Predicting in-hospital mortality after hip fracture

Rowan de Jong

Student Number: 2034275

R.deJong_2@tilburguniversity.edu

Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

Master of Science in Data Science & Society

Department of Cognitive Science & Artificial Intelligence

School of Humanities and Digital Sciences

Tilburg University

Thesis committee:

Supervisor – dr. Marijn van Wingerden

Second Reader – dr. Giacomo Spigler

Tilburg University

School of Humanities and Digital Sciences

Department of Cognitive Science & Artificial Intelligence

Tilburg, the Netherlands

June 2020

Preface

The basis for this study is a collaboration between the Elisabeth-Tweesteden Hospital in Tilburg and Tilburg University. The goal of this study is to give more insight into the prediction of mortality for hip fracture trauma patients. These insights might help to improve the evaluation of trauma care. The dataset I got to work with was certainly interesting and motivated me to work hard. In the end, I am rather pleased with the outcome of this study. I truly believe that this study yields some interesting results and conclusions.

As I would not be able to achieve these results alone, I would like to thank my supervisors for their guidance, support, and sharing their knowledge with me. Especially Marijn van Wingerden and Nanne Jansen for providing feedback, giving advice, and just generally serving as a point of contact in this special time. Furthermore, I would like to thank my fellow students in our thesis committee. Particularly Lisa Janssen, for peer-reviewing my study on multiple occasions.

Abstract

Trauma is one of the leading causes of death around the globe. Thus improving trauma care is extremely important. Improving trauma care starts with a systematic approach of evaluating it. In the Netherlands, this is done by collecting the data of trauma patients in the DTR, calculating a probability of survival for every individual using the TRISS model, and comparing the actual mortality by the expected mortality using the sum of the probabilities of survival (called the Standardized Mortality Ratio or SMR). This approach has its limitations. A hospital that has relatively more patients of a certain subset where the TRISS model structurally underpredict the survival probabilities, is prone to getting a poor evaluation. The hip fracture is such a subset and thus this study tries to create a better mortality prediction model for this subset using machine learning techniques when evaluating on SMR.

The results of this study show that mortality prediction with hip fracture patients could be improved and that using training data that closely represents the test set is the most important part of creating a model with a good SMR score. This supports the literature stating the importance of the observed cohort matching the reference group. Additionally, it shows that the use of variables does not matter that much when trying to optimize for SMR. A model that is biased enough, even to a point that its predictive abilities are heavily impaired, can, and will, still get a good SMR on a test set with similar properties. Thus raising the concerns about the TRISS model being biased to a point that its individual predictions are untrustworthy. This study provides arguments about this statement by showing the lack of predictive and discriminative abilities of the TRISS model and its ineffective way of coding the variables.

Table of Contents

1. Introduction

Trauma is one of the leading causes of death around the globe (Alberdi, Garcia, Atutxa & Zabarte, 2014; WHO, 2014). In 2018, around 80.000 trauma patients were treated in hospitals and trauma centres in the Netherlands (LNAZ, 2019). Out of those 80.000 trauma patients, 4.700 patients were heavily injured and 3% of the trauma patients passed away. These numbers display the importance of trauma care and therefore the importance of the evaluation. Although throughout the years this evaluation has been criticized.

The Elisabeth-TweeSteden Hospital (ETZ) in Tilburg, the Netherlands is a level 1 traumacentre, which roughly means that the hospital has all the resources in order to help injured patients 24 hours a day (ETZ, 2020).

World population has been increasing every year, which also indirectly means that there are more trauma patients. However, since the initiation of trauma-centres the mortality rate in the Netherlands has gone down by 50% for the most severely injured patients of almost all of level 1 hospitals (Hietbrink et al., 2019). The risk of death is shown to be significantly lower when trauma patients receive the care that they need in trauma-centres, compared to trauma patients that were not treated in trauma-centres (MacKenzie et al., 2006). In research done by MacKenzie and colleagues, numbers showed that trauma patients that received treatment in trauma-centres had a lower mortality rate than trauma patients that received treatment in non-trauma centres (7.6% versus 9.5%). Thus, the presence of trauma-centres is extremely important, and thereby the relevance of improving trauma care and trauma-centres is shown.

Improving trauma care starts with a systematic approach of evaluating it. This is done differently in every country. In the Netherlands, data of trauma patients that are treated in emergency departments or trauma centres is gathered in the Dutch Trauma Registry (DTR). The data consists of various variables, such as day of birth, systolic blood pressure and age (LNAZ, 2018). Furthermore, this data is subsequently used to calculate an expected mortality rate, the Probability of Survival (PSNL15).

$$PSNL15 = \frac{1}{(1 + \exp^{-b})}$$

The PSNL15 is a percentage with a range from 0 up to 1, and the variable shows the likelihood of a patient to survive (e.g. the higher the percentage, the more likely a patient is to survive) (LNAZ, 2019). The PSNL15 is calculated by using the predictions of the Trauma and Injury Severity Score (TRISS) (Boyd et al., 1987; Tlson & Copes, 1987; Champion, 1981). The PSNL15 is calculated for every admitted patient that enters trauma care, also when a patient re-enters with other trauma symptoms. The PSNL15 can also be calculated again at

another moment of the patient's hospitalization, and therefore a trauma patient can have multiple PSNL15 scores from different moments.

In the PSNL15 variable formula shown above, *b* is the outcome of a multivariable regression, calculated using the (Revised) Trauma Score, the Injury Severity Score, and the variable Age (Boyd et al., 1987). The Trauma Score was first described in 1981 and then revised in 1989 (Champion, 1981; Champion, 1989). The Revised Trauma Score consists of coded variables based on Systolic Blood Pressure (SBP), Respiratory Rate (RR) and the Glasgow Coma Scale (GCS) which is the sum of Eye Response, Motor Response and Verbal Response (Champion, 1989; Teasdale, 1974; Jennet, 1974).

The Trauma Score seems to correlate with the probability of survival (Boyd et al., 1987). The Injury Severity Score (ISS) is an overall injury severity score calculated from the different injuries. The ISS is proven to perform inadequately when used alone, however, in combination with the (Revised) Trauma Score, both performances increased (Boyd et al., 1987). Furthermore, the formula differs for blunt injuries and penetrating injuries. The coefficients of the formula are shown in table 1 (LNAZ, 2018). Schluter and colleagues (2010) estimated these coefficients using a logistic regression on data from the National Trauma Data Bank and the NTDB National Sample Project.

Table 1

Coefficients of the PSNL15 variable

	Variable	Blunt	Penetrating
β_0	Intercept	1.5090	0.6460
β_1	Respiratory Rate	0.2372	0.2114
β_2	Systolic Blood Pressure	0.6460	0.6806
β_3	Glasgow Coma Score (EMV)	0.4008	0.6333
β_4	Injury Severity Score	-0.1087	-0.0922
β_5	Age	-2.2091	-1.5366

Predictions of the PSNL15 variable are compared to the actual mortality rate by using the Standardized Mortality Ratio (SMR). The SMR is the outcome of the actual mortality rate, divided by the expected mortality rate (LNAZ, 2019). Would the TRISS model be calibrated right, then it would result in a SMR of around 1 (Rogers et al., 2021).

$$SMR = \frac{actual\ mortality}{expected\ mortality} = \frac{actual\ mortality}{\sum(1 - PSNL15)}$$

However, when the actual mortality rate is significantly higher than the expected mortality rate, then that particular hospital is obligated to explain their performance. Although this seems like a reasonable conclusion, there are some limitations in the evaluation process. For instance, table 2 shows that the TRISS model seems to perform well on all the data of the DTR (SMR of 1.087). However, the performance of the TRISS model can differ greatly on different subsets. Subsets that seem especially hard to predict, are those that include older patients, or patients with severe head injuries (de Jongh, Verhofstad & Leenen, 2010). This could result in an unwanted situation where a hospital gets a poor evaluation, which is only due to the fact that their treated patients deviate in characteristics in comparison to the DTR.

A relevant example is the data of patients with hip fractures. This particular subset is a large subset, that contains around 20 - 25% of the total amount of observations in the DTR. This number is expected to rise even more, as the number of people aged over 65 years old is increasing. Another aspect that makes this subset interesting, is that some variables that the TRISS model uses show a different distribution in comparison to the total dataset. The average age is 21 years higher (57 years of age in the total dataset versus 78 years of age in the subset). Next to that, the mortality rate is also greater (2.6% in the total dataset versus 4.1% in the subset). Unsurprisingly, the TRISS model performs poorly on this subset, with a SMR of only 1.293. This suggests that the performance of the TRISS model is dependent on the characteristics of the subset. It thus seems that a hospital that has a patient base with relatively more hip fracture patients than the DTR, is therefore prone to getting a bad evaluation. As this is not desired, it is relevant to create a model that can get a better SMR on this subset.

Table 2

SMR on the entire DTR and on the Hip fracture subset when missing data is imputed with maximum values (Max Values) or is multiple imputed (MI)

	Actual mortality	Expected mortality	SMR
DTR (Max Values)	1251	1148.06	1.090
DTR (MI)	1251	1150.64	1.087
Hip subset (Max Values)	487	376.36	1.294
Hip subset (MI)	487	376.48	1.293

Note. The data from 2015 to and including 2019 from the DTR is used

1.1 Research Questions

In order to improve predictions of mortality on the subset of patients with hip fractures, a number of research questions (RQ) have been formulated below. Each RQ addresses another aspect of the issue, and therefore will each have a contribution in this research.

- RQ 1: Can the predictions of the supervised machine learning models, K-Nearest Neighbors, Random Forest and Support Vector Machines, achieve a SMR closer to 1 than the TRISS model when predicting mortality rate for hip fracture cases using the same variables in the Dutch Trauma Registry as the TRISS model?
  - 1.1: Can the models achieve this using the same coded variables as the TRISS model?
  - 1.2: Can the models achieve this using the uncoded/raw variables?
- RQ 2: Can the predictions of the supervised machine learning models, K-Nearest Neighbors, Random Forest and Support Vector Machines, achieve a SMR closer to 1 than the TRISS model when predicting mortality rate for hip fracture cases while using all the variables in the Dutch Trauma Registry?
  - 2.1: Can the model achieve this using all the variables?
  - 2.2: What is the minimum number of variables needed to achieve a better SMR than the TRISS model?
- RQ 3: Can the best performing supervised machine learning models from the last two research questions, achieve a SMR closer to 1 than the TRISS model when predicting mortality rate for all the different cases in the Dutch Trauma Registry?
- RQ 4: Can the predictions of the supervised machine learning models, K-Nearest Neighbors, Random Forest and Support Vector Machines, achieve a SMR closer to 1 than the TRISS model when predicting mortality rate for hip fracture cases while using all the variables in the Dutch Trauma Registry and being selected on Balanced Accuracy?

2. Related Works

This chapter provides a theoretical background on various aspects that are important for this research. More information is given about the TRISS model and its predictions. Over the years, a lot of research has been done in order to improve the way that professionals predict mortality, this includes some machine learning studies.

2.1 TRISS Model

As mentioned before, the TRISS model has been used in many medical studies. Even though it is widely used, there are multiple limitations. Next to that, there are also a few alternatives to the TRISS model that could work better. This will be discussed in the next subchapters.

2.1.1 Limitations of the TRISS model and SMR

There are several limitations to the use of the TRISS model and the SMR when it comes to evaluating performance even though it has been developed and revised by several researchers. Firstly, regarding SMR, not only is this evaluation metric difficult to interpret due to the non-comparability of the observed cohort and the reference group (Richardson, Keil, Tchetgen & Cooper, 2015), but the SMR is also often biased. For example, if the assumption that the reference group accurately represents the observed cohort does not hold, then there can be a high bias (Kim, Lim, Kang & Khang, 2019). Furthermore, the SMR uses the sum of all expected mortality rates and the results of individual predictions are therefore not as important.

