Bachelor thesis
Computer Science

Radboud University

Predicting Pelvic Floor Surgery Outcomes

Author:
Thomas Welten
s4350456

First supervisor/assessor:
Prof. Dr. Tom Heskes
`t.heskes@science.ru.nl`

Second assessor:
Dr. Jesse Krijthe
`J.Krijthe@cs.ru.nl`

April 1, 2019

Abstract

Predictive data mining can be used to estimate the effects of medical treatment, especially for datasets with a large number of dimensions. In this thesis, random forest regression is utilized to predict the outcome of pelvic floor surgeries. We also apply an adaptation of random forests, a generalized random forest to the data to estimate the heterogenous treatment effect of the different types of surgeries involved. A dataset was provided by the department of Obstetrics and Gynaecology of the RadboudUMC. After processing, it contained usable pre- and post- operative data about 730 pelvic floor surgeries. 70 variables were used for predicting outcomes and treatment effects. The random forest regression predictor was able to predict the difference between pre- and post- operation Urinary Distress Inventory with a mean absolute error of 10.01 on a scale of -100 to 100. We provide a feature importance ranking to identify factors determining surgery success. Finally, we discuss the average treatment effects estimated by the generalized random forest.

Contents

1 Introduction 2

2 Preliminaries 4
2.1 Classification 4
2.2 Methodology 4
2.2.1 Random forest regressor 5
2.2.2 Treatment effect estimation with generalized random forests 5

3 Research 11
3.1 Data preprocessing 11
3.2 Variables 12
3.3 Data Processing 14
3.4 Results 15
3.4.1 RandomForestRegressor 15
3.4.2 Causal Forest 20

4 Related Work 23

5 Conclusions 24

Chapter 1

Introduction

Data mining techniques are commonly applied to medical problems. This involves the use of machine learning to understand medical data. One of the uses is for predictive data mining, where the effect of treatments on the patient is modeled.

The department of Obstetrics and Gynaecology of the RadboudUMC provided a dataset of 2500 pelvic floor surgeries. Of these, 730 were suitable for this research. This dataset consisted of pre- and post operation patient questionnaire and case reports. The main questionnaire used was the Urinary Distress Inventory (UDI) questionnaire. The answers to this questionnaire were transformed into five scores, which were averaged. The difference between pre- and post operational score is the variable which indicates the success of the surgery.

The goal of this thesis is twofold: one goal is to predict the outcome of pelvic floor surgeries. This should help to discriminate in which cases surgery will lead to improvement in patient outcomes. The second goal is to increase understanding of underlying factors which determine surgery success.

For the first goal we choose a random forest regressor. This is a suitable choice to predict outcomes for the large number (70) of input features and for the relatively large size of the data set. We train this regressor to predict the difference in UDI score before and after a surgery for a patient, based on the given data of a patient before the surgery.

The second goal is partly achieved by examining the feature importances for the random forest regressor. Additionally we let a generalized random forest estimate the average treatment effect for different surgery types. In this case this is an estimate of the difference in outcomes between applying and not applying a specific treatment (here surgery) [1]. The average treatment effect estimate thus gives information about the effectiveness of each type of treatment. It does not predict the outcome for a single patient case. A generalized random forest uses a variation of a random forest to estimate this heterogeneous average treatment effect. We use this generalized random forest because the basic random forest is not suited for treatment effect estimation. It is used for regression and classification.

Pelvic floor disorders frequently occur in women [7]. The pelvic floor for women is a series of muscles which support the pelvic organs, among them the vagina, bladder and uterus. When damaged, women can suffer from incontinence, obstructive bowel disease or pelvic pain. This can be treated with various types of pelvic floor surgeries. Women face a 10 to 25% lifetime risk of pelvic floor surgery [7]. At least 15% of pelvic floor surgeries do not lead to improvement in conditions [8]. The exact factors influencing surgery success are not well understood.

Machine learning in the form of feed-forward artificial neural networks has been used to identify patients with pelvic organ prolapse from a data set containing healthy and affected patients [5]. There is also extensive data available on average pre- and post operative UDI scores for the types of pelvic floor surgeries involved in this thesis. However, to the best of our knowledge, machine learning has not been applied to pelvic floor surgery data so far.

Chapter 2 will provide the background for classification, the random forest algorithm as applied to regression, feature importances and the estimation of average treatment effects with generalized random forests. Chapter 3 details the preparation of the dataset, the application of the data mining algorithms and the results found. Chapter 4 discusses related work. The final chapter concludes and discusses this thesis.

Chapter 2

Preliminaries

2.1 Classification

In this thesis, we will be using a classification and regression pipeline to estimate medical treatment effects and outcomes. This will be based upon a data set. The data set will contain a number of features X. These are variables that may or may not effect the outcome Y. The outcome Y is in our case the difference in health satisfaction before and after a pelvic floor surgery. This pipeline consists of first selecting relevant features, their values and those of the outcome variable Y from the data set and splitting the data into a training and a test set [10]. A machine learning algorithm is applied to the training set, resulting in a model. This model can predict the class or outcome Y in case of regression for a given set of input features. The final step in the pipeline is applying the model to a new, unknown set of data. This is generally first done to the test set, to estimate the accuracy or error rate of the classifier [10].

For our research we use medical history, surgery type, POP-Q scores and the Urinary Distress inventory(UDI) questionnaire score before the operation as features X. We use the difference between UDI-score before and after the surgery as our outcome variable Y.

We choose to apply the random forest algorithm to our problem. We made this choice because the random forest has been found to perform well [4] compared to other classification/regression algorithms.

2.2 Methodology

We will first examine the algorithm for a random forest regressor. We train this regressor on our data set to predict the difference in UDI-score before and after surgery for a patient. This will be discussed in subsection 2.2.1 . After that we explore in subsection 2.2.2 generalized random forest. We use this generalized random forest to estimate the treatment effect for the different types of surgery. The treatment effect is the difference between the case where treatment (surgery for us) and no treatment is used.

