Delft University of Technology
Faculty of EEMCS

When Can National-Level Colorectal Cancer Patient Data Be Pooled Across Years: A Statistical Analysis of Temporal Stability

by
M. A. (Merel) van Gruijthuijsen

to obtain the degree of Master of Science in Applied Mathematics
at the Delft University of Technology,
to be defended publicly on July 16, 2026, at 10:00 AM.

Student number: 5263247

Thesis committee:	Dr.	Ö. Şahin,	TU Delft,	Daily Supervisor
	Dr.	G.F. Nane,	TU Delft,	Responsible Supervisor
	Dr.	F. Mies,	TU Delft,	External Examiner

TU Delft

Contents

Acknowledgments

First of all, I want to thank my supervisors Özge Şahin and Tina Nane for guiding me throughout this thesis. Thank you for giving me freedom, for checking in on me, for setting boundaries when it was enough, for the continued feedback throughout the process, for the shared perplexity on the data extraction, for the intriguing questions you asked that improved the quality of this thesis so much, for the time you always had for me, for the valuable feedback on the writing, for the effort of trying to get my diploma on time, for the many coffees, for the many talks, both about this thesis, and about the future. It was both inspiring and fun to work with you. Thank you!

I also want to thank my committee member Fabian Mies, for making the time to go through my quite lengthy thesis and bringing in his expertise.

Without the data of the Dutch Institute for Clinical Auditing (DICA) this whole thesis would not have been possible, so I want to thank them for that, and especially Jacqueline for answering my questions. Also the medical and data experts from MST that I have worked with, Eline, Annemieke, Annemiek, Daan and Reini, I want to thank for their knowledge and time. The ability to combine their (medical) knowledge with mathematical knowledge has definitely enhanced the content of this thesis.

As this thesis marks the end of my student time at TU Delft, I am looking back with a lot of joy on the many experiences that I had and the wonderful friendships I made during this period. When I started this masters I expected my most memorable student years to be already behind me, but that was definitely not the case. In the master I still met so many more great people, made a ton of amazing memories, and learned so much. I want to thank everyone that made my student time so memorable and fun, you know who you are.

That being said, I want to give a special thanks to my boyfriend and family for always being there for me, for listening, supporting, encouraging, and many many other things. Thank you.

Abstract

Combining clinical registry data collected over several years can increase statistical power, but may also introduce bias when the underlying data generating process changes over time. This thesis investigates which years of the Dutch ColoRectal Audit (DCRA) can be pooled for statistical analysis without introducing practically relevant temporal heterogeneity.

The analysis uses national colorectal cancer surgery data from 2018 to 2024. The data are harmonised using the registry documentation and input from a medical expert. This has resulted in a comprehensive, refined dataset, including 47242 observations and 123 variables.

Temporal stability was assessed at three levels. Marginal changes in individual variables were quantified using effect size measures. Changes in the dependence structure were examined through year specific pairwise correlation matrices and their differences across consecutive years. Finally, a random forest analysis was used as a multivariate analysis to identify variables contributing to differences between years.

Since the effect sizes were classified as small, the corresponding marginal distributions in the univariate setting were sufficiently similar to pool the data of different years of one variable. Combining the analysis of the marginal distributions and the found evidence that the pairwise dependence structure among the selected variables remained stable between consecutive years, we concluded that these results support pooling the selected variables across the compared consecutive years. The outcomes of the random forest analysis are consistent with the effect size analysis.

As an illustration, the data of individual years and data with the consecutive years pooled were used in exploratory analysis of severe postoperative complications, in which the time between diagnosis and surgery emerged as a potentially relevant variable.

The framework described is applied to DCRA data, but the underlying problem occurs more broadly in evolving medical and administrative registries. The methods used in this thesis can therefore be expanded to other registries as well as to other sectors.

1 | Introduction

A conventional way of working with data to improve decision making in healthcare is by collecting data over multiple years. To draw conclusions from the collected data, the dataset is analysed as a whole to determine whether there are certain variables related to certain outcomes. However, combining several years of registry data without first evaluating temporal stability[1] may produce conclusions that reflect changes in data collection rather than changes in patient care. If that is the case, adjusting decision-making to the drawn conclusions might lead to undesired outcomes.

This problem is particularly relevant for the Dutch ColoRectal Audit (DCRA). This audit was established in 2009 to decrease the occurrence of severe complications following colorectal cancer surgery. Today, all Dutch hospitals performing colorectal cancer surgery participate in this audit, which is managed by the Dutch Institute for Clinical Auditing (DICA) [1]. In DCRA, data is collected on primary colorectal cancer surgeries. Colorectal cancer accounts for nearly one in ten of all annual cancer diagnoses in the Netherlands, resulting in over 12.000 new cases each year [2]. As of 2025, colorectal cancer ranks as the third most common cancer among females and the fourth among males [3]. The five-year survival rate for colorectal cancer is currently 69% [4].

Healthcare professionals use the insights from DCRA to evaluate the quality and effectiveness of their care. The goal is to identify areas for improvement and ultimately enhance care for colorectal patients [1]. However, if the produced conclusions reflect changes in data collection rather than changes in patient care, their interpretation may be misleading.

The DCRA consists of data over multiple years, namely from 2009 to 2024. Pooling[2] several years of DCRA data can substantially increase the sample size available for studying uncommon outcomes, including severe postoperative complications. At the same time, temporal heterogeneity may introduce bias. This could lead to unfounded conclusions which healthcare specialists could base treatment decisions on.

The central question of this thesis is: Which years of the DCRA registry can be pooled for statistical analysis without introducing practically relevant temporal heterogeneity?

Answering this question requires more than testing whether annual distributions are exactly equal. In a large national registry, even negligible differences may be statistically significant. This thesis therefore focuses on the magnitude and structure of differences across years, using effect-size measures, changes in associations, and multivariate classification, rather than relying on conventional p-values.

To achieve the answer, first, the data are harmonised using the registry documentation and input from a medical expert. This step addresses changes in variable definitions, registration rules, missingness patterns, and clinical applicability. The original dataset contains 221 variables

[1] Stability means that there are no differences that are practically important and sufficiently large to affect further analyses. With stability, it is still possible that there are some changes across years that are statistically detectable, especially with a very large registry.

[2] Pooling is combining multiple independent datasets or samples into a single, larger dataset.

and 147935 observations. Since the quality of any research output is inherently tied to the quality of its input, a significant portion of this work is dedicated to a thorough examination of the dataset itself with its contents, limitations, and potential gaps. This has resulted in a comprehensive, refined dataset, designed to serve as a valuable resource for both this study and future research[3].

Second, marginal temporal stability is assessed by comparing the distributions of individual variables across years. Effect size measures are used to quantify the magnitude of the observed differences. Small effect sizes are interpreted as evidence that the corresponding marginal distributions are sufficiently similar for the intended analysis. Note that this conclusion is subject to the limitations of the selected measures and thresholds.

Third, temporal stability in the dependence structure is investigated. This analysis examines whether variables may have similar marginal distributions while their joint structure changes over time. This is achieved with a pairwise correlation analysis and constructing pairwise correlation matrices per year. To quantify the change in association over the years, difference matrices are constructed from the pairwise correlation matrices per two sequential years. If the change in correlation is small, we conclude that the data of the subsequent years of multiple variables can be pooled.

Finally, multivariate temporal stability is assessed by examining whether observations from two different years can be distinguished using the selected variables. A random forest classifier is fitted with the year of observation as the class label. Then, variable importance measures indicate which variables contribute most to differences between years. These results are compared with the conclusions from the marginal and the dependence effect size analyses.
Additionally, by combining the evidence from the marginal, dependence, and multivariate analyses, we formulate recommendations on which years can be pooled. Namely, the variables `waitingTime` (waiting time), `severeCompl` (severe complications), `scopdistmm` (distance in mm from the lower edge of the first tumour to the anorectal junction on sagittal MRI (or CT, if applicable)), `leeftijd` (age), and `asascore` (ASA score) can be pooled for consecutive years from 2018 to 2024. Continuing with data of individual years and data with the consecutive years pooled, to inspect whether the selected variables contain potentially relevant variables for a binary outcome, e.g., severe complications, we use the random forest classifier as well. This approach revealed the time between diagnosis and surgery as a potentially relevant variable for severe postoperative complications. The exact relation between the variables can be investigated in further research. This analysis illustrates why the pooling decision matters: conclusions about clinical outcomes are credible only when the underlying observations are sufficiently comparable over time.

The framework described is applied to DCRA data, but the underlying problem occurs more broadly in evolving medical and administrative registries. The methods used in this thesis can therefore be expanded to other registries as well as to other sectors.

In Chapter 2 the preliminaries are given to provide the reader with the correct foreknowledge. In Chapter 3 a literature review is conducted which gives insights into the earlier studies done with this dataset, and into relevant mathematical studies. In Chapter 4 a description of the data is given. In Chapter 5 the imputations of the data are done and the final dataset is retrieved. In the final section of this chapter a univariate data analysis is conducted to find out how the selected variables behave over the years. In Chapter 6 a temporal stability analysis is performed, consisting of an effect size analysis and analyses with a random forest classifier. Chapter 7 contains the conclusion of the thesis and the discussion and recommendations for future research.

[3]This dataset is not publicly available, due to the medical nature of the dataset.

2 | Preliminaries

Since this thesis requires both prior medical knowledge and prior mathematical knowledge, this chapter is split into two sections. In Section 2.1 the medical preliminaries are treated. In Section 2.2 the mathematical preliminaries are handled.

2.1 Medical preliminaries

Since this thesis and the data on which this thesis is conducted are highly medical, some medical prior knowledge is also necessary for understanding the thesis. In this section we will discuss the common treatment that colorectal[1] cancer patients get and relevant medical definitions.

2.1.1 Treatment process

We will go through the process of colorectal cancer treatment to clarify the medical preliminaries. A flow chart of this process is in Figure 2.1. The solid lines show the ideal progress of the treatment. The dotted lines are lines that are not the ideal situation, but happen as well.

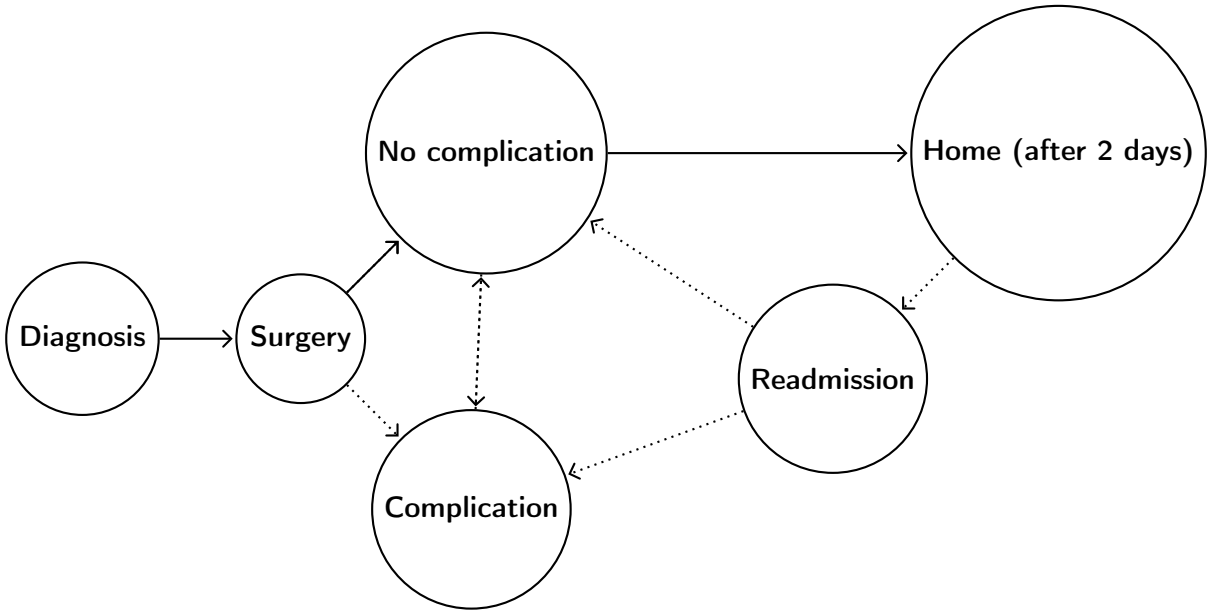


Figure 2.1: Flow chart of process of colorectal cancer treatment. The solid lines show the ideal progress of the treatment. The dotted lines are lines that are not the ideal situation, but happen as well.

The first part in the treatment of colorectal cancer is diagnosing the colorectal cancer. This

[1] Concerning the colon and rectum [5], see Table 2.4.

diagnostic procedure is also known as an endoscopy [6]. Generally, during the diagnosis or before the surgery, the overall health of the patient is estimated. One common way to classify a patients overall health is via the ASA (American Society of Anaesthesiologists) score. There are five ASA scores, and their meanings are described in Table 2.1.

ASA score	Meaning
I	Normal healthy patient
II	Mild systemic disease
III	Severe systemic disease
IV	Constantly life-threatening systemic disease
V	Moribund

Table 2.1: The possible ASA scores and their meaning [7].

Before the surgery takes place, the location of the tumour has to be determined. The colon exists of multiple areas which are the possible locations for the tumour. The areas we describe are the caecum (cecum), appendix, colon ascendens, flexura hepatica, colon transversum, flexura splenica, colon descendens, colon sigmoideum, and rectum. A visualisation of these possible locations of colorectal carcinoma (tumour) is in Figure 2.2.

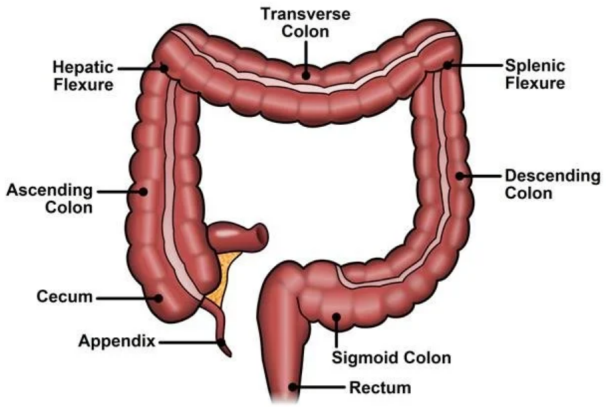


Figure 2.2: Figure from [8]. The different areas of the colon, used to describe the location of the tumour.

It is also possible for a patient to have a stoma prior to the surgery, or created during the surgery. A stoma is a 'surgically created opening in the skin that connects to one of the organs' [9]. In the context of colorectal cancer, the organ that the opening in the skin is connected to is the colon. This is visible in Figure 2.3.

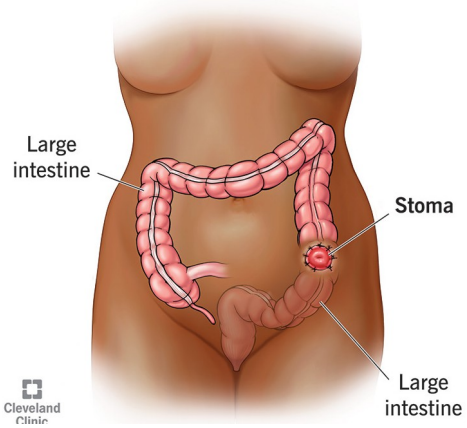


Figure 2.3: Figure from [10]. Visualisation of a stoma.

After the diagnosis, most of the times a surgery will take place. In principle, all colorectal cancers must be surgically removed. A large proportion of colorectal cancers can be treated with surgery alone. Some patient also receive chemotherapy or radiotherapy in addition to the resection. In the data we are looking at, patients received a primary resection due to a colorectal carcinoma. A primary resection is a medical procedure in which the surgeon removes the part of the colon that contains the tumour [11]. The surgeon also removes the lymph nodes close to the tumour. After that, the surgeon closes the end points of the colon again to each other. That is called an intestinal suture.

The stage of the cancer has a big influence into how to approach the treatment and the surgery. Depending on the size and stage of the tumour or whether the cancer has spread to other tissues outside the colon, a decision may be made to administer radiotherapy first, followed by surgery. Chemotherapy may also be given after the surgery. In previous versions of the Dutch guideline, adjuvant chemotherapy was recommended for all patients without lymph node metastases who had one or more of the *risk factors*[2] [2]. There are four stages of colorectal cancer (I-IV). The stages and their meaning are in Table 2.2 and visualised in Figure 2.4. These stages are classified with the TNM classification, see [2].

Stage	Meaning
I	The tumour is confined to the bowel wall itself.
II	The tumour has grown through the wall, but has not spread to the lymph nodes.
III	The tumour has spread to the local lymph nodes.
IV	The tumour has spread to distant lymph nodes or other organs and/or tissues in the body.

Table 2.2: The four stages of colorectal cancer and their meaning [12].

[2]Risk factors include a poorly differentiated or undifferentiated tumour, pathological T4 stage, vascular invasion, obstructing or perforated tumours, and fewer than ten lymph nodes assessed.

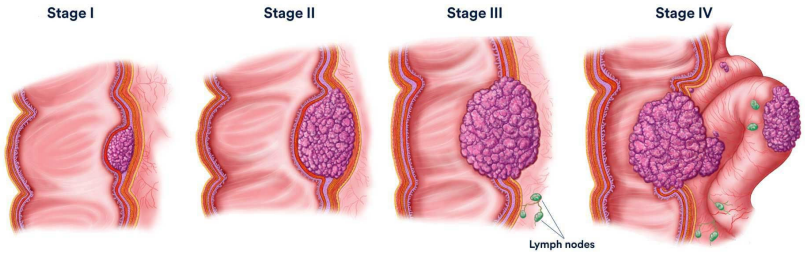


Figure 2.4: Figure from [13]. The different stages of colorectal cancer.

After the surgery, it is possible that a complication occurs. If there does not occur a complication, patients are sent home after two days in the hospital (according to our medical experts). In the hospital, no blood samples are taken for lab tests after this surgery, as the levels are always elevated and therefore don't really tell much. As long as the pain is under control, patients feel well and have no fever or elevated heart rate, they can go home. When they are sent home, they are provided with a good support network so they can raise alarm if problems arise. It is then still possible that they get a complication, and then, if the complication is severe enough, they will be readmitted to the hospital. Often home monitoring takes place to help identify complications. Patients come back to the hospital if their pain increases, they feel unwell or have a fever.

During the treatment of a patient with colorectal cancer, different specialists and people are involved. These professionals each have their own tasks and responsibilities. An overview of this is given in Table 2.3.

Person	Task
Anaesthesiologist	Coordinates perioperative pain treatment.
Assistant physician or specified nurse	Assists the surgeon and decides on a plan of action.
Case manager oncology	Informs patients and other nurses.
Holding/Recovery Nurse	Takes care of patients when going in or out of surgery.
Multidisciplinary team	A team that consists of all mentioned specialists: a surgeon, an oncologist, a radiologist, a radiotherapist, and a gastroenterologist [14].
Nurse	Informs other staff about the recovery of a patient.
Patient	Undergoes surgery and recovery.
Stoma nurse, oncology nurse	Informs patients and other nurses.
Surgeon	Decides plan of action and performs the surgery .

Table 2.3: The professionals involved in the treatment of colorectal cancer and their tasks.

2.1.2 Definitions

In this subsection we treat the definitions that will be used throughout the thesis. These are in Table 2.4.

Term	**Meaning/Definition**
Abdomen	Medical term for the belly [5].
Abdominal	Concerning the abdomen [5].
Abdominal cavity	Inside of the abdomen [5].

Abscess	Preoperative abscess formation in the intraperitoneal or extraperitoneal spaces [14].
Anaemia	Condition of not having enough functioning blood cells [5].
Anaesthesia	Medication administered to patient to be unaware of the surgery [5].
Anaesthesiologist	Medical doctor in charge of anaesthesia [5].
Anastomic leak	Leakage of bowel contents through anastomosis to the abdominal cavity [5].
Anastomic leakage	Clinically relevant anastomic leak requiring a radiological or surgical reintervention [14].
Anastomosis	The surgically made connection between the bowel, bowels are reconnected to themselves after removing part of the bowel [5].
Bleeding	Preoperative tumour related blood loss that requires an intervention or leads to anaemia [14].
Bowel	The small and large intestine and the rectum [5].
Cardiac complications	A complication which cause is related to the heart, e.g., cor pulmonary or myocardial infarction [15].
Colloids	Type of fluid that can be administered through Intra Venous (IV).
Colorectal	Concerning the colon and rectum [5].
Comorbidity	The presence of one or more other conditions influencing health besides the disease of interest [5].
Conversion of surgery	Changing the surgery from laparoscopic or robotic to open surgery during the surgery [5].
Epidural	An injection in the spine, often an anaesthesia [5].
Failure to rescue	Percentage of patients with severe complication who die [3] [16].
Gastrointestinal	Concerning the stomach and/or the intestines [5].
Ileus	Preoperative presence of (partial) mechanical bowel obstruction with symptoms of abdominal cramping, abdominal distension, nausea, vomiting or failure to pass gas or stool [14].
Infectious complications	Any infection other than surgical infection or pneumonia, e.g., urinary tract infection [15].
(Low) anterior resection	Rectosigmoid or rectal resection according to the TME principle with anastomosis of the colon to the intra- or extraperitoneal rectum or anal canal [14].
Metastasis	(Cancer) cells that have moved from their origin [5].
Mild complication	Any complication occurring within 90 days after resection and not being a severe complication [15].
Morbidity	Condition of being sick [5].
Mortality	Death within 90 days after the surgery or within the same admission [16].
Neurologic complications	A complication which cause is related to the nerve system or brain, e.g., cerebrovascular accident [15].
Perioperative	Throughout the whole process that is surgery (before, during and after surgery) [5].
Positive CRM	A circumferential resection margin of 1 mm or less [14].
Postoperative mortality	Mortality during the same hospital admission of the surgical resection or within 90 days after resection [17].
Pre- inter- and postoperative	Before, during, or after surgery [5].
Pulmonary complications	A complication mainly related to the lung (except for pulmonary embolism), e.g., bacterial pneumonia [15].

[3] Calculated as (number of patients who died as a result of severe complication)/(total number of patients who experienced a severe complication).

Rectal carcinoma	Carcinoma located within 150 mm of the anal verge [18].
Reintervention	An invasive (surgical, radiological or endoscopical) measure to treat a complication[4] [14].
Severe complication	Complication leading to a surgical endoscopic, or radiological reintervention or to a in-hospital stay of more than 14 days, or death [16].
Stoma	An artificial connection made by surgery, connecting part of the bowel to the outside directly through a hole in the belly [5].
Thromboembolic complications	A complication which cause is related to the blocking of a blood vessel due to the formation of a blood clot (except for cerebral accidents), e.g., pulmonary embolism or deep vein thrombosis [15].
Thrombosis	The complication due to the solidification of blood (blood clots) [5].
Total colonoscopy	Preoperative visualisation of the entire colon including the ascending colon by colonoscopy or CT colonography [14].
Tumour perforation	Preoperative tumour perforation with clinical signs of faecal peritonitis[5] [14].
Urgent procedure	Non-elective colorectal resection that was required and performed within 24 hours of admission [14].

Table 2.4: Medical definitions.

2.2 Mathematical preliminaries

In this section we describe the mathematical preliminaries. We discuss in detail the decision tree and the random forest classifier.

We want to reach the definition of a *decision tree*. In order to do so, we first need to define other concepts. These are given in Definitions 2.1-2.10.

Definition 2.1. *A **graph** G is a pair of sets, $G = (V, E)$. V is the set of vertices (also called nodes) and must be non-empty. These vertices are the points in a graph. The set E are the edges and is a subset of the set $\{\{u, v\} : u, v \in V, u \neq v\}$. These edges are the lines in a graph that connect the different vertices [19].*

Definition 2.2. *A **multigraph** is a triple (V, E, I) where V is non-empty, V and E are disjoint, and I is a relation between V and E, such that every element of E is related to one or two elements of V by I.*

Definition 2.3. *A **walk** in a multigraph is an alternating sequence $v_0, e_1, ..., v_{k-1}, e_k, v_k$ of vertices and edges such that for all $1 \leq i \leq k$ the edge e_i has endpoints v_{i-1} and v_i. A **trail** is a walk that has no repeated edges and a **path** is a trail that has no repeated vertices. A $u - v$ **walk** is a walk where $u = v_0$ and $v = v_k$, which are known as the endpoints of the walk and the other vertices in this walk are known as **internal vertices**. The length of a walk, trail, or a path, is the total number of edges in it. A walk or trail is called **closed** if $v_0 = v_k$. A **cycle** is a closed walk where $v_1, ..., v_{k-1}$ are distinct.*

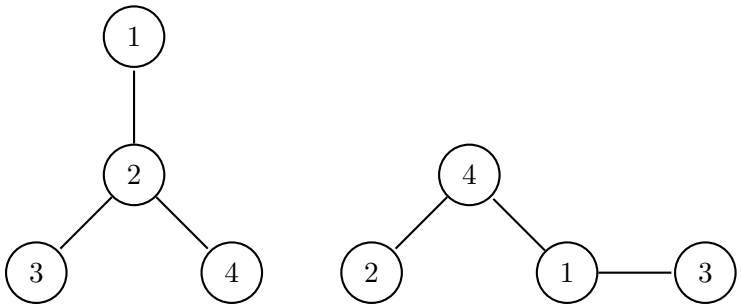
Definition 2.4. *A graph G is called **connected** if it either has just one vertex or between any two distinct vertices in G there is a path.*

[4] Excluding superficial drainage abscess of a wound abscess on the patient ward, introduction of a nasogastric tube, a central venous catheter, or tracheotomy.

