among others, length and diameter are reported. Additionally, non-device related features as medication before and during surgery are documented. After PCI, the vessel's stenosis is measured again and it is reported whether any adverse events happened.

To judge the patient's health status before hospital discharge, the EQ5d questionnaire is filled again. Also, a 12-lead ECG is performed, and the patient's ischemic status and cardiac enzymes are measured. Furthermore, it is reported whether any adverse events happened since the procedure.

After hospital discharge each patient remains in the study for five years. During this time there are regular follow-up appointments, namely after six months, twelve months and thereafter annually. At these follow-ups, it is checked whether any adverse events occurred since the last appointment. Additionally, the EQ5D questionnaire is filled, and the patient's ischemic status is assessed.

The final dataset is divided into more than forty different subsets, each corresponding to a specific study segment. In total, it contains more than 500 features per patient. At the time of this project all 2,066 patients had been enrolled. However, hardly any patient had completed the study. Thus, not all features are available for all patients. Additionally, there is occasional missingness. That is, sometimes features are not filled for no obvious reason.

We infer the TLF label from adverse events recorded during and after PCI. It is important to note that for a patient still enrolled in the study, the only statement that can be made is whether he has or has not *yet* suffered TLF.

2.2 Preprocessing

The used preprocessing pipeline consists of seven different steps.

It starts by loading all available data, given in more than 40 different subsets, provided in *.csv format.

The remaining steps are split up into the following.

2.2.1 Feature Selection

Closely following previous work [12], different features, available until the state of Hospital Discharge, are used and taken from the respective original *.csv files. Different sets of relevant features can be used to be plugged into the pipeline. The here used feature set

that lead to the gained results can be found in Appendix A. This includes overall 72 different raw features.

2.2.2 Nested Questions Handling

Much of the data collected is not independent. This leads to inconsistencies that have to be handled during preprocessing. Most commonly, features are missing conditioned on the value of a previous feature. The reason for this are nested questions. Figure 3 exemplary shows features collected when measuring creatine kinase myocardial band (CK-MB) prior to the stent implantation and when administering heparin. In both cases, if the first question is answered with "no", none of the following features are recorded. Although this makes sense, missing feature values pose a problem for most machine learning approaches. Because of this, missing features are filled as well as possible. For the administration of heparin, there exists an obvious solution: If no heparin was given, one can assume a value of zero of arbitrary unit. However, if no CK-MB measurement was done, no such thing can be done. Either patients with missing entries have be dropped or, if possible, the features have to be one-hot encoded as described in 2.2.6.

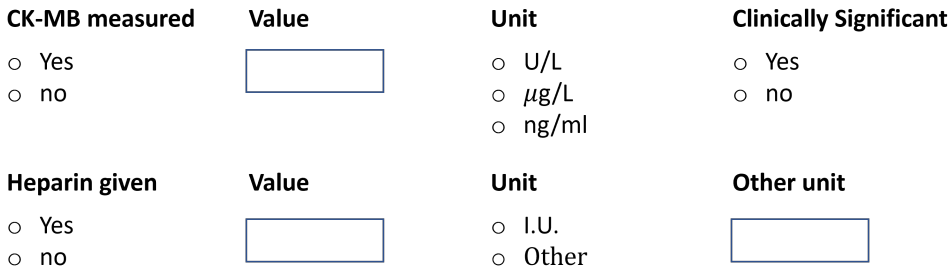

Figure 3: Features for CK-MB and Heparin in BIOSOLVE-IV

2.2.3 Unit Conversion

Another type of inter-feature dependency appears in form of measurements. Commonly, quantities, for example the dose of administered medicine, are recorded in different units. Therefore, it is necessary to convert the values of those quantities according to the used unit to a baseline unit. If this is not possible, e.g. the given unit cannot be converted into the base unit, the respective value gets replaced with *NaN*.

2.2.4 Multiple Patient Occurrences

Occasionally, features are filled several times for the same patient. An example for this is a patient having multiple lesions and thereafter getting multiple stents. All lesion- and stent- related features are not unique for this patient. To get a resulting homogeneous dataset, those multiple occurrences have to be tackled. Different approaches are implemented and used for this.

1. *Keep first*: Keeping the first sample for duplicate features and dropping all further appearances and information.

2. *Keep least NaN*: Some samples per patient are only sparsely filled. For every patient, this approach keeps the sample with the least amount of *NaN*s and excludes all other appearances. A major drawback of this approach, as well as of the above mentioned, is a loss of data. If underlying causes for developing TLF are located in these multiple occurrences, dropping them deletes this information and the hidden structures may not be found.

3. *Widen dataset*: To prevent dropping any information, this approach aims to keep all samples of a patient. Therefore, all further occurring samples are taken and their information get appended to the first one. A major drawback is that this approach leads to massive - but imbalanced - increase of features for some patients. All patients with less samples gain a huge amount of new *NaN*s. Even though this approach is implemented in the pipeline, because of its major drawback, it was not used.

A further approach, to keep the information of multiple samples per patient without increasing the amount of *NaN*s could be a mixture: Mixing/Aggregating the information of all occurring samples and keeping the mean, sum or similar as resulting information for that feature.

2.2.5 Label Generation

In order to use later in binary classification, survival analysis, and causal inference, labels have to be derived manually. Two different labels were derived, namely TLF label and TimeToTLF, to be later used for the aforementioned methods respectively.

For the binary TLF label, four different subcategories under the adverse events are considered: CD, TV-MI, TLR and emergent CABG. After different lists of patients belonging to each of the TLF cases were created, those lists are merged into one big, single list. In order to have a set of unique patients, the duplicates were removed from the list. Lastly, the patients in the respective set were assigned the label 1, which indicates that they developed TLF, while all others were assigned 0.

In order to derive the TimeToTLF label, an algorithm was needed to overcome the right-censored structured of the data: There were patients who already developed TLF, some terminating the study early and some that haven't yet developed TLF (while it is not known how their health status will be at a later point). Besides, the enrolment dates of the patients were also different from each other. Therefore, the following logic was implemented: If a patient already developed TLF, then the time interval between the procedure time and TLF event time was assigned as TimeToTLF label of the respective patients. If a patient terminated his/her study early (which might include also death), then the time interval between the procedure time and early termination time (or time of death) was assigned as TimeToTLF label of the respective patients. Lastly, if none of the aforementioned cases were valid, then the time of the whole study (1825 days) is assigned for the respective patients. All values of the TimeToTLF features are in terms of days.

2.2.6 Feature Encoding

Many machine learning methods, as well as causal inference and survival analysis, require input data to be numeric. Therefore, before applying these methods, features have to be

encoded by conversion into numeric values.

As a first step, each feature is investigated and classified manually in terms of the following scheme:

- Numeric: Features that are already numeric. Here no further transformation is made.
- One-hot: Features, containing nominal values, are one-hot encoded. This includes for example the feature which contains the diseased vessel's location. Besides, one-hot encoding is applied to features with high missingness in order to get rid of NaN values.
- Categorical: Features with values that can be mapped to integers as their values are ordinal. One example for such a feature is the calcification of the target vessel, which takes the values "None", "Mild", "Moderate" and "Severe". This class also includes binary features that have values of either "yes" or "no".

After manually classifying the features, conversion is done according to the previous scheme. Using the presented encoding techniques causes the number of features to increase up to 111.

2.3 Statistical Analysis

Before implementing an approach to learn the data, either for binary classification or survival analysis, it is crucial to gain substantive insight into patients who develop TLF and into those who do not. Therefore, we provide some descriptive statistics and visualizations before engaging into prediction tasks. Unsurprisingly, only few patients develop TLF, resulting in a highly unbalanced dataset - as is very common in medical diagnostic data. Here, approximately 9% of the patients experience TLF, while approximately 91% do not suffer from this.

To start off, Figure 4 demonstrates marginal densities of chosen continuous features, from those chosen in 2.2.1, separated by patients who developed TLF and those who did not. Even though both subcohorts are somehow of different shape for all shown features, they overlap massively. Both mean and standard deviation tend to be similar for both groups of patients. Table 1 provides the numerical analysis of mean and standard deviation of the chosen features. The distributions of other features are very similar to the shown ones.

Another approach to test significant differences between the subcohorts of TLF and Non-TLF patients is the calculation of the p-value. In order to see whether the differences in the means demonstrated in Table 1 are in fact negligible we employ hypothesis tests. We set the significance level $\alpha = 0.05$ and reject the null-hypothesis for p-value smaller or equal to α. To implement a suitable test to calculate the associated pairwise p-value for each feature, one has to distinguish between numerical/continuous data and categorical/binary data. Point-biserial correlation between the TLF outcome and the respective feature is calculated for continuous data where the null-hypothesis assumes no correlation. Chi-squared test - null-hypothesis assuming independence of both variables - is used for categorical and binary features. Before performing the chi-squared test, the data gets

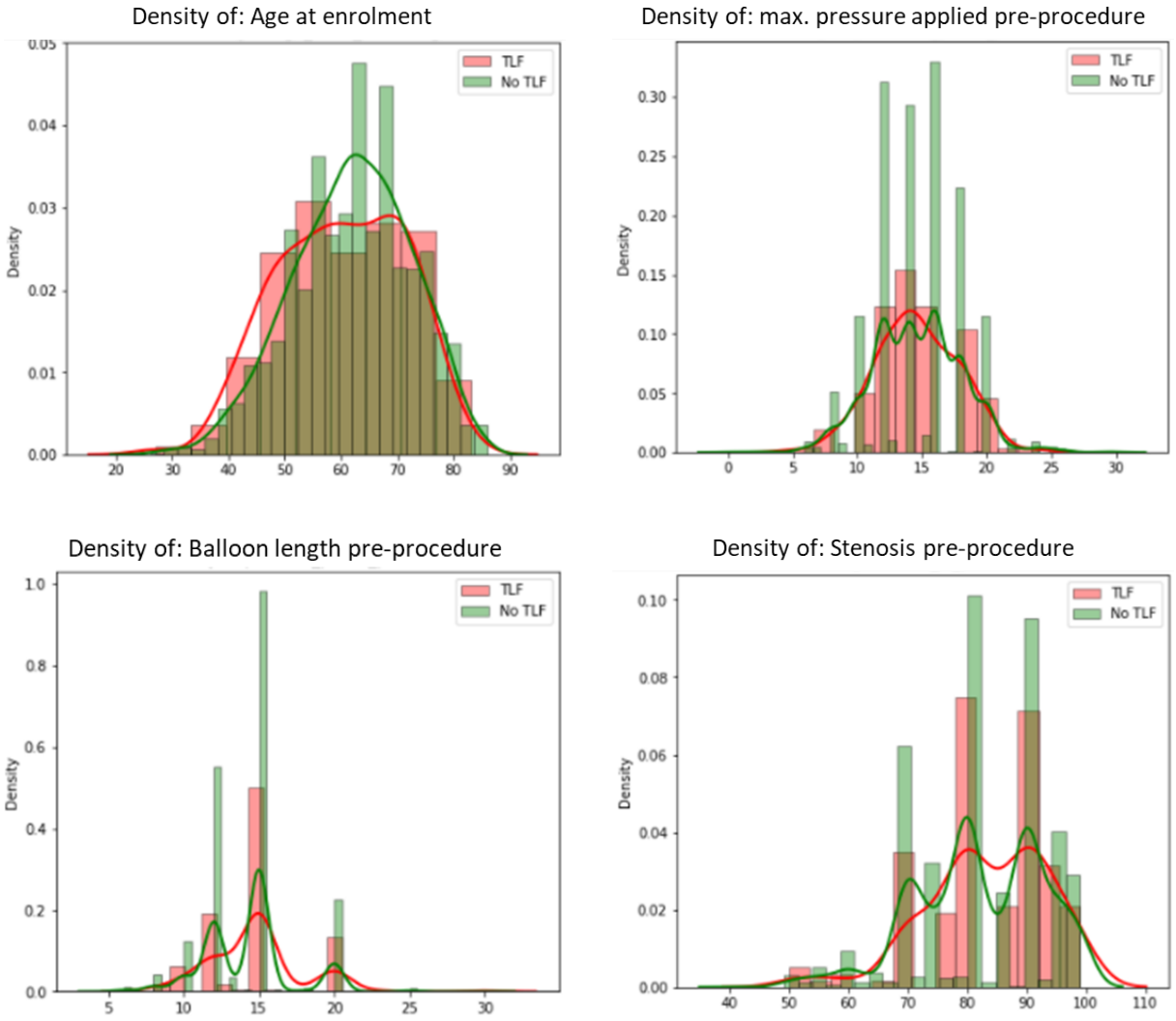

Figure 4: Density of continuous data for TLF patients and Non-TLF patients

splitted into both cohorts to calculate the p-value between the contingency tables of both groups. The resulting values claim 13 of the used features to be significantly different for both patient cohorts. Although this cannot be confirmed visually, because 7 of those 13 are non-numeric, the results are promising for later binary classification.