Algorithm 1 Random Forest Regressor (Adapted from Hastie, Tibshirani, p.588)[6]

All tuning parameters are pre-specified, including the number of trees B and the size of the sample s rate used in Bootstrap

```
procedure RandomForestRegressor(training set $\mathcal{S}$)
    for $b = 1$ to total number of trees $B$ do
        set of samples $\mathcal{I} \leftarrow$ Bootstrap($\mathcal{S}$, $s$)
        tree $\mathcal{T} \leftarrow$ RegressionTree($\mathcal{I}$)
                    ▷ Regression tree using CART algorithm (In this case they use random feature subset selection) [3]
    output $\{T_b\}_1^B$, the ensemble of trees
    prediction $\hat{f}_{rf}^B = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$        ▷ To make a prediction for sample $x$
```

Bootstrap draws a random subsample of size s from $\mathcal{S}$ with replacement (bootstrap aggregating)

2.2.1 Random forest regressor

A random forest regressor combines a set of randomly sampled regression trees to predict a regression target value for a given set of input features. Every trained regression tree predicts an output value for the given input features, which are then averaged for the total tree prediction[2]. In our case we are using this to predict the outcome, the difference in UDI-score before and after surgery for a given patient.

Breiman's Random Forest algorithm starts with a training set $\mathcal{S}$, which is used to train our regressor. Every element of set $\mathcal{S}$ consists of features X and outcome Y. From this training set $\mathcal{S}$ we generate a number of training sets $\mathcal{I}$ by bagging. Bagging, bootstrap aggregating, is randomly drawing with replacement from the training set $\mathcal{S}$. We do this B times to create B subsamples. Next we fit a regression tree modified with random feature selection on each of these bootstrapped training sets $\mathcal{I}$. A regression tree recursively partitions the space of a training set into halves, forming a binary tree. These splits are on the features X of a set. The leaves of this tree contain a range of predictions Y of the training set space. A fitted tree gives a prediction for a new given input. The output of the initial random forest algorithm application is an ensemble of regression trees T_b. In order to make a prediction for a sample of input features x, every regression tree in the ensemble gives a prediction for x. These predictions are averaged to give the result for the forest.

2.2.2 Treatment effect estimation with generalized random forests

In the previous paragraph we were exploring a random forest for predicting surgery outcomes based on variables and surgery types. In this paragraph we are using a different type of random forest, a generalized random forest for estimating the treatment effect of each type of pelvic floor surgery. The treatment effect is the difference between the situation where treatment W (surgery for us) is applied and the situation where no treatment is applied.

A treatment effect is estimated based on input variables X, treatment assignment W, and outcome Y for each situation. For our case: Input variables X contain medical history, surgery type excluding treatment W, POP-Q scores and the UDI-score before surgery. Treatment assignment W consists of the surgery for which the treatment effect estimation is done. Y is the difference in UDI-score before and after the surgery.

We will first examine the problem of treatment effect estimation. Next, we examine the general random forest algorithm and finally its implementation for treatment effect estimation, the so called causal forests.

Defining the problem: Heterogeneous treatment estimation under unconfoundedness

Take a number of independent training cases $i = 1, ...n$ with a feature vector $X_i \in [0,1]^d$, outcome $Y_i \in \mathbb{R}$ and treatment $W_i \in 0, 1$ each [11]. The heterogeneous treatment effect at x is then the function $\tau(x)$ for outcomes $Y_i^{(1)}$ (with treatment) and $Y_i^{(0)}$ (no treatment) as given by Wager and Atley [11] based on Rubin:

$$\tau(x) = \mathbb{E}[Y_i^{(1)} - Y_i^{(0)} | X_i = x].$$

The problem with estimating $\tau(x)$ here is that we only have information about one of the outcomes $Y_i^{(1)}$ and $Y_i^{(0)}$. This is because a patient can receive treatment or not, but not both. We can therefore not train a machine learning algorithm on the difference between $Y_i^{(1)}$ and $Y_i^{(0)}$. In order to estimate the treatment effect a further assumption is necessary: unconfoundedness [9]. The unconfoundedness assumption means that on condition of features X, the outcomes $Y^{(1)}$ and $Y^{(0)}$ are independent of treatment assignment W:

$$\{Y^{(0)}, Y^{(1)}\} \perp\!\!\!\perp W | X.$$

Under this assumption, as shown in figure 2.1, any confounding variable C which might influence both treatment W and outcome Y is already controlled for in X. An example of a confounding variable in our case would be a variable C which influences whether a patient does or does not receive treatment W which is not contained in X.

A counterexample where unconfoundedness does not hold, is shown in figure 2.2. Imagine a patient who requires surgery. Outcome Y is the improvement to his medical condition. Relative C advises the patient to take treatment W because he read a horoscope predicting disaster otherwise. C also ensures that the patient follows all the doctors advice before and after surgery, positively affecting the outcome Y. The treatment effect we measure for W now includes the effect of treatment W and the effect of closely following the doctors instructions. This is obviously not what we want to estimate.

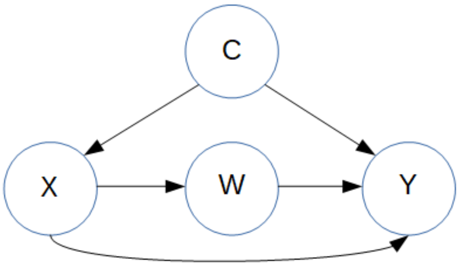

Figure 2.1: Unconfoundedness assumption: Confounding variable C will not influence the results, despite affecting treatment W and outcome Y. This is because C's influence on W goes through feature vector X, our control variable.

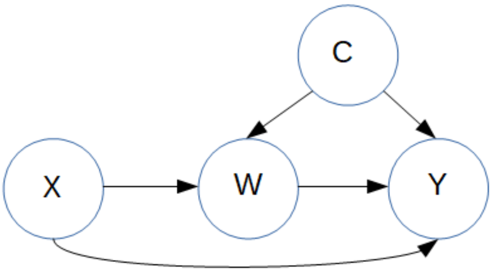

Figure 2.2: Unconfoundedness does not hold: Here confounding variable C influences both treatment assignment W and outcome Y, but the effect on W is not mediated by X.

Generalized random forests

Generalized random forests are an adaptation of Breiman's random forest concept adapted for non-parametric statistical estimation by Athey and Wager[1]. In our case, they are used for estimating heterogeneous treatment estimation. This makes it a causal forest, as it estimates the treatment effect in its trees. Generalized random forests keep a number of basic principles of random forests, specifically random selection of splits, subsampling and recursive partitioning of the sample space [1]. The main difference lies in the usage of forest based local estimation. An adaptive nearest neighbor estimator is used instead of using the average result of the forests.

The generalized random forest is an adaptation of the random forest for a number of statistical tasks, each with their own estimation equation. It works as follows (2) :

As input a set of examples S and a test point x are used. The algorithm follows the normal random forest procedure by growing an ensemble of trees B. For every tree we select a subsample s from the training set. This is split into to 2 evenly-sized, random halves.