[5] A faecal peritonitis is an inflammation of the peritoneum.

With these definitions, we can reach the definition of a *tree*. In Figure 2.5 there are two examples of trees. We need more definitions to reach the definition of a *decision tree*.

Definition 2.5. *A **forest** is a graph that does not contain any cycles. A **tree** is a forest that is connected.*



(a) Star. **(b)** Path.

Figure 2.5: Two trees on 4 vertices.

Definition 2.6. *A graph is **directed** when the edges have a direction. These directed edges are called **arcs**. The node from which the arc originates is the **tail**. The node to which the arc is directed is the **head**.*

Definition 2.7. *For $v \in V(G)$, the number of arcs with their direction towards v is referred to as the **indegree** of v, denoted by $indeg(v)$. The number of arcs with their direction away from v is referred to as the **outdegree** of v, denoted by $outdeg(v)$ [20]. The **parents** of a vertex v are the vertices that are the tails of the incoming arcs of v. The **children** of a vertex v are the head of the outgoing arcs of v.*

Definition 2.8. *Let $G = (V, E)$ be a graph and $u \in V$. Then the set $N(u) = \{v \in V : uv \in E \text{ or } vu \in E\}$ is known as the **neighbourhood** of u. The size of this set is known as the **degree** of the vertex u, denoted by $deg(u) = \#N(u)$.*

Definition 2.9. *A **binary** graph is a graph where all nodes have a maximum indegree and outdegree of 2 and total degree at most 3.*

Definition 2.10. *A **leaf** in a tree is a vertex v of indegree 1 and outdegree 0. A **pure leaf node** is a leaf that only contains data of one type of class.*

2.2.1 Decision tree

In Definition 2.11 the definition of a decision tree is given. Definitions that are needed as prerequisite are Definitions 2.1-2.10. Our goal for using a decision tree is to optimally split given data into different classes.

Definition 2.11. *A **decision tree** is a binary tree that recursively splits the dataset until we are left with pure leaf nodes.*

We will show the use and construction of a decision tree with an example. In Table 2.5 our dataset for this example is given, we call this dataset $D_{example}$. Dataset $D_{example}$ has six instances (0, 1, 2, 3, 4, 5) and five variables (also referred to as features) (x_0, x_1, x_2, x_3, x_4). The target variable l is the binary classification. We work towards retrieving the corresponding decision tree, by explaining step by step how this decision tree is built.

id	x_0	x_1	x_2	x_3	x_4	l
0	4,3	4,9	4,1	4,7	5,5	0
1	3,9	6,1	5,9	5,5	5,9	0
2	2,7	4,8	4,1	5,0	5,6	0
3	6,6	4,4	4,5	3,9	5,9	1
4	6,5	2,9	4,7	4,6	6,1	1
5	2,7	6,7	4,2	5,3	4,8	1

Table 2.5: Dataset $D_{example}$.

How is a decision tree built?

To build a decision tree, the decision tree finds the best split (reflected in the decision rule) by maximizing the *information gain (IG)*, defined as $IG = E(parent) - \sum w_i E(child_i)$ for $i = 0, 1$. The information gain is calculated per parent node. Here, w_i is the *weight*, the relative size of the child with respect to the parent, and E is the *entropy*, the measure of information contained in a state, which is $E = \sum -p_i \log_2(p_i)$ for $i = 0, 1$. In the calculation of entropy, p_i is the probability of class i, for $i = 0, 1$. The minimum value of entropy is 0 and the maximum value of entropy is 1. If entropy is high, it means we are very certain about a randomly picked point. The entropy of a node is equal to 0 if and only if this node is a leaf node.

For the sake of clarity, we continue with our example using dataset $D_{example}$. Suppose we are building this decision tree without knowing its final structure. We will define the nodes with letters (even though we pretend to not yet know the shape of the decision tree) to be clear about which node we are talking about. This is in Figure 2.6.

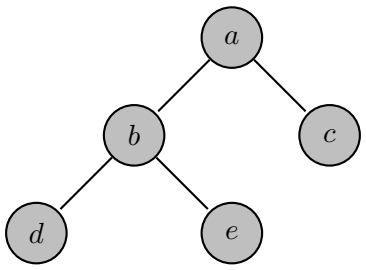


Figure 2.6: The empty decision tree that belongs to dataset $D_{example}$.

Node a is the parent node when we are considering node a, node b, and node c. The first step is to determine what should be in node a. To do so, we need to calculate the information gain IG for parent node a, and thus the entropy E for each of the possible splits. In equation (2.1) we calculate the entropy of node a, our parent node, so $E(parent)$. In dataset $D_{example}$ we have six instances in total. At node a we have not yet made any splits, and hence we also have these six instances at node a. Of these six instances, three of the instances correspond to class 0, and three of the instances correspond to class 1. Hence, $p_i = \frac{3}{6}$ for $i \in \{0, 1\}$.

$$\begin{aligned} E_a &= \sum -p_i \log_2(p_i) \\ &= -p_0 \log_2(p_0) + -p_1 \log_2(p_1) \\ &= -\frac{3}{6} \log_2\left(\frac{3}{6}\right) + -\frac{3}{6} \log_2\left(\frac{3}{6}\right) \\ &= -\log_2\left(\frac{1}{2}\right) \\ &= 1 \end{aligned} \tag{2.1}$$

There are 10 possibilities to split the data of dataset $D_{example}$ at node a, and thus make the division of instances between node b and node c. These possibilities are registered in Table 2.6. Rows 2 and 3 are equivalent, as are row 4 and 6, and row 7 and 8, and row 9 and 10. The table shows that row 4 (and row 6, due to equivalency) give the biggest information gain. The exact calculations of all the possibilities for the entropy and information gain are in Appendix A.

	division of instances ((in node b after split) - (in node c after split))	in node b (instance belonging to class 0, instance belonging to class 1)	in node c (instance belonging to class 0, instance belonging to class 1)	E_b	E_c	IG
1	6-0	(3, 3)	(0, 0)	1	0	0
2	5-1	(2, 3)	(1, 0)	0,971	0	0,191
3		(3, 2)	(0, 1)	0,971	0	0,191
4	4-2	(3, 1)	(0, 2)	0,811	0	0,459
5		(2, 2)	(1, 1)	1	1	0
6		(1, 3)	(2, 0)	0,811	0	0,459
7	3-3	(2, 1)	(1, 2)	0,918	0,918	0,082
8		(1, 2)	(2, 1)	0,918	0,918	0,082
9		(3, 0)	(0, 3)	0	0	0
10		(0, 3)	(3, 0)	0	0	0

Table 2.6: Possibilities to split the data at node a of dataset $D_{example}$.

Let us assume we decide to go for row 4 of Table 2.6 (it could as well have been row 6 because of equivalency). The calculation of the entropy of b is in equation (2.2). The calculation of the entropy of c is in equation (2.3). The corresponding information gain calculation is in equation (2.4).

$$\begin{aligned} E_b &= \sum -p_i \log_2(p_i) \\ &= -\frac{3}{4}\log_2\left(\frac{3}{4}\right) + -\frac{1}{4}\log_2\left(\frac{1}{4}\right) \\ &\approx 0,31 + 0,5 \\ &\approx 0,81 \end{aligned} \tag{2.2}$$

$$\begin{aligned} E_c &= \sum -p_i \log_2(p_i) \\ &= -\frac{2}{2}\log_2\left(\frac{2}{2}\right) \\ &= 0 \end{aligned} \tag{2.3}$$

$$\begin{aligned} IG &= E(parent) - \sum w_i E(child_i) \\ &= E_a - w_b \cdot E_b - w_c \cdot E_c \\ &\approx 1 - \frac{4}{6} \cdot 0,81 - \frac{2}{6} \cdot 0 \\ &\approx 1 - 0,54 - 0 \\ &\approx 0,46 \end{aligned} \tag{2.4}$$

Hence, the decision rule in node a is such that (assuming we look at row 4) it will divide the nodes where there are three instances of class 0 and one instance of class 1 in node b, and zero

instances of class 0 and two instances of class 1 in node c. Considering the data in Table 2.5, this results in the decision rule $x_0 \leq 4,3$.
Having calculated the information gain and the entropy, and knowing what decision rule results in the highest information gain, we can start building the decision tree, and we reach the decision tree in Figure 2.7. With this decision tree the instances with id 3 and id 4 of dataset $D_{example}$ are classified in node c as class 1, since these instances have $x_0 \nleq 4,3$. The instances that are left to classify are in node b and are the instances with id 0, 1, 2, and 5.

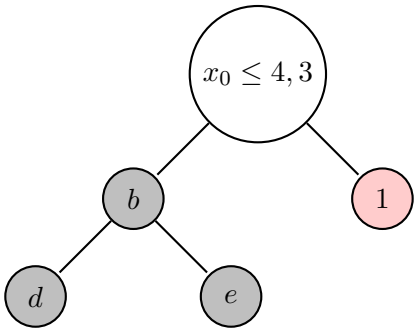


Figure 2.7: The partly filled in decision tree that belongs to the dataset in Table 2.5.

To classify the instances with id 0, 1, 2, and 5, we need to continue building the decision tree. Now node b is the parent node and nodes d and e are the child nodes. We thus need to look at what decision rule to give node b. Of the four instances that are at node b, we have three instances in class 0 and one instance in class 1. The entropy of node b is calculated in equation (2.5).

$$\begin{aligned} E_b &= \sum -p_i \log_2(p_i) \\ &= -\frac{3}{4}\log_2\left(\frac{3}{4}\right) + -\frac{1}{4}\log_2\left(\frac{1}{4}\right) \\ &\approx 0,31 + 0,5 \\ &\approx 0,81 \end{aligned} \tag{2.5}$$

There are three possibilities to split the data of dataset $D_{example}$ at node b, and thus make the division of instances between node d and node e. These possibilities are registered in Table 2.7. The exact calculations of all the possibilities for the entropy and information gain are in Appendix A.

	division of instances ((in node d after split) - (in node e after split))	in node d (instance belonging to class 0, instance belonging to class 1)	in node e (instance belonging to class 0, instance belonging to class 1)	E_d	E_e	IG
1	4-0	(3, 1)	(0, 0)	0,811	0	0
2	3-1	(2, 1)	(1, 0)	0,918	0	0,12
3		(3, 0)	(0, 1)	0	0	0,81

Table 2.7: Possibilities to split the data at node b of dataset $D_{example}$.

Since Table 2.7 shows that row 3 results in the biggest information gain, we continue with row 3 of the table. The calculation of the entropy of d is in equation (2.6). The calculation of the entropy of e is in equation (2.7). The corresponding information gain calculation is in equation (2.8).

$$\begin{aligned} E_d &= \sum -p_i \log_2(p_i) \\ &= -\frac{3}{3}\log_2\left(\frac{3}{3}\right) \\ &= 0 \end{aligned} \tag{2.6}$$

$$\begin{aligned} E_e &= \sum -p_i \log_2(p_i) \\ &= -\frac{1}{1}\log_2\left(\frac{1}{1}\right) \\ &= 0 \end{aligned} \tag{2.7}$$

$$\begin{aligned} IG &= E(parent) - \sum w_i E(child_i) \\ &= E_b - w_d \cdot E_d - w_e \cdot E_e \\ &\approx 0,81 - \frac{3}{4}\cdot 0 - \frac{1}{4}\cdot 0 \\ &\approx 0,81 - 0 - 0 \\ &\approx 0,81 \end{aligned} \tag{2.8}$$

The decision rule in node b is thus such that it will divide the nodes where there are three instances of class 0 in node d and one instance of class 1 in node e. Considering the data in Table 2.5, this results in decision rule $x_1 \leq 6,1$. With this split, the instances of different classes are split into separate nodes, and hence the decision tree has classified the data. Having calculated the information gain and the entropy, and knowing what decision rule results in the highest information gain, we can continue with the build of the decision tree. We then reach the decision tree in Figure 2.8. With this decision tree the instances with id 0, id 1, and id 2 of dataset $D_{example}$ are classified in node d as class 0, since these instances have $x_1 \leq 6,1$. The instance with id 5 is classified as class 1, since this instance has $x_1 \nleq 6,1$. Since instances 0, 1, and 2 belong to the same class, and all other nodes (d and c) also only have instances belonging to the same class, we have obtained only pure leaf nodes and thus this is the final decision tree corresponding to dataset $D_{example}$.

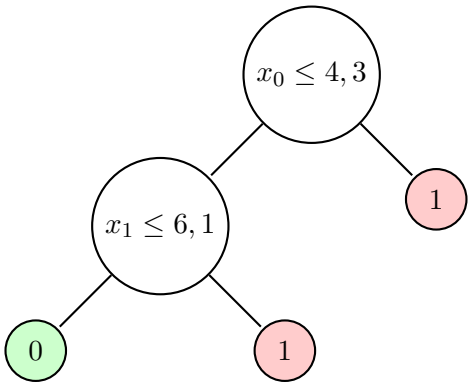


Figure 2.8: The decision tree corresponding to $D_{example}$.

How to classify a new data point?

The use of this tree is in classifying a new data point, so assigning a class to a data point. Suppose we would add the instance in Table 2.8 to dataset $D_{example}$. We do not yet know the label/class of this new instance with id 6. We determine its label/class by going through the constructed decision tree in Figure 2.8.

id	x_0	x_1	x_2	x_3	x_4	l
6	1,2	3,2	8,3	4,5	5,4	?

Table 2.8: Instance with unknown label/class to add to dataset $D_{example}$.

Suppose we are at node a. We see that for id 6 we have $x_0 = 1,2$ so indeed $x_0 \leq 4,3$. Hence, we go to node b. Then, for id 6 we have $x_1 = 3,2$ so $x_1 \leq 6,1$. Hence, we are in leaf node e, which contains the instances with id 0, 1, and 2, which belong to class 0. Hence we classify the instance with id 6 as the same class, namely class 0.

In this section we have covered the decision trees, how to make them, and how to use them to classify a new data point. To make a decision tree, the given data plays an important role. When changing the data slightly, it could already result in a completely different decision tree. Decision trees are thus highly sensitive to the training data and this could result in high variance. Hence, this model of decision trees might fail to generalize. The solution to this is the random forest, since random forests are much less sensitive to the data it is trained on.

2.2.2 Random forest

A random forest is essentially a collection of decision trees. The exact definition of a random forest is given in Definition 2.12.

Definition 2.12. *A **random forest** is a collection of multiple random decision trees.*

How to create a random forest?

Creating a random forest consist of two steps.

The first step to create a random forest is building multiple new datasets using the original data. This procedure is called *bootstrapping*. By performing bootstrapping, we ensure we are *not* using the same data for every tree, which results in different decision trees. This makes a random forest less sensitive to original training data than a decision tree. The bootstrapped dataset consists of rows from the original dataset. This is done by randomly selecting rows from the original dataset and putting it into the bootstrapped dataset. The new bootstrapped dataset must consist of the same number of rows as the original dataset, and it is allowed to pick a row multiple times.

An example of a bootstrapped dataset is given in Table 2.9, and we call this dataset $D_{example_{bootstrapped}}$. The original dataset that belongs to this bootstrapped dataset is dataset $D_{example}$ which is in Table 2.5. As is visible in Table 2.9, the rows with id 2, 4, 0, 4, 1, and 3 are selected from dataset $D_{example}$. The amount of rows in dataset $D_{example_{bootstrapped}}$, namely six, is the same as the amount of rows in dataset $D_{example}$ in Table 2.5. Note that this is only one example of a bootstrapped dataset and that to build a random forest, multiple bootstrapped datasets are constructed.

id	x_0	x_1	x_2	x_3	x_4	l
2	2,7	4,8	4,1	5,0	5,6	0
4	6,5	2,9	4,7	4,6	6,1	1
0	4,3	4,9	4,1	4,79	5,5	0
4	6,5	2,9	4,7	4,6	6,1	1
1	3,9	6,1	5,9	5,5	5,9	0
3	6,6	4,4	4,5	3,9	5,9	1

Table 2.9: Bootstrapped dataset $D_{example_{bootstrapped}}$ with the data from dataset $D_{example}$.

The second step to create a random forest is creating a decision tree using each of the bootstrapped datasets independently, while only using a random subset of variables/columns at each

step. We will thus not use every variable for training the decision trees. Instead, we will randomly select a *subset of variables* for each tree and use only these for training. We can for example decide to only use variables x_1 and x_3 to construct the decision tree, and not use variables x_0, x_2, and x_4. The ideal size of the subset of variables is about the logarithm or square root of the total. An example of this is in Table 2.10. This dataset will lead to a different decision tree than the dataset in 2.9. By randomly selecting a subset of variables, the correlation between trees is reduced, and bad predictions will be balanced out.

id	x_1	x_3
2	4,8	5,0
4	2,9	4,6
0	4,9	4,79
4	2,9	4,6
1	6,1	5,5
3	4,4	3,9

Table 2.10: Bootstrapped dataset $D_{example_{bootstrapped}}$ with variable selection with the data from dataset $D_{example}$.

Performing these two steps while building a decision tree, leads to different decision trees. The collection of these trees is the random forest.

How to classify a (new) data point using random forests?

Suppose we have a new data point of which we do not yet know the label/class and we want to predict this label/class using the random forest. This could for example be the data point with id 6 in Table 2.8. To determine the label/class of this data point, we pass the new data point through each constructed decision tree and note down whether the decision tree places the new data point in class 0 or in class 1. Then, we combine all the predictions and we take the majority voting (so the class that most decision trees have as outcomes), and assign this class as the class of the new data point. This is called *aggregation*[6]. For regression, we take the average instead of performing majority voting.

How to measure how good a random forest is?

To determine whether the random forest is a 'good' random forest, we estimate the accuracy of the random forest. To estimate the accuracy of the random forest, we need the *out-of-bag dataset*. The set of entries that did not end up in the bootstrapped dataset because they did not get randomly selected is defined as the out-of-bag (OOB) dataset. Typically around one third of the original data does *not* end up in the bootstrapped dataset.

Since the data from the out-of-bag dataset was not used to make a certain tree, we can run this data through this tree and see if the tree correctly classifies the data. The accuracy is defined as the proportion of OOB samples that were correctly classified by the random forest.

To estimate the accuracy of the random forest, the following steps need to be taken. First, we run an OOB sample through all trees that were built *without* this sample. Second, we perform majority voting and see whether the OOB sample is correctly labelled by the random forest. Then, we repeat these two steps for all other OOB samples for all of the trees.

Ultimately, we can measure the *accuracy* of our random forest by the proportion of OOB samples that were correctly classified by the random forest. The proportion of OOB samples that were incorrectly classified by the random forest is the *OOB error*.

[6]The process of bootstrapping and aggregation together is called bagging.

To make a good random forest with high accuracy, it is necessary to rebuild the random forest by estimating the accuracy and then changing the number of variables used per step to choose the most accurate random forest. In general, to find the optimal random forest, it is handy to start with the square root of the number of variables, and then try some values above and below that value.

2.2.3 Permuted random forests

A *permuted random forest* is a random forest that is constructed with data for which the labels are 'shuffled'. This shuffling of labels can best be explained with an example. Suppose we first have dataset $D_{example}$ from Table 2.5. We shuffle the labels and we, for example, obtain the new dataset in Table 2.11. There are of course also other possibilities for datasets that can be obtained after reshuffling. With the shuffling, the amount of rows that belong to each class is preserved, and the rest of the data is the same. Only the label/class that a row is assigned to is different.

id	x_0	x_1	x_2	x_3	x_4	l
0	4,3	4,9	4,1	4,79	5,5	1
1	3,9	6,1	5,9	5,5	5,9	0
2	2,7	4,8	4,1	5,0	5,6	1
3	6,6	4,4	4,5	3,9	5,9	0
4	6,5	2,9	4,7	4,6	6,1	1
5	2,7	6,7	4,2	5,3	4,8	0

Table 2.11: Our shuffled labels test dataset for dataset $D_{example}$. The highlighted labels represent labels that are different than in the unshuffled dataset (Table 2.5) due to shuffling.

We do this shuffling for K permutations, where K can be chosen. Hence, we make K of these datasets where the labels are shuffled in a certain way. These datasets are used to construct random forests, and the out-of-bag dataset is run through these random forests as described in the previous section to obtain the OOB error.

The purpose for constructing these permuted random forests and running the data through them is to overcome the arbitrary split in training and testing data. By shuffling the labels of the data to construct the random forest with, another element of randomness is added. If you shuffle the labels and the classes are similar, it does not matter that the labels are shuffled. Because even if you shuffle, you cannot distinguish between the two classes. If you shuffle the labels and the classes are different, it does matter that the labels are shuffled since then it will become more difficult to correctly classify a row and hence the OOB error will be larger.

2.2.4 Two sample problem

One of the reasons we are looking at these permuted random forests is because we want to solve a two sample problem. We namely want to know whether the datasets of the different years have the same distributions of the data or whether a change in distribution occurred since then we could not one-to-one compare the datasets.

The two sample problem that we thus have is the following. Let $X_1, ..., X_{n_0}$ and $Y_1, ..., Y_{n_1}$ be a collection of $\mathbb{R}^d$-valued random vectors, such that $X_i \overset{iid}{\sim} P$ for $i = 1, 2, ..., n_0$ and $Y_i \overset{iid}{\sim} Q$ for $i = 1, 2, ..., n_1$, where P and Q are some Borel probability measure on $\mathbb{R}^d$. The goal is to test

$$H_0 : P = Q, \qquad H_A : P \neq Q.$$

Given these i.i.d. samples of vectors, we define labels $l_i = 1$ for each X_i and $l_i = 0$ for each Y_i

to obtain the data (Z_j, l_j), $j = 1, ..., N$, for $N = n_0 + n_1$, and $Z_j = X_i$ or $Z_j = Y_i$. On this data we train a classifier $g : \mathbb{R}^d \to \{0, 1\}$. If g predicts the labels l better than expected under H_0, it is taken as evidence against H_0 [21]. Also [22] uses random forests as a classifier and they show how well random forests perform in classifying, also compared to other methods.

2.2.5 Random forest for two sample problem

In this section the random forest and the two sample problem are connected.

If g, the random forest classifier, predicts the labels l better than expected under H_0, it is taken as evidence against $H_0 : P = Q$. The significance level is $\alpha = 0,05$. This means that if g can distinguish between the two classes (labels) (which are related to the different distributions), then the classes (so distributions) are too different and hence $H_0 : P = Q$ gets rejected. This corresponds with a low p-value, lower than α. On the other hand, if g can **not** distinguish between the classes, then the classes (and hence the distributions) are very alike and $H_0 : P = Q$ is accepted. This corresponds with a high p-value, bigger than α.

To perform this hypothesis testing, the data is split on the variable that you want to investigate, variable l in our example. Then, the data is run through R using the R package hypoRF that is developed by [21]. This R package constructs the permuted random forests. In Table 2.12 the arguments that the function hypoRF takes are given. In Table 2.13 the values that the function hypoRF returns are given.

Before you can actually run the data with the hypoRF function, you need to remove *perfect separators*. Perfect separators are columns/variables that perfectly separate the observations belonging to class 0 and the observations belonging to class 1. You can identify perfect separators by iterating over the different variables in the split data and checking whether a value is unique for one of the classes.

Variable	Description	Type	Default
`data1`	The first sample	data.frame	x
`data2`	The second sample	data.frame	x
`K`	Number of times the created label is permuted	numeric	$K = 100$
`statistic`	Statistic for permutation testing	character	"PerClassOOB"
`normalapprox`	Asking for the use of a normal approximation	logical	FALSE
`seed`	Value for reproducibility	numeric	NULL
`alpha`	The level of the test	x	0,05
`...`	Arguments to be passed to ranger	x	x

Table 2.12: Description of the arguments of the hypoRF package [23].

Variable	Description	
	Description for $K > 1$	Description for $K \leq 1$
p-`value`	p-value of the test	
`obs`	OOB statistic	out-of-sample error
`val`	OOB-statistic of permuted random forests	NULL
`varest`	estimated variance of permuted RF OOB-statistics	NULL
`statistic`	used OOB-statistics	
`importance_ranking`	variable importance measure, when importance == 'impurity'	
`cutoff`	quantile of importance distribution at level α	

`call`	call to the function

Table 2.13: Description of the values of the hypoRF package [23].

Not in any of the times in this thesis that a hypoRF function was ran, was the OOB statistic (denoted by variable `obs`) bigger than the OOB-statistic of permuted random forests (denoted by variable `val`[7]).

Hence the p-values were always $\frac{1}{K+1}$, with K the number of permutations. The p-value, and therefore the acceptance or rejection of H_0 was thus very dependable on K. For example, one can want to accept their H_0, and then set K such that the p-value is bigger than $0,05$. It is also possible the other way around, to set K such that the p-value is always smaller than $0,05$ if we want to reject it. This is highly remarkable. For this reason, and with the knowledge from theory that p-values are not always the best fit when working with large sample sizes (described in more detail in Chapter 6), we will not focus on the p-value during the analysis of the hypoRF function outcome.

[7] Also differing the values of the number of trees or of K did not influence this. Only if the number of trees became very small (less than 3), the p-value would be Not Available (NA) instead of $\frac{1}{k+1}$. The explanation for this is that if the number of trees used in a random forest is small compared to the dataset, there won't be any training data for some parts of the domain, which might result in "NA" values in the out-of-sample. Likewise, it might give the same prediction values for some observations in the out-of-sample, such as those in permuted datasets in the hypoRF package's logic.