3 Machine Learning

3.1 Binary Classification

To gain an overview about the suitability of the chosen variables to express the huge amount of available information, binary classification was introduced. From clinical perspective, it is of special interest if doctors would have an idea who is likely to develop TLF. The properties of the given dataset suggest two different proceedings to construct a binary classification model: (1) Assuming the available time information not to be relevant for a

Feature	TLF patients	Non-TLF patients
Age at enrolment	60.08 ± 11.08	62.02 ± 10.48
Stenosis pre-procedure	14.76 ± 3.13	14.25 ± 2.94
Max. Pressure applied pre-procedure	14.62 ± 3.21	14.62 ± 3.44
Balloon length pre-procedure	1.74 ± 1.02	1.88 ± 1.44

Table 1: Mean and standard deviation of chosen features for TLF cases compared to Non-TLF cases

valid prediction. As a result bagging all chosen features together to afterwards fit a corresponding binary classifier; (2) On the other hand, additionally taking time information into account to implement a time series classification algorithm.

3.1.1 Bag Of Features Approach

To bag all features for binary classification, all time information was discarded and the dataset was treated as if it had been recorded at the same step in time for all features. Widely-used different Machine Learning models were implemented independently. Overall, five different models were used, being K-Nearest Neighbors, Random Forest, Logistic Regression, XGBoost and a Neural Network consisting of several Dense Layers.

The **Neural Network** consists of four hidden (dense-) layers with decreasing number of neurons per layer. To prevent the network from overfitting, numerous regularization techniques between the individual layers [20] are used. Starting with Gaussian noise on the input, each layer has $L2$ activation regularization, batch normalization and dropout. To evaluate the performance of the model, cross validation with five folds is used, making sure that every patient is exactly once in the test dataset for the model prediction. Before performing a train-test split, patients and features with more than 20% missingness are excluded. Thereafter, 13 patients are not included in the resulting dataset. All further remaining *NaN's* are being ignored by the network due to a masking layer. Then, 80% of the dataset is used in each fold to train the network while 10% is used for validation and 10% for testing. To ensure that every patient is exactly once in the test dataset, nested cross-validation is used, resulting in ten different models. Each two of the networks share the same training dataset with switched validation and test data. In the end, each of the ten networks performs a prediction on the respective, unique test dataset. The results are stacked together and result in a confusion matrix containing a prediction for every considered patient.

Next to this, the **classical Machine Learning** techniques are implemented. The main difference to the neural network is that a Random Search is performed for each of these algorithms before evaluating. The used parameter ranges were kept small around the respective default value when performing the Random Search. A five fold cross validation is done as well, resulting in the same kind of stacked confusion matrix as for the Neural Network. In contrast to the Deep Learning approach, different techniques for *NaN* imputation are applied. This includes KNN-Imputer [18] for numerical data, Simple-Imputer, using the most frequent value, for categorical and binary data, and the

oversampling technique *SMOTE* [3] for tackling the class imbalance. For categorical and binary data, the Simple-Imputer is necessary to make sure to only impute values that are already given in the data, avoiding to impute floating points or artificial data that implies wrong information. KNN-imputation for numerical data was used, assuming more realistic imputations. Instead of imputing a static value for all *NaN's* of a given feature as the Simple-Imputer does, KNN-imputation is dynamic with respect to the nearest appearing values.

Performing this, mainly two findings were observed:

1. The dataset as a whole is hard to learn for all kinds of implemented classification algorithms. Especially the Neural Network performs very poorly: While the (balanced) accuracy on the train dataset is almost 100% for all folds, accuracy drops down to 50% during test time.

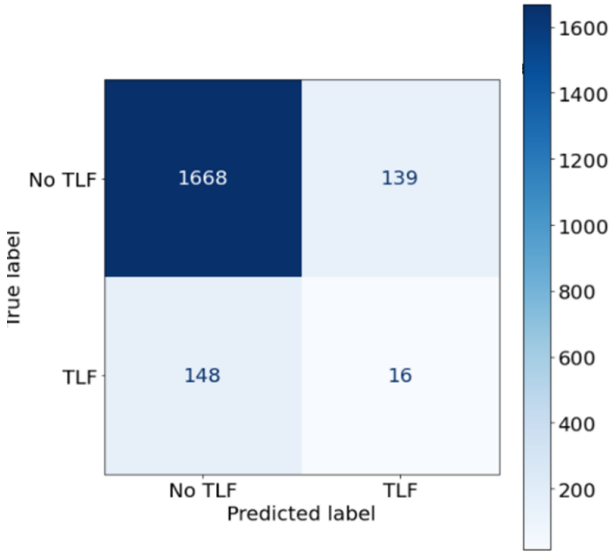

Figure 5: Result of evaluated Neural Network - stacked Confusion Matrix

Figure 5 shows that the Neural Network is randomly guessing the label during prediction. The proportion of predicted TLF cases is almost the same as in the original dataset. With approximately 9%, the proportion of predicted TLF cases is almost the same as in the original dataset. Even high regularization cannot prevent a rather low-dimensional model, approximately 8300 times more training points than trainable parameters, from overfitting.

Similar to the Neural Network, the classical Machine Learning algorithms perform poorly on the whole dataset. All of them achieve only low (balanced) accuracies, approximately 50% (see Figure 6). However, most of them mainly classify all patients as TLF, in contrast to the Neural Network that classifies most patients as Non-TLF. Overall, all models suffer from high generalization gaps.

2. Using the classical Machine Learning algorithms by majority vote, it is possible to find certain subcohorts in the dataset, referred to as *easy* and *hard* cases for both classes - TLF and Non-TLF. Patients are denoted as *easy*, if at least three of the four models (excluding the Neural Network because of random guessing behaviour) **correctly** classify

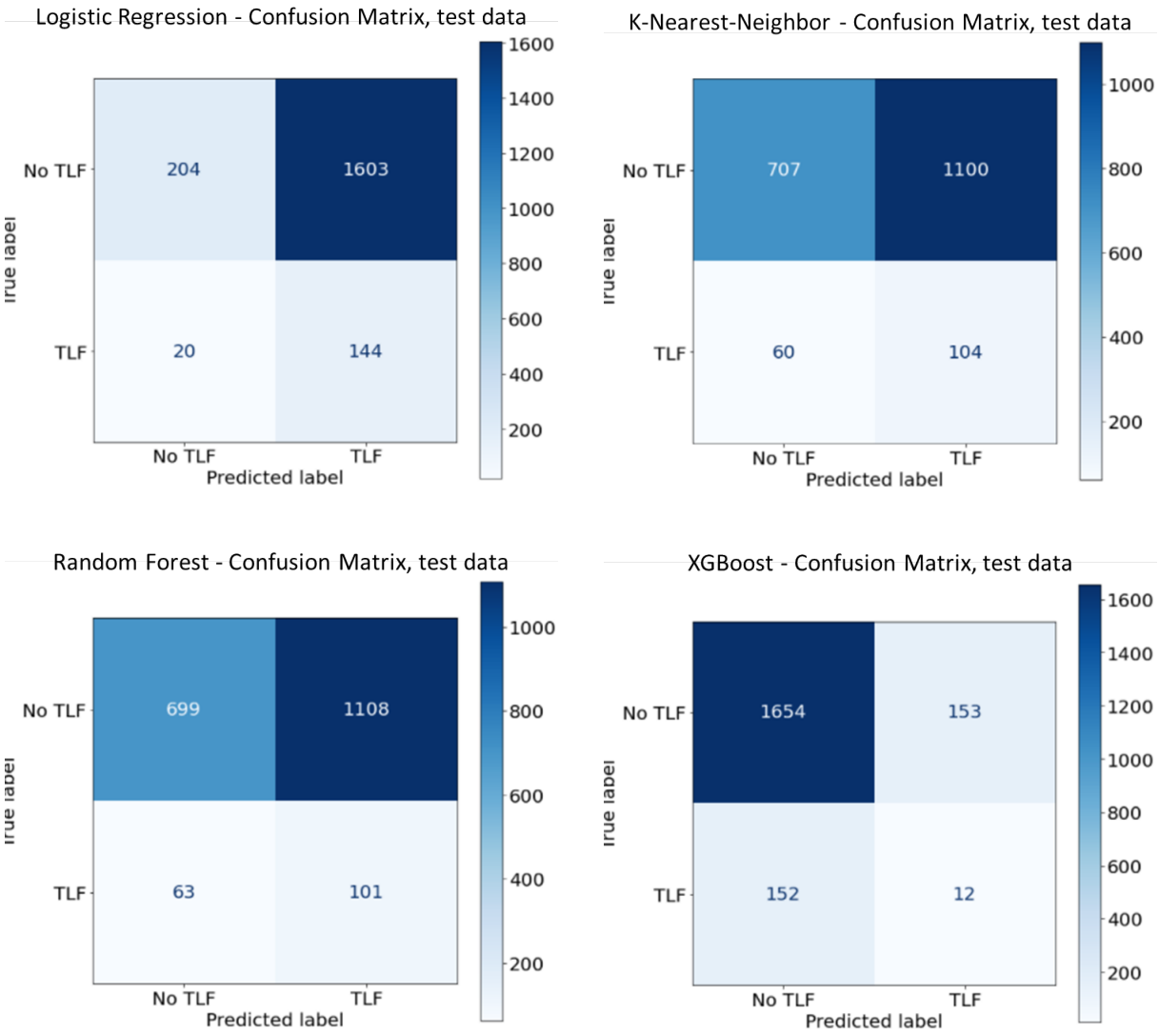

Figure 6: Stacked Confusion Matrices of classic Machine Learning approaches

them as TLF/Non-TLF. By applying this method, we can separate both classes each into two subcohorts - 72 *easy*, 92 *hard* TLF cases and 381 *easy*, 1426 *hard* Non-TLF cases. Further investigations were done to evaluate why most of the models are able to correctly classify those *easy* cases while failing on the *hard* ones. SWe are especially interested in the comparison between *easy* TLF cases, *hard* TLF cases and all Non-TLF cases. To evaluate this, all features were compared for *easy* TLF and Non-TLF patients using *independent two-sample two-sided t-test with assuming non-equal variance* for numerical features and *chi-square test of independence of variables in a contingency table* for categorical or binary features[1]. Deriving a significant difference between both subcohorts for a p-value ≤ 0.05, leads to a significant difference for the following features as summarized in Table 2.