A gradient tree algorithm is applied to samples in the first half of the split, growing a tree T using only this half. The algorithm differs from a normal CART regression tree algorithm in the way it chooses how to make splits. The gradient tree split increases the heterogeneity of estimates θ in the tree quickly [1]. This contrasts with a normal regression tree where the chosen split minimizes the prediction error.

From the second half of the split all those elements that fall into the same leaf as x in tree T are taken. These are the neighbors of x. The neighbors are used to adjust a number of similarity weights $\alpha[e]$ of every training sample i to equal positive values. Weights $\alpha[e]$ are later used to calculate a solution to the problem our algorithm has to solve. The split into two halves where one half is used for growing a tree and the other half for finding examples in the same leaf as x satisfies a condition called honesty. An

Algorithm 2 Generalized random forest with honesty and subsampling (by Athey and Wager) [1]

All tuning parameters are pre-specified, including the number of trees B and the subsampling s rate used in SUBSAMPLE. This function is implemented in the package `grf` for `R` and `C++`.

```
1: procedure GENERALIZEDRANDOMFOREST(training set $\mathcal{S}$, test point $x$)
2:     weight vector $\alpha \leftarrow$ ZEROS($|\mathcal{S}|$)
3:     for $b = 1$ to total number of trees $B$ do
4:         set of samples $\mathcal{I} \leftarrow$ SUBSAMPLE($\mathcal{S}$, $s$)
5:         sets of samples $\mathcal{J}_1$, $\mathcal{J}_2 \leftarrow$ SPLITSAMPLE($\mathcal{I}$)
6:         tree $\mathcal{T} \leftarrow$ GRADIENTTREE($\mathcal{J}_1$, $\mathcal{X}$)
7:         $\mathcal{N} \leftarrow$NEIGHBORS($x$, $\mathcal{T}$, $\mathcal{J}_2$)
8:         for all example $e \in \mathcal{N}$ do
9:             $\alpha[e]$ += $1/|\mathcal{N}|$
10:    output $\hat{\theta}(x)$, the solution to the estimating equation with weights $\alpha/B$
```

The function ZEROS creates a vector of zeros of length $|\mathcal{S}|$; SUBSAMPLE draws a subsample of size s from $\mathcal{S}$ without replacement; and SPLITSAMPLE randomly divides a set into two evenly-sized, non-overlapping halves. $\mathcal{X}$ is the domain of the feature vector X_i. NEIGHBOR returns those elements of $\mathcal{J}_2$ that fall into the same leaf as x in the tree $\mathcal{T}$. Note that only line 7, 8 and 9 are run for every test point x, as the sample split and the resulting gradient tree are reused.

honest tree uses a response Y only to calculate the treatment effect or to determine how to make a split, but not both [11].

The solution $\hat{\theta}(x)$ to the estimating equation for x is found by first averaging all the found weights for use in the estimating equation as seen in figure 2.3. Next the equation is solved with these weights. This is a nearest neighbor approach, because we use the values found for neighbors of x. The approach therefore differs from the averaging of solutions of a tree as in the earlier discussed random forest regressor.

Causal Forests

Finally we need to adapt the generalized random forest to our problem of heterogeneous treatment effect estimation under assumption of unconfoundedness. This application is called a causal forest, as we estimate causal relationships of treatments. For this Athey et. al [1] used treatment effect estimate $\theta(x)$ (2.1), assuming unconfoundedness:

$$\hat{\theta}(x) = \frac{Cov[W_i, Y_i | X_i = x]}{Var[W_i | X_i = x]} \tag{2.1}$$

W_i is the treatment assignment, X_i the feature, Y_i the outcome with $i = 1, 2, ..n$ training samples and test point x. We solve $\theta(x)$ with generalized random forest estimator $\hat{\theta}(x)$

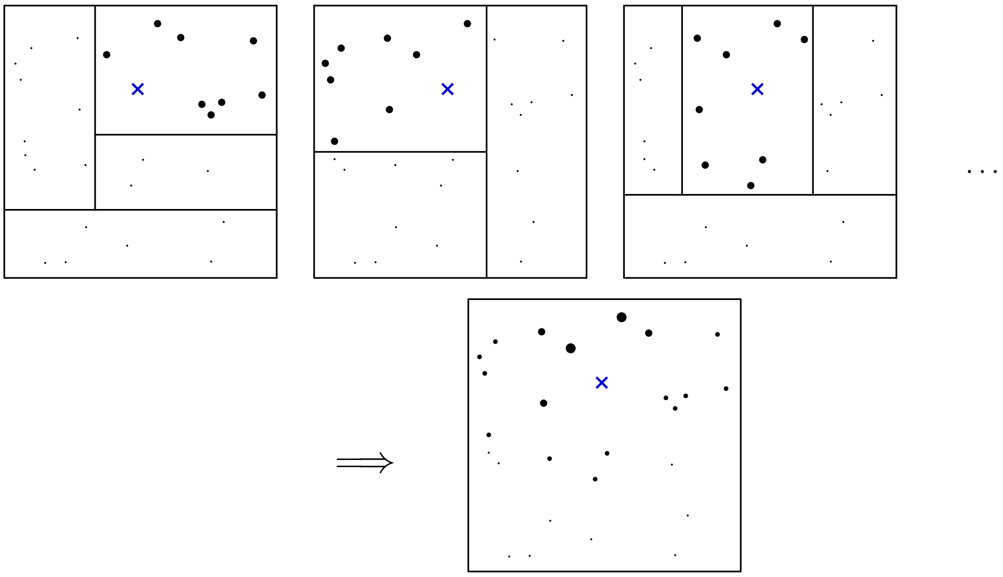

Figure 2.3: Illustration of the general random forest weighting function. The rectangles depicted above correspond to terminal nodes. Each tree starts by giving equal (positive) weight to the training examples in the same leaf as our test point x of interest, and zero weight to all the other training examples. Finally, the forest averages all these tree-based weightings, and effectively measures how often each training example falls into the same leaf as x (Figure and description by Athey and Wager).[1]

in the estimating equation in 2 . $\hat{\theta}(x)$ gives us a solution for the treatment effect τ, using the weights found in the gradient trees. This can be seen as a case of fitting a simple linear regression line in 2.1 as seen in figure 2.4 . Further details can be found in Athey et al [1].