3 | Literature review

In this chapter, we review current literature on the analysis of data on colorectal cancer. We cover studies that are conducted with the dataset of the Dutch ColoRectal Audit (DCRA). DCRA exists since 2009 and is managed by the Dutch Institute for Clinical Auditing (DICA). For more information on DICA itself, [14] describes the organisation extensively.

Multiple studies have been conducted with the DCRA dataset. This dataset contains data only from hospitals based in the Netherlands. All the patients in the dataset are 18 years or older. Surgery means the surgical resection of a tumour in the colon or the rectum. A picture of the colon and the rectum can be found in Figure 2.2. The 'appendix', 'cecum', 'ascending colon', 'hepatic flexure', 'transverse colon', 'splenic flexure', 'descending colon', and 'sigmoid colon' are part of the colon.

In Table 3.1 an overview of the studies conducted with the DCRA dataset is given. In this overview it is clarified what the outcome variable of the study is, the timeframe that the used data is from, whether there is a split in the training and the test set, what they have identified as risk factors for the outcome variable, whether they used additional data, and the modelling methods that were used.

In 2011, Kolfschoten et al. published a paper [17] based on the DCRA data of the year 2010. This study focuses on determining the expected mortality in different types of hospitals. There were variations in patients treated in (i) non-teaching, (ii) teaching, and (iii) university hospitals, and in (a) high, (b) low, and (c) intermediate-volume hospitals. To determine the factors to predict mortality after surgery, logistic regression models were used. The study concluded that 'high risk patients are not evenly distributed between hospitals'. So to be able to compare hospital performances it is necessary to adjust to this difference in risks. In 2013, there was also a focus on differences in hospitals in [16]. This was centred around the failure to rescue (FTR), the ability to let patients with a serious complication survive. The goal of this study was evaluating the relation between hospital performance on postoperative mortality and severe complications or FTR. To achieve these outcomes logistic regression models were applied to retrieve (i) adjusted mortality, (ii) complications, and (iii) FTR rates. The results showed that the rates of severe complications were slightly higher in high-mortality hospitals than in low-mortality hospitals. Nevertheless, high-mortality hospitals had a three times higher FTR than low-mortality hospitals. So there was no clear conclusion drawn.

In 2015, [15] offers a more economical view on the dataset. The authors state that colorectal cancer surgery comes with relatively many complications compared to other surgeries. These complications come with a cost and hence the paper wants to 'explicitly discuss the costs of complications and the risk factors for high costs after colorectal surgery'. There are 29 participating hospitals in this study. The selection for these hospitals was 'based on the availability of detailed cost-price information for two consecutive years'. Around one third of the total performed colorectal cancer surgeries was captured with this way of selecting. Of the total hospital costs for colorectal cancer, 31% is spent on the complications. The paper concludes that decrease

of complications will lead to substantial cost savings.

There are multiple papers with DCRA data focused on anastomotic leakage. An anastomotic leak is when the colon is first cut to get the tumour out of it. Then the two parts of the colon are connected again. However, this can start to leak, and that is called anastomotic leakage. A picture to illustrate this is in Figure 3.1.

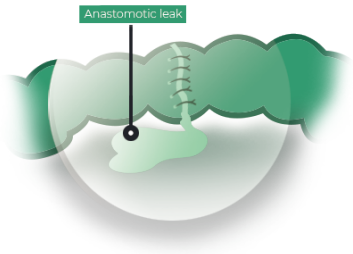


Figure 3.1: Figure from [24]. Anastomotic leakage.

In 2017, [25] used DCRA data from the year 2011 and extended this with data on 'additional treatment and long-term outcome data' for '71 out of potential 94 hospitals'. The focus of this study was on 'late detected anastomotic leakage after low anterior resection (a resection low in the rectum on the forepart of the body) for rectal cancer'. They also investigated the magnitude of leakages developing into a chronic presacral sinus. 'Anastomotic leakage was diagnosed in 200 of 998 (around 20%) of the patients during complete follow-up'. It presents that of the total anastomotic leakages, one third is diagnosed beyond 30 days after surgery. On top of that, almost half (48%) of the anastomotic leakages were not healed after 12 months. The days to anastomotic leakage after surgery for the DCRA data is in Figure 3.2, generated with the DCRA dataset obtained in Section 5.6. We see that the median differs quite a bit from year to year, but is between 60 and 150 days. The upper quartile differs between 150 and 260 days, and the upper whisker is around 260 to 340 days.

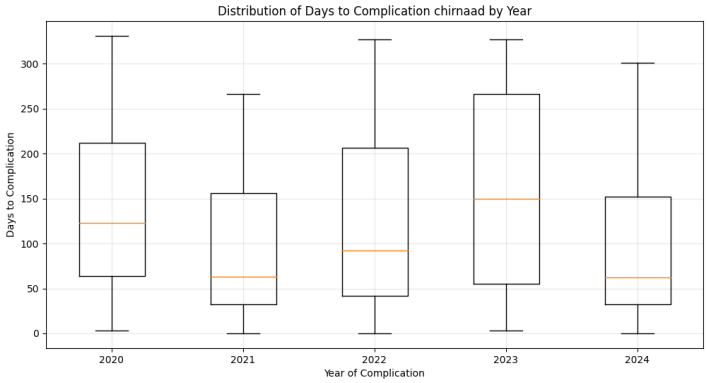


Figure 3.2: The days to complication of anastomotic leakage per year, from 2020 to 2024. With the dataset obtained in Section 5.6.

In 2022, [18] published a paper focused on developing a robust model to preoperatively predict anastomotic leakage after surgery. They have used the DCRA data from 2011 to 2019. To build this prediction model multivariable logistic regression was used. To validate the model, cross-validation was used. The study developed a specific prediction model that is based on variables that are preoperatively available. The suggested score is appropriate for discussing with a patient

before surgery for expectation management. The score can also be used for doctors to determine the risks before the surgery. The risk factors they identified for the risk are gender, age, BMI, ASA classification, neoadjuvant (chemo)radiotherapy, cT stage, distance of the tumour from anal verge, and deviating ileostomy. In 2023, [26] focusses on stoma construction and anastomotic leakage, with anastomotic leakage as their primary outcome. The data they have used is from 2011 to 2020. To analyse the differences between years for the construction of a defunctioning stoma, logistic regression was used. Multivariate analysis and propensity score matching were used to analyse postoperative outcomes. Linear or logistic regression were used in the multivariate analysis. The conclusion is that the decline in defunctioning stomas corresponds with an increase in anastomotic leakage, since with a stoma there is no anastomosis, so there also cannot be an anastomotic leakage. This decline in defunctioning stomas is due to less stomas that are created. At the same time, a defunctioning stoma corresponded to 'more minor complications, a prolonged length of stay, more intensive care admissions, and more readmissions'. This did thus not give one clear outcome of how to act and hence they recommend to look at each case individually to consider what to do.

In 2021, [27] conducted a research on the patients from 2011 up until 2016. The goal was to inspect the benefit of applying machine learning to (i) predict quality indicators for colorectal cancer surgery and (ii) to identify predictors of 30-day mortality that were not yet recognized. They have constructed different machine learning models (multivariable logistic regression, elastic net regression, support vector machine, random forest, and gradient boosting) to reach the goals. Furthermore, the risk factors were determined with logistic regression analyses and Shapley additive explanations. The variable of interest was 30-day mortality. Machine learning models 'identified specific patient groups for adverse outcomes and provided directions to optimize benchmarking in clinical audits'. The study thus concludes that to predict 30-day mortality after colorectal cancer surgery, machine learning methods were better than conventional scores. 'Machine learning models based on a colorectal cancer registry including 103 preoperative variables showed better performance for predicting 30-day mortality than the ASA score, the preoperative score to predict postoperative mortality, Charlson Comorbidity Index, and DCRA case-mix regression model'. This conclusion is contradicting with [28] since they state that machine learning methods perform worse than conventional scores. In the same year there was another study with a focus on 30-day mortality [28]. They conducted a study with the aim of analysing postoperative outcomes and creating a risk model for 30-day mortality. They used DCRA data from 2009 to 2017. The data from 2009 to 2014 was used to base the 30-day mortality prediction model on. This model was made with multivariable regression. The data from 2015 to 2017 was used to test the predictive performance of the model. They state that their developed model can accurately predict postoperative mortality risk and is therefore valuable in decision-making. The model is available online at https://crcsurgery.shinyapps.io/predict30daymortality/.

In 2021, [29] published a paper in which the aim was to develop a preoperative prediction model for severe complications specifically for patients aged 70 or older. They tested this with the DCRA data from 2014 up until 2017. This was the patient demographic data of the study. The study also used geriatric data that was provided by five Dutch hospitals. The model they developed is called the 'GerCRC model'. 'Potential predictors included demographics, comorbidity, tumour location, activities of daily living (ADL), history of falls, malnutrition, risk factors for delirium, use of mobility aid and polypharmacy'. This is 'the first prediction model specifically developed for older patients expected to undergo colorectal cancer surgery'.

In 2023, [30] published a paper with the goal to evaluate the completeness and reliability of a certain method to identify early tumours suitable for treatment by surgery alone (so without radiotherapy) and identify locally advanced tumours that require (radio)therapy to reduce the size of the tumour such that it is smaller for the surgery. The patients that were included were 'patients who underwent elective oncological resection of colon cancer without neoadjuvant treat-

ment (treatment before the surgery to shrink the tumour) in 77 Dutch hospitals between 2011 and 2021.' The method that they look at is CT-based (Computed Tomography) preoperative locoregional colon cancer staging (cTN), so trying to determine the stage of the tumour preoperatively with the CT scan[1]. The conclusion is that the completeness of T-stage of cTN improved during the years they have looked at. The completeness of cTN also varied between hospitals, independent from the number of cases.

In 2024, [31] published a paper with the goal to 'develop and externally validate two clinical prediction models'. 'Aerobic fitness, sex, age, ASA, tumour location, and surgical approach were included in the final models'. The data used was from DCRA for the perioperative data and from PROCLINA for the preoperative physical fitness data. After testing, both models could 'identify patients with and without an increased risk of any postoperative complications or a prolonged time to in-hospital physical recovery'. They state that the following characteristics give a higher probability of developing postoperative complications: male sex, a higher ASA classification, a lower preoperative aerobic fitness, a tumour located in the rectum, and surgery performed via laparotomy.

[1] More on the stage of the tumour is in Section 2.1.

Study	Outcome	Timeframe	Train / test split	Risk factors	Additional data	Modelling method
[17]	Postoperative mortality	2010	No	additional resection for tumour invasion, age, American Society of Anaesthesiologists (ASA) classification, Body Mass Index, comorbidity (Charlson-comorbidity index), gender, metastasis, neo-adjuvant (chemo) radiation therapy, preoperative tumour complications, previous abdominal surgery, TNM stage, type of resection, urgency of the resection	No	Logistic regression
[16]	Failure to Rescue (FTR) (serious complication and survival)	2009-2011	No	age, approach, body mass index, Charlson comorbidity index, mortality, pathological TNM stage, procedure, sex, urgency of the procedure	No	Logistic regression (backward stepwise, and mixed)
[15]	Any complication	2011-2012	No	age, anastomosis or stoma , ASA score, BMI (Body Mass Index (kg/m^2)), Charlson score, distant metastases, double tumour, extended resections, metastasectomy, preoperative radiotherapy, sex, surgical approach, tumour location, tumour stage (TNM), urgency of resection	No	Multivariate regression
[25]	Anastomotic leakage	2011	No	age, ASA, BMI, , distance to anorectal junction, diverting stoma, laparoscopic approach, neoadjuvant treatment, physical status, sex	Yes	Binary logistic model
[18]	Anastomotic leakage	2011-2019	No	age, ASA classification, BMI, cT stage, , deviating ileostomy, distance of the tumour from anal verge, gender, neoadjuvant (chemo)radiotherapy	No	Multivariable logistic regression
[26]	Anastomotic leakage	2011-2020	No	age, AJCC stage, ASA classification, BMI, defunctioning stoma, distance from tumour to anal junction, metastatic disease, neoadjuvant therapy, sex	No	Logistic regression
[27]	30-day mortality	2011-2016	2011-2015 training set; 2016 test set	age, approach, ASA score, atrial fibrillation or atrial flutter, BMI, cerebrovascular attack, COPD or asthma, decal peritonitis due to preoperative colorectal perforation, emergency direct procedure, (extended) right hemicolectomy procedure, history of bladder prostate urine or ovarian surgery, hypertension, liver disease or failure, M-sage, myocardial infarction, neoadjuvant radiotherapy, neoadjuvant radiotherapy short course, Parkinson disease or dementia, patient discussed in a MDT, patient received neoadjuvant chemotherapy, peripheral vascular disease, preoperative bowel obstruction or ileus owing to malignant neoplasm, sex, urgent, valvular heart disease	No	multivariable logistic regression, elastic net regression, support vector machine, random forest, and gradient boosting
[28]	30-day mortality	2009-2017	2009-2014 training set; 2015-2017 test set	age, approach (laparoscopic vs open), ASA grade, BMI, emergency surgery, sex, tumour location, tumour stage	No	Univariable and multivariable logistic regression
[29]	Preoperative prediction model for patients aged 70 or older	2014-2017	No	activities of daily living (ADL), comorbidity, demographics, history of falls, malnutrition, polypharmacy, risk factors for delirium, tumour location, use of mobility aid	Yes	univariable logistic regression
[30]	Method to identify early tumours	2011-2021	No	cTN-stage, pTN-stage, cT-stage, cN-stage, pT-stage, pN-stage	No	Fisher's F-test
[31]	Postoperative complications and prolonged in-hospital physical recovery time	Jan 2016 - Mar 2020	UMC data training set; Gelderse Vallei hospital, Nij Smellinghe hospital, Deventer hospital test set	aerobic fitness, age, American Society of Anesthesiologists (ASA) classification, body mass index, neoadjuvant treatment, preoperative haemoglobin level, sex, surgical approach, tumour location	Yes	Multivariable logistic regression

Table 3.1: An overview of the studies done with the DCRA dataset. In this overview it is clarified what the outcome variable is, the timeframe that the used data is from, whether there is a split in the training and the test set, what they have identified as risk factors for the outcome variable, whether they used additional data, and the modelling methods that were used.

4 | Data description

This chapter is meant to give an overview of what the dataset is about and how the dataset is constructed. In Section 4.1 the difficulties we have gone through with the data extraction are mentioned. In Section 4.2 we describe the patients that are registered in this dataset. In Section 4.3 we describe the content and the structure of the dataset. In Section 4.4 the process of data collection is described to give a complete understanding of where the data is coming from and how the data comes about. In Section 4.5 the number of patients per year that is registered is shown and discussed. In Section 4.6 the changes over the years in the data that is collected are mentioned and visualised.

4.1 Data extraction

The data that is worked with in this thesis is retrieved from the Dutch Institute for Clinical Auditing (DICA). DICA hires the data processing company MRDM [32]. 'MRDM designs, develops and manages registration systems for DICA's quality registrations. MRDM processes the data from the hospital so that DICA receives only coded (pseudonymous) data' [33]. This is the data that we have received.

We did not immediately receive the correct dataset, and the datasets that were incorrect were a big limitation during this thesis. Seen the incredible amount of observations and variables, and the multiple ways in which data gets delivered, it is not hard to imagine that merging this data is a difficult task and it does not always go as planned. It took around two years for my supervisors to retrieve the data from DICA. After these two years, I started my thesis in December 2025. Since then, in total it turned out four times that there were inconsistencies in the dataset that we retrieved, which made them unsuitable for research. These inconsistencies were brought up by me. The correct dataset we retrieved on May 15th 2026. In total, we have thus received 5 different datasets. Underneath, it is described in more detail what the timeline was, and what was wrong with the datasets.

The first dataset contained 120807 observations. My supervisors received this dataset on 8-10-2025, and I started working with it mid December 2025. This dataset contained 120807 observations. On 22-12-2025 it was said by DICA that this dataset was not useful since the data of *Integraal Kankercentrum Nederland* (IKNL) was not taken into account and the hospitals were not anonymized.

On 9-1-2026 I received dataset 2. This dataset contained 153043 observations. Because of this difference in observations, I asked MRDM and DICA how the difference in observations could be explained. On 16-2-2026 it was said by MRDM that it was not recommended any more to use this dataset for academic research since MRDM saw 'inconsistencies in the different extraditions (dataset 1 and dataset 2)'.

On 12-3-2026 I received dataset 3. This dataset contained 147916 observations. I decided to conduct an additional analysis on this dataset and I discovered differences in the data that were

hard to explain. I namely saw that the number of differences in values was bigger for some variables than the total number of different patients for dataset 2 and dataset 3. I asked MRDM and DICA about this possible mistake and it was indeed incorrect since, according to MRDM, 'in the join script to combine the data of IKNL and MRDM we see that the *radiotherapy*, *comorbidities*, *pathology*, and *systemic therapy* datasets of IKNL do not get picked up'. So then MRDM said I would retrieve a new dataset (dataset 4). By then, I had decided together with my supervisors for the sake of my thesis to not wait for dataset 4 anymore and continue the research with dataset 3.

On 16-4-2026 I received dataset 4. This dataset contained 148132 observations. I ran an analysis on this and on 17-4-2026 I let Medisch Spectrum Twente (MST) know that there were inconsistencies again and asked them to explain it. They forwarded this to DICA and MRDM on 24-4-2026. It turned out that it was wrong again.

On 15-5-2026 I received dataset 5. This dataset contained 147935 observations. Even though DICA has said that it is 'difficult to give absolute certainty for this request', it seems that this data is actually correct when comparing it to all of the other datasets I received. Even though it was then 2 months till me thesis defence when I received this dataset, I decided to go with the correct data, since the work done in this thesis is much more valuable with correct data.

This process is described to give a full overview and insight into what it was like to work with these datasets during the thesis and how this added challenges in the process. As mentioned before, the amount of variables and observations that these datasets contain is incredible. The data also gets delivered in multiple ways. Seen those circumstances it is highly understandable that the data extraction does not always go as planned.

4.2 Patients that are registered

The patients that are included in this dataset are all the patients aged 18 years or older who underwent a primary resection of colorectal carcinoma (tumour) in the hospitals in the Netherlands from 2009 to 2024 in the Dutch ColoRectal Audit (DCRA) [34] [27] [18]. Some patients occur multiple times in the dataset. These are patients that are diagnosed with more than one primary colorectal tumour. [34]

The exact inclusion criteria according to DICA itself are [1]:

- primary tumours of the colon and rectum for which part of the colon or rectum has been surgically removed;
- primary rectal tumours for which a 'watchful waiting' strategy [1] has been agreed (including cases without surgery);
- If, following a 'watchful waiting' approach, new growth is diagnosed at any point and surgery is performed as a result, this surgery is registered. The time elapsed since the initial diagnosis is irrelevant in this context.

Some diseases are not taken into account since they are not primary colorectal cancer. The patients that are not in the scope of this registration are [1]:

- Endoscopic resections, sarcomas, neuroendocrine tumours, melanomas, GISTs and lymphomas. Those are a different type of cancer that just happens to grow in the bowel, but the cancer does not originate in the bowel tissue. Therefore, it often requires a different type of treatment, as it can also grow in other parts of the body or have different

[1] A watchful waiting strategy is a strategy in which the patient does not get a surgery, but the (growth of the) tumour is closely observed and intensively checked [35].

characteristics.

- Dysplastic polyps. Dysplastic polyps are pre-malignant and therefore not yet cancer.
- Locoregional or distant recurrences of colorectal cancer. Those occur when the cancer was not entirely removed in the first treatment and hence came back.

4.3 Subdatasets and corresponding variables

The dataset of DCRA consists of multiple subdatasets. The subdatasets are *patient*, *comorbidities*, *presentation*, *diagnostics*, *surgery*, *pathology*, *radiotherapy*, and *systemic therapy*. For the description of the variables, we consult the 'datadictionary'. A datadictionary is an Excel file that describes the structure and composition of a registration [36]. These variables all have specific values that they can take, the options of the variables. The exact options of the variables are given in Appendix B.

The subdataset *patient* contains some basic patient characteristics, like date of birth and gender. The variables in this subdataset are collected before surgery. This subdataset contains 5 variables in the dataset. An overview of the variables with their translated description, category, and data type is in Table 4.1.

Variable	**Description**	**Category**	Type
Patient			
`patient_uri`			
`id`	Healthcare institution	Patient identification	
`gebdat`	Date of birth	Personal data	Date
`geslacht`	Gender	Personal data	Nominal
`datovl`	Date of death (if applicable)	Personal data	Date

Table 4.1: An overview of the variables of the subdataset *patient* with their description, category and data type.

The subdataset *comorbidities* contains the comorbidities that a patient might have before the surgery. The variables in this subdataset are thus all collected before surgery. Comorbidities are other diseases that the patient has, for example asthma or diabetes. They can be of influence for how well a patient will go through the process of the surgery and the development of cancer or complications after it. This subdataset contains 15 variables in the dataset. An overview of the variables with their translated description, category, and data type is in Table 4.2.

Variable	**Description**	**Category**	**Type**
Comorbidities			
`datcom`	Date of comorbidities	General	Date
`commyo`	Myocardial infarction[2]	Charlson Comorbidity Index	Nominal
`comhartfaal`	Congestive heart failure	Charlson Comorbidity Index	Nominal
`comperif`	Peripheral vascular disease/aortic aneurysm[3]	Charlson Comorbidity Index	Nominal
`comcva`	Cerebrovascular disorder (including CVA/TIA)[4]	Charlson Comorbidity Index	Nominal

[2] A myocardial infarction is a heart attack [37].
[3] This is an enlarged or weakened area in an artery other than the aorta [38].
[4] This is a term for conditions that affect blood flow to the brain [10].

comdem	Dementia	Charlson Comorbidity Index	Nominal
comlong	Chronic lung disease	Charlson Comorbidity Index	Nominal
combind	Connective tissue disease (including rheumatoid diseases)	Charlson Comorbidity Index	Nominal
comgiulcus	Gastrointestinal ulcer disease	Charlson Comorbidity Index	Nominal
comlever	Liver disease	Charlson Comorbidity Index	Nominal
comdiam	Diabetes mellitus	Charlson Comorbidity Index	Nominal
comparalys	Paraplegia/hemiplegia[5]	Charlson Comorbidity Index	Nominal
comnier	Kidney disease	Charlson Comorbidity Index	Nominal
commalig	Malignancy (excluding PCC or BCC of the skin)	Charlson Comorbidity Index	Nominal
comhiv	HIV/AIDS	Charlson Comorbidity Index	Nominal

Table 4.2: An overview of the variables of the subdataset *comorbidities* with their description, category and data type.

The subdataset *presentation* contains patient characteristics that are more about their general health and previous diseases or surgeries that might be relevant, like BMI and the question whether a patient has previously had a colon resection. The variables in this subdataset are collected before surgery. This subdataset contains 9 variables in the dataset. An overview of the variables with their translated description and data type is in Table 4.3. There is no categories for these variables, so that column is not in the table.

Variable	Description	Type
Presentation		
tumor_id[6]	Healthcare institution	
idaaba	Tumour serial number	
lengte	Length (in cm)	Continuous
gewicht	Weight (in kg)	Continuous
asascore	ASA score	Ordinal
resdarm	Is there a history of bowel resection (colon/rectum)?	Nominal
stomavg	Does the patient have a stoma prior to the commencement of treatment for colorectal carcinoma?	Nominal
datpa1	Date of diagnosis of first colorectal carcinoma	Date
presentatie_uri		

Table 4.3: An overview of the variables of the subdataset *presentation* with their description and data type.

The subdataset *diagnostics* contains all the information on the current tumour(s) of the patient and what is done in preparation for the surgical removal of the tumour(s). The variables in this subdataset are collected before surgery. This subdataset contains 32 variables in the dataset. An overview of the variables with their translated description, category, and data type is in Table 4.4.

[5] These are types of paralysis [39].
[6] The variable id from subdataset *patient* and the variable tumor_id from subdataset *presentation* are different.

Variable	Description	Category	Type
Diagnostics			
primcrc	Is it a first or second primary tumour?		Nominal
scopnumb	Number of diagnosed carcinomas		Ordinal
diagn	Reason for diagnosis		Nominal
locprim	Localisation of primary colorectal carcinoma		Nominal
histd	Histology based on biopsy/polypectomy		Nominal
complpre	Tumour-related complications?		Nominal
preanemie	Anaemia	Pre-operative tumour complications	Nominal
preobstruct	Obstruction / Ileus (without perforation)	Pre-operative tumour complications	Nominal
preabces	Abscess at the site of the tumour	Pre-operative tumour complications	Nominal
scorect	cT[7] 1st colorectal carcinoma		Ordinal
scorecn	cN[8] 1st colorectal carcinoma		Ordinal
locprim2	Localisation of second colorectal carcinoma		Nominal
scorect2	cT 2nd colorectal carcinoma		Ordinal
scorecn2	cN 2nd colorectal carcinoma		Ordinal
scorecm	cM stage[9]		Ordinal
prelocmri	Was an MRI pelvis performed for the first rectal carcinoma?[10]		Nominal
prelocmri2	Was an MRI pelvis performed for the second rectal carcinoma?		Nominal
scopdistmm	Distance in mm from the lower edge of the first tumour to the anorectal junction on sagittal MRI (or CT, if applicable)[11]		Continuous
scopdist2mm	Distance in mm from the lower edge of the second tumour to the anorectal junction on sagittal MRI (or CT, if applicable)		Continuous
afstmrf	Distance in mm from the first tumour to the mesorectal fascia on transverse MRI (or CT, if applicable)		Continuous
afstmrf2	Distance in mm from second tumour to mesorectal fascia on transverse MRI (or CT if necessary)		Continuous
prelocstad	Was preoperative imaging of the pelvis performed?		Nominal
emimri	Extramural invasion depth of the first tumour in the mesorectum on transverse MRI		Nominal
emimri2	Extramural invasion depth of second tumour in the mesorectum or transverse MRI		Nominal
genan	MSI determination[12]		Nominal

[7] The cT classifies the primary tumour itself [40].
[8] The cN classifies the regional lymph nodes [40].
[9] The cM classifies the distant metastasis [40].
[10] MRI is the standard imaging modality for primary staging [40].
[11] This matters, because based on the MRI a division is made into proximal rectal carcinoma and distal rectal carcinoma, based on the guidelines [40]. Whether the rectal carcinoma is proximal or rectal matters for the surgery. For example, in the case of a rectal resection for distal rectal cancer, the entire mesorectum must always be included, whereas for proximal rectal cancer located more than 5 cm from the anorectal junction, a partial mesorectal excision may be considered [40].
[12] 'MSI results from the inactivation of one of the four key MMR proteins (MLH1, PMS2, MSH2 and MSH6). MSI is characterised by variations in the lengths of microsatellites (short segments of DNA composed of repeated sequences) in tumour tissue' [41].

imuun	Has an immunohistochemical determination of the mismatch repair proteins been performed?		Nominal
mlh1	Mismatch repair proteins (MMR) MLH1 loss?		Nominal
pms2	Mismatch repair proteins (MMR) PMS2 loss?		Nominal
hyperm	Hypermethylation of the MLH1 promoter?		Nominal
msh2	Mismatch repair proteins (MMR) MSH2 loss?		Nominal
msh6	Mismatch repair proteins (MMR) MSH6 loss?		Nominal
diagnostiek_uri			

Table 4.4: An overview of the variables of the subdataset *diagnostics* with their description, category, and data type.