Secondly, there are limitations regarding TRISS model itself. The TRISS model is used to compute the PSNL15. As seen before, the PSNL15 is the base for the expected mortality in the SMR. Therefore, the TRISS model can be seen as the most important element in calculating the so-called reference group in the SMR. The TRISS model has been used widely around the globe in the past 40 years, despite the awareness of the numerous limitations that the model has (Cayten, Stahl, Murphy, Agarwal & Byrne,1991; de Munter, Polinder, Lansink, Cnossen, Steyerberg & de Jongh, 2017; Gabbe, Cameron & Wolfe, 2004).

The TRISS model is proven to lack homogeneity, especially with the subcategory of penetrating injuries (Cayten et al., 1991). This could be explained by the fact that there are not as many patients with penetrating injuries in many countries, which is also the case in the Netherlands. Next to that, the TRISS model lacks predictive ability when predicting for patients with small injuries or patients with severe injuries to just one body part. Additionally, multiple studies have concluded that the TRISS model can discriminate between survivors

and non-survivors, but that it does not have predictive reliability (de Munter et al., 2017; Kuhls, Malone, McCarter & Napolitano, 2002; Hannan et al., 1999).

When looking into the DTR, it was already clear that the TRISS model has difficulties predicting some subsets (SMR of hip fractures is 1.293). Furthermore, its ability to discriminate in the hip fracture subset is also not evident. Table 3 shows the mean PSNL15 for the hip fracture subset, for the patients that passed away and the patients that survived. It shows almost no discrimination between the patients that passed away or survived. Combining this with the probability distribution of the TRISS model, shown in appendix D, there seems to be some good arguments for the discussion that the TRISS model has difficulties discriminating.

Table 3

Mean Probability of Survival for different subsets in the DTR

	Hip subset	Hip and survived subset	Hip and passed away subset
Mean PSNL15	0.970	0.970	0.963

Note. The PSNL15 is calculated using Multiple Imputation for missing data

2.1.2 Alternatives to the TRISS model

The limitations caused an absence of consensus on what model to use in the evaluation of trauma care which caused the development of many more models. Each model has its own strengths and weaknesses and is often specialised for a certain demographic. For instance, the Norwegian Survival Prediction Model in Trauma (NORMIT), which is a model specifically created and validated on Norwegian data (Ghorbani et al., 2014). Furthermore, Glefering and colleagues (2014) created the Revised Injury Severity Score version II (RISC II), a revised RISC model that was developed by using German trauma data.

Additionally, some studies also try to revise the TRISS model in a more general manner. Weeks and colleagues (2014) developed the Kampala Trauma Score (KTS). The KTS is a model that can be used in countries with resource-limited settings, such as third world countries. The KTS model uses information that is typical for patients in resource-limited countries.

Finally, De Munter and colleagues (2017) searched through 90 articles, leading to a literature review of 258 different models. The research of De Munter concluded that most models were based on the same variables the TRISS model uses. Adding to that, most models perform acceptably on the general population. However, their performance on different subsets of the data differs greatly. Especially the probability of survival on subsets

with older patients, such as hip fractures, is commonly overestimated with the TRISS model (de Munter et al., 2018).

2.2 Patients with hip fractures

There is also a discussion on the inclusion of hip fractures in trauma registries. For now, this data is frequently excluded as it seems to have different characteristics than other trauma data (Bergeron et al., 2005; Gomez et al., 2010). The mortality rate for patients with a hip fracture is higher and these patients spend more days in the hospital. Therefore, excluding patients with a hip fracture could have a significant effect on the ranking of a hospital's performance.

Various studies have given insights into the predictions of a subset of patients with hip fracture (Henderson & Ryan, 2010; de Munter et al., 2018; Groff et al., 2020). These studies commonly argue over the importance of the comorbidities of the patients. A significant association between the Charlson Comorbidity Index (CCI) and in-hospital mortality was found, especially for patients with hip fractures (Roffman et al., 2016). For these reasons, models that are developed specifically for the prediction of mortality on patients with hip fractures often include variables stating the comorbidity (Maxwell, Moran & Moppett, 2018).

2.3 Predicting mortality using machine learning techniques

Formerly, regression techniques were primarily used when creating models for predicting mortality. However, when trying to accurately predict an outcome variable, machine learning models can be advantageous over traditional regression models and more studies in medicine have used them (Goldstein, Navar & Carter, 2017). Studies using machine learning or deep learning models for predicting mortality in trauma patients are still rare, however.

Studies that did use machine learning or deep learning techniques have mixed results with them. Rau and colleagues (2019) concluded that their Logistic Regression, Support Vector Machine and Neural Network performed similarly to the TRISS when evaluated on Balanced Accuracy and Sensitivity. The Neural Network created by DiRusso and colleagues (2000) only slightly outperformed the TRISS on ROC. Other studies often focused on specific demographics and did not always include a baseline (Taylor et al., 2016; Pao-Jen et al., 2018).

3. Methods

This chapter shows how this research has been set up, and which methods have been used in order to answer the research questions. Different models have been selected, based on their characteristics suited for this research. The experimental setup with these models is further described in chapter four.

For this research, supervised machine learning models are used. This choice was made because predicting mortality is a classification problem. Therefore, supervised machine learning models are the most appropriate (Brownlee, 2017). The algorithms of interest are K-Nearest Neighbors (KNN), Random Forests (RF) and Support Vector Machines (SVM). The first two are transparent, which is desirable. Support Vector Machines could help to give more accurate predictions. The models are imported from the Scikit-learn library (Pedregosa et al., 2011). The hyperparameter settings that are tuned in this study are displayed in table 4.

3.1 K-Nearest Neighbors

The K-Nearest Neighbors is used as a classification method. The method is known to be one of the simplest classification methods (Hechenbichler & Schliep, 2004; Peterson, 2009). Furthermore, a KNN can offer computational advantages over other classification methods since it only requires little information (Liao & Vemuri, 2002). The KNN classifier takes the k nearest points and makes the prediction of the outcome variable based on the majority vote (Islam, Wu, Ahmadi & Sid-Ahmed, 2007). The k value needs to be tuned, and the choice of k is often made by the use of cross-validation.

3.2 Random Forest

The Random Forest algorithm has been developed by Breiman (2001), in order to improve classification problems by using random sampling. In many cases, datasets contain imbalanced data which can result in poorly performing machine learning algorithms (Livingston, 2005). Since Random Forests use random sampling and attributes selection, imbalanced data can still be classified in a good way. Random Forests simply consists of an ensemble of decision trees, where the decision trees vote for a class. The class that gets the majority vote is chosen as the output and this can significantly increase the score in the desired metric such as accuracy or precision (Pal, 2007).

The most important hyperparameter is the number of estimators, which is the number of trees (Scikit-learn, n.d.). Therefore, this is one of the hyperparameters to be tuned in this study. The other hyperparameter is the maximum depth. This can retain the model from overfitting and reduce the computational time.

3.3 Support Vector Machine

Support Vector Machines (SVM) are a discriminative classifier that is defined by a separating hyperplane, meaning that a SVM outputs an optimal hyperplane that can categorize new examples (Fletcher, 2009). Furthermore, a SVM tries to minimize the classification error and maximize the geometric margin (Vapnik, 1995). SVM's have been used quite frequently in bio-informatics and natural language processing, and are known for their good generalization performance (Burges, 1998). SVM's are also often preferred because they have both a linear and a non-linear function (Furey et al., 2000). The algorithm can, therefore, work well on sparse and high dimensional data.

The hyperparameters used in this study are C and Kernel. C trades off misclassification of training examples against the simplicity of the decision surface (Scikit-learn, n.d). Kernels are a set of mathematical functions that the SVM can use to transform its inputs. Additionally, the linear Kernel is not used in this study because it took too long to compute.

Table 4

Hyperparameters used for the different algorithms

Algorithm	Hyperparameter 1	Hyperparameter 2
K-Nearest Neighbors	K	n/a
Random Forest	Max Depth	Number of Estimators
Support Vector Machine	C	Kernel

4. Experimental Setup

In this chapter, the experimental setup is explained. First, the dataset that is used for this study is described, as well as the subsets. Later on, experiments are discussed for each research question.

4.1 Description of Raw Dataset

The dataset that is being analysed in this study, is extracted from the DTR. The dataset consists of multiple smaller linked datasets, with all the trauma registrations from 2015 up to and including 2019. The data used in this research are the so-called 'findings'. The findings consist of the values of vital parameters in combination with many other variables (108 in total). These vital parameters are then subsequently used for analysis, such as the calculation of the variable PSNL15. The data consists of observations (N = 85135) indexed by patient (case), date of arrival at the emergency department, and whether the measurement of the vital parameters was done in the ambulance or at the emergency department. Thus, a specific case can have multiple occurrences in the dataset, for example when this person has had multiple accidents leading to a trauma registration at different times (less common), or when the vital parameters have been measured in the ambulance and then again at the emergency department (more common).

The PSNL15 is calculated based on unique accidents. Therefore, only one finding is kept for a specific case, with a specific arrival time. Often, this one finding is the measurement at the emergency department, as it is seen as a more precise measurement. Only when this finding is unavailable, the measurements in the ambulance is used. This reduces the size of our observations (N = 55404).

The dataset now has 55404 observations and 108 variables. However, many of these variables are indexes, derived from other variables, or meaningless for prediction. Therefore, it is necessary to create a subset with the outcome variable and variables that are meaningful for predictions. This is done differently for every research question and for the reconstruction of the TRISS model.

4.1.1 Reconstruction of the TRISS model

First, a subset has been created with the TRISS variables and the outcome variable (i.e. Eye Movement, Motor Response, Verbal Response, Respiratory Rate, Systolic Blood Pressure Injury Severity Score, Age, Type of injury, Passed Away). Illegal values are set to NA and all values of '888' and '999', as these values indicate missing data (LNAZ, 2018). Missing data is further analysed and the missing data percentage of every column is extracted (table 5) Then, the missing data is imputed using the MICE package in R. This results in five imputed

datasets that are combined into a single dataset using the average value rounded to the nearest integer.

New variables are created because the TRISS model uses coded variables. EMV (Glasgow Coma Scale) is created by adding the values of Eye Response, Motor Response and Verbal Response. The other variables are coded following table 6. Lastly, the PSNL15 of this multiple imputed dataset is calculated using the coefficients in table 1.

Table 5

Percentage of missing data for every variable

Variable	TRISS model	RQ1	RQ2	RQ3
Eye Response	2.5	3.5	3.5	2.5
Motor Response	2.6	3.5	3.5	2.6
Verbal Response	2.6	3.5	3.5	2.6
Respiratory Rate	33.6	32.1	32	33.6
Systolic Blood Pressure	14.2	5	5	14.2
Injury Severity Score	0.2	0	0	5
Age	0	0	0	0
Type of Injury	0.2	0	0	0.2
Passed Away	0	0	0	0
Cause of injury	n/a	n/a	4	5
Comorbidity (ASA)	n/a	n/a	4.1	5.5
Referred From	n/a	n/a	0.5	0.7
Referrer	n/a	n/a	4.5	5.3
Length of Stay IC	n/a	n/a	9.5	13.4
Level of Hospital Care	n/a	n/a	4.1	4.9
Number of AIS codes	n/a	n/a	0	0
Revised Trauma Score	n/a	n/a	0	0
Level Pre Hospital Care	n/a	n/a	0.5	0.6
Length of Stay ED	n/a	n/a	0.9	0.9
Length of Stay Hospital	n/a	n/a	0.1	0.2
Sex	n/a	n/a	0	0
Time Ambulance	n/a	n/a	37.7	49.7

Table 6

Coding the variables for the reconstruction of the TRISS model

Coded Variable	Systolic Blood Pressure	Respiratory Rate	EMV	Age
0	0	0	3	<54
1	1-49	1-5	4-5	>54
2	50-75	6-9	6-8	n/a
3	76-89	>29	9-12	n/a
4	>89	10-29	13-15	n/a

4.1.2 Research Question 1

The pre-processing of the data for the first research question follows a similar approach as the one recreating the TRISS model. Again, the same subset of variables is used, but now only the observations of patients with hip fractures (*N* = 12408). Next, illegal values and missing data are set to 'NA'. The missing data is analysed and imputed following the same procedure as in the recreation of the TRISS model. Furthermore, coded variables are created and the PSNL15 is calculated. Lastly, the data is split into a training and test set. The test set consists of the 2019 data (*N* = 2256) while the training set consists of the 2015 to 2018 data (*N* = 10147). The choice was made to exclude the penetrating injuries in the training set, as the TRISS model creates the impression that their behaviour is quite different, and with only 5 observations it seems unlikely that the used models are going to pick up on these nuances.