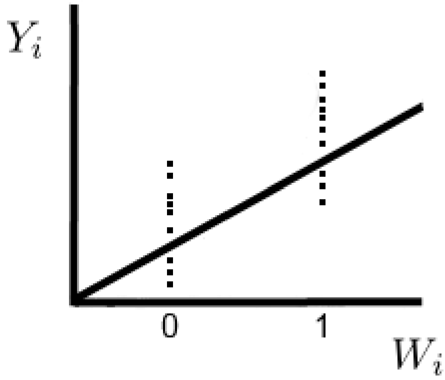

Figure 2.4: Linear regression for treatment effect estimation for sample i with treatment assignment W_i, outcome Y_i and test point (treatment) x. Random forest weights found by the generalized random forest algorithm 2 are used in the regression line calculation. $\hat{\theta}(x)$ is the fitted regression line in the figure.

Chapter 3

Research

3.1 Data preprocessing

For our research, a dataset of originally 2500 pelvic floor surgeries is used. The dataset has been provided by the department Obstetrics and Gynaecology of the RadboudUMC. It contains a combination of data, consisting of variables derived from pre- and post operation patient questionaires and case reports.The most relevant questionaire was the Urinary distress inventory (UDI). The dataset contained entries for pre- and post operation answers on the UDI. These answers were converted into 5 main UDI scores for each questionaire. The scores ranged from 0 to 100, 0 being the best, 100 the worst. If a value was missing it was, if possible, compensated for in the UDI-calcuation formula used. If too many values were missing to calculate a working UDI-score, the entry was not used in our final training/test set. We calculated the average score over the 5 UDI-scores for each questionaire. Next, the difference between the average UDI-score before the operation and 6 months after the operation was calculated for each patient (3.1). The difference ranges from -100 to 100. We flipped this from negative to positive, so that improvement is a positive number. We took the average progression/regression score after 6 months as our value Y. This is the value we wanted to predict with our machine learning algorithm.

We replaced missing values with an out of range value. Outliers and wrong entries were identified and reclassified as missing values or fixed in the case of typos. Duplicate variables were merged. For a number of missing pre-operation values, mainly height, weight and other general health information we used the post operation value instead.

Of the 2500 entries, only 730 were useful for prediction of treatment outcomes. There were multiple reasons for this culling of data entries. The main issue was missing values in one of the pre- or post operative UDI-questionnaires In such a case it was impossible to calculate the difference between pre- and post operative UDI score. The entries from 2 experimental studies were also removed.

Table 3.1: UDI-scores

Name	**UDI genital prolapse**	**UDI OAB**	**UDI incontinence**	**UDI obstructive micturition**	**UDI pain**	**Mean of all UDI-scores**
Missing entries	0	88	22	37	21	0
Mean pre-op	52.19	32.09	27.91	27.56	31.45	34.57
Median pre-op	66.66	22.22	16.67	16.67	33.33	31.11
Pre-op std. deviation	31.80	26.14	26.65	27.56	27.30	17.86
Mean post-op	7.74	19.01	18.84	15.20	17.30	15.64
Median post-op	0.00	11.11	18.84	0.00	0.00	12.22
Post-op std. deviation	17.41	22.58	22.46	22.22	22.66	14.98
Mean improvement	44.45	13.26	8.79	11.87	14.12	19.18
Median improvement	41.67	11.11	0.00	0.00	16.67	17.78
Improvement std. deviation	33.91	22.58	25.16	28.35	26.72	17.55

3.2 Variables

After pruning non-relevant and duplicate variables, 70 features were used for training the machine learning algorithm for predicting the UDI score. These can be aggregated into 4 basic categories: medical history, surgery type and pelvic organ prolapse quantification (POP-Q) scores, UDI-score pre-operation(3.2). It should be noted that one patient can have multiple types of surgeries at the same time.

Table 3.2: Variables for the random forest

Variable	**Description**	**Variable type**
age	Age of patient	Medical history
menopauze	Patient is post menopauze	Medical history
smoking	Patient smokes	Medical history
prev_prolsurgery_n	No of previous pelvic prolapse surgeries	Medical history
prev_incosurgery_n	No of previous pelvic incontinence surgeries	Medical history
SuHi_vaghyst	Previous vaginal hysterectomy	Medical history
SuHi_abhyst	Abdominal hysterectomy	Medical history
SuHi_vwn	Previous anterior colporrhaphy	Medical history
SuHi_aw	Previous posterior colporrhaphy	Medical history
SuHi_portamp	Previous portio amputation/Manchester operation	Medical history
SuHi_ssf	Previous sacrospinal fixation	Medical history

Continued on next page

Table 3.2 – continued from previous page

Variable	Description	Variable type
SuHi_burch	Previous colposuspension	Medical history
SuHi_stamey	Previous needle suspensions (Stamey)	Medical history
SuHi_raz	Previous needle suspensions (Raz)	Medical history
SuHi_v.enteroc	Previous vaginal enterocele resection	Medical history
SuHi_a.enteroc	Previous abdominal enterocele resection	Medical history
SuHi_fixvv	Previous fixation vaginal vault	Medical history
CoMo_1a	CNS disease	Medical history
CoMo_1ac	Still present?	Medical history
CoMo_2a	Cardiovascular disease	Medical history
CoMo_2ac	Still present?	Medical history
CoMo_3a	Respiratory disease	Medical history
CoMo_3ac	Still present?	Medical history
CoMo_4a	Gastrointestinal disease	Medical history
CoMo_4ac	Still present?	Medical history
CoMo_5a	Endocrine disease	Medical history
CoMo_5ac	Still present?	Medical history
CoMo_6a	Musculoskelatal disease	Medical history
CoMo_6ac	Still present?	Medical history
CoMo_7a	Allergy	Medical history
CoMo_7ac	Still present?	Medical history
CoMo_8a	Other	Medical history
CoMo_8ac	Still present?	Medical history
fysiovoor	Physiotherapy preoperative	Medical history
vagpessvoor	Vaginale pressary preoperative	Medical history
Aa	POP-Q	pelvic organ prolapse quantification
Ba	POP-Q	
C	POP-Q	
HG	POP-Q	
PB	POP-Q	
TVL	POP-Q	
Ap	POP-Q	
Bp	POP-Q	
D	POP-Q	
ass_oper	Surgery performed by resident	surgery type