Whereas the feature distributions of *easy* TLF cases mostly differ visibly from Non-

[1]Chi-square test was only performed if a binary feature contains at least five positive samples. Otherwise, this features does not contain enough information to assume significance.

Feature	**P Value**	***easy* TLFs 72 patients**	***hard* TLFs 92 patients**	**Non-TLFs 1807 patients**
Anti-platelet medication I loading dose	≤ 0.0001	93.06%	58.70%	63.59%
Age at time of enrolment	0.00139	58.17 ± 9.57	62.00 ± 11.92	61.99 ± 10.51
Lesion length	0.00569	16.12 ± 4.00	14.90 ± 5.54	14.76 ± 3.89
Stenosis pre-procedure	0.00311	85.14 ± 8.17	81.92 ± 11.57	82.11 ± 10.67
Magmaris scaffold length	0.00613	20.69 ± 3.78	19.57 ± 4.04	19.41 ± 3.87
History of previous myocardial infarction	0.01032	8.33%	28.26%	21.64%
Number of previous interventions	≤ 0.0001	0.18 ± 0.51	0.72 ± 1.40	0.48 ± 1.04
Gender = Male	0.00608	88.89%	81.52%	73.77%
Vessel = LAD	0.00090	72.22%	53.26%	48.42%
Lesion location = Prox LAD	≤ 0.0001	41.67%	28.26%	19.92%

Table 2: Distributions of *easy* and *hard* TLF cases compared to all Non-TLF cases for all features with significant difference between *easy* TLF and Non-TLF patients

TLF cases, the distributions of *hard* TLF cases are very similar to Non-TLF cases, see Table 2. This could explain why all models perform poorly on those *hard* TLF patients and why these models often misclassify Non-TLF patients as TLFs. These results may help doctors in forecasting patients to develop TLF, e.g. a younger, male patient with a longer lesion with no previous interventions that never had a myocardial infarct before.

3.1.2 Time Series Approach

As shown in subsection 2.1, patient data is recorded at multiple steps in time (e.g. at Screening, Procedure, Hospital Discharge, Follow-Up) sustaining the intuition that the evolution of features over time may indicate the risk of developing TLF. Essential for every kind of time series classification algorithm is the availability of (at least one) features over multiple steps in time. This dataset contains multiple features at various different steps in time. However, except features from the Follow-Up visits, none of those features occur at different visits. Furthermore, the features measured at the Follow-Up visits are only sparsely available. Due to the fact that the given dataset is the result of an still ongoing long-term study, many Follow-Up visits are not yet performed for most of the patients. Taking only the first three (of six possible) Follow-Up visits into account already cuts off half of the available patients. For those two reasons - only Follow-Up features occurring for at least three different steps in time; and even those being sparsely available - time series classification was not further pursued.

3.2 Survival Analysis

In classification, we were interested in studying how risk factors were associated with presence or absence of TLF. Sometimes, though, we are interested in how a risk factor or treatment affects time to TLF. Or we may have study dropout, and therefore subjects who we are not sure if they had TLF or not. In these cases, classification is not appropriate.

Survival analysis is used to analyze data in which the time until the event is of interest. It is especially suitable for analysing right censored data, which often occurs in medical studies. The response is often referred to as a *failure time*, *survival time*, or *event time*. In our case, we want to estimate the time to TLF after procedure.

3.2.1 Theoretical background

Here, we start by defining fundamental terms of survival analysis, including *survival time* and *event*, *censoring*, *survival function* and *hazard function*.

Event is something we want to observe, in our case the occurrence of TLF. The measure of interest in our study is the time until the event, commonly called *survival time*, which in our study is the time from the procedure to the occurrence of TLF.

As mentioned above, survival analysis focuses on the expected duration of time until the occurrence of an event of interest. However, the event may not be observed for some individuals within the study time period, producing the so-called *censored observations*. Censoring may arise in the following ways: a patient has not (yet) experienced the event within the study time period, or a patient is lost to follow-up during the study period. This type of censoring, called *right censoring*, is handled in survival analysis.

The *survival probability*, also known as the *survival function* $S(t)$, is the probability that an individual survives, in our case not getting TLF, from the time origin (e.g. procedure) to a specified future time t. The *hazard*, denoted by $h(t)$, is the instantaneous rate of getting TLF at time t. Considering T to be a random variable denoting the time of event, $S(t)$ and $h(t)$ can be expressed as follows

$$S(t) = P(T > t), \qquad h(t) = \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{\Delta t \cdot S(t)} \tag{1}$$

where P is the probability measure of T.

When we predict TimeToTLF ahead of time, all other meaningful features during screening and procedure are considered as *covariates*.

3.2.2 Kaplan-Meier Curve

A plot of the **Kaplan-Meier estimator** [8] is a series of declining horizontal steps which, with a large enough sample size, approaches the true survival function for that population. The value of the survival function between successive distinct sampled observations is assumed to be constant.

An important advantage of the Kaplan-Meier curve is that the method can take into account some types of censored data, particularly right-censoring, which occurs if a patient withdraws from a study, is lost to follow-up, or is alive without event occurrence at the last follow-up. When no truncation or censoring occurs, the Kaplan-Meier curve is the complement of the empirical survival function.

Time	n.risk	n.event	Survival	std.error	Lower 95% CI	Upper 95% CI
0	2063	9	0.996	0.00145	0.993	0.998
365	1916	124	0.935	0.00542	0.925	0.946
730	1876	36	0.918	0.00606	0.906	0.930
1095	1867	7	0.914	0.00618	0.902	0.927
1460	1865	1	0.914	0.00619	0.902	0.926
1825	1864	0	0.914	0.00619	0.902	0.926

Table 3: Result of Kaplan-Meier estimation without grouping - TLF by year

By plotting the Kaplan-Meier curves, each grouped by one feature from screening or procedure, survival time is shown as a curve for different groups. Further, performing a log-rank test quantifies the correlation between the observed feature and TimeToTLF.

Table 3 shows that among the 2063 patients, from whom the TimeToTLF can be computed, there are 91.4% who did not experience TLF before the start of our study. Most TLF events happen in the first year after procedure, some happen in the second year and very few happen after three to five years. In the plot, small vertical tick-marks state individual patients whose survival times have been right-censored. Performing log-rank tests shown in Figure 7, we see that the hazard functions between young and old patients are significantly different with p-value 0.002, and the hazard functions of males and females are also significantly different with p-value 0.001. In other words, young people and males have significantly higher risk of getting TLF. This surprising result regarding age is very counterintuitive at first glance. An explanation might be that young people have more active immune systems and therefore might have stronger reactions against an implanted stent.

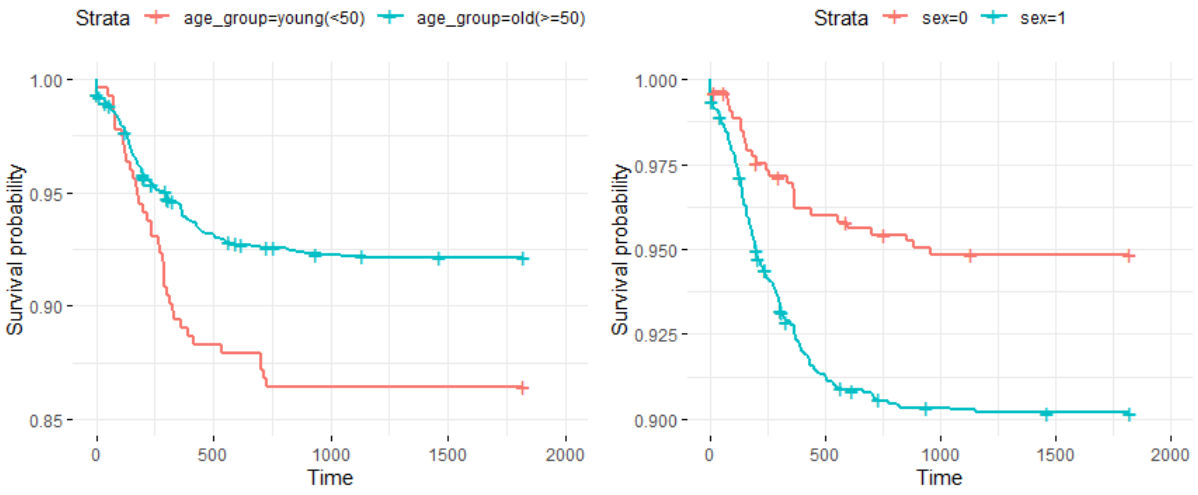

Figure 7: TimeToTLF by age (left) and sex (right)

Only observing simple features such as sex and age is apparently not enough for offering any medical help. After performing log-rank test for all features from screening and procedure, we obtained a list of features that have p-value less than 0.05, i.e. significantly different hazard functions among different subgroups.

Feature	Grouping	P-value	Risk of TLF
Age	young: 25-49 old: 50-89	0.002	old < young
Gender	1: male 0: female	0.001	female < male
AE during procedure	1: yes; 0: no	0.09	no AEs < have AEs
Device deficiency	1: yes; 0: no	0.02	good device < bad device
Dose-area product	low: 0-788 medium: 789-2543 high: 2544-5548	0.003	low < high < medium
Contrast product	low: 0-159 medium: 160-209 high: 210-2400	0.005	medium < low < high
Different vessels	LAD; LCA; LCX; RI	≤ 0.001	LCX < RI < LCA < LAD
Number of pre-dilation balloons	1; 2 or more	0.007	1 < 2 or more
Number of post-inflations	(≤ 2); (≥ 3)	0.03	(≥ 3) < (≤ 2)
Stent inflation time (sec)	fast: 1-14 slow: ≥ 15	0.004	slow < fast
Magmaris scaffold length (mm)	short: 0-24 long: 25-30	0.03	short < long
ASA prior to procedure (daily dose) (mg)	few: 0-75 some: 75-100 many: 100-1000	0.009	some < many < few
Patient received ASA (loading dose) (mg)	few: 0-499 many: 500-1000	0.01	few < many
Anti-platelet medication I prior to procedure (daily dose)	few: 0-74 some: 75-179 many: 180-600	0.009	some < many < some
Anti-platelet medication I (loading dose) (mg)	1: yes; 0: no	0.04	no < yes
Stent diameter vs. pre-dilation balloon diameter	small: 0-0.99 large: 1-3.5	0.009	large < small
Maximum pressure applied vs. stent inflation time	small: 0-0.99 large: 1-15	0.002	small < large

Table 4: Features with different hazard functions between different groups

3.2.3 Cox Proportional Hazard Regression

The above mentioned Kaplan-Meier estimations are examples of univariate analysis. They describe the survival according to one factor under investigation, but ignore the impact of any others. The Cox proportional-hazards model [4] is essentially a regression model commonly used in medical research for investigating the association between the survival time of patients and one or more predictor variables.

The purpose of the model is to evaluate simultaneously the effect of several factors on survival. In other words, it allows us to examine how specified factors influence the rate of a particular event happening (e.g., getting TLF) at a particular time point. The hazard function (1) can be estimated as

$$h(t) = h_0(t) \times exp(b_1x_1 + b_2x_2 + \ldots + b_px_p) \tag{2}$$

where $(x_1, x_2, \ldots, x_p)$ is a set of p covariates, the coefficients $(b_1, b_2, \ldots, b_p)$ measure the

impact of covariates, and h_0 is the baseline hazard, which is the hazard value if all x_i's are equal to zero.