4.1.3 Research Question 2 and 4

Research question 2 and 4 uses the same approach for pre-processing and are using the subset of patients with hip fractures (*N* = 12408). However, the variables used are different which led to a selection of chosen variables. Out of the 108 variables, many were not useful for this research question. Variables that have over 70% of missing data, variables that were indices, and variables that were deemed meaningless for prediction were excluded from the subset. In appendix A, a list with the dropped variables and the reason is published. When variables correlated, the variable that predicted best in a univariate logistic regression was kept. Furthermore, all the categorical variables were individually assessed on their predictive ability and class (in)balance. Most of the categorical variables were recoded to compromise for their class imbalances, but some had to be deleted. More information can be found in appendix A and B.

This resulted in 22 variables of which their missing data is analysed following the same procedure as in the first research question. After creating five imputed datasets using MICE, the different datasets were combined into one. For ordinal variables, the mean of the values was rounded to the nearest integer. For categorical variables, the mode of the values was chosen. Furthermore, the categorical variables that only had two classes left were binary coded and the categorical variables with more than two classes were one-hot coded.

Additionally, the data was split into a training and test set. Again, the test set consists of the 2019 data (*N* = 2256) and the training set of the data from 2015 up to and including 2018 (*N* = 10147), excluding the penetrating injuries.

4.1.4 Research Question 3

The pre-processing of the data for the third Research Question follows the exact approach as the second. The only difference is that the used observations are the entire DTR dataset (*N* = 55404), and in the end, the data is only split into a test split of the 2019 data (*N* = 8501).

4.2 Experimental Procedure

The creation, training, validation and testing of the different models for all the research questions follow a similar approach. Firstly, the selection of the supervised machine learning algorithm. Secondly, choosing the models hyperparameter settings and sampling hyperparameter settings. Thirdly, creating a train/validation split. Lastly, creating a subset for training and validating the model.

4.2.1 Algorithms and Hyperparameters

As mentioned in chapter three, the three algorithms that are used in this study are K-Nearest Neighbors, Random Forest and Support Vector Machine. For every algorithm, there are a number of hyperparameters chosen, shown in table 4.

After the set of hyperparameters are chosen, the data will be split into a training and validation set using 3-fold cross-validation, stratified on all the variables in the model. 3-fold cross-validation was chosen as using more folds resulted in more variance on the scores between the difference splits, and it increases processing time.

After splitting the data into a training and validation set, the data that is used for training is sampled into a subset. This sampling is done to help overcome the class imbalance in the outcome variable. This sampling is combined with Synthetic Minority Over-sampling Technique (SMOTE) following the guidelines of Chawla and colleagues (2011). The idea is to create augmented data in the minority class relative to the majority class and then resampling the majority class (Chawla et al., 2011; Browlee, 2020). SMOTE is regularly

combined with random under-sampling of the majority class but is in this research both undersampling and oversampling of the majority class is used.

In summary, the data is split into a training and validation set. Additionally, the training set is split into the majority and the minority class, synthetic data is created in the minority class, and the majority class is randomly resampled (with replacement) to a specific ratio of the minority class (with synthetic data). The data is combined and shuffled into a single training subset which is used for training the model. The trained model is then used to calculate the SMR of the validation set.

Subsets within the same training/validation split, with the same amount of synthetic data, and the same amount of resampling of the majority class, can still differ from each other. Therefore, multiple subsets are necessary for a fair model selection.

As the difference in the structure in a subset matters greatly for the SMR of the trained model, the amount of synthetic data created and the amount of resampling of the majority class are treated as hyperparameters in the analysis.

4.2.2 Research Question 1

First off, the three algorithms loop through many different combinations of hyperparameters, resulting in the creation of about 1000 different models each trained on three subsets for every of the three training splits (nine subsets total). The nine different SMR validation scores are combined and the 95% quantile confidence interval is calculated. The best five models for each algorithm are extracted based on the error function below.

$$Error = (CI_{lower} - 1)^2 + (CI_{upper} - 1)^2$$

This error function calculates the squared deviation from one. This metric is chosen because a SMR score of one is the best achievable result. So the closer a model's performance approaches one, the better the model. Additionally, the deviation is squared because a confidence interval surrounding one is arguably more favourable than a confidence interval barely including one (e.g. 0.85 – 1.15 is better than 0.70 – 1.00). Therefore, exclusively the SMR is the criteria for the model selection as this is also the metric used in the evaluation of trauma care in the Netherlands. However, individual models of a specific algorithm are fitting the data using different methods. The Random Forest uses cross-entropy, the Support Vector Machine aims to maximize the geometric margin between classes, and the K-Nearest Neigbors does not explicitly learn but it just stores the training dataset.

The five best models of each algorithm explain what combination of hyperparameters works. So these hyperparameters are extracted from the model and used to run the same

code again, but now using 15 subsets for each training/validation split. Again, these 45 SMR scores on the validation tests are combined into a confidence interval and the best performing model for each algorithm is extracted using the same approach as above.

Additionally, these final models are trained using the same hyperparameter settings, on all the training data, for 50 subsets, and then tested on the independent test set. Again, the confidence interval of these 50 SMR scores is extracted and compared to the performance of the TRISS model. Appendix C contains a schematic of this approach. If both of the deviations in the confidence interval is lower than the deviation from one of the TRISS model, it can be said that the model is outperforming the TRISS model. For example, when the TRISS model has an absolute deviation from one of 0.20 then the compared model is only performing better if the lower confidence interval is higher than 0.8 and the higher confidence interval is lower than 1.2.

4.2.3 Research Question 2

For the first part of the second research question, the same approach as in research question 1 is used for all the variables in appendix B. Appendix B also shows the descriptives of the variables.

For the second part, the best performing Random Forest from the first part is used to create a feature ranking for every subset. This feature ranking is pooled over all the subsets and the least important feature is dropped. This approach continues until only one variable is left. This will result in a feature ranking. This feature ranking is used to try to create a model with the least amount of variables possible that can still outperform the TRISS model on the test set. After choosing the variables, the same approach of model selection as in the first part and the first research question is used.

4.2.4 Research Question 3

This research question simply consists of selecting the best performing model of each research question and then testing it over 50 subsets on the 2019 data of the entire DTR.

4.2.4 Research Question 4

Research questions 4 is similar to the second research question. The difference is that in this research question Balanced Accuracy is used for the model selection and not SMR (Brodersen et al., 2010)

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

The models are still evaluated using SMR on the test set. The results will show the effect on SMR, an evaluation metric that does not care for individual prediction, when the model selection uses a relatively simple evaluation metric that is based on individual predictions.

5. Results

One of the earliest findings is that the hyperparameter settings for the model (number of estimators, depth, k, etc) are less important than the hyperparameter settings of the subset (SMOTE and ratio of resampling). Especially the ratio of resampling of the majority class seemed to have the biggest impact. This can be the result of the SMR needing to have a well-represented reference group as training data. Additionally, the majority/minority class ratio is also the most important variable for the stratification of the test/validation sets. Moreover, creating synthetic data in the minority class using SMOTE resulted in models with less variance, probably due to the increase in bias or/and the increase of training data in individual subsets.

5.1 Research Question 1

The (multiple imputed) TRISS model scored a SMR of 1.298 on the test set (2019 hip fracture subset). This is consistent with the SMR of the entire hip fracture subset (2015 – 2019), which is 1.293.

Although the TRISS model uses the coded variables, there was no other model that performed adequately using these variables (table 9). An analysis of the data showed that one specific feature row of these coded variables, shown in table 7, has 10,457 occurrences (total observations in the hip subset is 12,408). Furthermore, table 8 shows the distribution of the outcome variable with this specific row, there seems to be no discrimination in the classes. This makes it extremely difficult for models to make predictions with these variables. Therefore, the performance of these models was poor on the validation set and not tested on the test set (results are displayed in table 9).

The models with the uncoded variables did better. The parameters and the performance of the best performing models are shown in table 10. The Random Forest and the K-Nearest Neighbors, perform adequate and are outperforming the TRISS model. Additionally, the Support Vector Machine makes the worst predictions. Lastly, the SMR is lower on the test set for all of the models. Thus all of the models seem to overpredict the mortality on the test set in comparison to the validation set.

Table 7

Feature row that occurs more than 84% of the time

ISS	Type of Injury	EMV (coded)	Systolic Blood Pressure (coded)	Respiratory Rate (coded)	Age (coded)
9	0	4	4	4	1

Table 8

Occurrences of the feature row of table 7 in the hip fracture subset

Subset	Frequency of observed feature row	Total Frequency of rows	Relative
Survived	10044	11921	84.3%
Passed away	412	487	84.6%
Total	10457	12408	84.3%

Table 9

Results of Research Question 1 for the coded TRISS variables

Model	Ratio	Smote	Hyperparameter 1	Hyperparameter 2	CI val	
RF	1	0.06	50	50	0.042	1.309
KNN	1	0.15	3	None	0.041	1.644
SVM	1	0.15	12	rbf	0.041	0.853

Table 10

Results of Research Question 1 for the uncoded TRISS variables

Model	Ratio	Smote	Hyperparameter 1	Hyperparameter 2	CI val		CI test	
RF	12	0.15	100	50	0.852	0.999	0.742	0.848
KNN	8	0.15	9	None	0.912	1.109	0.820	1.018
SVM	2	0.1	12	rbf	0.495	1.073	0.561	0.881

5.2 Research question 2

When all the selected variables are used for predicting, the SMR of the models did not drastically improve. As is shown in table 11, the Random Forest and the K-Nearest Neighbors are the best predictor and are outperforming the TRISS model. The performance of the Support Vector Machine is, once again, underwhelming. These results suggest that the TRISS model is already using some of the best predictors, or that the selection of variables is not that important.

Table 11

Results of Research Question 2 part 1 (all the selected variables)

Model	Ratio	Smote	Hyperparameter 1	Hyperparameter 2	CI val		CI test	
RF	5	0.12	100	50	0.801	1.167	0.949	1.107
KNN	8	0.10	4	None	0.88	1.118	0.847	1.138
SVM	2	0.15	12	rbf	0.493	1.099	0.507	0.887

Therefore, for part 2, a feature ranking is created by repeatedly dropping the least important variable. The results are a feature ranking that can give an insight into the importance of the variables. The results are in table 12. Three of the last five dropped variables are variables that are also used in the TRISS mode, suggesting that the TRISS model uses already some highly predictive variables.

Table 12

Feature ranking based on the best performing Random Forest of RQ2 part 1. The upper variables are dropped first (and thus less important).

Type of Injury
Other (Cause of Injury)*
Motor Response
Revised Trauma Score
Eye Response
Level Pre Hospital Care
Lowfall (Cause of Injury)*
Injury Severity Score
Traffic (Cause of Injury)*
Verbal Response
Number of AIS codes
Referrer
Referred From
Length of Stay IC
Comorbidity (ASA)
Sex
Length of Stay Hospital
Time Ambulance
Respiratory Rate
Age

Systolic Blood Pressure
Length of Stay ED
Level of Hospital Care

Note. Cause of Injury is one hot coded into “Other”, “Lowfall” and “Traffic”

For part 2, Systolic Blood Pressure and Length of Stay ED are chosen. Level of Hospital Care is not used as it did not improve the predictions. Furthermore, the decision was made to not include the SVM, as its performance in the previous part was already inadequate. The results are shown in table 13. Again, the Random Forest and K-Nearest Neighbors outperform the TRISS model on the test set.

Table 13

Results of Research Question 2 part 2

Model	Ratio	Smote	Hyperparameter 1	Hyperparameter 2	CI val		CI test	
RF	15	0.08	50	50	0.863	1.136	0.849	1.051
KNN	7	0.12	9	None	0.815	1.062	0.915	1.250

5.3 Research Question 3

The models that were tested on the 2019 hip fracture subset are also tested on all the observations of the 2019 DTR. The results are shown in table 14 with the result of the TRISS model.