Continued on next page

Table 3.2 – continued from previous page

Variable	Description	Variable type
COK_vw	Anterior colporrhaphy	surgery type
COK_aw	Posterior colporrhaphy	surgery type
COK_peri	Perineoplasty	surgery type
COK_portamp	Portio amputation/Manchester operation	surgery type
COK_vue	Vaginal hysterectomy	surgery type
COK_ssf	Sacrospinal fixation	surgery type
COK_klaw	Classical posterior wall surgery	surgery type
COK_levator	Needle suspensions (Stamey)	surgery type
COK_labhardt	Needle suspensions (Raz)	surgery type
COK_v.enteroc	Vaginal enterocele resection	surgery type
COK_a.enteroc	Abdominal enterocele resection	surgery type
COK_shull	Shull vaginal vault fixation	surgery type
COK_Mccall	Mccall vaginal vault fixation	surgery type
COK_rectprol	Rectal prolapse	surgery type
COK_analsph	Anal sphincter repair	surgery type
Hoogste_opl	Highest education patient	Medical history
Kind_n	Number of children	Medical history
Kind_sct	No of previous c-sections	Medical history
Kind_tan	No of previous forceps delivery	Medical history
Kind_vac	No of previous vacuum delivery	Medical history
Kind_kni	No of previous episiotomy	Medical history
Kind_sch	No of previous perineal tear	Medical history
Lft_1bev	Age of first delivery	Medical history
bmi_1	Patient BMI	Medical history
UKLgpr_1	Pre-operative UDI genital prolapse score	UDI-score pre-operation
UKL_preop	Average pre-operative UDI score	UDI-score pre-operation

3.3 Data Processing

For the surgery outcomes prediction, we used the RandomForestRegressor function from the scikit-learn machine learning library, written in Python. In the random forest regressor we optimized the parameters for best prediction performance using the sci-kit learn RandomizedSearchCV function. For the estimation of average treatment effects, we used the grf (generalized random forest) package in the language R. Here we trained a causal forest on our data set with the causal_forest function, made predictions for individual cases and used the average_treatment_effect function to estimate the average treatment effect.

3.4 Results

3.4.1 RandomForestRegressor

We first randomly split our datset of 730 entries into a training and a test set. This split was 75% training set and 25% test set. Next, we trained a random forest regressor on the training set. The trained forest was used on the test set to evaluate its performance. We used three metrics to examine the performance of our random forest regressor in predicting: the mean absolute error, the explained variance and the explained absolute deviation. For continuous values the mean absolute error calculates the difference between the value of paired observations y_i and $\hat{y}_i$. In our case this are the actual patient UDI-score difference y_i and the predicted difference $\hat{y}_i$:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|. \tag{3.1}$$

The explained variance gives an estimate of how well a model explains the variation of a given set of observations:

$$\text{explained variance}(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}}. \tag{3.2}$$

Var is the square of the standard deviation, with y the actual output and $\hat{y}$ the predicted output. 1 is the best score for explained variance with lower values indicating worse performance:

$$\text{Explained absolute deviation} = 1 - \frac{MAE}{Mean\{|y - median\{y\})|\}}. \tag{3.3}$$

MAE is the mean average error and y the actual output. This metric adjusts the MAE error to the deviation of the y value. This is to ensure predictions for the UDI and UDI_GPR can be compared, even if the y values for them have a different standard deviation.

We fitted a random forest regressor to our training set. We used two separate regressors: one for the averaged UDI-improvement as y-variable and one for the UDI genital prolapse score. Afterwards we applied each regressor to our test set. This led to the results in table 3.3.

For the UDI improvement the mean absolute error was 10.01. As the mean improvement is 19.18 with a standard deviation of 17.55, the prediction error of the regressor is large in comparison with the standard deviation. One reason for this might be the relatively large standard deviation of the averaged UDI-improvement in the data set. The explained variance was relatively low at 0.32. This might be an effect of averaging the UDI-scores. For the UDI genital prolapse improvement(UDI_GPR) score, the mean

Table 3.3: Random forest regressor test results

Output	Output description	MAE	Explained variance	Explained absolute deviation
UKL	Urinary distress inventory improvement	10.01	0.32	0.29
UKL_GPR	UDI genital prolapse improvement	13.54	0.70	0.52

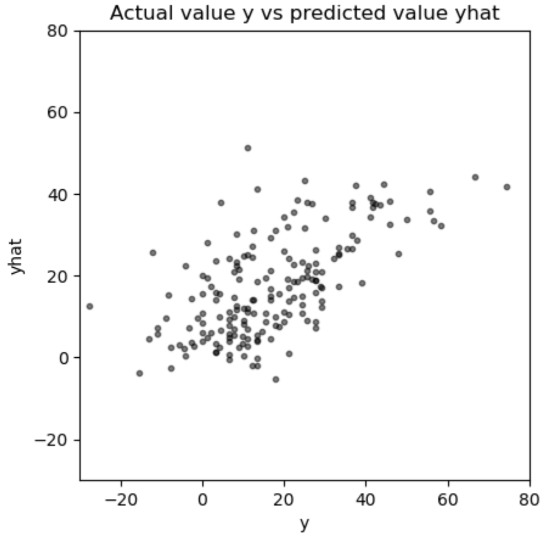

Figure 3.1: Random forest regression prediction $\hat{y}$ vs actual value y. They represent the difference between pre- and post operation average UDI in the test set.

absolute error of 13.53 was significantly higher than for the averaged UDI improvement score. However the explained variance was higher at 0.70. It appears that the predictor is better at accounting for dispersion of a single score in the UDI. The same is true for our third metric, were the MAE is adjusted for standard deviation. This can also be seen in the regression plot of the prediction $\hat{y}$ vs the actual value y, figure 3.1 and 3.2. One reason for this difference may be that the UDI variable is an average of all UDI scores, where the UDI_GPR is not an average.

Finally, we examine the feature importances of our regressor for average UDI and the UDI genital prolapse improvement. The feature importances are a ranking of which feature (variable) decreases the impurity the most. The impurity is a measure of the importance of each variable as the splitting criteria in the regression trees of the random forest. This is ranked on a scale on a scale from 0 to 1, summing up to 1.