By performing **Cox Proportional Hazard Regression (Cox-PH)**, we first pick all features with low p-values from the Kaplan-Meier estimate to efficiently limit the number of features and add efficiency, and we add the ratios of interest as features as well. Then we perform a Cox regression, and test the proportional hazards assumption for each covariate, i.e. whether the hazard rate is relatively constant over time. The assumption is satisfied globally and individually with respected to all except for two features: *the occurrence of AEs during procedure* and *Anti-platelet medication I prior to procedure (daily dose)*. We simply get rid of these two features, and use AIC [2], which estimates the quality of a model considering both the goodness of fit and the simplicity of the model, to perform feature selection.

Finally, we find a suitable model as shown in Figure 8 that can be used to predict TimeToTLF, based only on several features from screening and procedure. The goodness of the model is be estimated by *Harrell's C-index* (Concordance index) [6], which tells the proportion of observations that the model can order correctly in terms of survival times. A C-index of 0.68 is although not as perfect as 1, but tells that the model is much better than a coin flip with C-index 0.5.

We visualize the hazard ratios (HR) by creating a graphical summary of a Cox model using forest plot. For each covariate, it displays HRs and their 95% confidence intervals. P-values on the right denote the significance of a covariate when less than 0.05. Briefly, an HR > 1 indicates an increased risk of getting TLF (according to the definition of $h(t)$) if a specific condition is met by a patient, while an HR < 1 indicates a decreased risk.

It is interesting to notice that when the ratio stent diameter / pre-dilation balloon diameter is larger or equal to 1, the risk of getting TLF decreases up to more than 50% comparing to when the rate is less than 1. However, the conclusion can be questioned, since only in very few cases the rate is less than 1. It is also noticeable that if the rate maximum pressure applied / stent inflation time is greater or equal to 1, the patient is 2 times more at risk of getting TLF than those with rates less than 1. The result coincides with previous studies, saying that longer inflation time ensures a good extension of the stent [7], while the pressure does not have significant effects [5].

4 Causal Inference

An alternative approach to better understand the occurrence of TLF is using methods from causal inference, since standard methods from machine learning and statistics are only able to discover association but not causality. *Causal inference* describes the process of drawing conclusions from data about causal effects and quantifying them [1]. This is especially relevant for the investigation of treatment effects in healthcare, based on data from observational studies: Even if we find a statistically significant association between a treatment variable and the outcome, for example with approaches as in section 3, we do not know whether it was the treatment that led to the outcome or some other, perhaps unobserved factor.

Consequently, the goal of this part of the project is to quantify the direct causal effect that different treatment decisions, taken during the implantation of stents, have

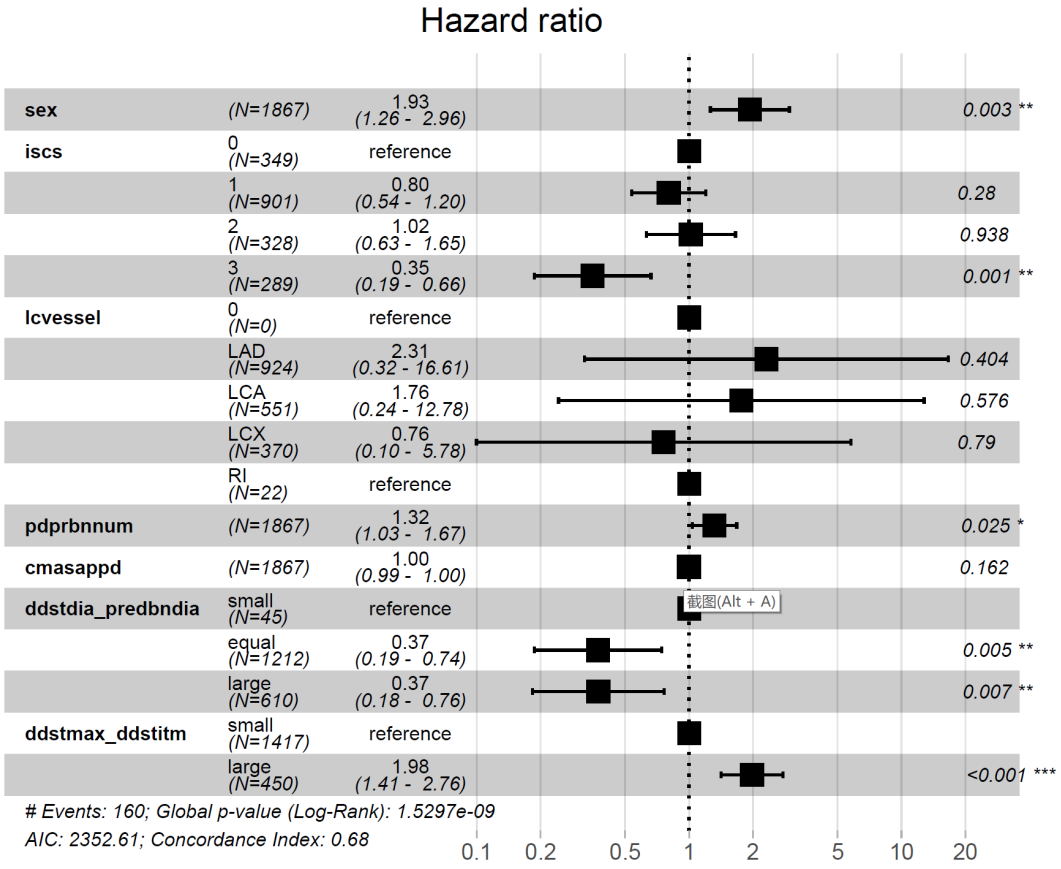

Figure 8: Result of Cox-PH regression. *sex* = {0: female; 1: male}, *iscs* = {0: AMI; 1: Stable angina; 2: Unstable angina; 3: Documented silent ischemia}, *lcvessel* = different vessels, *pdprbnnum* = number of pre-dilation balloons, *cmasappd* = ASA prior to procedure (daily dose) (mg), *ddstdia_predbndia* = Stent diameter / pre-dilation balloon diameter, *ddstmax_ddstim* = max. pressure applied / stent inflation time.

on the outcome of the procedure. We consider eight ratios of procedure parameters, like balloon diameter to stent diameter, that were rated as useful by doctors who conduct the procedure. Then we estimate the average causal effect of varying these ratios on the TLF outcome for the patients. This is done based on the given patient data and on our assumptions about their causal structure. Before specifying our approach, we give a short introduction to the theoretical background.

4.1 Motivation

It is well-known that causation is not the same as association or correlation. For example, higher ice cream consumption is positively correlated with an increase in violent crime in New York, but neither does eating ice cream cause violent crime nor the other way around. Instead, higher temperatures lead to both higher ice cream consumption and an increase in violent crime [14], causing this so-called *spurious association*. Such a factor that influences both the assumed cause and the assumed effect is called a *confounder* [13].

In medical research this is a common situation that can be illustrated with this project:

We consider data of patients containing their medical history, the treatment they received, and information about their recovery. We presume that on the one hand doctors make treatment decisions based on their knowledge about the patients' medical history, and on the other hand the patients' medical history influences the outcome of treatment. This poses a problem when trying to determine the effect of the treatment on the outcome based on the data: when a patient gets better after treatment, it is possible that only the specific treatment they received made their situation better and they would have had a worse outcome with another treatment. It is also possible that they would have felt the same or even better without the treatment due to other factors of their medical history.

This problem exists in general, as it is not possible to simultaneously observe the outcome of two different treatments on the same patient. The unobserved outcome is called *counterfactual*. The resulting challenge is known as the *fundamental problem of causal inference* [15]. One possible solution to circumvent this and still be able to accurately quantify the causal effect of a treatment are randomized controlled trials. By randomly assigning the different treatment options to the patients and comparing the outcomes to those of a control group, possible confounders can be controlled [14]. Often this is not possible due to ethical reasons or feasibility issues, and therefore only observational data is available for inferring causality [17].

In that case, we need to make additional assumptions about the underlying causal structure of the data generating process to be able to correctly quantify effects. Taking only statistical association between two variables into account might lead to wrong conclusions, because any statistical relationship between two variables can be reversed by additionally considering other factors [13].

In a classical example, sick patients are given the option of trying a new drug. Among the patients who took the drug, a lower percentage recovered than among those who did not take the drug, so the drug seems to have a negative effect. But, when additionally grouping the patients by gender, the percentage of recovery is in both groups higher for patients taking the drug, so the effect seems to be positive. This is known as *Simpson's paradox* and can be resolved by additional knowledge about the causal structure of the data. For example, we might know that men recover more easily from the condition but that they are less likely to take the drug. Then it is clear why the effect without taking gender into account seems to be negative, while it is in fact positive: A randomly picked drug user is more likely to be female and therefore less likely to recover. In other words, being female is a common cause of drug taking and not recovering in this example [14]. Therefore, a mathematical framework to formalize assumptions about causal structure is needed.

4.2 Theoretical Background

There are different approaches to mathematically formulate a theory of cause and effect: the Neyman-Rubin *Potential Outcomes Framework* [15], *Structural Causal Models*, and *Causal Graphical Models* established by Pearl [13]. These theories provide different languages for causality, but they are logically equivalent [15]. It can be useful to combine individual aspects of for example potential outcomes and causal graphical models. In this project we mostly rely on the theory of causal graphical models since they allow us to formally encode cause and effect relationships in an intuitive way.

We define a variable X to be the cause of a variable Y if Y relies on X for its value, i.e., changes in X lead to changes in Y [14]. Given a set of random variables, a *causal graphical model* is a pair consisting of a directed acyclic graph (DAG) with the variables as vertices and a joint distribution of the variables that together satisfy additional properties known as the causal Markov condition[2] [13], [15]. This condition requires that the model does not only encode probabilistic information about conditional independence but also causal information. The directed edges in the DAG correspond to direct causal relationships: If X is a direct cause of Y, there is a directed edge from X to Y [14]. In the example of Simpson's paradox this means that there are two edges from a vertex *Gender* to *Drug* and *Recovery* and an edge from *Drug* to *Recovery* as shown in Figure 9. The absence of an edge between two vertices is an even stronger assumption since it rules out any direct causal relationship between them for every individual of the considered population.

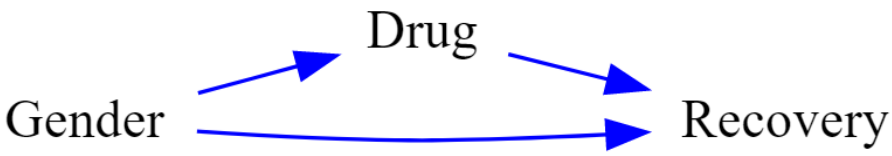

Figure 9: A graphical model, representing the relationship between Gender, Drug and Recovery

Unmeasured confounders of two variables are sometimes represented as bi-directed edges between the two variables. This makes the resulting graph an acyclic directed mixed graph (ADMG) by extending the previous definition. The resulting properties of the causal model and the adaptation of the theory can be found in [1].

To assess the effect of one variable on another we want to know how the system behaves under an intervention on, for example, a treatment variable X, while keeping everything else constant, similar to a randomized experiment. To be able to formalize the idea of interventions, Pearl established the notion of *do-calculus* that allows formulating interventions mathematically as do(variable = action) [13]. For instance, we can formulate the intervention of giving each patient treatment A as $do(X = A)$. It is important to note that this is fundamentally different to conditioning on the observation $X = A$ [14]. In this case we narrow our focus on a subset of patients where the condition of having received treatment A is fulfilled. Applying $do(X = A)$, i.e., intervening by fixing the treatment to be A for each patient, changes the system we are investigating. The intervention corresponds to a new graph, where every link from other vertices into the vertex that was intervened on is removed [14]. Therefore, $P(Y|do(X = A))$ is in general not the same as $P(Y|X = A)$.