There seems to be a systematic overprediction of mortality for the 2019 DTR. This is expected as the models are trained and optimized on data with a mortality of 3.9%, tested on the 2019 hip fracture subset with a mortality of 3.9%, and now tested on the 2019 DTR that has a mortality of 2.1%. It seems that the use of the SMR as an evaluation created a situation where models that are seriously biased, can still be seen as good predictors. Therefore, these results do support the literature stating the limitations of the use of SMR (Richardson et al., 2015; Kim et al., 2019).

The best example of this (and one specifically created for this point), are the models that only used two variables. These models are not able to make correct individual predictions but are trained to output a mortality rate of around 3.9%. Therefore, when the models are tested on the 2019 hip subset, the models seem to perform great. However, when tested on the 2019 DTR, which has around half the mortality, the SMR is only about half the amount it was. Following this logic, a lower drop in SMR between the test sets might indicate some predictive ability.

Table 14

All the test SMR scores

	2019 Hip fracture		2019 entire DTR	
TRISS	1.298		1.125	
RF	Lower CI	Higher CI	Lower CI	Higher CI
Trained on TRISS variables	0.742	0.848	0.693	0.814
Trained on all selected variables	0.949	1.107	0.305	0.351
Trained on Systolic Blood pressure and age	0.849	1.051	0.480	0.557
KNN				
Trained on TRISS variables	0.820	1.018	0.723	0.853
Trained on all selected variables	0.847	1.138	0.808	0.993
Trained on Systolic Blood pressure and age	0.915	1.250	0.418	0.524
SVM				
Trained on TRISS variables	0.495	1.073	0.485	0.724
Trained on all selected variables	0.507	0.887	0.507	0.887

Another indication of predictive ability could be the probability outputs of the models, they can be found in appendix D. This shows that the TRISS model barely discriminates between different observations. Furthermore, it shows that the models that are trained on the uncoded TRISS variables and all the variables have much more variance in their probability distribution than the models trained on two variables and the TRISS model.

Lastly, Appendix E shows all the accuracies of the different models on the two test sets and their confusion matrix. It shows that all the results of the models on the two different test sets always resulted in a lower accuracy than the majority class. Suggesting that optimizing on SMR can negatively affect the predictive abilities of a model.

5.3 Research Question 4

When all the selected variables are used again for predicting mortality and the model selection is based on Balanced Accuracy, the Balanced Accuracy score of the models on the test set is fair. However, the SMR of the models on the test set did decrease. Table 15 shows that both the Random Forest and the Support Vector Machine can get a decent score on Balanced Accuracy but that this negatively affected the SMR. Additionally, it shows that The KNN has the lowest score on Balanced Accuracy but that it does have the best SMR.

The low SMR score on the test set is due to overprediction of mortality and this overprediction is caused by the class imbalances. As the test set only has a low amount of cases that pass away, the penalty of missing such a case in the overall evaluation is greater than the penalty of misclassifying a case that survived. Resulting in the overprediction of mortality. This research question is, therefore, showing the difficulties in setting up a good evaluation matrix.

Table 15

Results of Research Question 4

Model	Ratio	Smote	Hyperparameter 1	Hyperparameter 2	CI BA* test		CI SMR test	
RF	1	0.05	100	50	0.726	0.771	0.149	0.179
KNN	7	0.12	5	None	0.544	0.571	0.564	0.664
SVM	1	0.30	12	poly	0.725	0.737	0.175	0.206

Note. BA = Balanced Accuracy

6. Discussion

In this chapter, the results will be evaluated in regard to the research questions and the overall goal of this study. Furthermore, the limitations of this study will be acknowledged.

As the results have shown, in the current evaluation process of Trauma Care, it is possible for a hospital to get a poor evaluation just because their structure in the client base is different than the DTR. This is the case when a hospital has relatively more patients of a subset where the mortality is structurally underpredicted by the TRISS model (e.g. hip fractures) than the entire DTR. Therefore, the goal of this study was to create a model that performs better in terms of SMR on the hip subset data than the TRISS model.

In research question 1, the same variables of the TRISS model (coded and uncoded) in combination with the chosen supervised machine learning models (KNN, RF, SVM) were used to improve the TRISS model. When the models used the coded variables, the results were poor. It seems that the coding of the variables made the variables less informative. It even led to a specific feature row occurring in more than 84% of the observations. Using the uncoded variables, the results were better. The best Performing Random Forest model and the K-Nearest Neigbors model were outperforming the TRISS model in terms of SMR. Unfortunately, the Support Vector Machine was not.

Part 1 of research question 2, followed a similar approach as the first research question, only with a different selection of variables. Once again, the Random Forest and the K-Nearest Neighbors performed better than the TRISS model. The Support Vector Machine was once again highly variate.

As part 1 raised questions about the importance of the variables, part 2 was set up to answer them. The feature ranking led to the conclusion that the TRISS model might be using some great predictors. More surprisingly, part 2 showed that it was possible to outperform the TRISS model on SMR while using only two variables.

Research question 3 and the literature gave more insights about the results and especially how a model with only two variables was able to a good SMR. Apparently, it is rather straightforward to create a model that can get a good SMR. Especially when it is trained on data that is similar to the observed cohort. The key is to impute much bias in the model. Even if this results in a model that has impaired predictive abilities. Therefore, the best performing models of research question 1 and 2 were not able to generalise well on the entire DTR.

This raises the concern that the current evaluation method, the TRISS model, is also just a biased model with impaired predictive abilities. The results concerning its SMR score on different subsets, the distribution of its predictions, its use of uninformatively coded variables, and other limitations brought up by the literature, suggests that the TRISS model

itself is not doing much more than just imputing the overall mortality rate as a probability of survival for each individual.

This has the significant clinical implication that the evaluation of the performance of a hospital is only accurate when the client base has the same structure as the DTR. However, as the DTR consists of highly heterogeneous data, this will rarely be the case. Even if the observed cohort of a hospital follows a similar structure as the DTR, the explanations of a hospital's performance will prove to be extremely difficult when the individual predictions are not trustworthy. Furthermore, as this study shows, it is difficult to develop a model that can make reliable individual predictions when using the SMR.

Additionally, it is not as straightforward to create a model that can make better individual predictions and still get a good SMR. This study showed that optimising models on Balanced Accuracy generally resulted in a decrease in SMR. Therefore, this study briefly shows the difficulties in developing an evaluation metric that will select models that can get a good SMR and still make accurate individual predictions.

With these results, this study supports the existing literature stating the importance of having a good reference group as training data, it emphasizes some to of the limitations of the TRISS model and the SMR, and gives some arguments why there have been such a widespread of trauma evaluation models. Lastly, it is one of the limited studies using machine learning techniques in de prediction of mortality in trauma patients and one of the first that primarily focusses on SMR.

6.1 Limitations

The first limitation is the choice to evaluate models on the SMR and not an individual evaluation metric. This choice was made because hospitals in the Netherlands are evaluated on SMR and using the same evaluation metric makes this study more relevant in this field. Unfortunately, this resulted in models with impaired predictive abilities Secondly, only a limited amount of different hyperparameters and hyperparameter settings were tried. Furthermore, no real effort was made to change the inputs (such as normalisation and standardisation) to make it easier for the algorithms to learn. This could have resulted in less variance and thus more stable results, which in turn could especially have helped the Support Vector Machine to get more interesting predictions, as its performance was now just underwhelming. Another limitation is the lack of knowledge about the individual variables. Better knowledge would have helped in choosing and recoding the variables. Now certain combinations of data might be controversial. Lastly, the assumption that all the missing data is missing at random is made in this study. Although looking at individual variables, this assumption is just naïve.

7. Conclusion

This study shows that it is possible to create a model that performs better on the hip fracture subset using SMR as an evaluation. However, as the model is optimized using SMR, it will probably be fairly biased. If the model is used as a reference group while trying to evaluate an observed cohort with the same structure/characteristics, this should not be a problem. If this assumption does not hold, the model should not be used.

Therefore, this study can be used as an argument to exclude the hip fracture subset from the DTR and perhaps be evaluated independently. However, this will raise the question which other subsets should also be excluded and what models should be used.

Furthermore, this study could be used to advocate against the use of the SMR. An evaluation metric that is prone to choose models that are highly biased and discard their predictive ability, might not be the best choice for the evaluation of trauma care. In this case, the overall goal of the evaluation should be clear. Is a simple comparison between different hospitals enough, or is more detailed evaluation considering individual subsets and even individual patients the goal? If the latter is the case, individual predictions should also be taken into account.

Lastly, a remark about the TRISS model. During this study, the TRISS model was often negatively mentioned for its limited predicting and discriminative abilities and the use of uninformatively coded variables. The TRISS model seems highly biased and might only be performing well (in its current state) because of the SMR. However, the TRISS model is also remarkably robust and uncommonly interpretable. It seems to use highly predictive variables that are easily gathered. This study also showed that a simple algorithm can create adequate predictions. The limited predictive and discriminative abilities of the TRISS model can probably be strongly increased when using differently coded variables. Therefore, the TRISS model might have a future in the prediction of trauma care.

References

Al Khady, M., Kambhampati, C. (2018). Resampling Imbalanced Class and the Effectiveness of Feature Selection Methods for Heart Failure Dataset. *International Robotics & Automation Journal, 4*(1), doi:/10.15406/iratj.2018.04.00090

Alberdi, F., Garcia, I., Atutxa, L., Zabarte, M. (2014). Epidemiology of severe trauma in Spain. Registry of trauma in the ICU (RETRAUCI). *Medicina Intsensiva, 38*(9), 580-588. doi: 10.1016/j.medin.2014.06.012.

Bergeron, E., Lavoie, A., Belcaid, A., Ratte, S., Clas, D. (2005). Should patients with isolated hip fractures be included in trauma registries? *Journal of Trauma, 58*(4). doi:10.1097/01.TA.0000158245.23772.0A

Brodersen, K.H.; Ong, C.S.; Stephan, K.E.; Buhmann, J.M. (2010). The balanced accuracy and its posterior distribution. Proceedings of the 20th International Conference on Pattern Recognition, 3121-24.

Boyd, C. R., Tlson, M. A., Copes, W. S. (1987). Evaluating Trauma Care: The TRISS Method. *Journa of Trauma and Acute Care Surgery, 27*(4), 370-378.

Breiman, L. (2001). Random Forest's. *Machine Learning. 45*(1). 5-32.

Browlee, J. (2017). *Machine Learning Mastery: Master Machine Learning Algorithms, Discover how they work and Implement them from Scratch*. Packt Publishing.