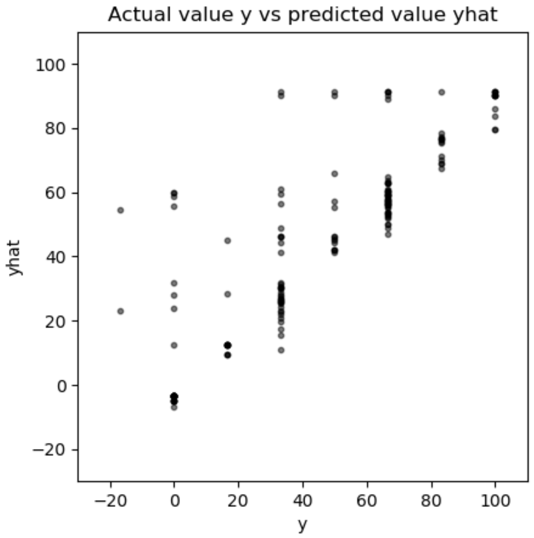

Figure 3.2: Random forest regression prediction $\hat{y}$ vs actual value y. They represent the difference between pre- and post operation UDI_GPR in the test set.

Feature Importances

From the feature importance ranking for the UDI-average improvement score 3.3, it is clear that the average UDI-score before the operation is the most important at over 0.55. The pelvic organ prolapse quantification (POP-Q) scores are also quite relevant. Age, body mass index and previous birth detail variables are ranked highly. The type of surgery seems to be irrelevant to the prediction. This may be because most patients show an improvement in UDI-scores after their operation. The condition of the patient before surgery seems to matter more than the exact type of surgery.

The results for the UDI genital prolapse score 3.4 are similar, but the UDI-genital prolapse pre-operation score is even a stronger predictor at 0.75. The other factors seem to weigh less, but are relatively comparable to their feature importance for the UDI-average score. The stronger feature importance of the pre-operation score may explain the better predictive score for the variance of the random forest regressor.

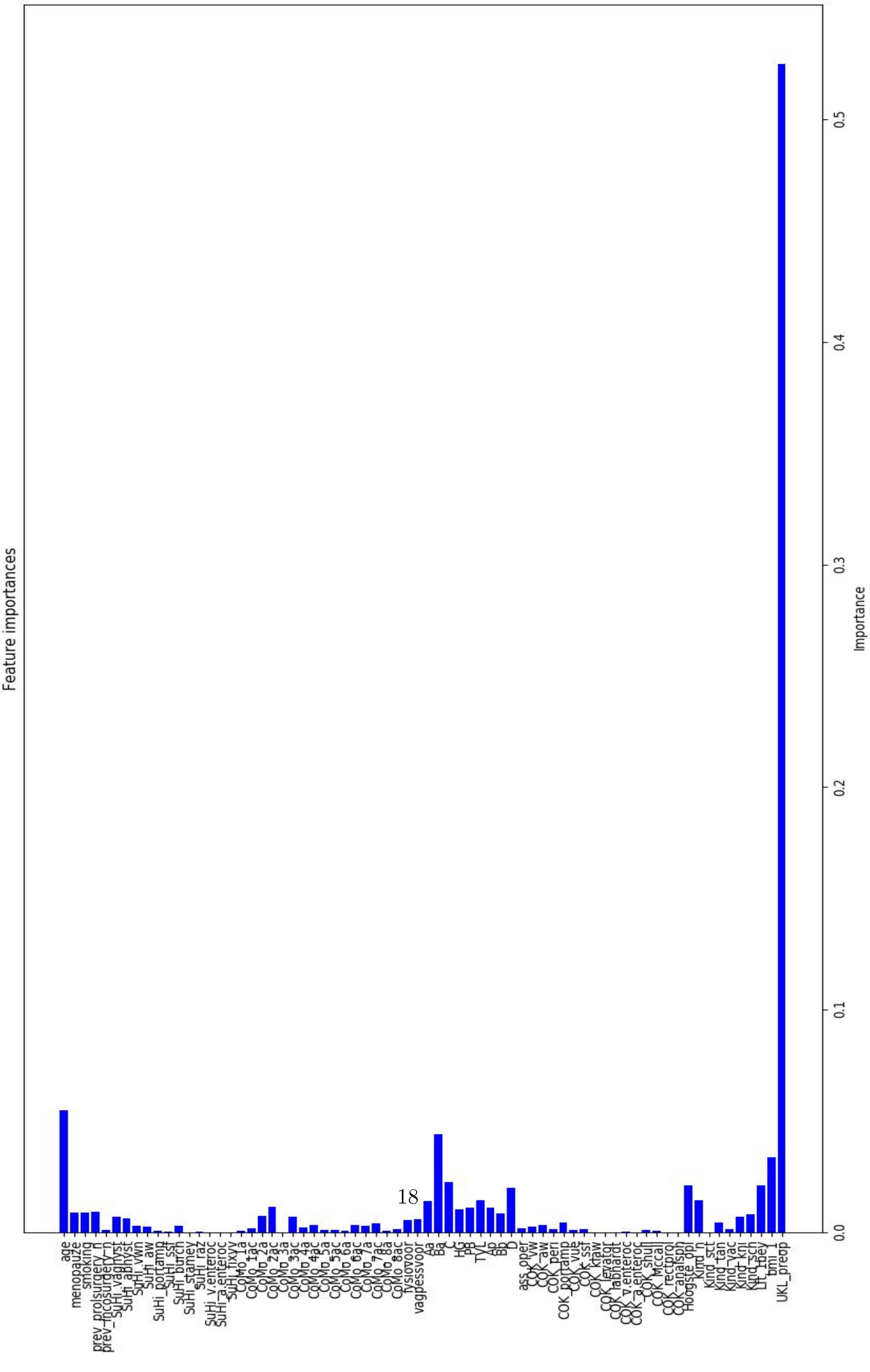

Figure 3.3: Feature importances for the random forest regressor which predicts the difference between pre- and post operation average UDI .