Now we can define the parameter of interest: the *average causal effect* (ACE) [1] of a binary treatment decision X on the outcome Y given by

$$ACE = \mathbb{E}[Y|do(X = 0)] - \mathbb{E}[Y|do(X = 1)]. \tag{3}$$

[2]See [13] and [15] for formal definitions.

This poses a challenge for computing the ACE, since we have not seen data from an *interventional distribution* $P(Y = y|do(X = x))$. But this probability is needed to compute the *counterfactual mean* $\mathbb{E}[Y|do(X = x)]$[15].

If the ACE is computable in terms of the observational distribution of the data given the graph structure, it is called *identifiable* [13]. For this calculation we need the three rules of do-calculus [13]. They guarantee, that under certain conditional independence assumptions, observations and actions can be inserted, deleted, and exchanged [13]. For example, by setting the treatment for every patient in Simpson's paradox to be the new drug, we remove the causal influence of gender on the decision to take the drug. For the graph in Figure 9, this means removing the edge from *Gender* to *Drug*. Given this intervention, the marginal distribution of gender stays the same, but drug use and gender become causally independent [14].

To compute the ACE with respect to a causal graph, we need to transform the counterfactual mean into statistical quantities that we can estimate from the data. Denoting the parents of a vertex X in the graph by $pa(X)$, the so-called *backdoor-adjustment formula*

$$\begin{aligned}\mathbb{E}[Y|do(X = x)] &= \sum_z P(Y = y|X = x, pa(X) = z)P(pa(X) = z) \\ &= \mathbb{E}[\mathbb{E}[Y|X = x, pa(X)]]\end{aligned} \tag{4}$$

can be derived [14], [1]. The procedure is called *adjusting* or *controlling* for the parents of X. It describes blocking every path between X and Y that has an edge into X, i.e., is a so-called *backdoor path*, by conditioning on $pa(X)$ [14]. Then intervening with $do(X = x)$ and conditioning on $X = x$ have the same effect on Y [13]. Intuitively, this means in Simpson's paradox to first calculate the effect for men and women separately. Then we average both individual effects taking the proportions of men and women in the population into account [14]. Sets of vertices like the parents of X that achieve this blocking and that do not contain children of X are said to satisfy the *backdoor criterion* [14]. In the case that X is multivariate by having multiple interventions, there exists a generalization of this formula, called *g-formula* [13].

When defining a causal structure, we often assume unmeasured confounders to exist. Therefore, we do not always have corresponding data to all parents of a vertex X [14]. Then the effect of X on Y is either not identifiable or it is possible to adjust for other sets of variables [14]. This can be done with a set Z of vertices satisfying the backdoor criterion by substituting Z for $pa(X)$ in Equation 4 [14]. If such a set does not exist, there sometimes is another solution called front-door adjustment. A set of vertices that intercepts all directed paths from X to Y with no un-blocked additional backdoor paths between this set and X and Y satisfies the *front-door criterion*. The effect can then be computed via the *front-door-adjustment formula*, as X affects Y only through other known variables [14].

After identifying the target parameter, standard methods from statistical inference can be used to estimate it based on the given data [1]. *Inverse probability weighting* (IPW) is a well-known technique that can be applied when the backdoor criterion is fulfilled for a set Z [14]. By using factorization properties of causal graphical models, we can transform

the adjustment formula in (4) to

$$\mathbb{E}[Y|do(X=x)] = \sum_z \frac{P(Y=y, X=x, Z=z)}{P(X=x|Z=z)}.$$

The function $g(x,z) := P(X=x|Z=z)$ is called the *propensity score* for each x and z [14]. If this function is known, artificial samples from the post-interventional distribution can be generated by re-weighting the probability $P(X=x, Y=y, Z=z)$ of each observed sample by the factor $1/P(X=x|Z=z)$ [14]. This means re-weighting the observed data of units truly assigned $X=x$ by the inverse of the normalized treatment assignment probability $P(X=x|Z)$ [1]. Then, we can count the frequency of the value $Y=y$ for each stratum $X=x$ and therefore do not have to sum over all strata $Z=z$ [14]. The propensity score can be estimated using standard regression techniques [14]. IPW is simple to implement but inefficient and not applicable to all identified effects [1]. Therefore, other estimators have been developed that are more efficient and applicable to causal structures with unmeasured variables. An example is *efficient augmented probability weighting* (eff-AIPW) for ADMGs [1]. A detailed description of eff-AIPW and other estimators can be found in [1].

4.3 Approach

There are several Python frameworks for causal inference based on causal graphical models available. Examples are *DoWhy* by Microsoft [16] or *Ananke-causal* developed at Johns Hopkins University [1]. In this project we use Ananke-causal since it provides an easy interface for estimating causal effects in three steps. First, reasonable assumptions about the causal structure of the observed data must be made by defining a causal graph. Then identifiability can be checked. If the effect is identifiable, Ananke provides different methods for estimating causal effects.

To quantify the direct causal effect that different treatment decisions taken during PCI have on the long-term outcome of the procedure, eight ratios of treatment parameters are considered. An overview is given in Table 5. In the dataset used for causal inference only features with at most 7% missingness are included. Then, all patients with missing values are excluded and resulting features that only take one value are removed. The resulting dataset contains 1789 patients including 154 TLF cases (11.6%). Checking the value ranges of the considered parameters shows that only the balloon diameter in post dilatation according to compliance chart (`pdbndiacc`) contains erroneous values outside a realistic range of 2mm to 5mm. This parameter is contained in ratio 3 and 4. Therefore, the investigation of those ratios considers a reduced dataset where patients with `pdbndiacc` value outside this range were not considered. The resulting reduced dataset contains 1766 patients including 151 TLF cases (11.7%).

Ananke requires the treatment variable to be binary. Therefore, we map the values of the ratios to $T=0$ or $T=1$ depending on whether they strictly lie below or above a pre-defined threshold. The median value of each ratio is chosen as threshold with the aim of producing a balanced distribution of the treatment values. This is not possible for all ratios, for example, ratio 1 takes the value 1 for about two thirds of the patients. Consequently, the distribution of the treatment values 0 and 1 depends on whether 1 lies below or above the threshold. All values strictly below the threshold get mapped

Ratio R	**Parameter 1** P_1	**Parameter 2** P_2	**Threshold**	**Binarization** $T = 1$ if
1	vessel diam.	stent diam.	median 1	$R \geq 1$
2	lesion length	stent length	median 0.8	$R \geq 0.8$
3	stent diam.	balloon diam. post dil. acc. to compl. chart	median 0.93	$R \geq 0.9$
4	vessel diam.	balloon diam. post dil. acc. to compl. chart	median 0.93	$R \geq 0.9$
5	vessel diam.	balloon diam. pre dil.	median 1.0	$R \geq 1$
6	max. pressure	inflation time	median 0.6	$R \geq 0.6$
7	stenosis post-proc.	stenosis pre-proc.	mean 0.02	$R \geq 0.02$
8	num. of balloons pre dil.	num. of inflations pre dil.	median 1.0	$R \geq 1$

Table 5: Ratios of treatment parameters with $R := P_1/P_2$

to 0, while all values at or above the threshold get mapped to 1. The median value of ratio 7 is already zero and it does not take values below zero since both parameters take only non-negative values and the stenosis post-procedure is for most of the patients zero percent. Therefore, the mean value 0.02 is used for a more refined result in this case.

We start with basic assumptions about the causal relationship and then make the assumed structure more complex by incorporating more features of the data. The first assumption is that the treatment, i.e., either of the eight ratios, is a direct cause for the outcome given by the TLF label. Additionally, we want to incorporate the screening information of the patients from enrolment containing information about sex, age, and medical history. We assume that this baseline information about each patient influences the treatment decision taken by the physicians as well as the outcome of the treatment. Based on that consideration, we add it as a measured confounder of treatment and outcome.

There are different ways to model this, either by treating each feature as an individual variable and therefore as an individual vertex, or by combining them to a multivariate variable in a single vertex. When not adding any additional edges between the individual vertices in the first model we implicitly assume those features to not have direct causal relationships, which is unrealistic. Even if we add some edges between individual vertices, it requires many assumptions about the causal structure that we are not confident to make. Hence, the second model, where no assumptions regarding causality between the component variables are taken, is preferred.

We also have information about the concomitant medication the patients received during procedure. Since this is based on the baseline data about each patient, we add it as a vertex with a directed edge from base to medication. It is conducted according to hospital protocol, so we assume it to be causally independent from the treatment decision.

Since according to the conducted survival analysis some medication features seem to be relevant for the TimeToTLF outcome, we assume there to be a direct causal influence of medication on the outcome. Because the baseline features can not capture every information the physicians base, perhaps unconsciously, their decisions on, we assume there to be an unmeasured confounder of treatment and medication. The resulting graph is shown in Figure 10.

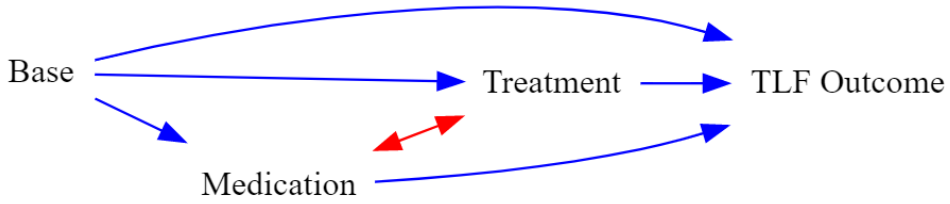

Figure 10: The hypothesised graphical model, depicting the effects on the TLF outcome

This graph is identifiable via backdoor adjustment. It can be argued that this graph does not capture all relevant causal relationships since it does not assume the existence of unmeasured confounding of treatment and outcome. Including this would render the effect unidentifiable. A mediating variable intercepting the direct causal relationship from treatment to outcome would be needed for identification. Since none of the variables in the given data can serve this purpose, we do not include this unmeasured confounding in the model.

Estimation of the causal effect with the framework provided by Ananke poses a challenge, since it does not support multivariate vertices. Providing tuples of several features as values for the vertices leads to erroneous results, because they are not handled as needed by the underlying regression models for estimating the propensity scores. Instead, different tuples are interpreted as different categorical values. Given that these tuples can contain more than 30 different features, this results in mostly unique tuples. The model then overfits and is not able to return meaningful results.

By modifying the Ananke framework this challenge can be solved. In Ananke, the required statistical quantities to compute the effect are estimated by fitting generalized linear models (GLM) to the data. Those quantities include the propensity score $P(T = t|Z)$ of treatment T given a suitable set $Z = \{Z_1, \ldots, Z_s\}$ of vertices[3] needed for standard IPW, but also expectations conditioned on different sets of variables needed for eff-AIPW. In our graph this set is provided by *Base* and *Medication*. The formula for fitting the GLM used in Ananke is $Treatment \sim Base + Medication$ for our graph. By replacing *Base* and *Medication* in the formula with all the features $B_1, \ldots, B_n, M_1, \ldots, M_m$ they contain, we receive the correct formula $Treatment \sim B_1 + \cdots + B_n + M_1 + \cdots + M_m$ for fitting a GLM to the data. Interestingly, for standard IPW, this coincides with the formula resulting from a graph with individual base vertices and medication vertices.

We use eff-AIPW for our experiments since it is recommended by Ananke due to improved robustness compared to IPW [1]. To obtain more robust results, we run bootstraps with $n = 2000$ iterations to calculate 95% percentile bootstrap confidence intervals.