Browlee, J. (2020). SMOTE for Imbalanced Classification with Python. Retrieved from https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

Burges, C.J.C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery. 2*(1), 121-167. doi:/ /10.1023/A:1009715923555

Cayten, C.G., Stahl, W.M., Murphy, J.G., Agarwal, N., Byrne, D.W. (1991). Limitations of the TRISS Method for Interhospital Comparisons: a Multihospital Study. *The Journal of Trauma 31(*4), 471-481. doi:/10.1097/00005373-199104000-00005

Champion, H.R., Copes, W.S., Sacco, W.J., Lawnick, M.M., Keast, S.L., Bain, L.W., Flanagan, M.E., Frey, C.F. (1990). The Major Trauma Outcome Study: establishing national norms for trauma care. *The Journal of Trauma 30*(11), 1356-1365. DOI:/2231804

Champion, H.R., Sacco, W.J., Carnazzo, A., Copes, W., Fouty, W. (1981). Trauma Score. *Critical Care Medicine*, *9*(9), 672-676. doi:/10.1097/00003246-198109000-00015

Champion, H.R., Sacco, W.J., Copes, W.S., Gann, D.S., Gennarelli, T.A., Flanagan, M. E. (1989). A revision of the Trauma Score. *The Journal of Trauma*, *29*(5), 623-629. Doi:/10.1097/00005373-198905000-00017

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W.P. (2011). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357. Doi:/10.1613/jair.953

de Jongh, M. A., Verhofstad, M. H. J., & Leenen, L. P. (2010). Accuracy of different survival prediction models in a trauma population*. British journal of surgery*, *97*(12), 1805-1813.

de Munter, L., Polinder, S., Lansink, K. W., Cnossen, M. C., Steyerberg, E. W., de Jongh, M. A. (2017). Mortality prediction models in the general trauma population: A systematic review. *Injury, 48*(2), 221-229. https://doi.org/10.1016/j.injury.2016.12.009

de Munter, L., Polinder, S., Nieboer, D., Lansink, K.W.W., Steyerber, E.W., de Jongh, M.A.C. (2018). Performance of the modified TRISS for evaluating trauma care in subpopulations: A cohort study. *Injury, 49*(9), doi.org/10.1016/j.injury.2018.03.036/

de Munter, L., ter Bogt, N.C.W., Polinder, S., Sewalt, C.A., Steyerberg, E.W., de Jongh, M.A.C. (2018). Improvement of the performance of survival prediction in the ageing blunt trauma population: A cohort study. *Plos One, 13*(12). doi:10.1371/journal.pone.0209099

DiRusso, S.M., Sullivan, T., Holly, C., Cuff, S.N., Savino, J. (2000) An artificial neural network as a model for prediction of survival in trauma patients: Validation for a regional trauma area. *The Journal of Trauma 49*(2), 212-223.

ETZ. (2020). Over ETZ, wie wij zijn. Retrieved from https://www.etz.nl/Over-ETZ.

Fletcher, T. (2009). Support Vector Machines Explained. Tutorial Paper.

Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Hausslet, D. (2000). Support Vector Machine Classification and Validation of Cancer Tissue Samples using Microarray Expression Data. *Bioinformatics. 16*(10). 906-914.

Gabbe, B.J., Cameron, P.A., Wolfe, R., (2004). TRISS: Does It Get Better than this? *ACAD EMERG MED* 11(2). doi.org/10.1111/j.1553-2712.2004.tb01432.x

Ghorbani, P., Troëng, T., Brattström, O., Ringdal, K.G., Eken, T., Ekbom, A., Strömmer, L. (2014) Validation of the Norwegian survival prediction model in trauma (NORMIT) in Swedish trauma populations, *Acta Anastheosiologica Scandinavica*, *58*(13), 381-390. https://doi.org/10.1111/aas.12256

Goldstein, B.A., Navar, A.M., Carter, R.E. (2017). Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European Heart Journal 23*(14), 1805-1814. https://doi.org/10.1093/eurheartj/ehw302

Gomez, D., Haas, B., Hemmila, M., Pasquale, M., Goble, S., Neal, M., Mann, N.C., Meredith, W., Cryer, H.G., Shafi, S., Nathens, A.B. (2010). Hips can lie: impact of excluding isolated hip fractures on external benchmarking of trauma center performance. *Journal of Trauma, 69*(5), 1037-1041.

Groff, H., Kheir, M.M., George, J., Azboy, I., Higuera, C.A., Parvizi, J. (2020). Causes of in-hospital mortality after hip fractures in the elderly. HIP International, *HIP International, 30*(2). https://doi.org/10.1177/1120700019835160

Hacibeyoglu, M., Ibrahim, M.H. (2018) The Effect of Over-Sampling and Under-sampling Techniques in Medical Datasets, presented at the International Conference on Advanced Technologies, Computer Engineering and Science, Safranbolu, 2018, Turkey: ICATCES'18

Hannan E., Farrell L., Gorthy S., et al. (1999). Predictors of mortality in adult patients with blunt injuries in New York State: a comparison of the Trauma and Injury Severity

Score (TRISS) and the International Classification of Disease, Ninth Revision-based Injury Severity Score (ICISS). *J Trauma Inj Infect CritCare*, *47*(1).

Hechenbichler, K., Schliep, K. (2004). Weighted k-Nearest-Neighbor Techniques and Ordinal Classification. *Collaborative Research Center. 386*(1). 399:. doi:/10.5282/ubm/epub.1769

Henderson, C. Y., Ryan, J. P. (2010). Predicting mortality following hip fracture: an analysis of comorbidities and complications. *Irish Journal of Medical Science, 184,* 667-671. doi:10.1007/s11845-015-1271-z https://doi.org/10.1016/S1072-7515(02)01211-5

Hietbrink, F., Houwert, R.M., van Wessem, K.J.P., Simmermacher, R.K.J., Govaert, A.M., de Jong, M.B., de Bruin, I.G.J., de Graaf, J., Leenen, L.P.H. (2019). The Evolution of Trauma Care in the Netherlands over 20 years. *European Journal of Trauma and Emergency Surgery. 46*(1). 329-335.

Islam, M.J., Wu, J.Q.M., Ahmadi, M., Sid-Ahmed, M.A. (2007). Investigating the Performance of Naïve-Bayes Classifiers and K-Nearest Neighbor Classifiers. *International Conference on Convergence Information Technology,* Gyeongju, 2007. 1541-1546. doi: 10.1109/ICCIT.2007.148.

Kim, I., Lim, H., Kang, H., & Khang, Y. (2019). Comparison of three small-area mortality metrics according to urbanity in Korea: the standardized mortality ratio, comparative mortality figure, and life expectancy*. Manuscript submitted for publication.* doi:10.21203/rs.2.19405/v2

Kuhls D., Malone D., McCarter R., Napolitano L. (2002) Predictors of mortality in adult trauma patients: The Physiologic TraumaScore is equivalent to the Trauma and Injury Severity Score. *J Am Coll Surg*. *194*(6), 695-704.

Kuo, P.J., Wu, S.C., Chien, P.C., Rau, C.S., Chen, Y.C., Hsieh, H.Y., Hsieh, C.H. (2018). Derivation and validation of different machine-learning models in mortality prediction of trauma in motorcycle riders: A cross-sectional retrospective study in southern Taiwan. *BMJ open 8(*5). doi:10.1136/bmjopen-2017-018252

Lefering, R., Huber-Wagner, S., Nienaber, U., Maegele, M., Bouillon, B. (2014) Update of the trauma risk adjustment model of the TraumaRegister DGU™: the Revised Injury

Severity Classification, version II. *Crit Care, 476*(18). https://doi.org/10.1186/s13054-014-0476-2

Liao, Y., Vemuri, V.R. (2002). Use of K-Nearest Neighbor classifier for intrustion detection. *Computers & Security. 21*(5). 439-448. doi:/10.1016/S0167-4048(02)00514-X.

Livingston, F. (2005). Implementation of Breiman's Random Forest Machine Learning Algorithm. *Machine Learning Journal Paper. 1*(1). 1-13.

LNAZ. (2018*). Datadictionary LTR European dataset: versie 2.6.* Tilburg, the Netherlands: LNAZ

LNAZ. (2018). *Documentatie Afgeleide variabelen 'LTR European dataset' Landelijke Traumaregistratie.* Tilburg, the Netherlands: LNAZ

LNAZ. (2019). *Traumazorg in beeld : LTR Regiorapport 2014-2018*. Tilburg, the Netherlands: LNAZ

MacKenzie, E.J., Rivara, F.P., Jurkovich, G.J., Nathens, A.B., Frey, K.P., Egleston, B.L., Salkever, D.S., Scharfstein, D.O. (2006). A National Evaluation of the Effect of Trauma-Center Care on Mortality. *The New England Journal of Medicine. 78*(4). 354-366. doi:

Maxwell, M.J. Moran, C.G., Moppett, I.K. (2008). Development and validation of a preoperative scoring system to predict 30 day mortality in patients undergoing hip fracture surgery. *British Journal of Anaesthesia 101*(4), 511-517. https://doi.org/10.1093/bja/aen236

Pal, M. (2007). Random Forest Classifier for Remote Sensing Classification. *International Journal of Remote Sensing. 26*(1). 217-222. doi: /10.1080/01431160412331269698.

Pedregosa, F., Varoquax, G., Gramfort. A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourd, V., Vanderplas, J., Passos, A., Cournapeau, D. Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research 12,* 2825-2830.

Peterson, L.E. (2009). K-Nearest Neighbor. *Scholarpedia. 4*(2). :1883. doi:10.4249/scholarpedia.1883

Rau, C. S., Wu, S. C., Chuang, J. F., Huang, C. Y., Liu, H. T., Chien, P. C., & Hsieh, C. H. (2019). Machine learning models of survival prediction in trauma patients. *Journal of clinical medicine*, *8*(6), 799.

Ratan, S.K., Anand, T., Ratan, J. (2019). Formulation of Research Question – Stepwise Approach. *J. Indian Association Pedriatr. Surgery, 24*(1), 15-20. doi:/10.4103/jiaps.JIAPS_76_18

Richardson, D.B., Keil, A., Tchetgen, E., Cooper, G.S. (2015). Negative control Outcomes and the Analysis of Standardized Mortality Ratios, *Epidemiology, 26*(5), 727-732. doi: 10.1097/EDE.0000000000000353

Roffman, C.E., Buchanan, J., Allison, G. T. (2016). Charlson Comorbidities Index. *Journal of Physiotherapy, 62*(3), 171. https://doi.org/10.1016/j.jphys.2016.05.008

Rogers, F. B., Osler, T., Krasne, M., Rogers, A., Bradburn, E. H., Lee, J. C., ... Horst, M. A. (2012). Has TRISS become an anachronism? A comparison of mortality between the National Trauma Data Bank and Major Trauma Outcome Study databases. *Journal of trauma and acute care surgery, 73*(2), 326-331. doi: 10.1097/TA.0b013e31825a7758.

Scikit-learn (n.d.) Ensemble methods. Retrieved form https://scikit-learn.org/stable/modules/ensemble.html#forest

Scikit-learn (n.d.) 1.4. Support Vector Machines. Retrieved form https://scikit-learn.org/stable/modules/svm.html#svm-classification

Schluter, P.J., Nathens, A., Neal, M.L., Goble, S., Cameron, C.M., Davey, T.M., McClure, R.J. (2010). Trauma and Injury Severity Score (TRISS) coefficients 2009 revision. *Journal of Trauma, 68*(4), 761-770. doi:10.1097/TA.0b013e3181d3223b.

Sternbach, L.S. (2000). The Glasgow Coma Scale. *The Journal of Emergency Medicine, 19*(1), 67-71. https://doi.org/10.1016/S0736-4679(00)00182-7

Teasdale G., Jennett B. (1974). Assessment of coma and impaired consciousness. A practical scale. *The Lancet 304*(7872), 81-85. https://doi.org/10.1016/S01406736(74)91639-0

Taylor, R.A., Pare, J.R., Venkatesh, A.K., Mowafi, H., Melnick, E.R., Fleischman, W., Hall, M.K. (2016). Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach. *Acad Emerg Med 23*(3), 269-278. doi:10.1111/acem.12876

Vapnik, V.N. (1995). The Nature of Statistical learning Theory. Springer-Verlag New York Inc. New York, USA, 1995.

Weeks, S.R., Juillard, C.J., Monono, M.E. *et al.* (2014) Is the Kampala Trauma Score an Effective Predictor of Mortality in Low-Resource Settings? A Comparison of Multiple Trauma Severity Scores. *World Journal of Surgery, 38***,** 1905–1911. https://doi.org/10.1007/s00268-014-2496-0

WHO. (2014). *Injuries and violence: The facts.* Geneva, Switzerland: WHO

Appendix A. Unused variables

Appendix A contains the variables that are not used in this study and the reason why.