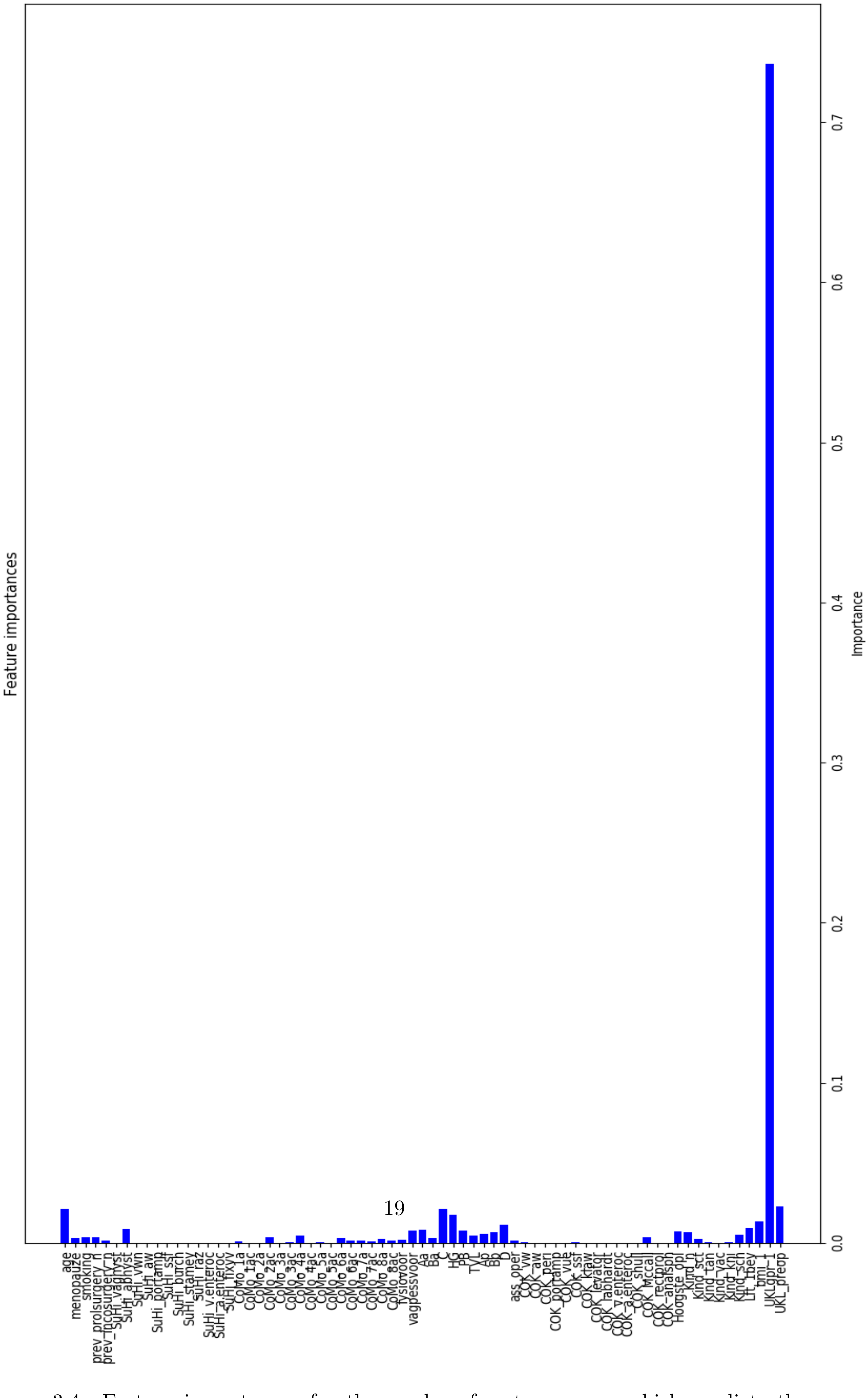

Figure 3.4: Feature importances for the random forest regressor which predicts the difference between pre- and post operation UDI_gpr.

3.4.2 Causal Forest

The causal forest calculated an estimate of the average heterogeneous treatment effect $\tau(x)$ for 6 different types of surgery for patients $i = 1, ..n$. The variables in table 3.2 were chosen as feature vectors X_i. The surgery type for which we estimate $\tau(x)$ is treatment W_i. Treatment W_i was left out of features X_i. Outcome Y_i was the mean difference in patient UDI score pre- and post operation. The higher the treatment effect value, the more positive the effect of the treatment. We trained a generalized random forest for each of the surgery types.

An important issue is that our data set only contained cases where at least one type surgery was performed. Our data set did not contain cases where no surgery was performed. This means there was no case where $W_i = 0$ for all possible surgery treatments. The treatment effect is thus the treatment effect of one type of surgery compared with all the other types of surgeries. This is a major limitation of our study, which needs to be taken into consideration when looking at the results. It is also occurred that one patient received multiple types of surgery at the same time.

The results of the causal forest in table 3.4 estimate the average treatment effect on the treated sample where $\mathbb{E}[Y_i^{(1)} - Y_i^{(0)} | X_i = x, W_i = 1]$. These are the samples where the patient receives treatment W_i. The number of cases with the respective treatment is also given. As the total number of cases exceeds the number of patients (730), it is obvious that some patients receive multiple treatments simultaneously.

Table 3.4: Average treatment effect on treated:

Treatment (W_i)	**Nr of cases($W_i = 1$)**	**Average treatment effect on treated**
COK_vw	537	0.24
COK_aw	485	1.11
COK_peri	108	0.61
COK_portamp	130	0.94
COK_vue	140	0.89
COK_ssf	137	-0.31

For a number of surgeries types there were too few cases to calculate valid treatment effect values. We did not include these in the results. Noticeable is also that each entry may have multiple surgeries conducted simultaneously. We were able to detect a strong treatment effect of the posterior colporrhaphy (COK_aw), portio amputation (COK_portamp) and vaginal hysterectomy (COK_vue surgery). The treatment effect for perineoplasty (COK_peri) was slightly less. We only found a small treatment effect for anterior colporrhaphy (COK_vw). This was somewhat strange. The reason for

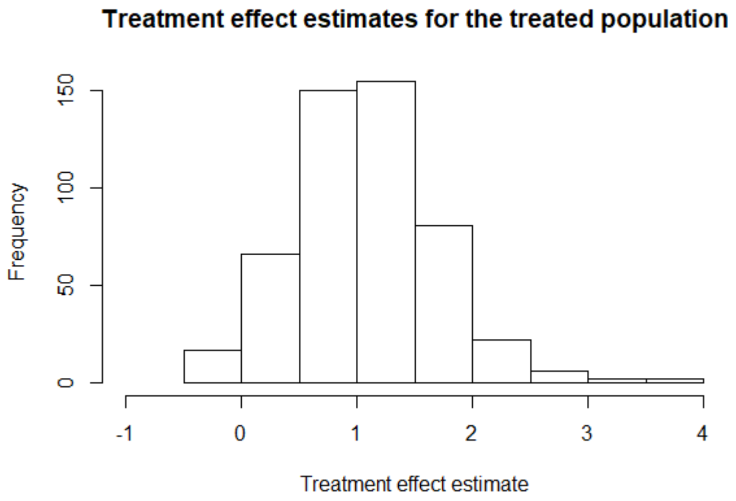

Figure 3.5: Treatment effect estimates as determined by the generalized random forest estimator for the treated part of the sample with treatment COK_aw.

it might be that almost all pelvic floor surgeries in the dataset include this surgery. This may reduce its own effect. Sacrospinal fixation (COK_sff) had no measurable positive treatment effect. This may be an anomaly in the data. Figures 3.5 and 3.6 depict histograms showing the distribution of treatment effect estimates found by the generalized random forest estimator for the treated and non-treated cases for COK_aw. When comparing the histograms it becomes clear that the treatment effect estimates for the treated population are distributed further to the right. This implies that the treatment effect estimate for the treated population is higher than for the non-treated population.