[3]The set used in Ananke is called *Markov pillow*. A formal definition can be found in [1].

Ratio R	LOR	0.025 quantile	0.975 quantile	median
1	-0.017	-1.895	1.147	0.138
2	-0.281	-0.762	0.171	-0.299
3	0.093	-0.314	0.501	0.080
4	0.214	-0.167	0.597	0.208
5	-0.220	-1.720	1.145	-0.108
6	-0.060	-0.439	0.329	-0.064
7	-0.140	-0.937	0.593	-0.147
8	-0.378	-1.060	0.321	-0.351

Table 6: Causal effect as LOR and 95% percentile bootstrap confidence interval

4.4 Results

To estimate the causal effect of different ratios (Table 5) as treatments X on the TLF outcome Y, several experiments are done as aforementioned. We calculate the causal effects as log of odds ratios (LOR) following the formula

$$LOR = \log(\frac{P(Y=1|do(X=1))/P(Y=0|do(X=1))}{P(Y=1|do(X=0))/P(Y=0|do(X=0))}).$$

This is the default behaviour of Ananke for binary outcome due to better interpretability, since $\mathbb{E}[Y|do(X=x)] = P(Y=1|do(X=x))$ for binary Y. Additionally, we use bootstrapping with $n = 2000$ to calculate 95% percentile bootstrap confidence intervals. The combined results of the calculated effects and the corresponding confidence intervals can be found in Table 6. Figure 11 shows a histogram of the bootstrapping results. In Figure 12, the bootstrapping results for computing the causal effect as ACE according to Equation 3 can be found for comparison.

It can be seen in Figure 11 that most of the bootstrap distributions of the LOR are centered around zero, which implies insignificant causal effect, while three of the ratios, namely, Ratio 2 (lesion length/ stent length), Ratio 4 (vessel diam./ balloon diam. post dil. acc. to compl. chart) and Ratio 8 (num. of balloons pre dil./ num. of inflations pre dil.), show a slight tendency to have a causal effect on the outcome, being approximately centered around the bootstrap median values -0.30, 0.21 and -0.35 respectively. The corresponding LOR values, calculated using all considered samples, are -0.28, 0.21, and -0.38.

In order to have a better intuition, the odds ratios (OR) of the respective individual LOR's are calculated as $OR = e^{LOR}$. This results in OR values of 0.76, 1.24, and 0.69 for Ratio 2, Ratio 4, and Ratio 8, respectively. Considering the OR for each ratio, we deduce the following:

- The probability of having a TLF outcome with a bigger value than 0.8 for Ratio 2 is 0.76 of the the probability of having a TLF outcome with a smaller value than 0.8, which might support the decision of using a bigger value than 0.8 for Ratio 2.
- The probability of having a TLF outcome with a bigger value than 0.9 for Ratio 4 is 1.24 of the the probability of having a TLF outcome with a smaller value than

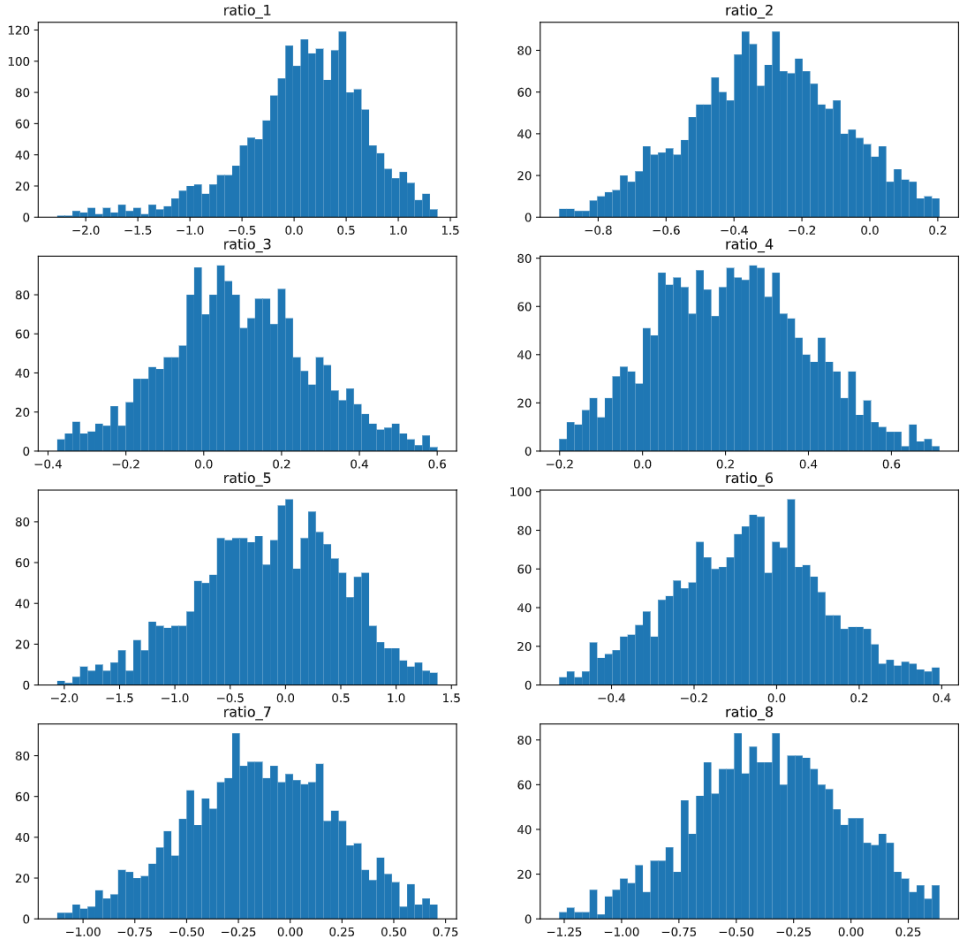

Figure 11: Distribution of log of odds ratios (LOR) for 8 different ratios on TLF outcome. X-axis of each plot is clipped to the lower and upper quantiles of the respective bootstrapping experiment for better visualization. $n_{bin} = 50$

0.9. This might indicate that using a smaller value than 0.9 for Ratio 4 can be a better approach.

- The probability of having a TLF outcome with a bigger value than 1.0 for Ratio 8 is 0.69 of the probability of having a TLF outcome with a smaller value than 1.0, which can promote using a bigger value than 1.0 for Ratio 8.

Even if we assume those findings to be significant enough for recommending specific treatment ratios, they do not always indicate clear action that should be taken. For example, to increase Ratio 8 above the threshold 1, we can in some cases either increase the numerator or reduce the denominator. It only gives a clear indication for action when one of the parameters is fixed by patient characteristics like lesion length.

However, considering that the confidence intervals for the LOR of these ratios still contain both a considerable amount of positive and negative estimates, we cannot assess those shifts to be significant.

Furthermore, it can be seen in Figure 12 that the corresponding ACE values for the different ratios are all centered closely around zero, indicating that the actual expectations

differ very little. Each individual expectation, namely $\mathbb{E}[Y|do(X = 1)]$ and $\mathbb{E}[Y|do(X = 0)]$ for the 8 different ratios can be found in Appendix B. Given that those expected values themselves are very small, this does not necessarily mean that there is no effect. However, considering the relative difference of the values yields the same result as for the LOR: even if we observe slight tendencies of the ACE of some ratios away from zero, given the bootstrap confidence intervals this is not significant enough.

Based on those results, we can therefore not deduce a strong indication for action regarding the considered treatment parameter ratios.

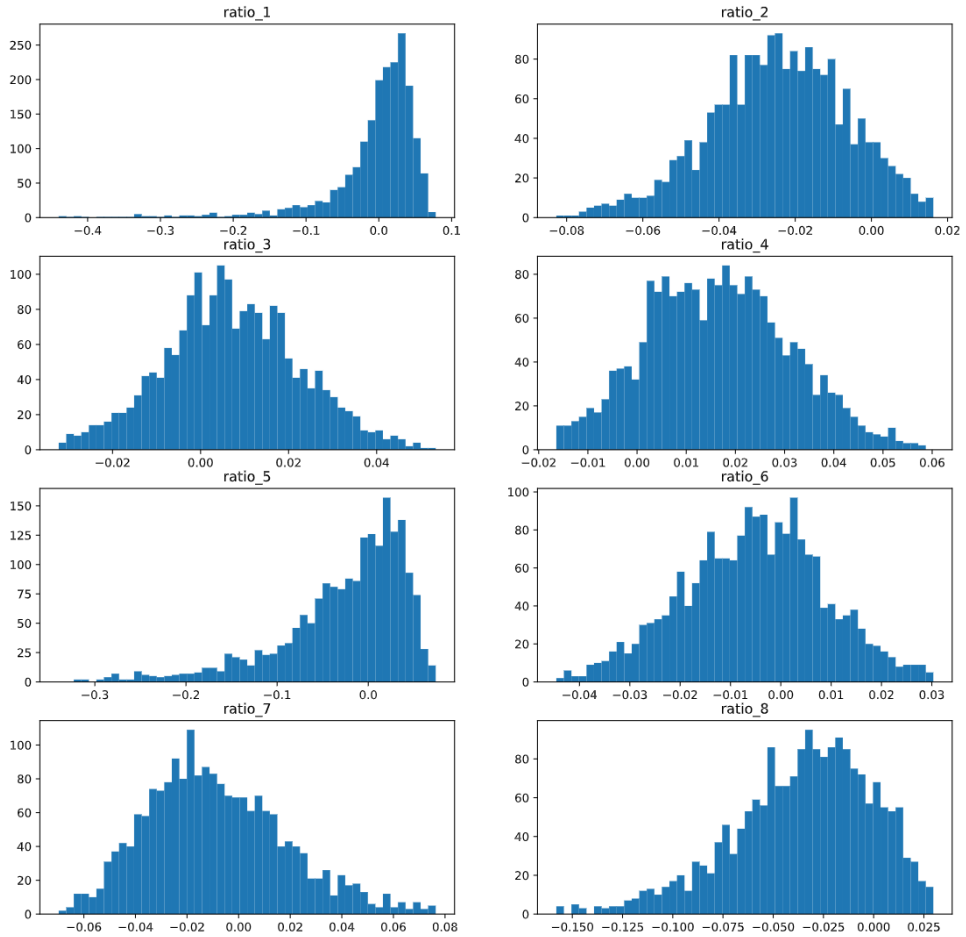

Figure 12: Distribution of average causal effect (ACE) for 8 different ratios on TLF outcome. X-axis of each plot is scaled by setting the limits as the lower and upper quantiles of the respective bootstrapping experiment for better visualization. $n_{bin} = 50$

5 Conclusion and Outlook

Learning hidden structures in medical data is a hard task. Even though BIOTRONIK tries to only include very similar patients in their study using various inclusion and exclusion criteria, every patient has his unique behaviour. Especially in medical diagnosis it is difficult to compare different patients. Various factors can be adjusted; every available

feature could be the reason to develop TLF. Challenges with this particular dataset are numerous. (1) The high amount of nested questions causes a variety of sparsely filled features. Depending on whether the patient or physician filled in the parent question or which answer was given, the respective feature is empty. (2) Almost all features were given in String format, bringing on a big amount of encoding work during preprocessing. (3) Patients with multiple occurrences per visit have to be tackled, resulting in a - possibly crucial - data loss due to the here favored handling approach. (4) Different medication, measured in distinct units, impedes to objectively compare patient characteristics and (5) right-censored data aggravates reliable statements for all different approaches.

The performed binary classification, especially the Deep Learning approach using a Neural Network, corroborates the assumption that this particular dataset is not easy to learn. A highly regularized low-dimensional Neural Network is as well not able to generalize as all other used classic Machine Learning approaches. Those are not even performing well during training.