The subdataset *surgery* is about the surgery itself and the tumours, and also about the possible complications after the surgery and when a patient is discharged from the hospital. The variables in this subdataset are collected before, during, and after surgery. This subdataset contains 96 variables in the dataset. An overview of the variables with their translated description, category, data type, and moment of collection is in Table 4.5.

Variable	**Description**	**Category**	**Type**	**Collected before, during, or after surgery?**
Surgery				
datbez1	Date of first preoperative visit (outpatient clinic) surgery		Date	Before
verwezen	Has the patient been referred by another centre?		Nominal	Before
darmperf	Was there an intestinal perforation?		Nominal	Before
ileusblow	Ileus with blow-out proximal to the tumour with peritonitis	Intestinal perforation	Nominal	Before
perfperi	Perforation at the site of the tumour with peritonitis	Intestinal perforation	Nominal	Before
perfscopie	Perforation as a result of colonoscopy	Intestinal perforation	Nominal	Before
prechirstomanew	Was a stoma created prior to the resection of the tumour?		Nominal	Before
datstoma	Date of stoma creation		Date	Before
prechirstent	Was a stent placed prior to resection of the tumour?	Preoperative surgical procedure	Nominal	Before
datstent	Was a stent placed prior to resection of the tumour? If so, please enter the date.	Preoperative surgical procedure	Date	Before
prechirmetanew	Was local treatment of the metastasis performed prior to resection of the tumour?	Preoperative surgical procedure	Nominal	Before
datprechirmetaeen	Date of first local treatment of metastasis	Preoperative surgical procedure	Date	Before
datprechirmetatwee	Date of second local treatment for metastasis	Preoperative surgical procedure	Date	Before
prechirappendix	Was an appendectomy performed prior to the resection?	Preoperative surgical procedure	Nominal	Before

datappendix	Date of appendectomy	Preoperative surgical procedure	Date	Before
prechircoloscex	Was a colonoscopic excision performed prior to resection of the tumour?	Preoperative surgical procedure	Nominal	Before
datcoloscex	Date of colonoscopic excision	Preoperative surgical procedure	Date	Before
chirenkel	Was the surgical treatment a single procedure?	Preoperative surgical procedure	Nominal	Before
datok	Date of first surgical procedure		Date	During
typok	Urgency of surgery		Nominal	During
procok	First surgical procedure		Nominal	During
ingreep	Approach to the first surgical procedure		Nominal	During
conversien	Has a conversion taken place?		Nominal	During
aprtyp	Type APR		Nominal	During
typprocok	Specification procedure that was carried out		Nominal	During
datok2	Date of second surgical procedure		Date	During
procok2	What procedure was used for the second tumour?		Nominal	During
procok3	What second procedure was performed?		Nominal	During
ingreep2	Approach to second surgical procedure		Nominal	During
conversien2	Conversion 2nd procedure		Nominal	Duromh
aprtyp2	Type APR 2nd procedure		Nominal	During
resadd	Additional resection due to tumour growth?		Nominal	During
locaddmaag	Stomach	Spread to which organs (localisation)?	Nominal	During
locaddmilt	Spleen	Spread to which organs (localisation)?	Nominal	During
locaddduod	Duodenum	Spread to which organs (localisation)?	Nominal	During
locadddoverig	Small intestine, other (jejunum/ileum)	Spread to which organs (localisation)?	Nominal	During
locaddoment	Omentum	Spread to which organs (localisation)?	Nominal	During
locaddbuikw	Abdominal wall	Spread to which organs (localisation)?	Nominal	During
locaddvag	Vagina	Spread to which organs (localisation)?	Nominal	During

locadduter	Uterus	Spread to which organs (localisation)?	Nominal	During
locaddovarl	Left ovary	Spread to which organs (localisation)?	Nominal	During
locaddovarr	Right ovary	Spread to which organs (localisation)?	Nominal	During
locadduretl	Left ureter	Spread to which organs (localisation)?	Nominal	During
locadduretr	Right ureter	Spread to which organs (localisation)?	Nominal	During
locaddblaas	Bladder	Spread to which organs (localisation)?	Nominal	During
locaddprost	Prostate	Spread to which organs (localisation)?	Nominal	During
locaddvesl	Left vesicle	Spread to which organs (localisation)?	Nominal	During
locaddvesr	Right vesicle	Spread to which organs (localisation)?	Nominal	During
locaddvesb	Both vesicles	Spread to which organs (localisation)?	Nominal	During
locadddund	Small intestine	Spread to which organs (localisation)?	Nominal	During
locaddsacr	Sacrum	Spread to which organs (localisation)?	Nominal	During
locaddoverig	Other	Spread to which organs (localisation)?	Nominal	During
resadm	Additional resection due to metastases?		Nominal	During
resomm	Omentum resection	Resection of metastases	Nominal	During
reslem	Liver resection or RFA liver	Resection of metastases	Nominal	During
respem	Peritoneum resection (e.g., as part of HIPEC)	Resection of metastases	Nominal	During

reslmm	Lymph node dissection (e.g., iliac)	Resection of metastases	Nominal	During
resbum	Resection of other abdominal organs	Resection of metastases	Nominal	During
perrtxchemo	Was additional radiotherapy or chemotherapy administered during surgery?		Nominal	During
anastom1	Was an anastomosis of the intestine performed after resection of the tumour?		Nominal	During
stoma	Was a stoma created?		Nominal	During
stomahef	Was a previously created stoma removed after resection of the tumour?		Nominal	During
compl90	Did any complications occur within 90 days of the resection?		Nominal	After
compchir	Did any surgical complications occur?	Surgical complication	Nominal	After
datcompchir	Date of first surgical complication	Surgical complication	Date	After
chirnaad	Suture leakage	Surgical complication	Nominal	After
chirnaadoccult	Were there any indications of occult suture leakage (within 90 days)?	Surgical complication	Nominal	After
chirabces	Abscess not at the suture	Surgical complication	Nominal	After
chirbloeding	Postoperative bleeding	Surgical complication	Nominal	After
chirileus	Ileus	Surgical complication	Nominal	After
chirgastro	Gastroparesis	Surgical complication	Nominal	After
chirfascie	Fascial dehiscence	Surgical complication	Nominal	After
chirdarmperf	Intestinal perforation	Surgical complication	Nominal	After
chirlekkage	Leakage from ureter/bladder	Surgical complication	Nominal	After
chirwondinf	Wound infection	Surgical complication	Nominal	After
datcompl	Date of first general complication	General complications	Date	After
comppul	Pulmonary complication	General complications	Nominal	After
compcar	Cardiac complication	General complications	Nominal	After
compthrom	Thromboembolic complication	General complications	Nominal	After
compinf	Infectious complication (other than pulmonary or surgical infection)	General complications	Nominal	After
compneu	Neurological complication	General complications	Nominal	After
compx	Other complication	General complications	Nominal	After

letsel	Is the patient likely to suffer permanent injury as a result of the complication?		Nominal	After
icdg	Number of days in intensive care		Continuous	After
icreden	What was the reason for the (extended) admission to the MC/IC and/or readmission to the MC/IC?		Nominal	After
transf	Blood transfusion during admission		Nominal	After
heropn90	Was the patient readmitted < 90 days after discharge?		Nominal	After
datheropn	Was the patient readmitted < 90 days after discharge? If so, enter the date.		Date	After
reintreq	Was there a re-intervention?	Re-intervention	Nominal	After
datreint	Date of re-intervention	Re-intervention	Date	After
reintaard	Nature of re-intervention	Re-intervention	Nominal	After
reintsts	Was a new stoma created during a reintervention?	Re-intervention	Nominal	After
datont	Date of discharge		Date	After
status90	Did the patient die during hospitalisation or within 90 days of the surgery?		Nominal	After
doodoorz	Cause of death		Nominal	After
klaar	Record final		Binary	After

Table 4.5: An overview of the variables of the subdataset *surgery* with their description, category, data type, and moment of collection.

The subdataset *pathology* is about the medical examination of the tumour that has been removed with the surgery. The variables in this subdataset are collected after surgery. This subdataset contains 39 variables in the dataset. An overview of the variables with their translated description, category, and data type is in Table 4.6.

Variable	Description	Category	Type
Pathology			
tumorrest	Was a tumour found in the specimen from this surgical resection?		Nominal
tumorrest2	Was a tumour found in the specimen from this second surgical resection?		Nominal
histtype	Histological type of 1st colorectal carcinoma		Nominal
stadpt	Pathological T classification of 1st colorectal carcinoma according to TNM system		Ordinal
stadpnnew	Pathological N classification of 1st colorectal carcinoma according to TNM system		Ordinal
stadpn	pN Stage		Ordinal
histtype2	Histological type of 2nd colorectal carcinoma		Nominal
stadpt2	Pathological T classification of second colorectal carcinoma according to TNM system		Ordinal
stadpn2	Pathological N classification of second colorectal carcinoma according to TNM system		Ordinal
stadpm	pM Stage		Ordinal
resectien	Radicality of first tumour		Nominal
cirmarge	Smallest circumferential margin of rectum of first tumour (non-peritoneal) in millimetres		Ordinal

dmarge	Distal resection margin (DRM), in millimetres		Continuous
resectie2n	Radicality 2nd procedure/2nd tumour		Nominal
cirmarge2	Smallest circumferential margin rectum 2nd procedure/2nd tumour (non-peritoneal) in millimetres		Ordinal
dmarge2	Distal resection margin 2nd procedure/2nd tumour (DRM), in millimetres		Continuous
cirmargelok	Distance between tumour and nearest basal resection margin (in depth) in millimetres		Ordinal
cirmargebasaal	Distance between tumour and nearest mucosal resection margin in millimetres		Ordinal
macrotme	Macroscopic assessment of TME specimen		Nominal
numlym	Number of lymph nodes found		Ordinal
numlympos	Number of positive lymph nodes found		Ordinal
cirmargeklier	Circumferential resection margin (CRM[13]) to the nearest positive lymph node rectum 1st tumour		Ordinal
cirmargeklier2	Circumferential resection margin (CRM) to the nearest positive lymph node rectum 2nd procedure/2nd tumour		Ordinal
tumdepositnew	Number of tumour deposits		Ordinal
anginvnew_0	Lymphangioinvasion found?	Lymphangio/venous invasion 1st tumour	Nominal
anginv_1	Extramural venous invasion	Lymphangio/venous invasion 1st tumour	Nominal
anginv_3	Intramural venous invasion	Lymphangio/venous invasion 1st tumour	Nominal
diffgraad	Degree of differentiation		Nominal
paperfdarm	Perforation of the intestine, 1st tumour		Nominal
regress	Tumour regression 1st tumour after neoadjuvant treatment		Nominal
anginvnew2_0	Lymphangioinvasion 2nd tumour found?	Lymphangio/venous invasion 2nd tumour	Nominal
anginv2_1	Extramural venous invasion	Lymphangio/venous invasion 2nd tumour	Nominal
anginv2_3	Intramural venous invasion	Lymphangio/venous invasion 2nd tumour	Nominal
diffgraad2	Degree of differentiation 2nd tumour		Nominal
paperfdarm2	Perforation of the intestine, 2nd tumour		Nominal
regress2	Tumour regression 2nd tumour after neoadjuvant treatment		Nominal
braf	BRAF mutation present		Nominal
ras	RAS mutation present		Nominal

Table 4.6: An overview of the variables of the subdataset *pathology* with their description, category, and data type.

The subdataset *radiotherapy* is about radiating the cancer cells specifically, and whether the patient has had this. The variables in this subdataset are collected before, and after surgery. This subdataset contains 15 variables in the dataset. An overview of the variables with their translated description, data type, and moment of collection. is in Table 4.7. There is no categories for these variables, so that column is not in the table.

[13]CRM is circumferential resection margin.

Variable	Description	Type	Collected before, during, or after surgery?
Radiotherapy			
radtiming	Was radiotherapy given (pre-operative, post-operative, as part of watchful waiting)?	Nominal	Before, during, after
datrad1	Start date of radiotherapy	Date	Before, during, after
datprertxstop	End date of radiotherapy	Date	Before, during, after
rtxrestad	Was restaging performed after completion of radiotherapy?	Nominal	Before, during, after
rtxct	ycT	Ordinal	Before, during, after
rtxcn	ycN	Ordinal	Before, during, after

Table 4.7: An overview of the variables of the subdataset *radiotherapy* with their description, data type, and moment of collection.

The subdataset *systemic therapy* is a treatment that is done for the whole body, often with a drip. This dataset contains information on whether the patient has had this and for how long. The variables in this subdataset are collected before and after surgery. This subdataset contains 6 variables in the dataset. An overview of the variables with their translated description, data type, and moment of collection is in Table 4.8. There is no categories for these variables, so that column is not in the table.

Variable	Description	Type	Collected before, during, or after surgery?
Systemic therapy			
datctx1	Start date of preoperative systemic therapy	Date	Before
prectx	Was systemic therapy administered prior to tumour resection (chemotherapy and/or targeted agents)?	Nominal	Before
prectxaard	Nature of preoperative systemic therapy	Nominal	Before
datctx1stop	End date of preoperative systemic therapy for first tumour	Date	Before
datchemo	Start date of postoperative systemic therapy	Date	After
chemopost	Was postoperative systemic therapy administered (chemotherapy and/or targeted agents)?	Nominal	After
chemo	Nature of postoperative systemic therapy	Nominal	After
chemo1	Schedule of postoperative systemic therapy	Nominal	After
datchemostop	End date of postoperative systemic therapy	Date	After
chemokuur	Number of courses received	Ordinal	After
chemokuuraanpas	Type of adjustment	Nominal	After
chemokuurreden	Reason for adjustment	Nominal	After
chemokuurgrad	Grading of adverse event	Nominal	After
systgeen	Reason for no systemic therapy	Nominal	After
tumorgergeen	Reason for no tumour-targeted therapy	Nominal	After

Table 4.8: An overview of the variables of the subdataset *systemic therapy* with their description, data type, and moment of collection.

4.4 Data collection process

The information in this section is based on meetings with medical experts and a data analyst of Medisch Spectrum Twente (MST).

The hospitals themselves bear ultimate responsibility for that the data of all the relevant patients is at DICA on time. Per audit there is also one person responsible for the registration. So for this audit, the colorectal cancer, there is one person in the hospital responsible. This is someone that does this on the side, for example a surgeon.

During the treatment of a patient, data is collected by the hospitals. It differs a lot *when* the data is registered in the hospital. A possibility is that for example the surgeons do it (partly) themselves. The surgeon registers the first part of the data quite soon after the surgery. Then a nurse fills in the rest of the data after the 90 days after the surgery. This data that is collected by the hospitals does not exist solely of the exact information that is needed for the registration of DICA. There is often information missing, and also information added that is not relevant for, or not collected by, the registration. So from the data that is collected by the hospitals, the relevant data for the registration has to be selected.

In general, there are three possibilities that the data is delivered to DICA. The first possibility is that a surgeon or nursing specialist fills in the data directly into the DICA system. The second possibility is that it is automatised by the hospital and a batch from the hospital is sent to DICA. The third possibility is that the hospital sends the data to *Integraal Kankercentrum Nederland* (IKNL) and a batch is sent from them to DICA. An advantage of IKNL registering is that the employees handle a uniform definition of which patients to register and how. In case of the colorectal cancer registration, so our dataset, employees of IKNL select the relevant data for the registration.

So after the registration in the hospital, IKNL does the remaining part before the data can go to DICA. In March the year closes for DICA, and the year runs from October to October for them. So before March all of the data from the period from October to October has to be known. During the registration by IKNL there are mandatory fields and optional fields to fill in.

After this registration, IKNL sends it to the hospitals to get the data checked by a doctor. It is common that there is one doctor per hospital that verifies this. After approval of the hospital, IKNL makes a batch from the data they registered in their own system. They prepare the data and send this to DICA. The data file gets sent by DICA to MRDM and they make sure that the data gets into the DICA system.

4.5 Number of patients per year

The number of patients per year that are registered with an elective[14] surgery is visible in Figure 4.1.

In 2014, an increase of the number of patients is visible in the figure. The colorectal cancer screening programme was introduced in the Netherlands in 2014 [42]. That could explain the increase in the number of patients with colorectal cancer in that year [12].

Due to the outbreak of the COVID-19 coronavirus, the healthcare system in the Netherlands came under severe strain in March 2020 and the period that followed. Consequently, the Dutch Ministry of Health, Welfare and Sport decided at that time to temporarily suspend the screening programmes for cervical cancer, breast cancer and **colorectal cancer** [43]. From 11 May 2020, the screening programme resumed in phases [44]. It was announced in 2022 that the backlog in

[14] What is meant by an 'elective' surgery is explained in Section 5.1.

screening programmes had been cleared in 2021 [45] [46] [47]. This could explain the drop in the number of patients in 2020 in Figure 4.1. Since 2019, the year in which the screening programme was fully rolled out, an average of 21% of all new cases of colorectal cancer have been detected through the screening programme. More details on this topic are given in [2].

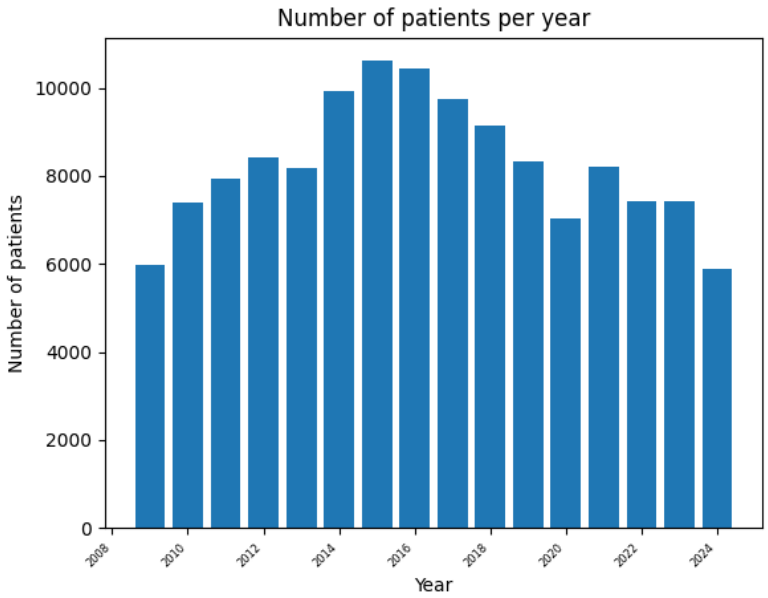


Figure 4.1: Number of patients that are registered for elective[14] surgery per year. The y-axis shows the number of registered patients and the x-axis shows the years.

4.6 Changes in data that is collected

Over the years there have been quite some changes in the data that was collected. For example, from 2018 onwards the complications were registered for up to 90 days instead of 30 days and the registration of comorbidities also changed in 2018 [34]. This is visible in Table 4.9. In this table you see per year the amount of variables that are or are not in the most recent (2024) datadictionary on which the dataset we used is based. We see that the count difference compared to the previous year is biggest in 2018, with 87 variables. In other years there were also slight changes in the variables that were collected, but not as big as in 2018. This is visualised in Figure 5.1. It is also good to mention that there is no datadictionary of 2023. The changes in the variables that are collected show the importance of determining temporal stability to ensure that conclusions reflect changes in patient care rather than changes in data collection. This is discussed in detail in Chapter 6.

Year of data-dictionary	Count of variables in datadictionary of that year	Count of variables not in datadictionary of that year	Count difference compared to previous year
2014	73	144	-
2015	72	145	1
2016	73	144	1
2017	100	117	27
2018	187	30	87
2019	194	23	7
2020	194	23	0
2021	203	14	9
2022	204	12	2

Table 4.9: The variables that we list whether they are in the datadictionary or not are the variables that are in the datadictionary of 2024.

5 | Data Engineering & Preprocessing

The original dataset that we have received still needs a lot of work before we can actually work with it. The goal of this chapter is to retrieve a dataset with useful predictor variables that have no missing values. As a result, we derive a dataset that is 'complete', so it has no missing values. To reach this goal, we need some careful considerations and many steps, namely nine. Here we provide you with a short summary of the nine steps and an overview of this chapter.

1. Select the observations with elective surgery only, discussed in Section 5.1.
2. Imputations from datadictionary[1], discussed in Section 5.2.
3. Imputation due to data collection, discussed in Section 5.3.
4. Select relevant years, discussed in Section 5.4.
5. Decide which preoperative variables to keep, discussed in Section 5.5.
6. Clinical imputation, discussed in Section 5.6.
7. Select relevant preoperative variables, discussed in Section 5.7.
8. Select and construct relevant outcome variables, discussed in Section 5.8.
9. Select observations with no remaining missing variables, discussed in Section 5.9.

After we have retrieved the final dataset, an exploratory data analysis with this dataset is in Section 5.10.

5.1 Select observations with elective surgery

The first step we take is selecting only the observations for patients that have had an elective surgery, so no emergency surgery. It is already known that the type of surgery influences the chances of survival and developing complications, where emergency surgery gives more chances of developing complications and lower chances of survival, compared to elective surgery. On top of that, the number of missing variables for the emergency surgery is much higher than for the elective surgery. This is not beneficial for our imputations of the Not Available (NA) values. Hence, we start by selecting only the observations that belong to an elective surgery. The variable `typok` tells us whether a patient has had elective or emergency surgery. The options that `typok` can take (also available in Appendix B), are 'Elective' (1), 'Elective after stent' (2), 'Urgent' (3), 'Urgent, direct' (4), 'Elective after decompressing stent or stoma' (5), and 'Unknown' (9). It is of course also possible that no value is filled in, so that it is NA. The original distribution of the `typok` values is in Table 5.1.

[1] A datadictionary is an Excel file in which the structure and the construction of the dataset is described. In this datadictionary, also the requirements (called 'conditions') are given for when a variable is visible or not during registration.

Value of `typok`	Number of observations that have this value
1	130936
2	536
3	7509
4	7648
5	1037
9	95
NA	174

Table 5.1: The possible values of `typok` and the amount of observations that have this value in the original dataset.

We have excluded the observations that have for variable `typok` the value 3 (Urgent) or 4 (Urgent, direct). We have thus included all the observations with another value for `typok`. In total, there were 7509 observations with value 3 and 7648 observations with value 4. In total we have 147935 observations. So after excluding the observations for the emergency surgeries we have 132778 remaining observations.

5.2 Imputations from datadictionary

Since the original data contains a lot of NA values (see Appendix C), one wonders whether there is a possible valid way to reduce the amount of NA values. An engineering solution to this can be found in the 'conditions' that the data has, given by the datadictionary. A complete list with conditions per variable is in Appendix D. Multiple variables have their own condition. When this condition is *not* met, it means that the corresponding variable is *not* visible when the data is registered. Hence, it will always turn out such that the value of this variable when this condition is *not* met, is NA. We will show this in the next paragraph with an example for the variable `datcompchir`.

It is important to make a distinction between (i) NA values that are not filled in by accident and (ii) NA values that are not filled in because they do not satisfy the given condition. The reason this is important is because this distinction gives information on the missingness of the data and hence on the usability of certain variables for further processing the data and even for the direction of the research. When the condition is not met and the value is NA, the value is *supposed* to be a value that is not filled in. It is *not* a value that was not filled in by accident and is therefore NA. It is an NA value because the corresponding variable is irrelevant since it is not satisfying the given condition. Hence by imputing these NA values with values that reflect that the value is not filled in because the condition was not met, we can find the 'true' NA values. We can then see how many values are not filled in by accident and are therefore indeed missing data instead of also including values that are not filled in on purpose since they do not satisfy the given condition.