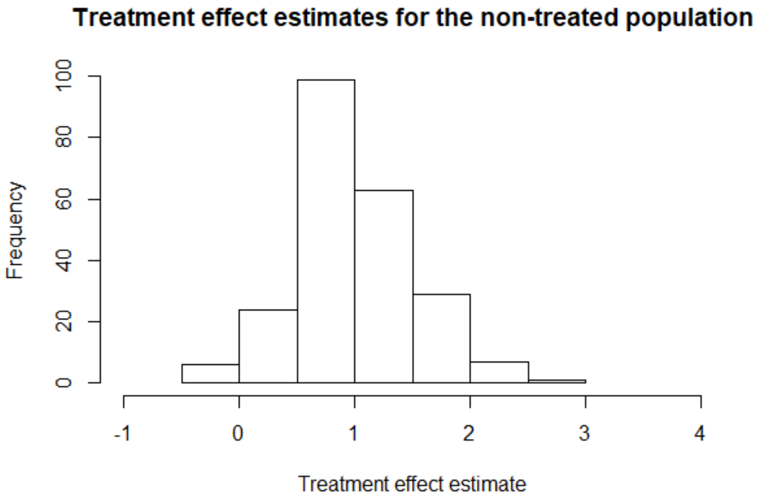

Figure 3.6: Treatment effect estimates as determined by the generalized random forest estimator for the non-treated part of the sample with treatment COK_aw.

Chapter 4

Related Work

In this chapter we describe related research and other approaches to prediction and treatment effect estimation with machine learning.

We have found no previous work involving the use of random forests or generalized random forests for the prediction of specific pelvic floor surgery outcomes. Machine learning approaches have however been used in related areas. Robinson, Swift and Johnson conducted research where they used artificial neural networks to predict pelvic organ prolapse from a data set containing healthy and affected patients [5].

There are many options for the prediction of target outcomes with machine learning. These options include regression trees, nearest neighbor approaches, artificial neural networks and the random forest which we used. Nearest neighbor algorithms use the closest training examples next to the sample for which a prediction is wanted to provide an estimate. Artificial neural networks are inspired by biological neurons and use layers of nodes with adjustable weights to learn the best output for a given set of input features [6]. They have the disadvantage of being somewhat of a black box, as they do provide scant information on the importance of specific input features on the prediction.

For the estimation of the average treatment effect we used the causal forest variant of the generalized random algorithm. Previously, Wager and Athey [11] used a causal forest method which they call Procedure 1. This closely matches the generalized random forest. Both methods use a random forest variant to predict the treatment effect. One difference is in the calculation of the final treatment estimate from all the trees. A generalized random forest uses nearest neighbor matching to calculate estimates. On the contrary, Wager and Athey's causal forest has each tree compute an estimate of the treatment effect and then averages the result of the trees. Another difference lies in the construction of the trees. Here the generalized random forest uses a gradient based criterion to split the feature space into leafs. The causal forest applies an exact loss criterion for the split.

Chapter 5

Conclusions

The goal of this thesis was to predict pelvic floor surgery outcomes and give insight in the variables influencing them. We used a random forest regressor for the outcome prediction. This regressor predicted improvement in patient score in the average urinary distress inventory(UDI) and the UDI genital prolapse score(UDI_GPR) on a scale from -100 to 100. The predictor for the mean UDI-improvement score achieved a mean absolute error of 10.01 and an explained variance of 0.32. For the UDI_GPR, our predictor achieved a mean absolute error of 13.53 and an explained variance of 0.70. The explained variance here was larger than for the UDI. The feature importances of the random forest revealed that the prior UDI-scores were the most important features for prediction, followed by POP-Q scores.

We tried to gain insight into the effect of the different surgery types. We estimated an average treatment effect for different types of pelvic floor surgeries with a generalized random forest. For a number of surgery types we did not have enough data to estimate the treatment effect. One major limitation of our research is that we did not have cases where no surgery was performed. Our treatment effect estimation could thus only use data where different types of pelvic floor surgery where performed. This means it could not use cases without treatment in its calculation of the treatment effect. For a first attempt this was adequate, but future efforts need to account for this.

In order to use a random forest predictor for clinical application, the performance could be improved in further research. A larger data set of pelvic floor surgeries, to train the random forests on, might be a suitable solution. This might be especially helpful for treatment effect estimation of rarer surgery types in our data set.

In order to improve the treatment effect estimates of the causal forest, the data set could be supplemented with pelvic floor disorder cases where no surgical intervention was performed. Another approach would be to decrease the number of missing values in the collected data. Aside from improving the quantity and quality of the data set, further research might also look at using different machine learning techniques than our random forest approach.

Bibliography

[1] Susan Athey, Julie Tibshirani, Stefan Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

[2] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[3] Olshen Breiman, Friedman and Stone. *Classification and Regression Trees*. CRC, 1984.

[4] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006.

[5] Robinson CJ, Swift S, Johnson DD, and et al. Prediction of pelvic organ prolapse using an artificial neural network. *American Journal of Obstetrics and Gynecology*, 199:2:193.e1–193.e6, 2008.

[6] Friedman Hastie, Tibshirani. *Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2009.

[7] Jean M Lawrence, Emily S Lukacz, Charles W Nager, Jin-Wen Y Hsu, and Karl M Luber. Prevalence and co-occurrence of pelvic floor disorders in community-dwelling women. *Obstetrics & Gynecology*, 111(3):678–685, 2008.

[8] Christopher M. Maher, Benny Feiner, Kaven Baessler, and Cathryn M. A. Glazener. Surgical management of pelvic organ prolapse in women: the updated summary version cochrane review. *International Urogynecology Journal*, 22(11):1445, Sep 2011.

[9] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[10] Pang-Ning Tan et al. *Introduction to data mining*. Pearson Education India, 2007.

[11] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.