Nevertheless, it seems that for some of the patients it is easier to predict the development of TLF than for others. These, here called *easy* patients, do have significantly differing univeriate characteristics than the other cohorts, indicating that there could be a reason for patients to be at higher risk to develop TLF than others.

Future work should be done including more and more of the available feature set to identify underlying coherences that could not yet be found. Further improvements can be accomplished once the study is heading to finish, providing a less sparse dataset, especially on the performed Follow-Up visits. Another interesting approach could be to pursue the latest step of the performed binary classification: Trying to distinguish between easy- and hard-to-predict patients to be able to provide more sophisticated advices to the operating physicians. Investigating on significant multivariate differences between those cohorts may lead to further understanding which patient characteristic could be a reason for developing TLF or which group of patients is at higher risk to do so. Even the information whether a patient is harder-to-predict than others may help the operator to focus on this specific group of patients.

Survival analysis provides additionally the risk of getting TLF over time. When making a medical decision, a doctor can take a the result from the Kaplan-Meier estimation and Cox model into consideration, and choose different treatments during procedure or give special attention to patients more at risk after hospital discharge. In particular, we select 17 feature that one can tell difference from and 7 features one can estimate from. Our Cox regression model for predicting time-to-TLF achieves a satisfying accuracy of 68%. However, one limitation of this method is that it cannot deal well with a huge amount of features. The Kaplan-Meier curve, for example, takes only one feature at a time. Hence, going through all features takes too much manual effort. The procedure should be able to be modified.

Further work can be done by using some state-of-the-art survival models rather than the classical ones. For example, the recently developed model DeepSurv [9], a Cox proportional hazards deep neural network, models interactions between a patient's covariates and treatment effectiveness in order to provide personalized treatment recommendations. It can especially model complex relationships between a patient's covariates and their risk of failure, and is said to outperform other survival models.

By employing methods from causal inference we considered not only association but

also causal relationships. We estimated the ACE of eight ratios of treatment parameters on the TLF outcome based on a causal graphical model we defined. The results are in general not strong enough to reject the null hypothesis of having no causal effect. Therefore, we cannot deduce recommendations for specific treatment ratios that might improve the TLF outcome for the patients.

However, three of the ratios seem to exhibit a slight tendency towards a positive or negative impact and should therefore be further investigated. A significant causal effect of a ratio of parameters on the outcome cannot always be directly mapped to a treatment recommendation. Therefore, future work could be done by investigating some treatment parameters directly and by testing different modifications of the ratios with clinicians. Another promising approach to take the causal structure of the data into account is employing methods from causal discovery to learn the causal structure of the data. This could give valuable insights into the development of TLF and thereby help doctors to adjust procedure parameters to reduce the risk for patients.

References

[1] Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. 2020.

[2] Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

[3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[4] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[5] Josef Dirschinger, Adnan Kastrati, Franz-Josef Neumann, Peter Boekstegers, Shpend Elezi, Julinda Mehilli, Helmut SchuÌˆhlen, JuÌˆrgen Pache, Eckhard Alt, Rudolf Blasini, et al. Influence of balloon pressure during stent placement in native coronary arteries on early and late angiographic and clinical outcome: a randomized evaluation of high-pressure inflation. *Circulation*, 100(9):918–923, 1999.

[6] Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

[7] Thomas Hovasse, Darren Mylotte, Philippe Garot, Neus Salvatella, Marie-Claude Morice, Bernard Chevalier, Augusto Pichard, and Thierry Lefèvre. Duration of balloon inflation for optimal stent deployment: five seconds is not enough. *Catheterization and Cardiovascular Interventions*, 81(3):446–453, 2013.

[8] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[9] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1–12, 2018.

[10] Yoshinobu Onuma, John Ormiston, and Patrick Serruys. Bioresorbable scaffold technologies. *Circulation journal : official journal of the Japanese Circulation Society*, 75:509–20, 02 2011. doi: 10.1253/circj.CJ-10-1135.

[11] World Health Organization. Cardiovascular diseases (cvds). `https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`, 2021. Accessed: 2021-07-15.

[12] Elisabeth Pachl. Prediction of heart surgery outcome based on patient data with machine learning. Master's thesis, Ludwig-Maximilian-UniversitÃt MÃ$\frac{1}{4}$nchen, 12 2020.

[13] Judea Pearl. *Causality*. Cambridge university press, 2009.

[14] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[15] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[16] Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

[17] Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

[18] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[19] Stefan Verheye, Adrian Wlodarczak, Piero Montorsi, Jan Torzewski, Johan Bennett, Michael Haude, Gregory Starmer, Thomas Buck, Marcus Wiemer, Amin, Amin Nuruddin, Bryan, P.-Y Yan, Michael, K.-Y Lee, and Stefan Correspondence. Biosolve-iv-registry: Safety and performance of the magmaris scaffold: 12-month outcomes of the first cohort of 1,075 patients. *Catheterization and Cardiovascular Interventions*, 09 2020. doi: 10.1002/ccd.29260.

[20] Xue Ying. An overview of overfitting and its solutions. In *Journal of Physics: Conference Series*, volume 1168, page 022022. IOP Publishing, 2019.

Appendices

Appendix A Used Feature Set

Feature	TLF (n=164)	No TLF (n=1807)	Total (n=1971)	P-Value	Missingness [%]
Patient received ASA loading dose, Dose	157.9 ± 213.15	128.03 ± 224.12	130.52 ± 223.33	0.1	0.1
Anti-platelet medication I loading dose	121 (73.8%)	1149 (63.6%)	1270 (64.4%)	**0.01**	0.1
Heparin during procedure	115 (70.1%)	1307 (72.3%)	1422 (72.1%)	0.61	0.0
Anti-platelet medication I prior to procedure (daily dose), Dose	34.57 ± 57.16	50.06 ± 77.79	48.75 ± 76.37	**0.01**	1.7
Patient on ASA prior to procedure (daily dose), Dose	69.57 ± 43.59	76.12 ± 52.33	75.57 ± 51.69	0.12	0.0
Heparin bolus injection prior to procedure, Dose	2873.77 ± 3299.61	2651.33 ± 3328.67	2669.78 ± 3326.0	0.42	0.9
Age at time of enrolment	60.32 ± 11.08	61.99 ± 10.51	61.85 ± 10.57	**0.05**	0.0
LVEF class	1	1	1	0.91	2.9
Thrombus present	0 (0.0%)	3 (0.2%)	3 (0.2%)	0.6	0.0
Lesion length	15.44 ± 4.95	14.76 ± 3.89	14.81 ± 3.99	**0.04**	0.0
Reference vessel diameter after Nitro/ISDN i.c.	3.22 ± 0.29	3.24 ± 0.28	3.24 ± 0.28	0.58	0.2
Ecccentricity	57 (34.8%)	601 (33.3%)	658 (33.4%)	0.76	0.0
Number of pre-dilatation balloons used	1.46 ± 0.63	1.34 ± 0.59	1.35 ± 0.59	**0.02**	0.0
Stenosis pre-procedure	83.34 ± 10.31	82.11 ± 10.67	82.21 ± 10.64	0.16	0.0
Number of post-dilatation ballons used	1.16 ± 0.47	1.11 ± 0.35	1.12 ± 0.36	0.07	0.0
Number of Magmaris scaffolds that were implanted	1.05 ± 0.24	1.03 ± 0.17	1.03 ± 0.18	0.18	0.0
Bifurcation lesion	5 (3.0%)	87 (4.8%)	92 (4.7%)	0.4	0.0
Pre-procedure TIMI flow distal to target lesion	3	3	3	0.14	0.1
ACC / AHA lesion characterization	2	2	2	0.42	0.0
Maximum pressure applied	14.62 ± 3.21	14.62 ± 3.44	14.62 ± 3.42	0.99	0.2
Balloon diameter	3.06 ± 0.38	3.07 ± 0.36	3.07 ± 0.36	0.73	0.0
Number of inflations	1.52 ± 0.75	1.59 ± 1.12	1.58 ± 1.09	0.43	0.2
Balloon length	14.76 ± 3.13	14.25 ± 2.94	14.29 ± 2.96	**0.03**	0.1
Maximum pressure applied	16.73 ± 3.24	17.0 ± 3.22	16.98 ± 3.22	0.29	0.2
Balloon diameter	3.47 ± 0.33	3.46 ± 0.35	3.46 ± 0.35	0.72	0.0
Number of inflations	1.74 ± 1.02	1.88 ± 1.44	1.87 ± 1.41	0.22	0.2
Balloon length	14.16 ± 3.61	14.08 ± 3.48	14.09 ± 3.49	0.78	0.1
Maximum pressure appliead	14.28 ± 2.64	14.36 ± 2.76	14.35 ± 2.75	0.72	0.8
Inflation time	21.94 ± 11.89	23.57 ± 11.05	23.43 ± 11.13	0.08	1.7
Magmaris scaffold length	20.06 ± 3.95	19.41 ± 3.87	19.47 ± 3.88	**0.04**	0.0
Residual % stenosis after Magmaris implantation	1.22 ± 4.17	1.41 ± 4.35	1.39 ± 4.34	0.59	0.2
Magmaris scaffold diameter	3.24 ± 0.25	3.25 ± 0.25	3.25 ± 0.25	0.78	0.0
Unter dialysis	0 (0.0%)	0 (0.0%)	0 (0.0%)	1.0	0.0
Renial disease	10 (6.1%)	111 (6.1%)	121 (6.1%)	0.88	0.0
History of stroke or TIA	8 (4.9%)	60 (3.3%)	68 (3.5%)	0.41	0.0
Hepatic disease	4 (2.4%)	36 (2.0%)	40 (2.0%)	0.92	0.0
History of previous myocardial infarction	32 (19.5%)	391 (21.6%)	423 (21.5%)	0.59	0.0
Hypertension	111 (67.7%)	1205 (66.7%)	1316 (66.8%)	0.86	0.0
Hypercholesteremia	108 (65.9%)	1182 (65.4%)	1290 (65.4%)	0.98	0.0
Number of previous interventions	0.48 ± 1.13	0.48 ± 1.04	0.48 ± 1.05	0.98	0.0
Cancer	10 (6.1%)	125 (6.9%)	135 (6.8%)	0.81	0.0
Diabetes	0	0	0	0.76	0.0