Dropped variable	Reason
AGB	>70% Missing Data
CTSCANDT	>70% Missing Data
CTSCANTIJD	>70% Missing Data
EMVMAX	>70% Missing Data
INTERVENTIEDT	>70% Missing Data
BASE	>70% Missing Data
BENORMAALTIJD	>70% Missing Data
INR	>70% Missing Data
BEADEMING	>70% Missing Data
DATUMOVERLEDEN	>70% Missing Data
RTSAMBR	>70% Missing Data
INTERVENTIETIJD	>70% Missing Data
RTSAMBR	>70% Missing Data
LUCHTWEG	>70% Missing Data
BEVERLOOPTIJD	>70% Missing Data
BASER	>70% Missing Data
INRR	>70% Missing Data
ICDUUR	>70% Missing Data
BEADEMINGR	>70% Missing Data
LEVELHOOGHOSPR	>70% Missing Data
MORTALITEITUUR	>70% Missing Data
MORTALITEITDGN	>70% Missing Data
AANRIJTIJD	AANRIJTIJD, BEHANDELTIJD VERVOERTIJD are combined
BEHANDELTIJD	AANRIJTIJD, BEHANDELTIJD VERVOERTIJD are combined
VERVOERTIJD	AANRIJTIJD, BEHANDELTIJD VERVOERTIJD are combined
PSXX98	Already a score
PSUS0998	Already a score
PSXX98R	Already a score
PSUS0998R	Already a score
OVERLEDENMETOBDUCTIE	Can indicate if a person passed away

OVERPLBESTID	Can indicate if a person passed away
EMVCODETOTAAL	Derived
EMVCODE	Derived
SBPCODE	Derived
ISSCAT	Derived
AISNUMR	Derived
NUMFRACT	Derived
LEEFTIJDSEHCAT	Derived
VERBLIJFSDUURSEHR	Derived
EMVCODETOTAALCORR	Derived and imputed with max values
EMVCODECORR	Derived and imputed with max values
ADEMFREQUENTIEKLASSEID	Inconsistent
TOTAALTIJDONGEVAL	Inconsistent
RECORD_	Index
CASEID_	Index
IDAABA	Index
IDAABACA	Index
ISSCAT16	Index to check for multitrauma
OORZAAKCATEGORIEID	Intentie is used
ICJANEE	Is mostly derived from Dagenopname
ISS98	ISS08 is used
BEPALINGVITALEPARMSDT	Meaningless for prediction
ONGEVALDT	Meaningless for prediction
INANA	Meaningless for prediction
DATUMAANKOMST	Meaningless for prediction
DATUMVERTREKSEH	Meaningless for prediction
DATUMONTSLAG	Meaningless for prediction
CHECKLEVEN	Meaningless for prediction
JAARONGEVAL	Meaningless for prediction
MAANDONGEVAL	Meaningless for prediction
UURONG	Meaningless for prediction
UURONGCAT	Meaningless for prediction
ONGEVALDT_FUZ	Meaningless for prediction
INTERVTYPEANDERS	Mostly empty
OVERLEDEN30D	Not the outcome variable used
MORTALITEIT	Not the outcome variable used

HARTSTILSTAND	Only valid till 2019
VERKEERWAARDEID	Only valid till 2019
DAGENOPNAME	Opnameduur is used
GOSONTSLAG	Patients that pass away will not have this value
ONTSLAGBESTEMMINGID	Patients that pass away will not have this value
RTSSEH	RTSSEHPS is used
RTS	RTSSEHPS is used
RTSSEHR	RTSSEHPS is used
RTSSCORE	RTSSEHPS is used
RTSSEHPSUP09	RTSSEHPS is used
TRAUMACENTRUM	The LNAZ tries to compare the performance to see if there is a difference
TRAUMATEAMSEH	The LNAZ tries to compare the performance to see if there is a difference
TOTAALTIJD	TIJDAMB is used
INTUBATIEPREHOSP	Too much class imbalance
INTERVENTIETYPE	Too much class imbalance
OORZAAK	Too much class imbalance
EMVQUALIFIERWAARDEID	Too much class imbalance
VERVOERINTERKL	uses ONTSLAGBESTEMMINGID

Appendix B. Descriptives

Appendix B contains the descriptives for the used variables for the hip fracture subset (“Hip Fractures”) and for the entire DTR (“DTR”).

	EYEOPENINGWAARDEID Eye Response		MOTORRESPONSEWAARDEID Motor Response		VERBALRESPONSEWAARDEID Verbal Response	
	DTR	Hip Fractures	DTR	Hip Fractures	DTR	Hip Fractures
Min. :	1	1	1	1	1	1
1st Qu.:	4	4	6	6	5	5
Median :	4	4	6	6	5	5
Mean :	3.932	3.988	5.938	5.99	4.877	4.955
3rd Qu.:	4	4	6	6	5	5
Max. :	4	4	6	6	5	5
NA's :	1403	431	1418	431	1437	430

	RRSYSTOLISCH Systolic Blood Pressure		ISS08 Injury Severity Score		ADEMFREQUENTIE Respiratory Rate	
	DTR	Hip Fractures	DTR	Hip Fractures	DTR	Hip Fractures
Min. :	0	17	1	9	0	0
1st Qu.:	120	132	2	9	14	14
Median :	139	150	4	9	16	16
Mean :	141.8	152.5	6.119	9.206	17.37	16.46
3rd Qu.:	159	170	9	9	20	18
Max. :	434	288	75	59	187	187
NA's :	7849	619	105		18606	3973

	RTSSEHPS Revised Trauma Score (for MTOS*)		LEEFTIJDSEH Age		VERBLIJFSDUURSEH Length of stay ED (Emergency Department)	
	DTR	Hip Fractures	DTR	Hip Fractures	DTR	Hip Fractures
Min. :	0	4.094	0	0	0	1
1st Qu.:	7.841	7.841	28	71	128	137
Median :	7.841	7.841	62	81	178	181
Mean :	7.778	7.827	54.45	77.88	191.5	194.7
3rd Qu.:	7.841	7.841	80	87	241	239
Max. :	7.841	7.841	105	105	1043	754
NA's :			9	1	505	111

Note. MTOS: Major Trauma Outcome Study (Champion et al., 1990)

	DAGENIC Length of Stay IC (Intensive Care)		AISNUM The number of AIS codes		TIJDAMB Time Ambulance*	
	DTR	Hip Fractures	DTR	Hip Fractures	DTR	Hip Fractures
Min. :	0	0	0	1	9	12
1st Qu.:	0	0	1	1	35	37
Median :	0	0	1	1	44	45
Mean :	7.8	0.1178	2.033	1.217	45.72	46.97
3rd Qu.:	0	0	2	1	54	54
Max. :	124	41	28	16	225	182
NA's :	7413	1175			27518	4672

Note. Time Ambulance is a combination of AANRIJTIJD, BEHANDELTIJD, and VERVOERTIJD

LEVELHOOGHOSP Level of Hospital Care			LEVELHOOGPREHOSP Level Pre Hospital Care		
Levels	DTR	Hip fracture	Levels	DTR	Hip Fracture
0 (Other)	31082	1191	0 (None and Basic)	14570	851
1 (Operating room)	21627	10711	1 (Other)	40497	11493
NA's	2695	506	NA's:	337	64

LETSELAARDWAARDEID Type of Injury			OVERLEDEN Passed Away		
Levels	DTR	Hip fracture	Levels	DTR	Hip Fracture
0 (Blunt)	53639	12402	0 (Survived)	1251	487
1 (Penetrating)	1650	5	1 (Passed away)	54153	11921
NA's	115	1			

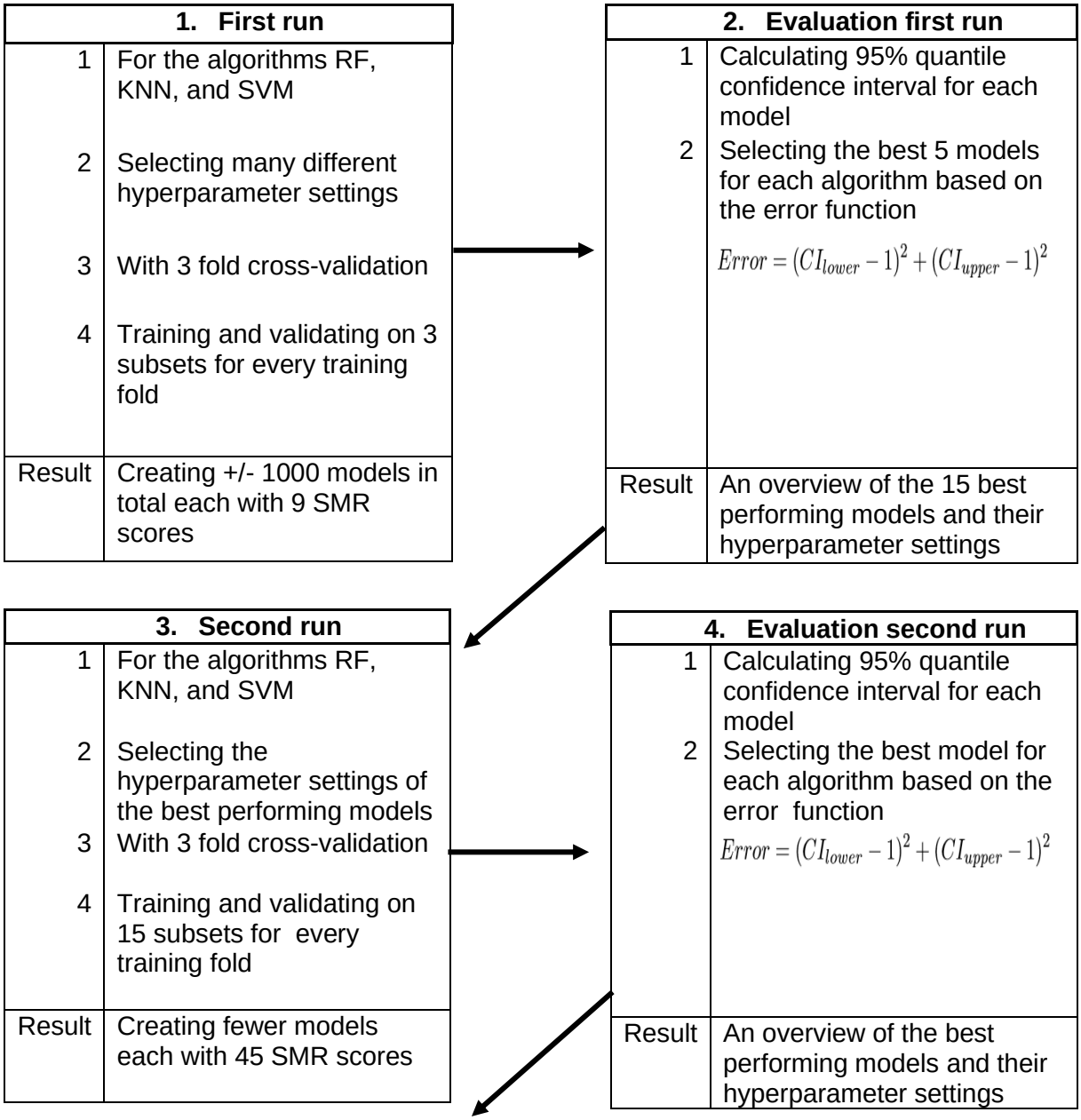
HERKOMSTWAARDEID Referred From			GESLACHTMAN Sex		
Levels	DTR	Hip fracture	Levels	DTR	Hip Fracture
0 (Place Trauma)	39488	9701	0 (Male)	28122	4166
1 (Other)	15535	2640	1 (Female)	27277	8241
NA's	381	67	NA's	5	1

VERWYZER Referrer			INTENTIE Cause of Injury		
Levels	DTR	Hip fracture	Levels	DTR	Hip Fracture
0 (112)	29841	7270	1 (Traffic)	12341	1264
1 (Other)	22653	4579	2 (Other)	13560	400
NA's	2910	559	3 (Low fall)	26743	10252
			NA's	2760	492

COMORB Comorbidity (ASA)			OPNAMEDUUR Length of Stay Hospital		
Levels	DTR	Hip fracture	Levels	DTR	Hip Fracture
1 (ASA 1)	20995	1085	1 (1 day)	5611	104
2 (ASA 2)	19278	5514	2 (2 days)	18171	406
3 (ASA 3)	11573	5088	3 (3-7 days)	18033	5887
4 (ASA 4 & 5)	500	216	4 (8 -14 days)	9204	4310
NA's	3058	505	5 (15-21 days)	2675	1130
			6 (>21 days)	1580	564
			NA's	130	7