Suppose we have variable `datcompchir` (date of first surgical complication) in the category `Surgical complication`. This variable has the condition `[compchir]==1`, where `compchir` is defined as *Did a surgical complication occur?*, and `[compchir]==1` means *Yes*. So we can then say that `IF datcompchir == NA AND (NOT ([compchir]==1), THEN datcompchir = 08-08-1808`, where `08-08-1808` is defined as *irrelevant (in Dutch: 'Niet van toepassing')*. This makes sense, since if no surgical complication occured, then there is also no date on which the surgical complication started. So if there is not yet a value for `datcompchir` and the condition of `did a surgical complication occur is 'yes'` is *not* met (so no surgical complication occured), then we say that the value of `datcompchir` is 08-08-1808, which means 'irrelevant'. It is thus

irrelevant since the given condition (which is the negative of the original condition) was not met.[2]

This example is visible in the code in Code Listing 5.1. In line 1 the name of the variable that we are looking at is given, so `datcompchir` in this example. In line 2, 3, and 4 the condition that the variable has to satisfy to be visible is written in the negative form, so what would be the case if the condition does *not* hold and the concerning value is an NA value. In line 5 the value that we want to insert for when the condition of the datadictionary does not hold is given. So in this case, it is the date '08-08-1808' which means 'irrelevant', see the variables with their options in Appendix B. In line 6 the row count is added with 1 up. This is necessary for making a sheet with all of the original and the new counted NA values. In line 7 the big function is given that imputes the NA values that we desire with the correct condition. This function is in the class that *I* constructed called 'DatasetCleaning'. The way this function is constructed is written in Code Listing 5.2. This function takes the arguments `column`, `condition`, `new_value`, `row_xlsx`, `df`, and `sheet`. The argument column stands for the variable that we are working on. In our example this is `datcompchir`. The argument condition gives the condition (so the negative of the condition in the datadictionary) that has to be satisfied in order to impute the NA value. The argument `new_value` is the new value that we want to put instead of the NA value. The `row_xlsx` describes the row that we want to be on in our sheet such that we can quickly have an overview of the NA values per variable. The argument `df` describes the pandas dataframe that we are working with. In this case, that is the original dataset, but with only the elective surgeries, so after step 1 (Section 5.1) is done. The `sheet` describes the sheet that we are making for the overview of the NA values. In line 3, the code uses another function in the class, namely `count_na_values_PERCENTAGE`. This function gives the percentages of the NA values of the variable. We will describe this function in Code Listing 5.3. In line 5 this function is used again so that we can see the difference before and after the following two functions are used. In line 4, the code uses another function in the class, namely `find_rows_condition`. We will describe this function in Code Listing 5.4. In line 4, the code uses another function in the class, namely `insert_missing_values_with_list`. We will describe this function in Code Listing 5.5. In lines 9, 10, 11 we write the column name, the NA value before the imputation and the NA value after the imputation to an excel file.

```
1 var = 'datcompchir'
2 cond = lambda i: (pd.isnull(mydataframe.at[i, 'datcompchir'])) and (not (
3                 mydataframe.at[i, 'compchir']==1
4 ))
5 newvalue = pd.Timestamp(1808, 8, 8)
6 row = row+1
7 DatasetCleaning.impute_na_values(var, cond, newvalue, row, mydataframe, sheet)
```

Code Listing 5.1: Code to check if the condition for the variable `datcompchir` is met and insert with a new value if it is not.

```
1 def impute_na_values(column, condition, new_value, row_xlsx, df, sheet):
2
3         NA_value_before = DatasetCleaning.count_na_values_PERCENTAGE([column],
    df)
4         count, lst = DatasetCleaning.find_rows_condition(df, condition)
5         DatasetCleaning.insert_missing_values_with_list(lst, column, new_value,
    df)
6         NA_value_after = DatasetCleaning.count_na_values_PERCENTAGE([column], df
    )
```

[2]It is important to pay attention here. This condition (let us call it condition b) is namely the negative of the original condition (we call this condition a) given by the datadictionary. It is also always included in condition b that the value of which this is about has to currently be an NA value. In Code Listing 5.2, in line 5 the observations are found for which this is the case. A row corresponds to the record of a patients. In a statistical sense, a record corresponds to an observation. In line 6 the obtained list is used to insert the missing values with a new value, 888 in this case.

```
7
8           # Write it to an excel file
9           sheet.cell(row = row_xlsx, column = 1).value = column
10          sheet.cell(row = row_xlsx, column = 2).value = NA_value_before
11          sheet.cell(row = row_xlsx, column = 3).value = NA_value_after
12
13          print(column, "is done")
```

Code Listing 5.2: Code of the function `impute_na_values`.

In Code Listing 5.3 the code is given of the function that counts the NA values, function `count_na_values_PERCENTAGE`. The function both counts the number of NA values and generates the percentage of NA values. For this function the package `pandas` is needed. This function returns only the percentage of the NA counts.

```
1  def count_na_values_PERCENTAGE(columns, df):
2      col_na_counts = {}
3      percentage_col_na_counts = {}
4      for col in columns:
5          col_na_counts[col] = df[col].isnull().sum()
6          percentage_col_na_counts[col] = ((df[col].isnull().sum())/(len(df)))*100
7      print(col_na_counts) # The number of NA counts of the selected columns
8      print(percentage_col_na_counts) # The percentage of NA counts of the
       selected columns
9      percentage_col_na_counts_percentage_only = percentage_col_na_counts.get(
       columns[0])
10     col_na_counts_count_only = col_na_counts.get(columns[0])
11     return percentage_col_na_counts_percentage_only
```

Code Listing 5.3: Code of the function `count_na_values_PERCENTAGE`.

In Code Listing 5.4 the code is given of the function that returns a list of the rows (records) that satisfy the given condition, the function `find_rows_condition`. The function also returns the amount of rows (records) satisfying the given condition, condition b. For this function the packages `pandas` and `numpy` are needed.

```
1  def find_rows_condition(df, condition_func):
2      '''
3      With this function we get a list of the indices for the rows for which
       nonempty_column is nonempty and empty_column is empty.
4      df: the dataframe on which to do this
5      condition_func: the condition that has to be satisfied. This needs to be in
       the following format; 'lambda i: (...)'
6      '''
7      count = 0
8      lst = []
9      for $I$ in range(len(df)):
10         if condition_func(i): # if condition is True
11             count = count + 1
12             lst.append(i)
13     print("The number of rows that satisy this condition, is: ", count)
14     # print("The indices on which the condition is satisfied, are: ", lst)
15     return count, lst
```

Code Listing 5.4: Code of the function `find_rows_condition`.

In Code Listing 5.5 the code is given of the function that inserts a new value at certain rows (records). This is the function `insert_missing_values_with_list`. The function `find_rows_condition` generates a list of the rows (records) that satisfy the given condition, condition b. This list is then given to function `insert_missing_values_with_list`. This function uses the list to know at which rows (records) to impute a new value instead of an NA value. When this given condition

is satisfied we know that the original condition, condition a, from the dataset was *not* satisfied and hence the variable was *not* visible which led to an NA value.

```
1 def insert_missing_values_with_list(lst, column_for_inserting, value_to_insert,
      df):
2     '''
3     The indices of the list generated by DatasetCleaning.find_rows_condition to
      use to fill in column_for_inserting = value_to_insert instead of empty
4     '''
5     for row in lst:
6         df.at[row, column_for_inserting] = value_to_insert
```

Code Listing 5.5: Code of the function `insert_missing_values_with_list`.

When running the code of Code Listing 5.1, the output is as in Code Listing 5.6. In line 1 is the original amount of NA values for the variable `datcompchir`, so this variable originally has 129960 NA values. In line 2 of Code Listing 5.6 the original corresponding percentage of NA values for the variable `datcompchir` is displayed, so this variable originally has 97,88% NA values. In line 3 we see the amount of observations that satisfy the 'new condition', condition b. So we see that 109903 observations did not have a surgical complications. Thus, when imputing the NA values corresponding to these observations, we get the new amount and percentage of NA values in line 4 and 5, respectively. In line 4 we see that the new amount of NA values for the variable `datcompchir` are 20057 observations. In line 5 is the corresponding new percentage of NA values for the variable `datcompchir`, namely around 15,11%. So there is a drop in percentage of NA values from 97,88% to 15,11%, which is around an 82% decrease in NA values.

```
1 {'datcompchir': 129960}
2 {'datcompchir': 97.8776604557984}
3 The number of rows that satisfy the condition, is: 109903
4 {'datcompchir': 20057}
5 {'datcompchir': 15.105665095121179}
```

Code Listing 5.6: Output from code of Code Listing 5.2.

The full code is available on GitHub via the link `github.com/merelvgr`. When the Dutch ColoRectal Audit (DCRA) data is delivered in CSV format with the rows reflecting different records/patients and the columns corresponding to the different variables, then this code can be used to impute the NA values where appropriate. One can fill in the path to the CSV file and a new file will be generated with the imputed NA values.

In total, 150 out of 217 variables have a condition. For subdataset *diagnostics* we have 20 variables that have a condition. For subdataset *surgery* we have 79 variables that have a condition. For subdataset *pathology* we have 34 variables that have a condition. For subdataset *radiotherapy* we have 4 variables that have a condition. For subdataset *systemic therapy* we have 13 variables that have a condition. The change of NA values per variable is visible in Appendix E. The values for which the missing rate is higher than 5% after imputing are highlighted. The variables with over 5% NA values after this imputation, are also in Table 5.2 with the percentage of NA values before and after imputation. For *diagnostics*, 3 out of 20 variables have NA values higher than 5% after imputing. For *surgery*, 27 out of 79 variables have NA values higher than 5% after imputing. For *pathology*, 9 out of 34 variables have NA values higher than 5% after imputing. For *radiotherapy*, 2 out of 4 variables have NA values higher than 5% after imputing. For *systemic therapy* 10 out of 13 variables have NA values higher than 5% after imputing.

Variable	Before imputing	After imputing
Diagnostics		
scopdistmm	71,15	19,87
afstmrf	79,16	32,96
prelocstad	81,47	6,84
emimri	90,38	17,21
Surgery		
prechirstent	59,96	54,57
conversien	33,58	7,64
procok3	99,70	96,07
locaddoment	97,58	5,44
locaddbuikw	97,57	5,43
locaddvag	97,58	5,44
locaddovarl	97,58	5,44
locaddovarr	97,58	5,44
perrtxchemo	31,34	30,11
datcompchir	97,88	15,11
chirnaad	90,49	7,72
chirnaadoccult	95,12	12,34
chirabces	93,40	10,63
chirbloeding	93,48	10,71
chirileus	93,27	10,50
chirgastro	98,11	15,34
chirfascie	93,55	10,78
chirdarmperf	93,59	10,82
chirlekkage	93,59	10,81
chirwondinf	93,42	10,65
datcompl	94,74	23,88
comppul	76,92	6,07
compcar	77,02	6,16
compthrom	77,30	6,44
compinf	76,89	6,04
⋮	⋮	⋮
compneu	77,23	6,37
compx	76,77	5,91
letsel	83,25	12,40
icdg	17,50	16,06
icreden	94,63	6,00
transf	7,17	5,74
heropn90	61,42	55,93
datreint	96,77	6,28
Pathology		
tumorrest	91,11	91,06
dmarge	97,17	25,75
macrotme	91,20	6,76
tumdepositnew	73,36	71,32
anginvnew_0	9,19	9,14
anginv_1	49,89	49,84
anginv_3	55,57	55,52
diffgraad	17,47	17,42
paperfdarm	61,07	59,86
Radiotherapy		
datprertxstop	97,79	17,03
rtxrestad	94,85	14,08
Systemic therapy		
prectx	44,49	44,49
datctx1stop	98,45	6,57
chemo1	84,81	38,56
datchemostop	95,61	49,31
chemokuur	98,82	15,79
chemokuuraanpas	98,97	15,94
chemokuurreden	99,03	16,00
chemokuurgrad	99,23	52,93
systgeen	100,00	44,77
tumorgergeen	100,00	44,77

Table 5.2: All the variables for which the percentage of NA values after imputation with datadictionary is over 5%.

Variable	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
scopdistmm	1615	1909	2156	2193	2097	2605	3051	2959	2675	2325	1974	223	141	129	158	171
afstmrf	3164	3885	4311	3055	2720	3194	3325	3247	2844	2317	1974	2182	2614	2368	2251	318
prelocstad	128	117	53	16	17	7	2	6	26	2410	2114	1924	1978	266	13	7
emimri	1449	1872	2208	2568	2474	2840	3248	997	201	568	606	947	1029	822	812	205
prechirstent	5004	6548	7098	7886	7542	9007	10211	9959	8862	302	1	0	4	3	1	25
conversien	1754	2468	141	335	173	104	305	267	404	267	228	107	160	1102	1153	1170
procok3	5251	6796	7320	8187	7839	9250	10411	10258	9367	8843	8148	6811	7808	7156	7237	6877
locaddoment	496	600	671	744	681	792	818	750	360	415	370	343	132	18	15	15
locaddbuikw	496	600	671	744	681	792	818	750	359	415	370	342	131	11	11	15
locaddvag	496	600	671	744	681	792	818	750	360	415	369	342	136	18	17	17
locaddovarl	496	600	671	744	681	792	818	750	359	415	369	342	132	17	16	17
locaddovarr	496	600	671	744	681	792	818	750	358	415	369	342	131	17	17	16
perrtxchemo	5749	7229	7606	8254	1014	1446	353	460	356	2968	3369	264	289	168	159	295
datcompchir	23	125	1397	1528	1535	1735	1972	1956	1716	1808	1535	868	1006	959	936	958
chirnaad	19	96	940	1024	1023	1171	1413	1393	1157	472	520	21	21	329	327	322
chirnaadoccult	23	125	1397	1528	1534	1735	1972	1954	1716	1808	1534	30	28	331	339	336
chirabces	23	125	1397	1528	1534	1735	1972	1953	1705	532	569	17	26	329	336	328
chirbloeding	23	125	1397	1528	1534	1735	1972	1954	1705	590	629	16	23	329	334	326
chirileus	23	125	1397	1528	1534	1735	1972	1954	1704	450	510	16	20	324	325	326
chirgastro	23	125	1397	1528	1535	1735	1972	1956	1716	1808	1537	1381	1588	1480	295	292
chirfascie	23	125	1397	1528	1534	1735	1972	1954	1705	617	669	19	24	332	346	336
chirdarmperf	23	125	1397	1528	1534	1735	1972	1954	1705	638	692	22	27	334	342	334
chirlekkage	23	125	1397	1528	1534	1735	1972	1954	1705	640	687	22	26	331	344	335
chirwondinf	23	125	1397	1528	1534	1735	1972	1954	1705	512	577	29	29	339	340	336
datcompl	1868	2406	2414	2550	2456	2865	3144	3126	2711	2674	2312	576	595	558	536	921
comppul	1825	2171	169	145	155	113	98	96	47	806	871	57	73	427	480	521
compcar	1825	2205	152	134	171	117	112	103	52	841	903	58	75	430	484	519
compthrom	1837	2216	192	223	192	127	128	102	55	929	973	60	78	429	491	524
compinf	1825	2206	157	125	160	118	113	96	49	783	816	55	76	428	483	525
compneu	1837	2215	195	183	187	136	130	102	53	901	942	60	77	432	491	523
compx	1825	2254	198	137	141	82	85	83	41	709	777	54	64	423	471	507
letsel	1861	2396	2384	2549	2448	2850	98	65	58	146	149	96	101	399	407	452
icdg	5727	7001	767	268	333	163	186	364	332	362	447	249	458	1638	1575	1457
icreden	10	70	1507	1690	1479	1645	308	312	265	215	209	23	35	51	87	67
transf	326	184	121	156	67	90	245	250	265	382	372	220	349	1488	1579	1524
heropn90	5279	6870	7478	7638	7247	8677	9873	9682	8676	27	5	11	15	1229	1237	322
datreint	11	70	816	667	584	596	641	599	553	667	509	455	537	569	560	503
tumorrest	5736	7251	7693	8492	8139	9606	10796	10596	9628	9189	8413	7080	8072	7375	1415	1425
dmarge	1879	2304	2532	2722	2580	2882	3190	2918	2659	2377	2100	1934	2097	741	655	621
macrotme	644	716	852	799	811	908	999	674	349	402	307	224	185	373	385	351
tumdepositnew	5736	7249	7690	8354	8043	9462	10620	10379	9395	6344	697	999	1418	2641	2896	2777
anginvnew_0	1087	1010	542	584	377	438	524	566	261	377	152	107	194	1791	2032	2090
anginv_1	5105	6439	6620	7373	7009	8351	9712	8220	312	310	327	171	290	1802	2042	2090
anginv_3	5736	7250	7687	8491	8138	9600	10418	8870	313	365	368	193	302	1821	2052	2114
diffgraad	5680	7049	7265	897	259	117	43	16	32	66	36	39	92	447	485	605
paperfdarm	5736	7248	7690	8492	8138	9603	10566	8118	7088	347	169	149	215	1794	2087	2043
datprertxstop	1535	1948	2235	2367	2186	1932	2067	1882	1627	921	794	674	676	643	594	534
rtxrestad	1535	1948	2235	2367	2185	1933	2067	1882	1611	81	83	67	49	256	216	185
prectx	4031	5104	5164	5707	5606	7291	8084	6269	5426	15	3	2	11	31	81	6252
datctx1stop	109	227	569	688	565	568	652	607	475	692	688	511	575	638	625	541
chemo1	1615	1966	566	419	364	344	375	234	166	6597	7008	5979	6788	6292	6352	6136
datchemostop	1623	2025	2015	2064	1915	2046	2283	2013	1716	6990	7466	6379	7193	6616	6697	6436
chemokur	1141	1415	1592	1742	1627	1754	1953	1861	1621	972	1040	803	1021	871	837	710
chemokuraanpas	1141	1415	1592	1742	1627	1754	1953	1861	1622	1054	1085	843	1044	875	836	722
chemokurreden	1141	1415	1592	1742	1627	1754	1953	1861	1622	1085	1102	849	1045	881	846	727
chemokurgrad	1623	2025	2015	2064	1915	2047	2283	2014	1733	7815	8260	6967	7959	7277	7309	6972
systgeen	1623	2025	2015	2064	1915	2047	2283	2014	1733	7815	8260	6967	7959	7277	7309	6972
tumorgergeen	3999	4943	5610	6391	6171	7507	8439	8531	7837	17	0	0	0	0	0	1

Table 5.3: An overview of the amount of missing values per year for the variables with conditions that have over 5% missingness. If there is over a 1000 missing values for a certain year, the amount is highlighted.

Table 5.3 shows the amount of missing values per year for the variables with conditions that have over 5% missingness. If there is over a 1000 missing values for a certain year, the amount is highlighted. When we look at Table 5.3 and focus on the variables that have highlighted values, we see the following. For quite some variables with the amount of NA values over 1000 in some years, the variable collection started from a certain year onwards, and hence in the years before that, there was a high amount of missing values. This is for example the case for the variable `datcompl` (Date of first general complication) which started to be collected in 2020, which explains the high NA amount from 2009 to 2019. Some other variables were always collected and still had high NA values in certain years. The variable `prectx` (Was systemic therapy administered prior to tumour resection (chemotherapy and/or targeted agents)?) is an

example of this. This variable was always collected, but has high NA counts from 2009-2017. There is no obvious explanation for this.

5.3 Imputation due to data collection

Some of the variables are in specific categories. For these categories we assume that if one of the variables is 'yes' (1) we assume that the others are 'no' (0). We call this *imputations due to data collection*.

The categories that we have are 'patient identification', 'Charlson Comorbidity Index', 'preoperative tumour complications', 'colon perforation', 'preoperative surgical procedure', 'continued growth in which organs (localisation)', 'resection metastases', 'surgical complication', 'general complication', 're-intervention', 'lymphatic/venous invasion 1st tumour', and 'lymphatic/venous invasion 2nd tumour'. For the categories 'Charlson Comorbidity Index', 'preoperative tumour complications', 'colon perforation', 'continued growth in which organs (localisation)', 'resection metastases', 'surgical complication', and 'general complication' we can assume that if one of the variables in these categories is 'yes' (1) and the others in that category are not filled in, then we can fill in 'no' (0) for the others. We reached this conclusion in consult with a medical expert. This is due to the definition of the variables in these categories. An example for this is that if we look at the category in comorbidities, and one of the comorbidities, say `comdem` (dementia), is 'yes' (so the patient has dementia), and the other comorbidities are all not filled in, then we assume that the patient does not have any of the other comorbidities, so we set the other comorbidities to 'no' (0). We make this assumption because of how the data is collected, namely that if a doctor registers only one of the comorbidities, they can continue to the next page to fill in. On top of that, if there are no comorbidities, then the comorbidity page for registering is not even showing, which we also discussed in Section 5.2. We have decided for which categories this is possible in consult with a medical expert. Also for 'umbrella variables' imputation due to data collection is applicable. An umbrella variable we define as a variable that has to have a specific value in order for the variables in its category to be visible. For example, for the category 'preoperative tumour complications' the umbrella variable is `complpre`. Namely, if `complpre` (Tumour-related complications?) is 'yes' (1), only then the variables `preanemie`, `preobstruct`, `preabces` are showing, since they all belong to its category (pre-operative tumour complications). This is also decided in consult with a medical expert.

In the category 'Charlson Comorbidity Index' are the variables `commyo` (Myocardial infarction), `comhartfaal` (Congestive heart failure), `comperif` (Peripheral vascular disease/aortic aneurysm), `comcva` (Cerebrovascular disorder (including CVA/TIA)), `comdem` (Dementia), `comlong` (Chronic lung disease), `combind` (Connective tissue disease (including rheumatoid diseases)), `comgiulcus` (Gastrointestinal ulcer disease), `comlever` (Liver disease), `comdiam` (Diabetes mellitus), `comparalys` (Paraplegia/hemiplegia), `comnier` (Kidney disease), `commalig` (Malignancy (excluding PCC or BCC of the skin)), and `comhiv` (HIV/AIDS).

In the category 'preoperative tumour complications' are the variables `preanemie`, `preobstruct`, `preabces`. The 'umbrella' variable of this category is `complpre`.

In the category 'colon perforation' are the variables `ileusblow`, `perfperi`, and `perfscopie`. There is no 'umbrella' variable of this category.

In the category 'continued growth in which organs (localisation)' are the variables `locaddmaag`, `locaddmilt`, `locaddduod`, `locadddoverig`, `locaddoment`, `locaddbuikw`, `locaddvag`, `locadduter`, `locaddovarl`, `locaddovarr`, `locadduretl`, `locadduretr`, `locaddblaas`, `locaddprost`, `locaddvesl`, `locaddvesr`, `locaddvesb`, `locadddund`, `locaddsacr`, and `locaddoverig`. The 'umbrella' variable of this category is `resadd`.

In the category 'resection metastases' are the variables `resomm`, `reslem`, `respem`, `reslmm`, and `resbum`. The 'umbrella' variable of this category is `resadm`.

In the category 'surgical complication' are the variables `chirnaad`, `chirnaadoccult`, `chirabces`, `chirbloeding`, `chirileus`, `chirgastro`, `chirfascie`, `chirdarmperf`, `chirlekkage`, and `chirwondinf`. There is no 'umbrella' variable of this category.

In the category 'general complication' are the variables `comppul`, `compcar`, `compthrom`, `compinf`, `compneu`, and `compx`. There is no 'umbrella' variable of this category.

For the preoperative variables, the variables for which the imputation due to data collection is applicable are thus `commyo`, `comhartfaal`, `comperif`, `comcva`, `comdem`, `comlong`, `combind`, `comgiulcus`, `comlever`, `comdiam`, `comparalys`, `comnier`, `commalig`, `comhivcomplpre`, `preanemie`, `preobstruct`, `preabces`, `darmperf`, `ileusblow`, `perfperi`, `perfscopie`, `resadd`, `locaddmaag`, `locaddmilt`, `locaddduod`, `locadddoverig`, `locaddoment`, `locaddbuikw`, `locaddvag`, `locadduter`, `locaddovarl`, `locaddovarr`, `locadduretl`, `locadduretr`, `locaddblaas`, `locaddprost`, `locaddvesl`, `locaddvesr`, `locaddvesb`, `locadddund`, `locaddsacr`, `locaddoverig`, `resadm`, `resomm`, `reslem`, `respem`, `reslmm`, and `resbum`. For the outcome variables, the variables for which the imputation due to data collection is applicable are thus `chirnaad`, `chirnaadoccult`, `chirabces`, `chirbloeding`, `chirileus`, `chirgastro`, `chirfascie`, `chirdarmperf`, `chirlekkage`, `chirwondinf`, `comppul`, `compcar`, `compthrom`, `compinf`, `compneu`, `compx`.

The performed imputation due to data collection for the preoperative variables is in Table 5.4. In this table the variables are displayed on the left and the corresponding condition with which we filled in the NA values per variable is on the right. The performed imputation due to data collection for the outcome variables is in Table 5.5. This table is similar to the previously described table.