Table 7: Feature characteristics part I

Feature	TLF (n=164)	No TLF (n=1807)	Total (n=1971)	P-Value	Missingness [%]
Smoking habits	1	1	1	0.69	0.3
Respiratory disease	13 (7.9%)	151 (8.4%)	164 (8.3%)	0.97	0.0
Did one or more AE, SAE, ADE or SADE event(s) occur during this procedure	15 (9.1%)	111 (6.1%)	126 (6.4%)	0.18	0.0
X-ray: Dose-area product in cGy*cm2	4308.25 ± 4474.9	5100.72 ± 9988.43	5033.93 ± 9647.95	0.31	1.3
Contrast product in mL	182.25 ± 123.96	183.01 ± 151.42	182.94 ± 149.3	0.95	0.5
Has a device deficiency occurred prior or during the procedure	3 (1.8%)	12 (0.7%)	15 (0.8%)	0.24	0.0
Gender = Female	25 (15.2%)	474 (26.2%)	499 (25.3%)	**0.0**	0.0
Gender = Male	139 (84.8%)	1333 (73.8%)	1472 (74.7%)	**0.0**	0.0
Vessel = LAD	101 (61.6%)	875 (48.4%)	976 (49.5%)	**0.02**	0.0
Vessel = LCX	15 (9.1%)	376 (20.8%)	391 (19.8%)	**0.0**	0.0
Vessel = RCA	47 (28.7%)	534 (29.6%)	581 (29.5%)	0.88	0.0
Vessel = RI	1 (0.6%)	22 (1.2%)	23 (1.2%)	0.75	0.0
Type of most recent MI = NSTEMI	12 (7.3%)	167 (9.2%)	179 (9.1%)	0.5	0.0
Type of most recent MI = STEMI	17 (10.4%)	191 (10.6%)	208 (10.6%)	0.96	0.0
Type of most recent MI = Unknown	3 (1.8%)	33 (1.8%)	36 (1.8%)	0.76	0.0
NYHA Class = I	3 (1.8%)	28 (1.5%)	31 (1.6%)	0.96	0.0
NYHA Class = II	12 (7.3%)	84 (4.6%)	96 (4.9%)	0.18	0.0
NYHA Class = III	3 (1.8%)	23 (1.3%)	26 (1.3%)	0.81	0.0
NYHA Class = IV	0 (0.0%)	3 (0.2%)	3 (0.2%)	0.6	0.0
If AMI, please specify = NSTEMI	36 (22.0%)	329 (18.2%)	365 (18.5%)	0.28	0.0
If AMI, please specify = STEMI	1 (0.6%)	6 (0.3%)	7 (0.4%)	0.91	0.0
CCS Class = I	18 (11.0%)	280 (15.5%)	298 (15.1%)	0.15	0.0
CCS Class = II	45 (27.4%)	458 (25.3%)	503 (25.5%)	0.62	0.0
CCS Class = III	16 (9.8%)	130 (7.2%)	146 (7.4%)	0.3	0.0
CCS Class = IV	1 (0.6%)	9 (0.5%)	10 (0.5%)	0.7	0.0
Braunwald classification = IA	3 (1.8%)	22 (1.2%)	25 (1.3%)	0.76	0.0
Braunwald classification = IB	4 (2.4%)	32 (1.8%)	36 (1.8%)	0.76	0.0
Braunwald classification = IIA	5 (3.0%)	28 (1.5%)	33 (1.7%)	0.26	0.0
Braunwald classification = IIB	6 (3.7%)	43 (2.4%)	49 (2.5%)	0.46	0.0
Braunwald classification = IIC	1 (0.6%)	7 (0.4%)	8 (0.4%)	0.83	0.0
Braunwald classification = IIIA	1 (0.6%)	3 (0.2%)	4 (0.2%)	0.76	0.0
Braunwald classification = IIIB-Tneg	12 (7.3%)	116 (6.4%)	128 (6.5%)	0.78	0.0
Braunwald classification = IIIB-Tpos	1 (0.6%)	53 (2.9%)	54 (2.7%)	0.13	0.0
Braunwald classification = IIIC	0 (0.0%)	1 (0.1%)	1 (0.1%)	0.13	0.0
Ischemic status = Documented silent ischemia	14 (8.5%)	283 (15.7%)	297 (15.1%)	**0.02**	0.0

Table 8: Feature characteristics part II

Feature	TLF (n=164)	No TLF (n=1807)	Total (n=1971)	P-Value	Missingness [%]
Ischemic status = Stable angina	80 (48.8%)	879 (48.6%)	959 (48.7%)	0.96	0.0
Ischemic status = Unstable angina	33 (20.1%)	309 (17.1%)	342 (17.4%)	0.38	0.0
Clinically relevant findings = No	117 (71.3%)	1261 (69.8%)	1378 (69.9%)	0.74	0.0
Clinically relevant findings = Yes	38 (23.2%)	356 (19.7%)	394 (20.0%)	0.34	0.0
Clinically relevant findings = No	115 (70.1%)	1363 (75.4%)	1478 (75.0%)	0.16	0.0
Clinically relevant findings = Yes	18 (11.0%)	179 (9.9%)	197 (10.0%)	0.76	0.0
Lesion location = 1st Diag	1 (0.6%)	19 (1.1%)	20 (1.0%)	0.89	0.0
Lesion location = 1st Ob Mar	3 (1.8%)	66 (3.7%)	69 (3.5%)	0.32	0.0
Lesion location = 1st RPL	0 (0.0%)	5 (0.3%)	5 (0.3%)	0.89	0.0
Lesion location = 1st Septal	0 (0.0%)	2 (0.1%)	2 (0.1%)	0.39	0.0
Lesion location = 2nd Ob Mar	0 (0.0%)	14 (0.8%)	14 (0.7%)	0.52	0.0
Lesion location = 3rd Ob Mar	0 (0.0%)	3 (0.2%)	3 (0.2%)	0.6	0.0
Lesion location = Dist CX	2 (1.2%)	55 (3.0%)	57 (2.9%)	0.28	0.0
Lesion location = Dist LAD	1 (0.6%)	20 (1.1%)	21 (1.1%)	0.84	0.0
Lesion location = Dist RCA	10 (6.1%)	102 (5.6%)	112 (5.7%)	0.95	0.0
Lesion location = Mid CX	2 (1.2%)	126 (7.0%)	128 (6.5%)	**0.01**	0.0
Lesion location = Mid LAD	43 (26.2%)	475 (26.3%)	518 (26.3%)	0.94	0.0
Lesion location = Mid RCA	21 (12.8%)	272 (15.1%)	293 (14.9%)	0.51	0.0
Lesion location = Prox CX	8 (4.9%)	111 (6.1%)	119 (6.0%)	0.63	0.0
Lesion location = Prox LAD	56 (34.1%)	360 (19.9%)	416 (21.1%)	**0.0**	0.0
Lesion location = Prox RCA	14 (8.5%)	145 (8.0%)	159 (8.1%)	0.94	0.0
Lesion location = R-PDA	2 (1.2%)	10 (0.6%)	12 (0.6%)	0.6	0.0
Lesion location = Ramus	1 (0.6%)	22 (1.2%)	23 (1.2%)	0.75	0.0
CCS Class = I	4 (2.4%)	64 (3.5%)	68 (3.5%)	0.6	0.0
CCS Class = II	4 (2.4%)	14 (0.8%)	18 (0.9%)	0.09	0.0
CCS Class = III	1 (0.6%)	6 (0.3%)	7 (0.4%)	0.91	0.0
Braunwald classification = IA	0 (0.0%)	1 (0.1%)	1 (0.1%)	0.13	0.0
Braunwald classification = IB	0 (0.0%)	3 (0.2%)	3 (0.2%)	0.6	0.0
Braunwald classification = IIB	0 (0.0%)	1 (0.1%)	1 (0.1%)	0.13	0.0
Braunwald classification = IIIB-Tpos	0 (0.0%)	1 (0.1%)	1 (0.1%)	0.13	0.0
Ischemic status = Documented silent ischemia	0 (0.0%)	4 (0.2%)	4 (0.2%)	0.76	0.0
Ischemic status = Stable angina	9 (5.5%)	84 (4.6%)	93 (4.7%)	0.77	0.0
Ischemic status = Unstable angina	0 (0.0%)	6 (0.3%)	6 (0.3%)	1.0	0.0
Ischemic status = Without pathological findings	155 (94.5%)	1713 (94.8%)	1868 (94.8%)	0.98	0.0

Table 9: Feature characteristics part III

Appendix B Individual Expectations

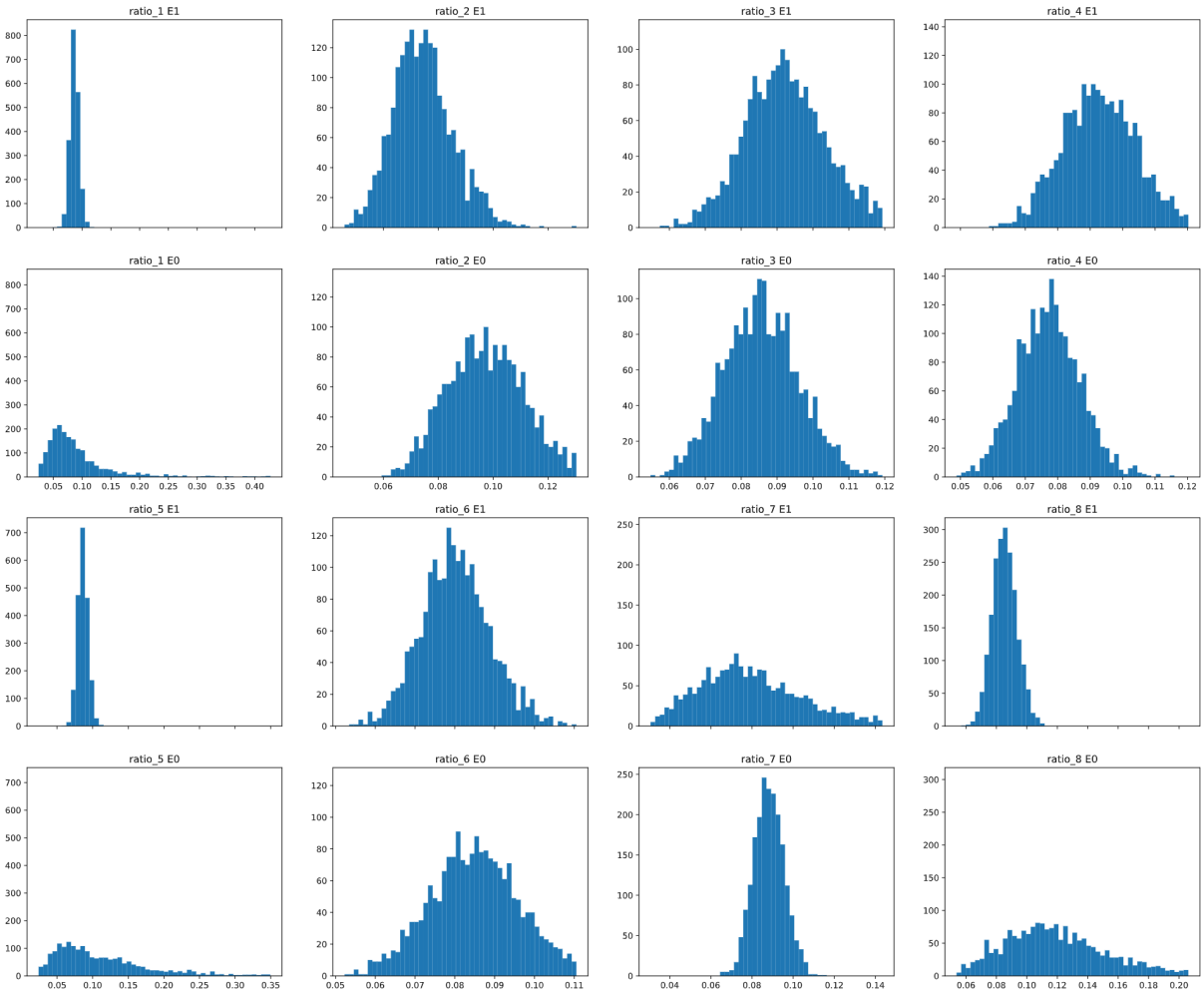

Figure 13: Individual expectations $\mathbb{E}[Y|do(X=1)]$ and $\mathbb{E}[Y|do(X=0)]$ for the 8 different ratios. First row shows the empirical distribution of $\mathbb{E}[Y|do(X=1)]$ for Ratios 1 to 4, whereas second row represents $\mathbb{E}[Y|do(X=0)]$ for the same set of ratios. Third and forth rows show the same quantities, this time for Ratios 5 to 8. For each of the ratios, X and Y axes are shared. X-Axes are scaled for better visualization. $n_{bin} = 50$