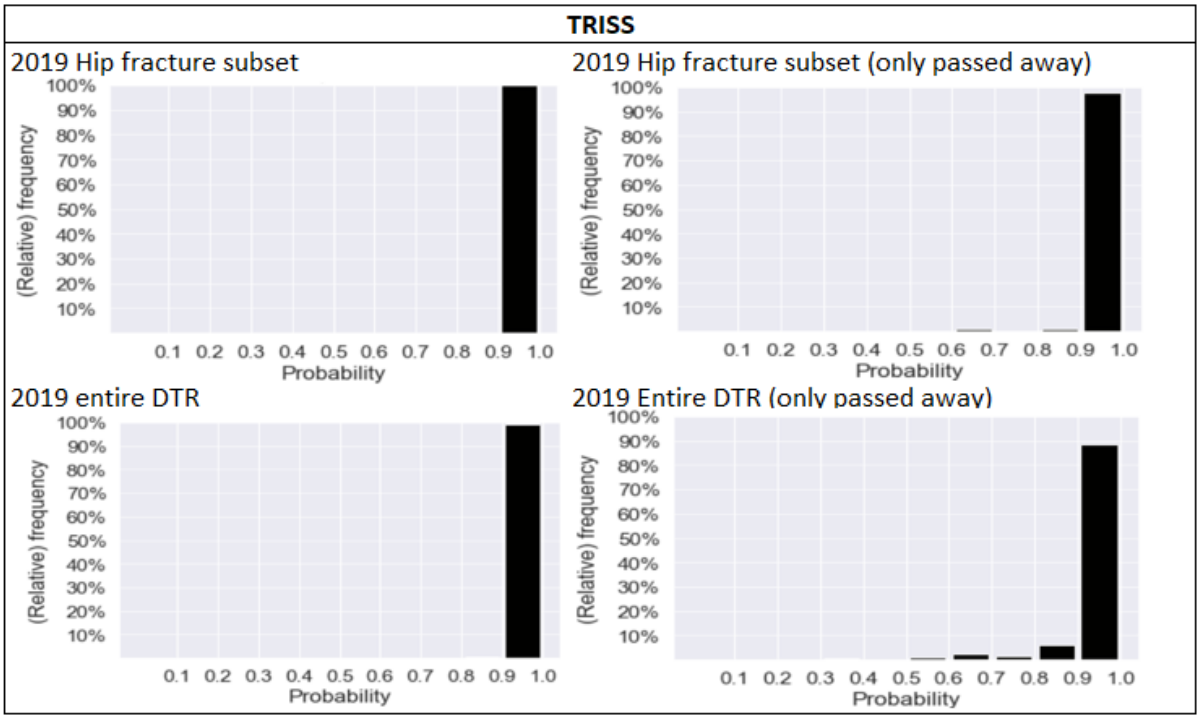
Appendix C. Schematic approach of model selection

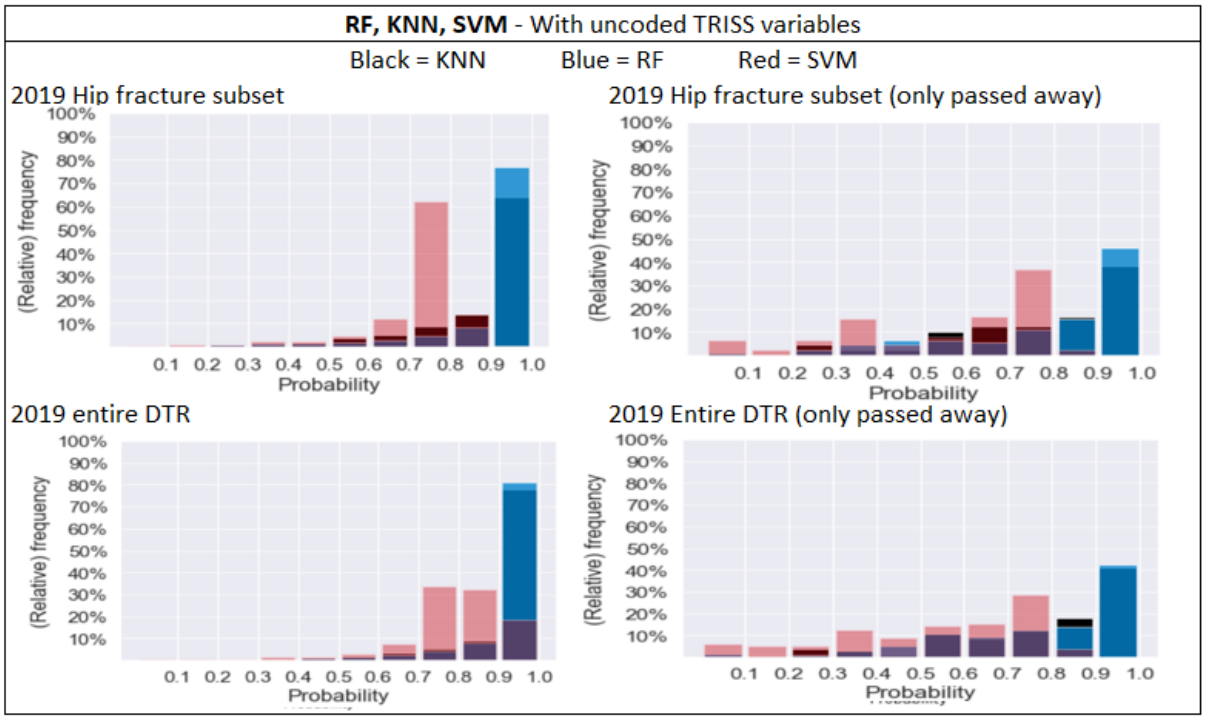
Appendix C contains a schematic of the model selection approach.



Appendix D. Probability distributions

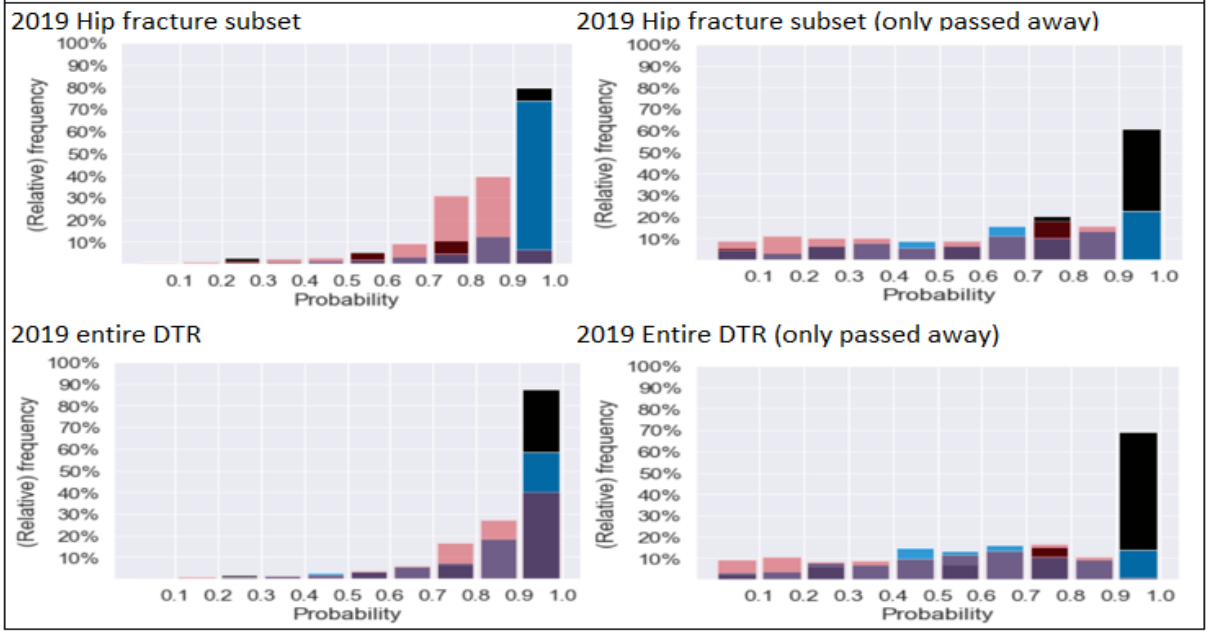
Appendix D contains the probability distribution of the best performing models on the specified test sets. For the TRISS model, the probability is just the PSNL15. For the KNN and RF, this is the percentage of neighbours or trees voting for a negative outcome (survived). For the SVM the probabilities are calibrated using Platt scaling (Platt, 1999).





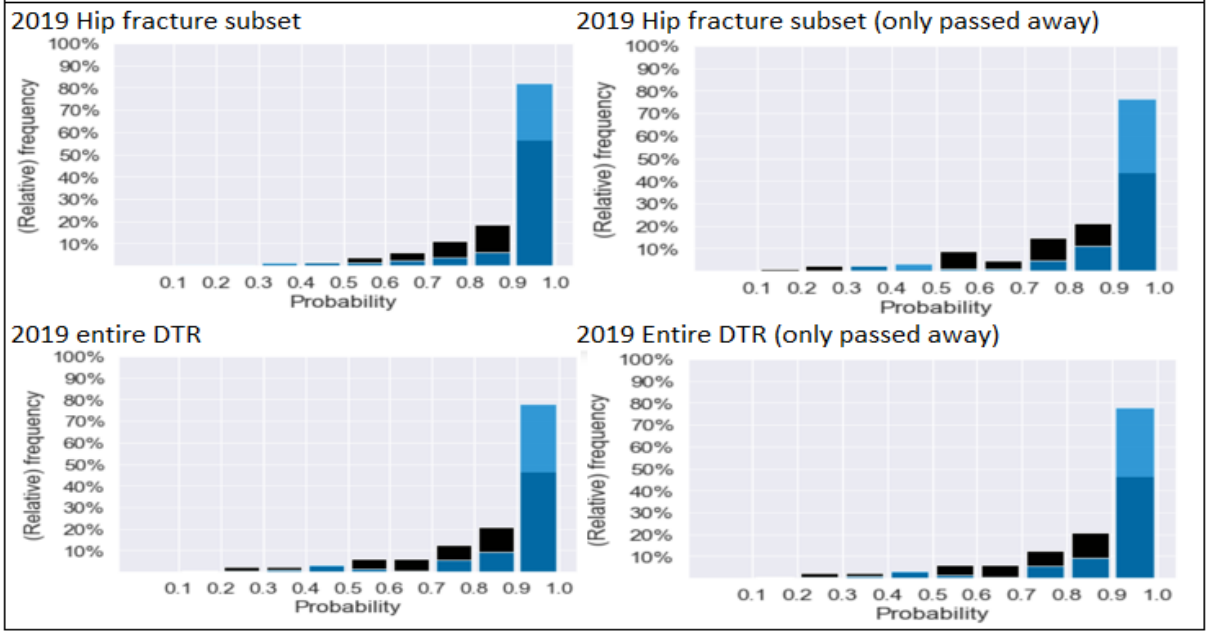
RF, KNN, SVM - With all selected variables

Black = KNN Blue = RF Red = SVM



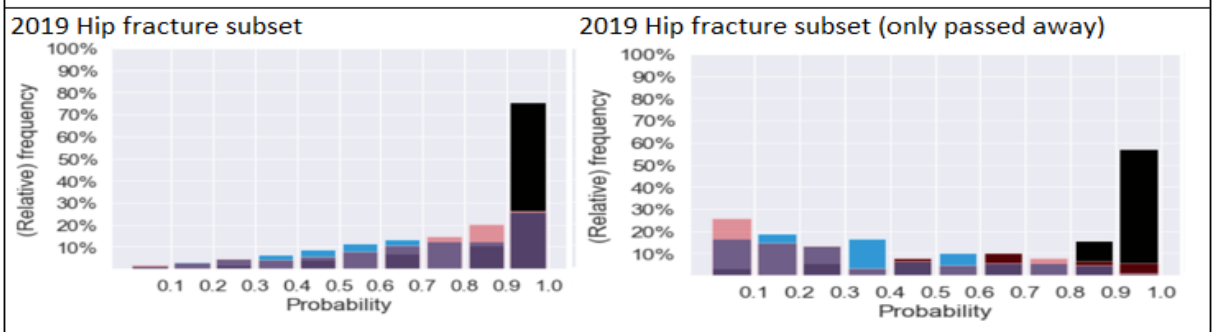
RF, KNN - with Systolic Blood Pressure and Length of Stay ED