Variable	**Condition**
Diagnostics	
`complpre`	Look at the complications that follow, `preanemie`, `preobstruct`, `preabces` . If all are NA, then we assume `complpre` = 0. 1) IF (`preanemie` == 0 or NA) AND (`preobstruct` == 0 or NA) AND (`preabces` == 0 or NA), THEN `complpre` == 0.
`preanemie`	If other complication (`preobstruct` or `preabces`) is filled in, and this is not, we can assume `preanemie` = 0.
`preobstruct`	If other complication (`preanemie` or `preabces`) is filled in, and this is not, we can assume `preobstruct` = 0.
`preabces`	If other complication (`preanemie` or `preobstruct`) is filled in, and this is not, we can assume `preabces` = 0.
Surgery	
`darmperf`	IF `darmperf` == NA AND (`ileusblow` != 1 AND `perfperi` != 1, AND `perfscopie` != 1) THEN `darmperf` == 0
`ileusblow`	1) IF `ileusblow` == NA AND (perfperi != NA, OR `perfscopie` != NA) THEN `ileusblow` == 0.
`perfperi`	IF `perfperi` == NA AND (`ileusblow` != NA, OR `perfscopie` != NA) THEN `perfperi` == 0.
`perfscopie`	IF `perfscopie` == NA AND (`ileusblow` != NA, OR `perfperi` e != NA) THEN `perfscopie` == 0.

resadd	If the rest of the 'doorgroei' is all 0, we put this also on 0. 1) IF resadd == NA AND locaddmaag != 1 AND locaddmilt != 1 AND locaddduod != 1 AND locaddddoverig != 1 AND locaddoment != 1 AND locaddbuikw != 1 AND locaddvag != 1 AND locadduter != 1 AND locaddovarl != 1 AND locaddovarr != 1 AND locadduretl != 1 AND locadduretr != 1 AND locaddblaas != 1 AND locaddprost != 1 AND locaddvesl != 1 AND locaddvesr != 1 AND locaddvesb != 1 AND locadddund != 1 AND locaddsacr != 1 AND locaddoverig != 1, THEN resadd == 0. 2) If one of the 'dooorgroei' is 1, we put this as 1. We can do this with what is left since this is the negate of what we did before.
locaddmaag	If another is already 1, we can just put this as 0. If the rest of the 'doorgroei' is all 0, we also put it on 0. 1) IF locaddmaag == NA AND (locaddmilt == 1 OR locaddduod == 1 OR locaddddoverig == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locaddmaag == 0.
locaddmilt	1) IF locaddmilt == NA AND (locaddmaag == 1 OR locaddduod == 1 OR locaddddoverig == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locaddmilt == 0.
locaddduod	1) IF locaddduod == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddddoverig == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locaddduod == 0.
locaddddoverig	1) IF locaddddoverig == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddduod == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locaddddoverig == 0.
locaddoment	1) IF locaddoment == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddduod == 1 OR locaddddoverig == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locaddoment == 0.
locaddbuikw	1) IF locaddbuikw == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddduod == 1 OR locaddddoverig == 1 OR locaddoment == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locaddbuikw == 0.
locaddvag	1) IF locaddvag == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddduod == 1 OR locaddddoverig == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locadduter == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locaddvag == 0.

locadduter	1) IF locadduter == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddduod == 1 OR locadddoverig == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locadduter == 0.
locaddovarl	1) IF locaddovarl == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddduod == 1 OR locadddoverig == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locaddovarl == 0.
locaddovarr	1) IF locaddovarr == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddduod == 1 OR locadddoverig == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarl == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locaddovarr == 0.
locadduretl	1) IF locadduretl == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddduod == 1 OR locadddoverig == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locadduretl == 0.
locadduretr	1) IF locadduretr == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddduod == 1 OR locadddoverig == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locadduretr == 0.
locaddblaas	1) IF locaddblaas == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddduod == 1 OR locadddoverig == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locaddblaas == 0.
locaddprost	1) IF locaddprost == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddduod == 1 OR locadddoverig == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locaddprost == 0.
locaddvesl	1) IF locaddvesl == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddduod == 1 OR locadddoverig == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locaddvesl == 0.

locaddvesr	1) IF locaddvesr == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddduod == 1 OR locaddddoverig == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locaddvesr == 0.
locaddvesb	1) IF locaddvesb == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddduod == 1 OR locaddddoverig == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locadddund == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locaddvesb == 0.
locadddund	1) IF locadddund == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddduod == 1 OR locaddddoverig == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locaddsacr == 1 OR locaddoverig == 1), THEN locadddund == 0.
locaddsacr	1) IF locaddsacr == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddduod == 1 OR locaddddoverig == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddoverig == 1), THEN locaddsacr == 0.
locaddoverig	1) IF locaddoverig == NA AND (locaddmaag == 1 OR locaddmilt == 1 OR locaddduod == 1 OR locaddddoverig == 1 OR locaddoment == 1 OR locaddbuikw == 1 OR locaddvag == 1 OR locadduter == 1 OR locaddovarl == 1 OR locaddovarr == 1 OR locadduretl == 1 OR locadduretr == 1 OR locaddblaas == 1 OR locaddprost == 1 OR locaddvesl == 1 OR locaddvesr == 1 OR locaddvesb == 1 OR locadddund == 1 OR locaddsacr == 1), THEN locaddoverig == 0.
resadm	1) If resomm till resbum are all not 1, we put this at 0.IF resadm == NA AND resomm != 1 AND reslem != 1 AND respem != 1 AND reslmm != 1 AND resbum != 1, THEN resadm == 0. 2) If one is 1, we also put this as 1. IF resadm == NA AND (resomm == 1 OR reslem == 1 OR respem == 1 OR reslmm == 1 OR resbum == 1), THEN resadm == 1.
resomm	1) If another is 1, we put this as 0. IF resomm == NA AND (reslem == 1 OR respem == 1 OR reslmm == 1 OR resbum == 1), THEN resomm == 0.
reslem	IF reslem == NA AND (resomm == 1 OR respem == 1 OR reslmm == 1 OR resbum == 1), THEN reslem == 0.
respem	IF respem == NA AND (resomm == 1 OR reslem == 1 OR reslmm == 1 OR resbum == 1), THEN respem == 0.
reslmm	IF reslmm == NA AND (resomm == 1 OR reslem == 1 OR respem == 1 OR resbum == 1), THEN reslmm == 0.
resbum	IF resbum == NA AND (resomm == 1 OR reslem == 1 OR respem == 1 OR reslmm == 1), THEN resbum == 0.

Table 5.4: The preoperative variables that are imputed due to data collection.

Variable	Condition
chirnaad	If another is 1, put this at 0. IF chirnaad == NA AND (chirnaadoccult == 1 OR chirabces == 1 OR chirbloeding == 1 OR chirileus == 1 OR chirgastro == 1 OR chirfascie == 1 OR chirdarmperf == 1 OR chirlekkage == 1 OR chirwondinf == 1) THEN chirnaad == 0 (No)

chirnaadoccult	If another is 1, put this at 0. IF chirnaadoccult == NA AND (chirnaad == 1 OR chirabces == 1 OR chirbloeding == 1 OR chirileus == 1 OR chirgastro == 1 OR chirfascie == 1 OR chirdarmperf == 1 OR chirlekkage == 1 OR chirwondinf == 1) THEN chirnaadoccult == 0 (No)
chirabces	If another is 1, put this at 0. IF chirabces == NA AND (chirnaad == 1 OR chirnaadoccult == 1 OR chirbloeding == 1 OR chirileus == 1 OR chirgastro == 1 OR chirfascie == 1 OR chirdarmperf == 1 OR chirlekkage == 1 OR chirwondinf == 1) THEN chirabces == 0 (No)
chirbloeding	If another is 1, put this at 0. IF chirbloeding == NA AND (chirnaad == 1 OR chirnaadoccult == 1 OR chirabces == 1 OR chirileus == 1 OR chirgastro == 1 OR chirfascie == 1 OR chirdarmperf == 1 OR chirlekkage == 1 OR chirwondinf == 1) THEN chirbloeding == 0 (No)
chirileus	If another is 1, put this at 0. IF chirileus == NA AND (chirnaad == 1 OR chirnaadoccult == 1 OR chirabces == 1 OR chirbloeding == 1 OR chirgastro == 1 OR chirfascie == 1 OR chirdarmperf == 1 OR chirlekkage == 1 OR chirwondinf == 1) THEN chirileus == 0 (No)
chirgastro	If another is 1, put this at 0. IF chirgastro == NA AND (chirnaad == 1 OR chirnaadoccult == 1 OR chirabces == 1 OR chirbloeding == 1 OR chirileus == 1 OR chirfascie == 1 OR chirdarmperf == 1 OR chirlekkage == 1 OR chirwondinf == 1) THEN chirgastro == 0 (No)
chirfascie	If another is 1, put this at 0. IF chirfascie == NA AND (chirnaad == 1 OR chirnaadoccult == 1 OR chirabces == 1 OR chirbloeding == 1 OR chirileus == 1 OR chirgastro == 1 OR chirdarmperf == 1 OR chirlekkage == 1 OR chirwondinf == 1) THEN chirfascie == 0 (No)
chirdarmperf	If another is 1, put this at 0. IF chirdarmperf == NA AND (chirnaad == 1 OR chirnaadoccult == 1 OR chirabces == 1 OR chirbloeding == 1 OR chirileus == 1 OR chirgastro == 1 OR chirfascie == 1 OR chirlekkage == 1 OR chirwondinf == 1) THEN chirdarmperf == 0 (No)
chirlekkage	If another is 1, put this at 0. IF chirlekkage == NA AND (chirnaad == 1 OR chirnaadoccult == 1 OR chirabces == 1 OR chirbloeding == 1 OR chirileus == 1 OR chirgastro == 1 OR chirfascie == 1 OR chirdarmperf == 1 OR chirwondinf == 1) THEN chirlekkage == 0 (No)
chirwondinf	If another is 1, put this at 0. IF chirwondinf == NA AND (chirnaad == 1 OR chirnaadoccult == 1 OR chirabces == 1 OR chirbloeding == 1 OR chirileus == 1 OR chirgastro == 1 OR chirfascie == 1 OR chirdarmperf == 1 OR chirlekkage == 1) THEN chirwondinf == 0 (No)
comppul	If another is 1, put this at 0. IF comppul == NA AND (compcar == 1 OR compthrom == 1 OR compinf == 1 OR compneu == 1 OR compx == 1) THEN comppul == 0 (No)
compcar	If another is 1, put this at 0. IF compcar == NA AND (comppul == 1 OR compthrom == 1 OR compinf == 1 OR compneu == 1 OR compx == 1) THEN compcar == 0 (No)
compthrom	If another is 1, put this at 0. IF compthrom == NA AND (comppul == 1 OR compcar == 1 OR compinf == 1 OR compneu == 1 OR compx == 1) THEN compthrom == 0 (No)
compinf	If another is 1, put this at 0. IF compinf == NA AND (comppul == 1 OR compcar == 1 OR compthrom == 1 OR compneu == 1 OR compx == 1) THEN compinf == 0 (No)
compneu	If another is 1, put this at 0. IF compneu == NA AND (comppul == 1 OR compcar == 1 OR compthrom == 1 OR compinf == 1 OR compx == 1) THEN compneu == 0 (No)
compx	If another is 1, put this at 0. IF compx == NA AND (comppul == 1 OR compcar == 1 OR compthrom == 1 OR compinf == 1 OR compneu == 1) THEN compx == 0 (No)

Table 5.5: The outcome variables that are imputed due to data collection.

5.4 Select relevant years

In this section we decide which years we are going to include. The conclusion is that we are going to include it from 2018 onwards. The reason for this is because a lot of data collection changes occurred from 2017 to 2018. In other years there were also data changes happening, but it was by far the most from 2017 to 2018, for example, from 2018 onwards the complications were registered for up to 90 days instead of 30 days and the registration of comorbidities also changed in 2018 [34]. This is visible in Table 4.9 and in Figure 5.1. In this figure you see visualised per year the amount of variables that are or are not in the most recent datadictionary (of 2024) on which the dataset we used is based. We see that the count difference compared to the previous year is biggest in 2018. In other years there were also slight changes in the variables that were collected, but not as big as in 2018. Note that there is no datadictionary of 2023. Hence, we conclude that we continue with data included from 2018 onwards.

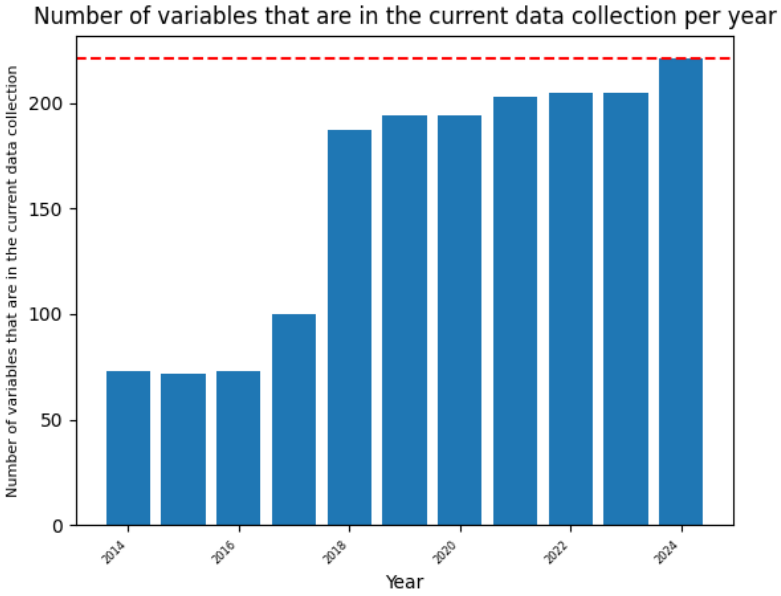


Figure 5.1: The amount of variables that were collected in that year and are still collected in the datadictionary of 2024.

5.5 Decide which preoperative variables to keep

We continue with the dataset that is retrieved with only the elective surgeries (Section 5.1), the imputations from datadictionary (Section 5.2), the imputations from data collection (Section 5.3), and data from 2018 onwards (Section 5.4). For a description of the variables, we refer to Chapter 4. For an overview of all of the options of the values of the variables and what they mean, we refer to Appendix B. For an overview of all the original percentages of missing values per variable, we refer to Appendix C.

In this step we check which variables can later be removed and decide which preoperative variables to keep. The variables that can later be removed are one or more of the following, (A) outcomes rather than predictors, (B) postoperative and hence could not be predictors, (C) only useful for administration, (D) said by the the medical expert they are not useful as predictors (E) containing over 15% missingness after the imputations from datadictionary and data collection. We call the variables of this group, group R (removed). The preoperative variables that will not be removed, we name group I (included).

Types A, B, C, and D are all quite similar in the sense that they remove variables for which the meaning of the variables is not useful for predicting the postoperative complications. Type E is a bit different since this excludes variables that have too much missingness, but that otherwise could have been possible useful predictors for postoperative complications. For type E we decided to go for variables with less than 15% missingness (*after* imputations from datadictionary and data collection and data from 2018 onwards), since otherwise we would have to impute too much and the results could be influenced a lot by the percentage of imputed values and variables. The variables that are of type A, B, C, D, or E we are in group R. The variables that are left and thus are not of type A, B, C, D, or E are in group I.

Variables of type A include `chirnaad`, `chirnaadoccult`, `chirabces`, `chirbloeding`, `chirileus`, `chirgastro`, `chirfascie`, `chirdarmperf`, `chirlekkage`, `chirwondinf` for the surgical complications, and `comppul`, `compcar`, `compthrom`, `compinf`, `compneu`, `compx` for general complications. So these variables are in group R.

Variables of type B include `compl90`, `compchir`, `datcompchir`, `datcompl`, `letsel`, `icdg`, `icreden`, `transf`, `heropn90`, `datheropn`, `reintreq`, `datreint`, `reintaard`, `reintst`, `datont`, `status90`, `doodoorz` for the *surgery* subdataset. These variables describe whether complications occurred, whether and why the patient was in the ICU, when the patient was discharged or if and when the patient was readmitted to the hospital. Variables `tumorrest`, `tumorrest2`, `histtype`, `stadpt`, `stadpnnew`, `stadpn`, `histtype2`, `stadpt2`, `stadpn2`, `stadpm`, `resectien`, `cirmarge`, `dmarge`, `resectie2n`, `cirmarge2`, `dmarge2`, `cirmargelok`, `cirmargebasaal`, `macrotme`, `numlym`, `numlympos`, `cirmargeklier`, `cirmargeklier2`, `tumdepositnew`, `anginvnew_0`, `anginv_1`, `anginv_3`, `diffgraad`, `paperfdarm`, `regress`, `anginvnew2_0`, `anginv2_1`, `anginv2_3`, `diffgraad2`, `paperfdarm2`, `regress2`, `braf`, `ras` for the *pathology* subdataset. These variables describe the tumour with its stadium and size. All these variables are collected around two weeks after the surgery occurred, because then the pathology research has taken place, with the excised tumour. Variables `rtxrestad`, `rtxct`, `rtxcn` for the *radiotherapy* subdataset. These variables are describing whether restaging of the tumour was done after the radiotherapy was completed, and what the newly determined stage then would be. According to our medical expert, this does not contribute to a possibility of predicting complications after surgery. Variables `prectxaard`, `datchemo`, `chemopost`, `chemo`, `chemo1`, `datchemostop`, `chemokuur`, `chemokuuraanpas`, `chemokuurreden`, `chemokuurgrad` for the *systemic therapy* subdataset. These variables give very detailed information on whether systemic therapy was done and in what form. This is too detailed information, and according to our expert, this is not useful in the prediction of postoperative complications. So these variables are in group R.

Variables of type C include `id`, `uri`, `patient_uri` for the *patient* subdataset. Variables `tumor_id`, `idaaba`, `presentatie_uri` for the *presentation* subdataset. Variable `diagnostiek_uri` for the *diagnostics* subdataset. Variables `klaar`, `chirurgie_uri` for the *surgery* subdataset. Variables `systgeen`, `tumorgergeen` for the *systemic therapy* subdataset. So these variables are in group R.

Variables of type D include `prelocstad`, `emimri`, `emimri2`, `mlh1`, `pms2`, `hyperm`, `msh2`, `msh6` of the subdataset *diagnostics*. So these variables are in group R.

Variables of type E include `datovl` for the *patient* subdataset. Variable `datcom` for the *comorbidities* subdataset. Variables `updated`, `genan`, `imuun` for the *diagnostics* subdataset. Variables `datstoma`, `datstent`, `datappendix`, `procok3` for the *surgery* subdataset. Variable `datrad1` for the *radiotherapy* subdataset. Variable `datctx1` for the *systemic therapy* subdataset. So these variables are in group R.

So the variables that are still included after variables from group R are removed (which will happen later on in Section 5.7), are `gebdat`, `geslacht`, `commyo`, `comhartfaal`, `comperif`, `comcva`,

comdem, comlong, combind, comgiulcus, comlever, comdiam, comparalys, comnier, commalig, comhiv, lengte, gewicht, asascore, resdarm, stomavg, datpa1, primcrc, scopnumb, diagn, locprim, histd, complpre, preanemie, preobstruct, preabces, scorect, scorecn, locprim2, scorect2, scorecn2, scorecm, prelocmri, prelocmri2, scopdistmm, scopdist2mm, datbez1, verwezen, darmperf, ileusblow, perfperi, perfscopie, prechirstomanew, prechirstent, prechirmetanew, datprechirmetaeen, datprechirmetatwee, prechirappendix, prechircoloscex, chirenkel, datok, typok, procok, ingreep, conversien, aprtyp, typprocok, datok2, procok2, ingreep2, conversien2, aprtyp2, resadd, locaddmaag, locaddmilt, locaddduod, locadddoverig, locaddoment, locaddbuikw, locaddvag, locadduter, locaddovarl, locaddovarr, locadduretl, locadduretr, locaddblaas, locaddprost, locaddvesl, locaddvesr, locaddvesb, locadddund, locaddsacr, locaddoverig, resadm, resomm, reslem, respem, reslmm, resbum, perrtxchemo, anastom1, stoma, stomahef, radtiming, prectx, charlgrp, type_ziekenhuis, leeftijd. These 103 variables form the group I. These are the variables on which we will perform clinical imputation in Section 5.6.

5.6 Clinical imputation

In this step we perform clinical imputations on group I which consist of 103 variables. With clinical imputation we mean imputation based on clinical or medical reasoning. So instead of "just" taking the mean or median and impute that, we also look at clinical conditioning imputation. Before we performed this imputation, we had multiple consults with a medical expert who told us when and how to assume a certain value and when to absolutely not to assume anything and leave the value NA. The variables for which the medical expert said it was possible to assume certain values with clinical reasoning and for which we thus perform this clinical imputation, are gebdat, geslacht, commyo, comhartfaal, comperif, comcva, comdem, comlong, combind, comgiulcus, comlever, comdiam, comparalys, comnier, commalig, comhiv. lengte, gewicht, asascore, resdarm, stomavg, datpa1, scopnumb, diagn, histd, scorect, scorecn, scorect2, scorecn2, scorecm, prelocmri, prelocmri2, scopdistmm, scopdist2mm, datbez1, datprechirmetaeen, datprechirmetatwee, prechirappendix, chirenkel, conversien, datok2, procok2, ingreep2, conversien2, aprtyp2, perrtxchemo, radtiming, prectx, type_ziekenhuis.

There are also some variables for which the medical expert said that these variables could not be estimated and that the responsible choice was to leave these variables as unknown. These are the variables geslacht, asascore, resdarm, locprim2, prechirstomanew, prechirmetanew, typok, procok, ingreep, aprtyp, typprocok, anastom1, stoma, stomahef, type_ziekenhuis. Sometimes there is something else that is not medically relevant but more for the administration that can be imputed first. This is the case with asascore and resdarm, where we first select whether the value was before 2018, since this variable has the option 10 as a value which means that the variable 'was not collected before 2018 and that this was still before 2018'.

Note that we have not yet removed the variables that are to be removed (group R, determined in Section 5.5) and we will work with all variables (both group R and group I), to make these clinical imputations happen. However, we do not do the imputations on all of the variables, only on the variables of group I. The reason for this is that later on, we will continue for the preoperative variables, with the variables in group I but we sometimes also need variables from group R (if they are available) to impute variables from group I.

We perform clinical imputation until all these variables (from group I) have as little missing values as possible without filling in that the missing values are unknown. Unless 'unknown' was already an option in the beginning and so also for the surgeons to fill in unknown. The comorbidities are a special case. We know from the 'unknown' values that are left that it does say that there are comorbidities, but that we just do not know which specific comorbidity it is. Hence, there is value and also possible predictor value in the 'unknown' values of the comorbidities.

In Table 5.6 the variables for which we will perform clinical imputation are shown with their corresponding percentage of missing values before the clinical imputation.

Variable	**% NA**
Patient	
gebdat	0,02
geslacht	0,01
Comorbidities	
commyo	57,79
comhartfaal	60,05
comperif	61,23
comcva	57,04
comdem	60,54
comlong	55,14
combind	59,68
comgiulcus	59,49
comlever	60,66
comdiam	52,54
comparalys	64,32
comnier	59,77
commalig	53,60
comhiv	61,20
Presentation	
lengte	0,91
gewicht	0,90
asascore	0,05
resdarm	0,90
stomavg	47,99
datpa1	1,07
Diagnostics	
scopnumb	0,41
diagn	53,16
histd	61,40
⋮	⋮
⋮	⋮
scorect	20,82
scorecn	21,17
scorect2	1,01
scorecn2	2,05
scorecm	20,58
prelocmri	0,16
prelocmri2	1,17
scopdistmm	19,87
scopdist2mm	1,02
Surgery	
datbez1	3,47
datprechirmetaeen	0,43
datprechirmetatwee	0,36
prechirappendix	4,43
chirenkel	59,62
conversien	7,64
datok2	2,28
procok2	0,58
ingreep2	2,03
conversien2	0,04
aprtyp2	0,08
perrtxchemo	30,11
Radiotherapy	
radtiming	4,94
Systemic Therapy	
prectx	44,49
Other	
type_ziekenhuis	1,39

Table 5.6: The variables that have to be imputed, with their percentage of missing values before the clinical imputation.

In Table 5.7 the variables and their clinical reasoning are displayed. There is a lot of different reasonings and all are created in consult with the medical expert. An example is the clinical reasoning for the variable `scorect` (cT 1st colorectal carcinoma). To impute this value we use the subdataset *pathology*. In this subdataset we have the variable `stadpt` (Pathological T classification of 1st colorectal carcinoma according to TNM system). This variable is essentially the same as the variable `scorect`, only `stadpt` is determined by pathologists after the tumour was resected. So if the classification was not done prior to the surgery and therefore `scorect` was NA, we use the corresponding value of `stadpt` determined after the surgery to fill in `scorect`. By

doing this imputation in consult with a medical expert, the data becomes much more powerful than if we would have imputed it with the most common value, the mean, or the median. For example, for the variable `lengte` (height) we first looked per gender at the average height and then we imputed the height corresponding to the gender of the patient for this patient.