Black = KNN Blue = RF



RF, KNN, SVM - With all selected variables and selected based on Balanced Accuracy

Black = KNN Blue = RF Red = SVM

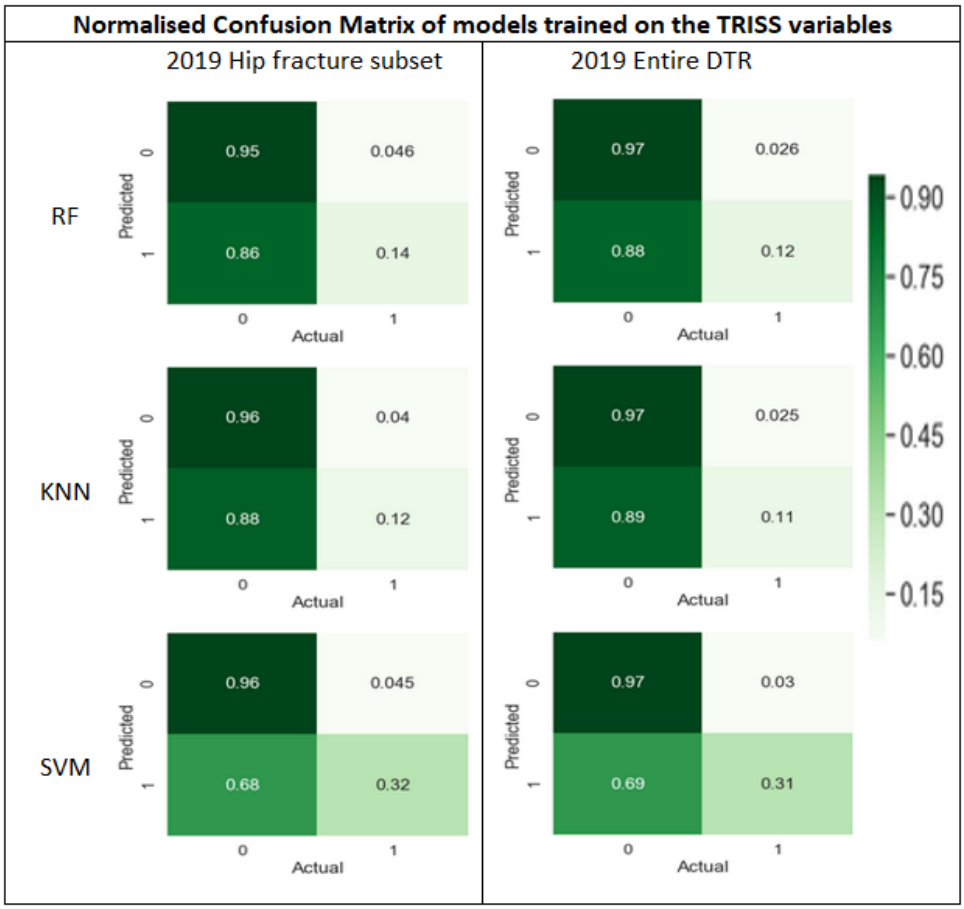


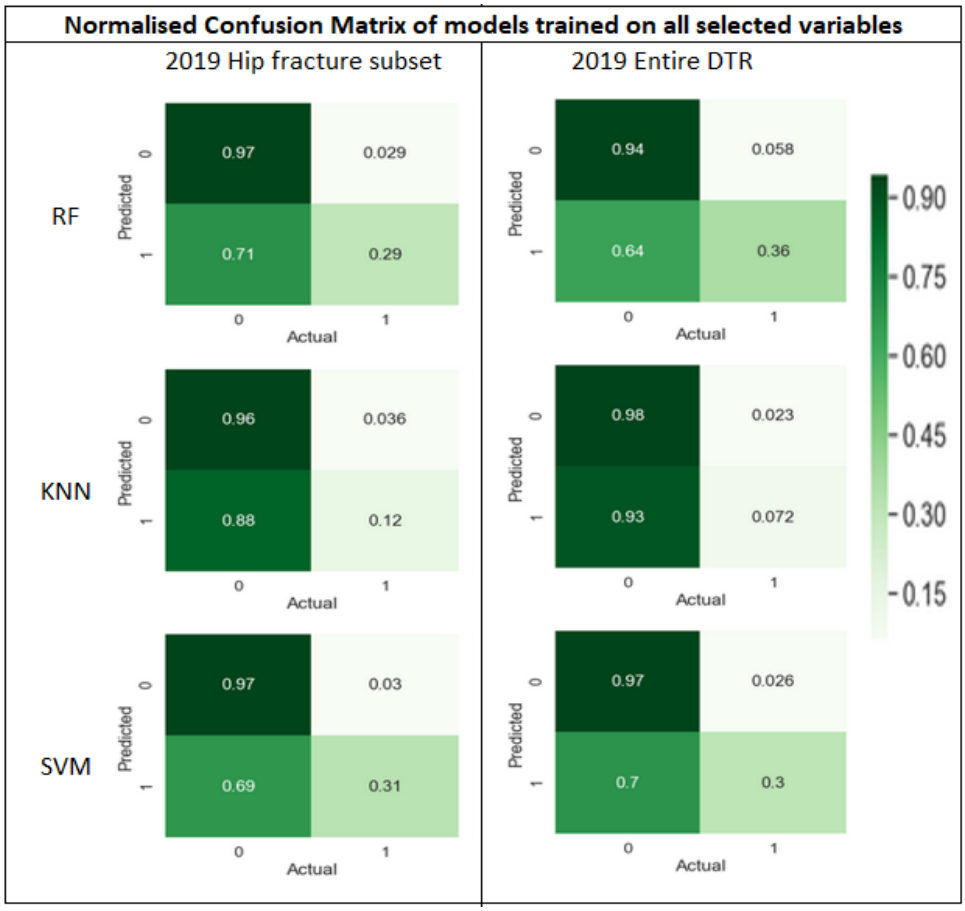
Appendix E. Accuracy and Confusion Matrices

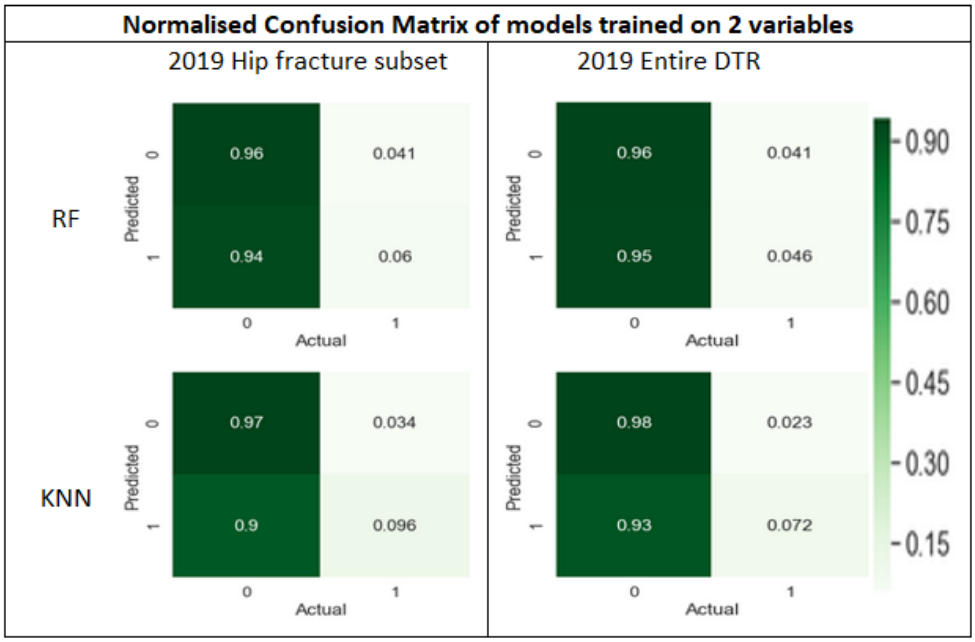
Appendix E contains the accuracy of the best performing models on the specified test sets and their confusion matrix.

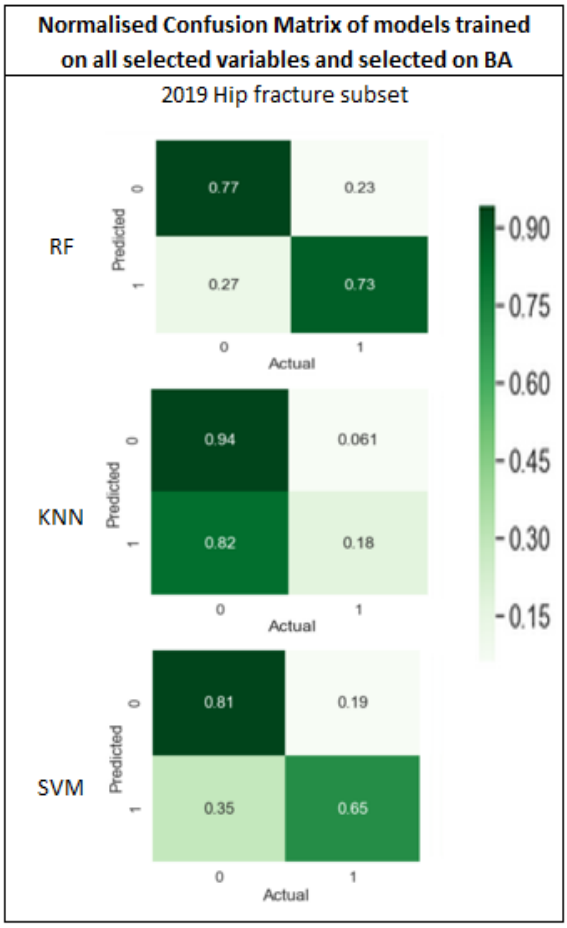
	2019 Hip fracture		2019 entire DTR	
RF	Lower CI	Higher CI	Lower CI	Higher CI
Trained on TRISS variables	0.912	0.925	0.954	0.958
Trained on all selected variables	0.940	0.948	0.924	0.933
Trained on 2 variables	0.918	0.927	0.936	0.942
KNN				
Trained on TRISS variables	0.923	0.939	0.954	0.959
Trained on all selected variables	0.925	0.936	0.955	0.960
Trained on 2 variables	0.926	0.936	0.931	0.941
SVM				
Trained on TRISS variables	0.917	0.938	0.950	0.960
Trained on all selected variables	0.934	0.952	0.952	0.966

Note. The numbers displayed is the accuracy, not the SMR.




Normalised Confusion Matrix of models trained on all selected variables
2019 Hip fracture subset
2019 Entire DTR
RF
KNN
SVM
Predicted
Actual
0.97
0.029
0.71
0.29
0.94
0.058
0.64
0.36
0.96
0.036
0.88
0.12
0.98
0.023
0.93
0.072
0.97
0.03
0.69
0.31
0.97
0.026
0.7
0.3
0.90
0.75
0.60
0.45
0.30
0.15


Normalised Confusion Matrix of models trained on 2 variables
2019 Hip fracture subset
2019 Entire DTR
RF
KNN
Predicted
Actual
0.96
0.041
0.94
0.06
0.96
0.041
0.95
0.046
0.97
0.034
0.9
0.096
0.98
0.023
0.93
0.072
0.90
0.75
0.60
0.45
0.30
0.15


Normalised Confusion Matrix of models trained on all selected variables and selected on BA
2019 Hip fracture subset
RF
KNN
SVM
Predicted
Actual
0
1
0.77
0.23
0.27
0.73
0.94
0.061
0.82
0.18
0.81
0.19
0.35
0.65
0.90
0.75
0.60
0.45
0.30
0.15