Variable	**Condition**
Patient	
`gebdat`	Take the average of age (= `gebdat` to `datok`) and impute with that
`geslacht`	Fill in unknown
Comorbidities	
`commyo`	Fill in unknown
`comhartfaal`	Fill in unknown
`comperif`	Fill in unknown
`comcva`	Fill in unknown
`comdem`	Fill in unknown
`comlong`	Fill in unknown
`combind`	Fill in unknown
`comgiulcus`	Fill in unknown
`comlever`	Fill in unknown
`comdiam`	Fill in unknown
`comparalys`	Fill in unknown
`comnier`	Fill in unknown
`commalig`	Fill in unknown
`comhiv`	Fill in unknown
Presentation	
`lengte`	IF `lengte` == NA AND (`geslacht` == x) THEN `lengte` = 'average height of patients with `geslacht` == x'
`gewicht`	IF `gewicht` == NA AND (`geslacht` == x) THEN `gewicht` = 'average height of patients with `geslacht` == x'
`asascore`	If before 2018, fill in 10.
`resdarm`	If before 2018, fill in 10.
`stomavg`	Fill in 0.
`datpa1`	1) If `datbez1` == `datok` AND `datpa1` == empty, THEN `datpa1` == `datok`, 2) If `typok` == 3 of 4 and `datbez1` == empty, AND `datpa1` == empty, THEN `datpa1` == `datok`, 3) If `typok` == NA and `datbez1` == empty, AND `datpa1` == empty, THEN `datpa1` == `datok`, 4) If `datbez1` == empty, AND `datpa1` == empty, THEN `datpa1` == `datok`, 5) If `datbez1` == empty, AND `typok`==9 AND `datpa1` == empty, THEN `datpa1` == `datok`, 6) If `datbez1` == not empty, AND `datpa1` == empty, THEN `datpa1` == `datbez1`
Diagnostics	
`scopnumb`	Look at all other variables that have to do with the second carcinoma. If that is all empty, we can assume that it is only one carcinoma. 1) IF (`locprim2` == 0 or 999 or NA), AND (`scorect2` == 0 or 9 or NA), AND (`scorecn2` == 10 or NA), AND (`prelocmri2` == 0 or 10 or NA), AND (`scopdist2mm` == 999 or 888 or NA), AND (`afstmrf2` == 999 or 888 or NA), AND (`emimri2` == 10 or NA), THEN `scopnumb` == 1
`diagn`	Fill in unknown.
`histd`	Fill in unknown.
`scorect`	We are going to use the subdataset *pathology*. 1) IF `scorect` == NA AND `stadpt` == 1, THEN `scorect` == 1 2)IF `scorect` == NA AND `stadpt` == 2, THEN `scorect` == 2 3) IF `scorect` == NA AND `stadpt` == 3, THEN `scorect` == 3 4) IF `scorect` == NA AND `stadpt` == 4, THEN `scorect` == 4 5) IF `scorect` == NA AND `stadpt` == 11, THEN `scorect` == 5 6) IF `scorect` == NA AND `stadpt` == 12, THEN `scorect` == 6 7) IF `scorect` == NA AND `stadpt` == 9, THEN `scorect` == 9.

scorecn	1) IF cT is 1 or 2, then this is most likely 0 (according to our medical expert). 2) Fill in with *pathology* stadpn. IF scorecn == NA AND stadpn == 0, THEN scorecn == 0. IF scorecn == NA AND stadpn == 1, THEN scorecn == 1. IF scorecn == NA AND stadpn == 2, THEN scorecn == 2. IF scorecn == NA AND stadpn == 5 or 9, THEN scorecn == 9 3) Fill in with stadpnnew. IF scorecn == NA AND stadpnnew == 0, THEN scorecn == 0. IF scorecn == NA AND stadpnnew == 1 or 2 or 3 or 7, THEN scorecn == 1. IF scorecn == NA AND stadpnnew == 4 or 5 or 8, THEN scorecn == 2. IF scorecn == NA AND stadpnnew == 6, THEN scorecn == 9.
scorect2	1) Start with scopnumb. IF scopnumb == 1 THEN scorect2 == 0. 2) Fill in with *pathology*. IF scorect2 == NA AND stadpt2 == 1, THEN scorect2 == 1. IF scorect2 == NA AND stadpt2 == 2, THEN scorect2 == 2. IF scorect2 == NA AND stadpt2 == 3, THEN scorect2 == 3. IF scorect2 == NA AND stadpt2 == 4, THEN scorect2 == 4. IF scorect2 == NA AND stadpt2 == 11, THEN scorect2 == 5. IF scorect2 == NA AND stadpt2 == 12, THEN scorect2 == 6. IF scorect2 == NA AND stadpt2 == 9 or 5, THEN scorect2 == 9.
scorecn2	1) Start with scopnumb. IF scopnumb == 1 THEN scorecn2 == 10. 2) IF cT is 1 or 2, then this is most likely 0 (according to our medical expert). IF scorecn2 == NA AND scorect2 == 1 or scorect2 == 2, THEN scorecn2 == 0. 3) Fill in with *pathology* stadpn2. IF scorecn2 == NA AND stadpn2 == 0, THEN scorecn2 == 0. IF scorecn2 == NA AND stadpn2 == 1 or 2 or 3, THEN scorecn2 == 1. IF scorecn2 == NA AND stadpn2 == 4 or 5, THEN scorecn2 == 2. IF scorecn2 == NA AND stadpn2 == 6, THEN scorecn2 == 9. IF scorecn2 == NA AND stadpn2 == 10, THEN scorecn2 == 10.
scorecm	1) If cN == 0, then this is also 0 (according to our medical expert). 2) With *pathology* stadpm. IFscorecm == NA ANDstadpm== 0, THENscorecm == 0. IFscorecm == NA ANDstadpm== 1, THENscorecm == 1. IFscorecm == NA ANDstadpm== 5, THENscorecm == 9.
prelocmri	IFemimri == NA, we assume no MRI scan was made.
prelocmri2	0) IF scopnumb == 1, THEN prelocmri2 == 0. 1) IF emimri2 == NA, we assume no MRI scan was made.
scopdistmm	1) IF mri is made (prelocmri==1), THEN scopdistmm == 999 (Unknown). 2) IF mri is not made (prelocmri!=1), THEN scopdistmm == 888 (irrelevant).
scopdist2mm	IF mri is made (prelocmri2==1), THEN scopdistmm == 999 (Unknown). 2) IF mri is not made (prelocmri2!=1), THEN scopdistmm == 888 (irrelevant).
Surgery	
datbez1	1) If datbez1 == NA AND datpa1 == datok, THEN datbez1 == datok. 2) If datbez1 == NA AND datpa1 == NA and typok == 3 or 4, THEN datbez1 == datok. 3) If datbez1 == NA AND datpa1 == NA and typok == NA, THEN datbez1 == datok. 4) If datbez1 == NA AND datpa1 == NA, THEN datbez1 == datok. 5) If datbez1 == NA AND datpa1 == NA, AND typok == 9, THEN datbez1 == datok. 6) If datbez1 == NA AND datpa1 == not NA, THEN datbez1 == datpa1
datprechirmetaeen	1) IF datprechirmetaeen == NA AND datok == 08-08-1808 or 09-09-1809 or 11-11-1811 THEN datprechirmetaeen == 09-09-1809 (Unknwon). 2) Take the middle of the two dates of datok and datbez1 if they are in it, so then the year is correct. IF datprechirmetaeen == NA AND datok != NA AND datbez1 != NA THEN datprechirmetaeen == datbez1 + (datok-datbez1)/2.
datprechirmetatwee	1) IF datprechirmetatwee == NA AND datok == 08-08-1808 or 09-09-1809 or 11-11-1811 THEN datprechirmetatwee == 10-10-1810 (Unknown). 2) Take the middle of the two dates of datok and datbez1 if they are in it, so then the year is correct. IF datprechirmetatwee == NA AND datok != NA AND datbez1 != NA THEN datprechirmetatwee == datbez1 + (datok-datbez1)/2.

prechirappendix	1) IF prechirappendix == NA AND datok < 2019 THEN prechirappendix == 10 (Not mandatory before 2019, so not filled in). 2) IF prechirappendix == NA AND datappendix == 08-08-1808 THEN prechirappendix == 0. 3) IF prechirappendix == NA AND datappendix == 09-09-1809 THEN prechirappendix == 1.
chirenkel	Fill in unknown.
conversien	IF the conversien is NA, we assume there is no conversion (according to our medical expert).
datok2	1) If all other variables that relate to the second surgery are NA or 'No', we assume datok is also 0. IF datok2 == NA AND (procok2 == NA or 888) AND (procok3 == NA or 888) AND (ingreep2 == NA or 888) AND (conversien2 == NA or 888) AND (aprtyp2 == NA or 888), THEN datok2 == 10-10-1810 (No second surgery took place).
procok2	1) If no 2nd surgery, then procok2 == 888. IF procok2 == NA AND (datok2 == 10-10-1810), THEN procok2 == 888 (irrelevant). 2) If all other variables that relate to the second surgery are NA or 'No', we assume this is also 0. IF procok2 == NA AND (datok2 == 9-9-1809) AND (procok3 == NA or 888) AND (ingreep2 == NA or 888) AND (conversien2 == NA or 888) AND (aprtyp2 == NA or 888), THEN procok == 888 (irrelevant).
ingreep2	1) If no 2nd surgery, then ingreep2 == 888. IF ingreep2 == NA AND (datok2 == 10-10-1810), THEN ingreep2 == 888 (irrelevant). 2) If all other variables that relate to the second surgery are NA or 'No', we assume this is also 0. IF ingreep2 == NA AND (datok2 == 9-9-1809) AND (procok3 == NA or 888) AND (procok2 == 888 or 999) AND (conversien2 == NA or 888) AND (aprtyp2 == NA or 888), THEN procok == 888 (irrelevant)
conversien2	1) If no 2nd surgery, then conversien2 == 888. IF conversien2 == NA AND (datok2 == 10-10-1810), THEN conversien2 == 888 (irrelevant). 2) If all other variables that relate to the second surgery are NA or 'No', we assume this is also 0. IF conversien2 == NA AND (datok2 == 9-9-1809) AND (procok3 == NA or 888) AND (procok2 == 888 or 999) AND (ingreep2 == 9 or 888) AND (aprtyp2 == NA or 888), THEN conversien2 == 888 (irrelevant).
aprtyp2	1) If no 2nd surgery, then aprtyp2 == 888. IF aprtyp2 == NA AND (datok2 == 10-10-1810), THEN aprtyp2 == 888 (irrelevant). 2) If all other variables that relate to the second surgery are NA or 'No', we assume this is also 0. IF aprtyp2 == NA AND (datok2 == 9-9-1809) AND (procok3 == NA or 888) AND (procok2 == 888 or 999) AND (ingreep2 == 9 or 888) AND (conversien2 == 9 or 888), THEN aprtyp2 == 888 (irrelevant).
perrtxchemo	Fill in no.
Radiotherapy	
radtiming	1) If datrad1, datprertxstop, rtxrestad, rtxct, rtxcn are all 0, then this is also 0. IF radtiming == NA AND datrad1 == NA AND datprertxstop == NA or 08-08-1808 AND rtxrestad != 1 AND rtxcn == NA or 888 AND rtxct == NA or 888, THEN radtiming == 0.
Systemic therapy	
prectx	Fill in no.
Other	
type_ziekenhuis	Fill in unknown.

Table 5.7: The variables that are imputed with clinical reasoning.

5.7 Select and construct relevant preoperative variables

We now have imputed all of the preoperative variables that we have selected in Section 5.5 that we want to impute (group I). In this step, we remove the variables that were selected by one of the ABCDE criteria from the dataset (group R). Hence, we are left with a dataset

with the variables of group I for the preoperative variables that have been imputed with clinical imputation when applicable in Section 5.6.

We are also going to add two preoperative variables that we construct ourselves from the existing data.

The first variable we construct is `leeftijd` (age). We construct this with the variable `gebdat` and `datok` so we know the age of the patient at the time of the surgery. Note that `gebdat` only shows the year the patient is born in, so the age is not that exact. We therefore round the variable `leeftijd` to years. After construction, we add this variable to our dataset.

The second variable we construct is `waitingTime` (waiting time). We construct this with the variable `datpa1` and `datok` so we know the number of days between the diagnosis and the surgery. After construction, we add this variable to our dataset.

5.8 Select and construct relevant outcome variables

In this step we select the relevant outcome variables. We have decided to go with all of the possible relevant complications that can occur, and also with the constructed variable severe complication. This variable captures a lot of other information, such as whether the patient died, whether they were readmitted and how long their stay was in the hospital and at the ICU. By selecting these 17 variables we capture a lot of the possible outcomes and the dataset can be used for a wide variety of future research. The variables that we have selected as outcome variables are `chirnaad`, `chirnaadoccult`, `chirabces`, `chirbloeding`, `chirileus`, `chirgastro`, `chirfascie`, `chirdarmperf`, `chirlekkage`, `chirwondinf`, `comppul`, `compcar`, `compthrom`, `compinf`, `compneu`, `compx`, and `severeCompl`.

The variable `severeCompl` stands for 'severe complications' and is a variable that I have constructed myself with the definition from [29]. According to [29] the definition of severe complications is 'a complication leading to ICU admission (more than 2 days), a reintervention (surgical or radiological), prolonged hospital stay (more than 14 days) or postoperative mortality'. On variable level this means the following. We have `severeCompl` $= 1$ if one or more of the following is met:

- `icdg` ≥ 2 and not `icdg` $= 888$ and not `icdg` $= 999$,
- `reintreq` $= 1$,
- `lengthOfStay`[3] ≥ 14,
- `status90` $= 1$.

Hence, I have constructed this variable such that if one or more of these requirements is met, then `severeCompl` $= 1$ and if none of these requirements is met, then `severeCompl` $= 0$. Note that the other selected outcome variables already had imputations for the datadictionary (Section 5.2) and imputations due to data collection (Section 5.3).

5.9 Select observations with no remaining missing variables

In this section we decide whether we are also going to include the observations with one or multiple variables that are filled in with 'unknown' or are unknown after all of the previous steps are taken.

[3] Length of (hospital) stay. This variable is constructed using variables `datok` and `datont`, so the days between surgery and discharge. We keep this variable in our dataset.

Before we do this, we also exclude the observations that have over 10 of the 120 variables[4] as missing variables, as these observations simply have too many unknowns. There are 1045 observations with more than 10 missing variables. This is visible in Table 5.8. After removing these observations, we are left with 53734 observations with 10 or less missing variables.

Number of missing variables	Number of observations that have exactly this amount of missing variables
0	47242
1	4315
2	805
3	362
4	85
5	605
6	105
7	95
8	31
9	20
10	69
11	8
12	3
15	5
16	605
17	297
18	55
19	29
20	19
21	13
22	4
23	2
24	2
26	1
29	1
38	1

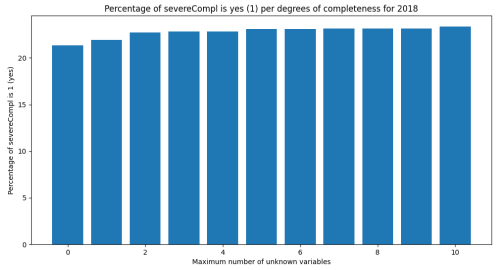
Table 5.8: The number of observations per number of missing variables for both the preoperative (103) and the outcome variables (17), hence for 120 variables in total.

With the remaining 53734 observations, we are doing an analysis to see whether a shift in severe complications occurred. To do so, we make an analysis of all of the observations and how many variables there are missing for this. Next, we make an analysis on what is the distribution in severe complications. If the distribution in severe complications stays quite similar[5] we say that distribution in severe complications does not significantly change and hence we can exclude the other observations that have multiple unknown variables.

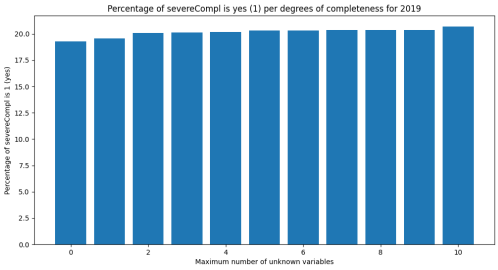
To perform the above described analysis, we make plots for the proportions of severe complications, per year and in total, per degree of completeness (the maximum number of missing variables). This is plotted in Figure 5.2. Plots 5.2a to 5.2g are per year, and plot 5.2h is over all the years. In plot 5.2h also the percentage of severe complications that are 1 are visible above the bars. This shows that when there is a maximum number of 0 variables missing, the proportion of severe complications that is 1 (yes) is 20,53%. When there is a maximum number of 1 variable

[4] We do not include the constructed variables `leeftijd`, `waitingTime`, and `lengthOfStay` here.
[5] We did it visually by looking at the empirical distribution of severe complications.

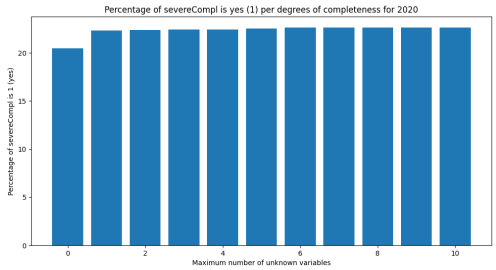
missing, so there is 1 or 0 variables missing, the proportion of severe complications that is 1 (yes) is 21,23%. An overview of this with all years combined is given in Table 5.9. The maximum difference is between 21,74% and 20,53%, so a difference of 1,21%. We accept this difference and continue with the dataset with 0 variables missing.



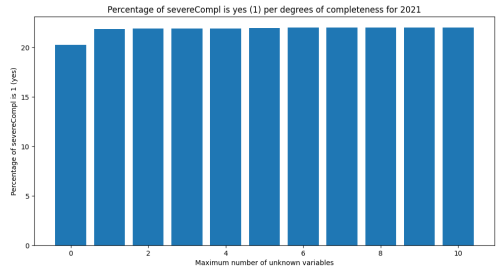
(a) 2018



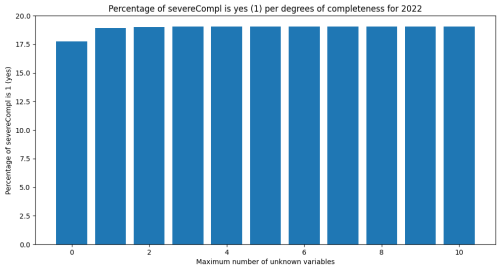
(b) 2019



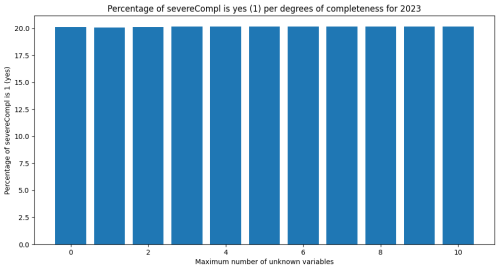
(c) 2020



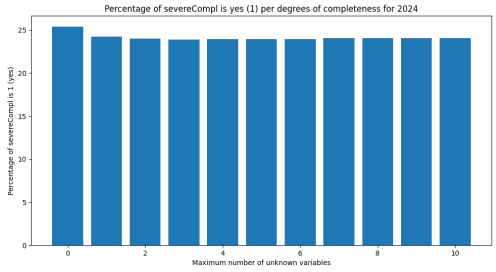
(d) 2021



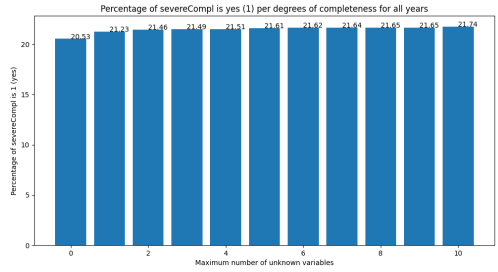
(e) 2022



(f) 2023



(g) 2024



(h) 2018-2024

Figure 5.2: Plots of the percentage of severe complications that is 'yes' per degree of completeness. Plots 5.2a to 5.2g are per year, and plot 5.2h is over all the years.

Maximum number of missing variables	Percentage of `severeCompl` is 'yes' (1)
0	20,53%
1	21,23%
2	21,46%
3	21,49%
4	21,51%
5	21,61%
6	21,62%
7	21,64%
8	21,65%
9	21,65%
10	21,74%

Table 5.9: Percentage of `severeCompl` is 'yes' (1) per degrees of completeness (maximum number of missing variables) for all years (2018-2024), belonging to Figure 5.2h.

Hence, in the end, we have a complete dataset with no unknown values of 123 variables (103 preoperative variables, 17 outcome variables and 3 constructed variables) and 47242 patients who underwent elective surgery from the years 2018 to 2024. This is our final dataset.

5.10 Exploratory final data analysis

After completing all the steps, we now have the final dataset, existing of 123 variables (group Y) and 47242 observations. Now that we have reached the final dataset we are going to conduct a univariate data analysis on this dataset to see the distributions in the variables over the years. We consider the variables of group Y and remove the date variables. Then we split this data into two categories, discrete variables and continuous variables. In Section 5.10.1 we handle the discrete variables, and in Section 5.10.2 we handle the continuous variables.

The goal of this univariate data analysis is to get more insight at the univariate level into whether there might be shifts occurring in the dataset or whether we can say that the variables over the years are all 'similar enough'. Because if there are shifts occurring, then you cannot just compare all the data from the different years together, since this may produce conclusions that reflect changes in data collection rather than changes in patient care.

5.10.1 Discrete variables

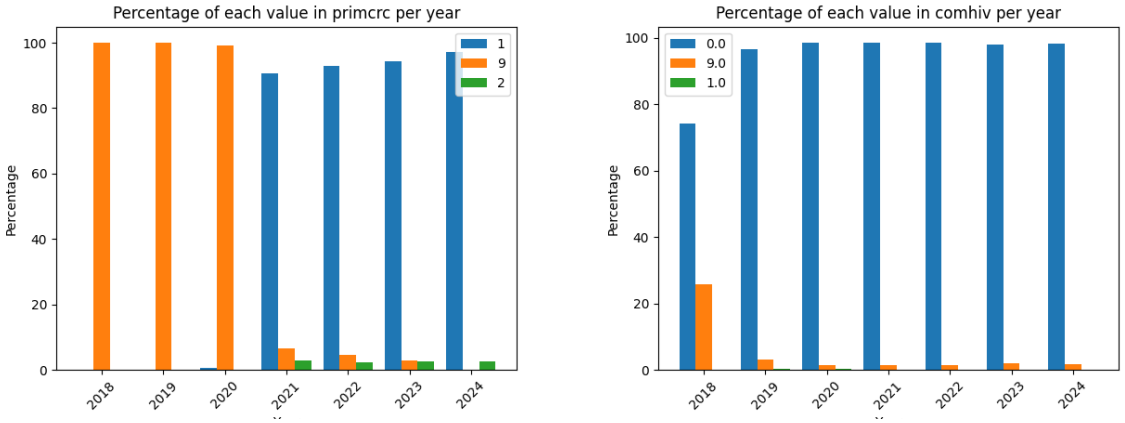
The 108 discrete variables are `geslacht`, `commyo`, `comhartfaal`, `comperif`, `comcva`, `comdem`, `comlong`, `combind`, `comgiulcus`, `comlever`, `comdiam`, `comparalys`, `comnier`, `commalig`, `comhiv`, `asascore`, `resdarm`, `stomavg`, `primcrc`, `scopnumb`, `diagn`, `locprim`, `histd`, `complpre`, `preanemie`, `preobstruct`, `preabces`, `scorect`, `scorecn`, `locprim2`, `scorect2`, `scorecn2`, `scorecm`, `prelocmri`, `prelocmri2`, `verwezen`, `darmperf`, `ileusblow`, `perfperi`, `perfscopie`, `prechirstomanew`, `prechirstent`, `prechirmetanew`, `prechirappendix`, `prechircoloscex`, `chirenkel`, `typok`, `procok`, `ingreep`, `conversien`, `aprtyp`, `typprocok`, `procok2`, `ingreep2`, `conversien2`, `aprtyp2`, `resadd`, `locaddmaag`, `locaddmilt`, `locaddduod`, `locadddoverig`, `locaddoment`, `locaddbuikw`, `locaddvag`, `locadduter`, `locaddovarl`, `locaddovarr`, `locadduretl`, `locadduretr`, `locaddblaas`, `locaddprost`, `locaddvesl`, `locaddvesr`, `locaddvesb`, `locadddund`, `locaddsacr`, `locaddoverig`, `resadm`, `resomm`, `reslem`, `respem`, `reslmm`, `resbum`, `perrtxchemo`, `anastom1`, `stoma`, `stomahef`, `radtiming`, `prectx`, `charlgrp`, `type_ziekenhuis`, `chirnaad`, `chirnaadoccult`, `chirabces`, `chirbloeding`, `chirileus`, `chirgastro`, `chirfascie`, `chirdarmperf`, `chirlekkage`, `chirwondinf`, `comppul`, `compcar`, `compthrom`, `compinf`, `compneu`, `compx`, and `severeCompl`. Of these variables, we have 8 variables that are ordinal

(`asascore`, `scopnumb`, `scorect`, `scorecn`, `scorect2`, `scorecn2`, `scorecm`, `charlgrp`), 1 variable that is binary (`severeCompl`), and 99 variables that are nominal.

The graphs of the distribution of the values per year of these variables are in Appendix F in Section F.1.1. For the sake of clarity, we pick two of the graphs as examples to explain how to read all of the graphs.

As an example, we take the variable `primcrc`. This variable has the description 'is it a first or second primary tumour?'. This description can be found in Chapter 4. The options for the value of this variable are 1 (First primary tumour), 2 (Second primary tumour), and 9 (Unknown). These options can be found in Appendix B. In Figure 5.3a the distribution of the values per year for `primcrc` is shown. It is visible that in the years up to 2020, the percentage of values that is 9 (Unknown) is near 100%. In the year 2021 a change in the distribution of the values occurs. In 2021 we have around 90% of the values equal to 1 (First primary tumour), around 7% of the values are 9 (Unknown), and around 3% of the values are 2 (Second primary tumour). In the years that follow we see that the percentage of values that are 2 (Second primary tumour) stays quite constant, the values that are 9 (Unknown) decrease, and the values that are 1 (First primary tumour) increase.

Another example is the variable `comhiv`. This variable describes whether a patient has HIV/AIDS. Its description can be found in Chapter 4. The options for the value of this variable are 0 (No), 1 (Yes), and 9 (Unknown). These options can be found in Appendix B. In Figure 5.3b the distribution of the values per year for `comhiv` is shown. It is visible that in 2018, the percentage of values that is 9 (Unknown) is around 25-30% and the percentage of values that is 0 (No) is around 70-75%. In the year 2019, we have around 95-100% of the values equal to 0 (No), and around 0-5% of the values are 9 (Unknown). In the years that follow we see that the distribution in the percentage of values stays quite constant, with a slight decrease for the percentage of values tat are 9, relative to 2019.



(a) `primcrc` **(b)** `comhiv`

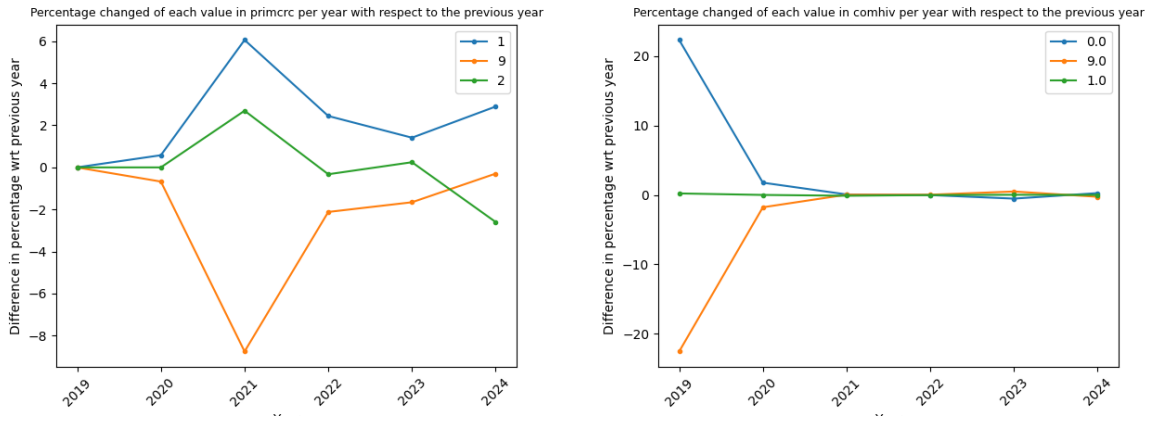
Figure 5.3: The distribution of the values per year for the variables `primcrc` and `comhiv`.

In addition, we also generated the plots with the percentage change in values compared to the previous year. The plots show per year how much (in percentage) a value changed compared to the previous year. These graphs can be found in Appendix F, in section F.1.2 for all of the variables. For clarity, we continue with the examples of the variables `primcrc` and `comhiv` to make clear how to read the graphs.

In Figure 5.4a the graph is shown for the variable `primcrc` for the percentual change of values compared to the previous year. As was also visible in Figure 5.3a, up to the year 2020, no big

changes in distribution occur. In 2021 we see a decrease for value 9 (Unknown), and an increase for value 1 (First primary tumour). Only now, also value 2 (Second primary tumour) shows an increase, and the increase of value 1 (First primary tumour) is smaller than the decrease of value 9 (Unknown). In 2022 there is again a decrease for value 9 (Unknown), and an increase for value 1 (First primary tumour), only smaller than before. And value 2 (Second primary tumour) does not show much change. In 2023 there is again a decrease for value 9 (Unknown), and an increase for value 1 (First primary tumour), only bit smaller than before. And value 2 (Second primary tumour) also shows a slight increase again. In 2024 there is a decrease for value 9 (Unknown), and a bigger increase for value 1 (First primary tumour). Value 2 now shows a decrease for the first time.

For the variable `comhiv`, Figure 5.4b shows the graph for the percentual change of values compared to the previous year. As was also visible in Figure 5.3b, there is a big shift from 2018 to 2019 in the distribution of the values 0 (No) and 9 (Unknown). From 2019 to 2020 we see again a slight shift in the distribution of these values, but far less than the year before. In the years that follow, there are relatively small changes in the difference of the percentages of the values.



(a) `primcrc` **(b)** `comhiv`

Figure 5.4: The percentage change of the values compared to the previous year per year for the variables `primcrc` and `comhiv`.

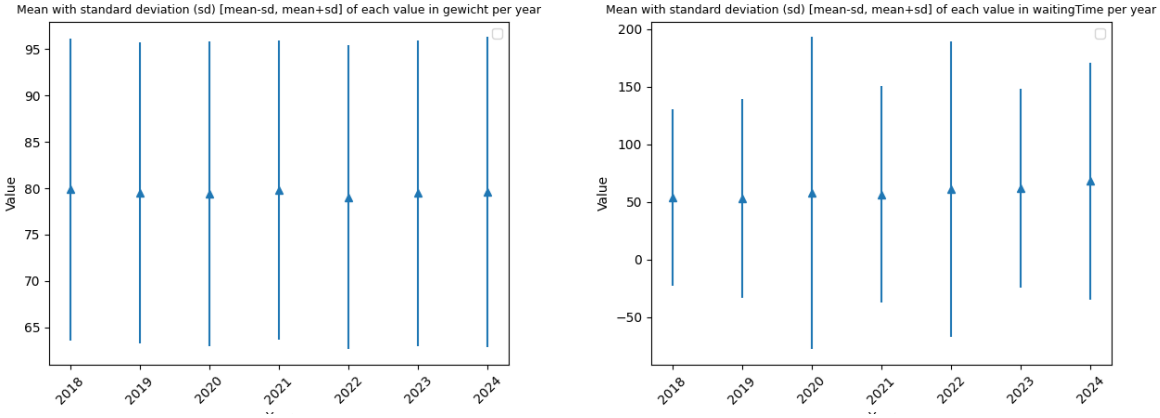
5.10.2 Continuous variables

The 7 continuous variables are `lengte`, `gewicht`, `scopdistmm`, `scopdist2mm`, `leeftijd`, `waitingTime`, `lengthOfStay`. In Appendix F in Section F.2 the plots are shown for these variables. These plots show the empirical mean with the empirical standard deviation (sd) [mean-sd, mean+sd] of each value in the concerning variable per year. For the sake of clarity, we pick two of the graphs as examples to explain how to read all of the graphs.

As an example, we look at the plot for the variable `gewicht` (weight) which is in Figure 5.5a. The variable `gewicht` is the weight of the patient in kilograms. The plot shows us the empirical mean (marked with an upward pointing arrow) and the empirical standard deviation (sd) (shown with the vertical lines above and below the mean). So on the y-axis, the values for [mean-sd, mean+sd] are marked in blue. In Figure 5.5a it is visible that the mean is not consistent over the years but that slight changes occur from year to year. The same holds for the standard deviation. We see for example from 2022 to 2023 that the mean slightly increases, and the standard deviation also increases. The mean in 2022 is around 79,02 kilograms and in 2023 around 79,47 kilograms. Also the standard deviation is in 2022 around 16,39 kilograms and in 2023 around 16,51 kilograms.

Another example is the variable `waitingTime` which is in Figure 5.5b. The variable `waitingTime`

is the time between the diagnosis and the surgery in days. The plot shows us the empirical mean (marked with an upward pointing arrow) and the empirical standard deviation (sd) (shown with the vertical lines above and below the mean). So on the y-axis, the values for [mean-sd, mean+sd] are marked in blue. In Figure 5.5b it is visible that the mean is not consistent over the years but that slight changes occur from year to year. The same holds for the standard deviation. We see for example from 2021 to 2022 that the mean slightly increases, and the standard deviation also increases.



(a) `gewicht` **(b)** `waitingTime`

Figure 5.5: The mean with standard deviation (sd) [mean-sd, mean+sd] for the variables `gewicht` and `waitingTime` per year.

In the next chapter, Chapter 6, we will investigate these changes over the years from a statistical point of view with effect size analyses.

6 | Temporal Stability Analysis

Data of the Dutch ColoRectal Audit (DCRA) dataset has been collected for over 15 years. We focus on the last 7 years from 2018 to 2024. Before we can draw conclusions from the collected data, we need to know whether we can use the data of all these years without introducing practically relevant temporal heterogeneity.

In these 15 years, there have been changes in data collection (see Section 5.4), e.g., the added variables in 2018. Also the surroundings and circumstances have gone through changes over the years, e.g., the COVID period. Another possibility is that protocols and processes changed, which could lead to shifts in distributions, e.g., the start of the population screening programme in 2014. Lastly, it is also possible that over the years changes occurred in the distribution of the data itself. When the distribution of the data is changing, we have to take that into account and decide what to do with this information. We might for example remove data for which the distribution changes, because this might add bias or noise.

It is necessary to account for all these possibilities in our decision of which data we are keeping to do further research and hence possibly draw conclusions with. The question of whether we can use the data of all the years is thus closely related to to the question of whether the data of all the years has a similar enough distribution. To gain insight into the data and its marginal and pairwise stability, we conduct an effect size analysis in Section 6.1. With the conclusions from the effect size analysis, we question in Section 6.2 whether the data of all the individual and consecutive years has a similar enough distribution by constructing a two-sample problem and evaluating this with a random forest classifier, which takes conditional relations in the data into account.

6.1 Effect size analysis

For deciding whether we have a statistically significant difference between two univariate distributions, it is common practice to use p-values. However, there are some disadvantages to using these, especially when working with large sample sizes. With the sample sizes that we are working with (around 6800 observations per year), the p-values resulting from some hypotheses tests, such as the equality of means of two samples (years), are likely to be small in many cases.

The reason for this occurrence is as follows. In the hypothesis testing framework, the null hypothesis usually specifies an exact value, for example $H_0 : \theta = \theta_0$. With a large number of observations, even very small deviations from θ_0 can be detected statistically. As the sample size increases, the standard error decreases, so the test becomes increasingly sensitive to small differences between the estimate and the null value. With a large number of trials, such extreme precision would make the testing procedure overly sensitive to trivial differences, making them appear like significant even when they are not. More details on this topic are given in [48].

To decide whether we have a statistically significant difference between two distributions, we will therefore use another approach: effect size analysis. Marginal temporal stability is assessed by

comparing the distributions of individual variables across years. Effect size measures are used to quantify the magnitude of the observed differences. Small effect sizes are interpreted as evidence that the corresponding marginal distributions are sufficiently similar for the intended analysis, subject to the limitations of the selected measures and thresholds.

Effect sizes 'facilitate cumulative science, can be used to determine the sample size for follow-up studies, and can be used to examine effects across studies'. [49] With the effect size you can find out whether a difference in two samples has an effect greater than zero, and if so, how big this effect is. This difference in two samples can for example be an intervention or an experimental manipulation, but it can also be a difference in the circumstances as is the case in what we want to investigate. Namely, whether the different years have an effect and if so, how big that effect is.

To calculate the effect size, we need to split the variables into discrete (categorical) and continuous variables. For the continuous variables we use Cohen's d_s. This is described in Section 6.1.1. For the categorical variables we use Cohen's h. This is described in Section 6.1.2. Furthermore, in Section 6.1.3 a pairwise effect size analysis is conducted to investigate pairwise stability for five selected variables.

6.1.1 Cohen's d_s for continuous variables

For the continuous variables, we are using Cohen's d_s to determine the effect sizes. By calculating the effect size, we can (for continuous variables) combine the distance between the *sample means* and the *sample standard deviations* into a single metric.

Cohen's d_s is defined as in equation (6.2) [49]. In this equation $\bar{X}_i$ is the sample mean of sample i, n_i is the sample size of sample i, and SD_i is the sample standard deviation of sample i, with $i = 1, 2$. For small samples ($n < 20$) Cohen's d_s gives a biased estimate of the population effect size [50]. With Hedges' d_s this bias can be corrected [51]. However, Hedges' g_s is not relevant to us seen our sample size.

The standard error of Cohen's d_s is defined as in equation (6.1).

$$SE_{d_s} = \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{d_s^2}{2(n_1 + n_2)}} \tag{6.1}$$

$$d_s = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)SD_1^2 + (n_2-1)SD_2^2}{n_1+n_2-2}}}. \tag{6.2}$$

The 95% confidence interval of Cohen's d_s is defined as in equation (6.3).

$$CI_{d_s} = d_s \pm 1,96 \cdot SE_{d_s}. \tag{6.3}$$

To interpret the outcome we apply the following guidelines. The effect size is small if it is around $|d_s| = 0,2$, the effect size is medium if it is around $|d_s| = 0,5$, and the effect size is large if it is around $|d_s| = 0,8$ [52]. For values of $|d_s|$ between $0,2$ and $0,5$ we call this a small to moderate effect size. For values of $|d_s|$ between $0,5$ and $0,8$ we call this a moderate to large effect size. If we have a negative effect size this means that the empirical mean of sample 2 is bigger than the empirical mean of sample 1.

Example of use

To illustrate how this is applicable to our dataset, we will show an example.

Suppose we are looking at continuous variable `lengte` (height). We want to investigate whether there are changes occurring in its distribution between the consecutive years. To do so, we look what the effect size is from year 1 to year 2. We take 2018 as year 1 and 2019 as year 2. The years do not have to be consecutive but in our analysis they are. This can be investigated further, but that is outside the scope of this thesis.

The empirical mean of `lengte` in 2018 is $\bar{X}_1 = 172,65$. The empirical standard deviation of `lengte` in 2018 is $SD_1 = 9,67$.

The empirical mean of `lengte` in 2019 is $\bar{X}_2 = 172,65$. The empirical standard deviation of `lengte` in 2019 is $SD_2 = 10,00$.

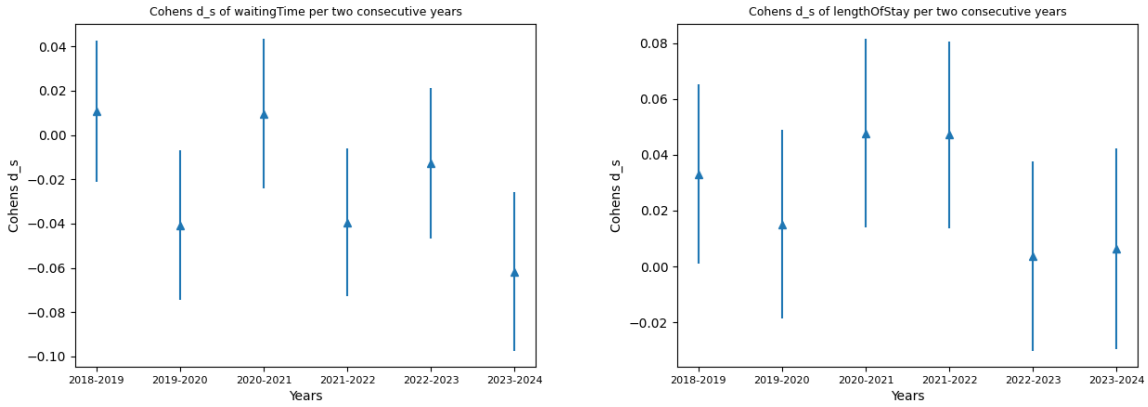
From here onwards the calculations are straightforward using equation (6.2). We do this for all the years in our final dataset (2018 to 2024).

Results

We have calculated the effect size for the continuous variables `leeftijd` (age), `gewicht` (weight), `lengte` (height), `scopdistmm` (Distance in mm from the lower edge of the first tumour to the anorectal junction on sagittal MRI (or CT, if applicable)), `scopdist2mm` (Distance in mm from the lower edge of the second tumour to the anorectal junction on sagittal MRI (or CT, if applicable)), `waitingTime`, and `lengthOfStay`.

For none of these variables there was an absolute effect size that was bigger than 0,2. Hence, according to Cohen's conventional thresholds, all of the effect sizes in all years for all variables were small.

Since the variables `waitingTime` and `lengthOfStay` are particularly interesting for future research, their plots for the effect sizes per year are visible in Figure 6.1. In each plot the value of Cohen's d_s (displayed with the triangle) with the corresponding confidence interval (displayed with the vertical lines) is on the y-axis, and the years are on the x-axis.



(a) `waitingTime` **(b)** `lengthOfStay`

Figure 6.1: Plots of Cohen's d_s for the continuous variables `waitingTime` and `lengthOfStay`. In each plot the value of Cohen's d_s (displayed with the triangle) with the corresponding confidence interval (displayed with the vertical lines) is on the y-axis, and the years are on the x-axis.

6.1.2 Cohen's h for categorical variables

For the categorical variables, we are using Cohen's h to determine the effect sizes. Cohen's h is the difference between the arcsine transformations of two proportions.

Cohen's h is defined as in equation (6.4) [52]. In this equation p_1 is the probability in sample 1 that the variable has a certain value, say value x. And p_2 is the probability in sample 2 that the variable has value x. This p_i can be calculated by taking the number of observations for which the value of the variable is x and dividing this over the sample size of sample $i = 1, 2$. The sample sizes are denoted by n_1, n_2 for samples 1 and 2, respectively.

$$h = 2 \cdot \arcsin \sqrt{p_1} - 2 \cdot \arcsin \sqrt{p_2} \tag{6.4}$$

The standard error of Cohen's h is defined as in equation (6.5).

$$SE_h = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{6.5}$$

The 95% confidence interval of Cohen's h is defined as in equation (6.6).

$$CI_h = h \pm 1,96 \cdot SE_h \tag{6.6}$$

To interpret the outcome we apply the same guidelines as for Cohen's d_s. If we have a negative effect size this means that, since $0 \leq p_i \leq 1$ for $i = 1, 2$, $p_1 < p_2$. We can also apply the same guidelines from Cohen's d_s in the negative setting.

The difficulty in this, is that we are dealing with categorical variables, which can be binary but they do not have to. To deal with this categorical data we use 'one-hot encoding' or 'dummy variables'. With this encoding, categorical variables are transformed to numeric variables. An example of this is the categorical variable `geslacht` (gender). This variable has four options, namely 1 (Male), 2 (Female), 7 (Undefined), and 9 (Unknown). The label encoding for this variable is in Table 6.1. The corresponding one-hot encoding for this variable is in Table 6.2.

Name	Category	Number of occurrences
Male	1	4371
Female	2	3399
Undefined	7	0
Unknown	9	0

Table 6.1: Label encoding of categorical variable `geslacht` for 2018.

Male	Female	Undefined	Unknown	Number of occurrences
1	0	0	0	4371
0	1	0	0	3399
0	0	1	0	0
0	0	0	1	0

Table 6.2: One-hot encoding of categorical variable `geslacht` for 2018.

From Table 6.2 it is clear to see that we can derive variables X_1, X_2, X_3, X_4. Then X_1 is defined as follows, and X_2, X_3, X_4 are defined similarly:

$$X_1 = \begin{cases} 1 & \text{if } \texttt{geslacht} = 1 \\ 0 & \text{if } \texttt{geslacht} = 2 \text{ or } \texttt{geslacht} = 7 \text{ or } \texttt{geslacht} = 9. \end{cases}$$

Note that we have kept the observations for which the value of the variable is "unknown" in the effect size analyses as well. The reason for this is to create a complete overview of the data and how it behaves over the years. One should be careful with the "unknown" values though, seen we do not want that they are given meaning, in the sense that an unknown value becomes a predictor. For some variables however, it might be possible that the 'unknown' values actually do capture a meaning. An example of this is with the comorbidities, see Section 5.6. Which other variables might also carry meaning in their 'unknown' values, can be investigated in further research. It is also possible in further research to make an overview of the unknown values and to again do an effect size analysis, but then without including the 'unknown' values that carry no meaning.

Example of use

To illustrate how this is applicable to our dataset, we will show an example continuing with categorical variable `geslacht` (gender).

We want to investigate whether there are changes occurring in the distribution of the values of the variable between the consecutive years. To do so, we look per possible value (1, 2, 7, or 9 for `geslacht`) what the effect size is from year 1 to year 2. So for example we look at the empirical distribution of males and then we decide on what the effect size is when comparing year 1 and year 2. We take 2018 as year 1 and 2019 as year 2. The years do not have to be consecutive but in our analysis they are. This can be investigated further, but that is outside the scope of this thesis.

The total number of patients in 2018 is $n_1 = 4371+3399+0+0 = 7770$. The number of patients in 2018 that are male (`geslacht`=1) is 4371. So $\hat{p_1} = \frac{4371}{7770} \approx 0,5625$.

The total number of patients in 2019 is $n_2 = 3912+3382+0+1 = 7295$. The number of patients in 2019 that are male (`geslacht`=1) is 3912. So $\hat{p_2} = \frac{3912}{7295} \approx 0,5363$.

From here onwards the calculations are straightforward using equations (6.4), (6.5), and (6.6). Note that n_1 and n_2 stay the same as long as year 1 and year 2 are the same. The $\hat{p_1}$ and $\hat{p_2}$ change when the value of the variable that we are interested in changes, so for example if we want to look at the female patients.

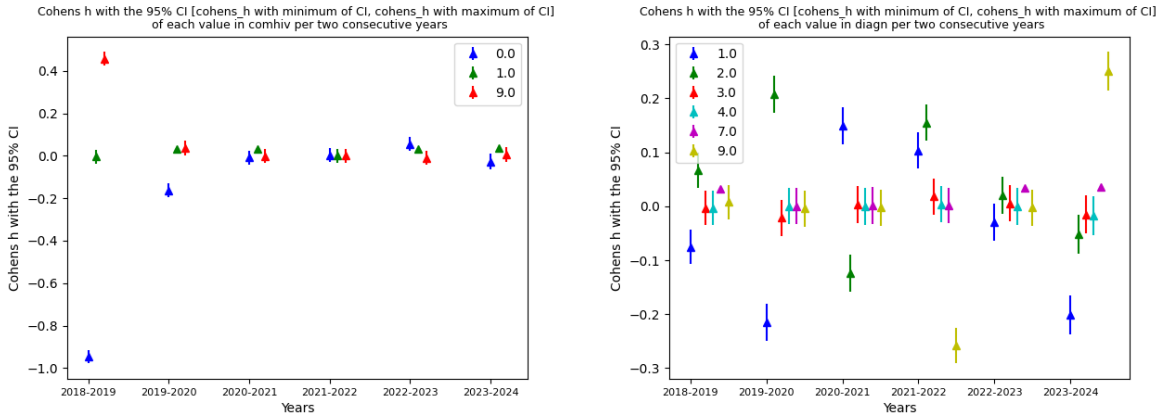
Results

We have calculated the effect size for the 108 categorical variables `geslacht`, `commyo`, `comhartfaal`, `comperif`, `comcva`, `comdem`, `comlong`, `combind`, `comgiulcus`, `comlever`, `comdiam`, `comparalys`, `comnier`, `commalig`, `comhiv`, `asascore`, `resdarm`, `stomavg`, `primcrc`, `scopnumb`, `diagn`, `locprim`, `histd`, `complpre`, `preanemie`, `preobstruct`, `preabces`, `scorect`, `scorecn`, `locprim2`, `scorect2`, `scorecn2`, `scorecm`, `prelocmri`, `prelocmri2`, `verwezen`, `darmperf`, `ileusblow`, `perfperi`, `perfscopie`, `prechirstomanew`, `prechirstent`, `prechirmetanew`, `prechirappendix`, `prechircoloscex`, `chirenkel`, `typok`, `procok`, `ingreep`, `conversien`, `aprtyp`, `typprocok`, `procok2`, `ingreep2`, `coversien2`, `resadd`, `locaddmaag`, `locaddmilt`, `locaddduod`, `locadddoverig`, `locaddoment`, `locaddbuikw`, `locaddvag`, `locadduter`, `locaddovarl`, `locaddovarr`, `locadduretl`, `locadduretr`, `locaddblaas`, `locaddprost`, `locaddvesl`, `locaddvesr`, `locaddvesb`, `locadddund`, `locaddsacr`, `locaddoverig`, `resadm`, `resomm`, `reslem`, `respem`, `reslmm`, `resbum`, `perrtxchemo`, `anastom1`, `stoma`, `stomahef`, `radtiming`, `prectx`, `charlgrp`, `type_ziekenhuis`, `chirnaad`, `chirnaadoccult`, `chirabces`, `chirbloeding`, `chirileus`, `chirgastro`, `chirfascie`, `chirdarmperf`, `chirlekkage`, `chirwondinf`, `comppul`, `compcar`, `compthrom`, `compinf`, `compneu`, `compx`, `severeCompl`.

For most of these variables there was no absolute effect size that was bigger than 0,2. We conclude that all of these variables have effect sizes that are small in all consecutive years. Note that this

also means that there is not a trend in general visible that the values differ significantly around 2020, the COVID year.

For the comorbidities we did see some interesting results with effect sizes bigger than 0,2. An example to illustrate the results we saw for multiple comorbidities, is `comhiv` (HIV/AIDS). The absolute effect size of the value 0 (No HIV/AIDS) was close to 1, so a large effect size, in 2018-2019. This can be seen in Figure 6.2a. In each plot the value of Cohen's h (displayed with the triangle) with the corresponding confidence interval (displayed with the vertical lines) is on the y-axis, and the years are on the x-axis. This indicates a large change, in terms of Cohen's h, in the proportion of patients records in this category between 2018 and 2019. A possible explanation for this could be in the data collection process, since in the registration a conversion of the registration of comorbidities took place from before 2018 and after 2018 [34]. However, this should then not have an effect from 2018 to 2019 if it is all registered as it should be. Other possible explanations for this observation can be explored and investigated in further research. Also the variable `diagn` (reason for diagnosis) shows some absolute effect sizes bigger than 0,2. This is visible in Figure 6.2b. From 2019 to 2020 the value 1 (symptoms) shows an absolute effect size of around 0,22. This means that the difference from 2019 to 2020 has a small (to moderate) effect on the empirical distribution of whether patients were diagnosed because they were having symptoms.



(a) `comhiv` **(b)** `diagn`

Figure 6.2: Plots of Cohen's h for the categorical variables `comhiv` and `diagn`. In each plot the value of Cohen's h (displayed with the triangle) with the corresponding confidence interval (displayed with the vertical lines) is on the y-axis, and the years are on the x-axis.

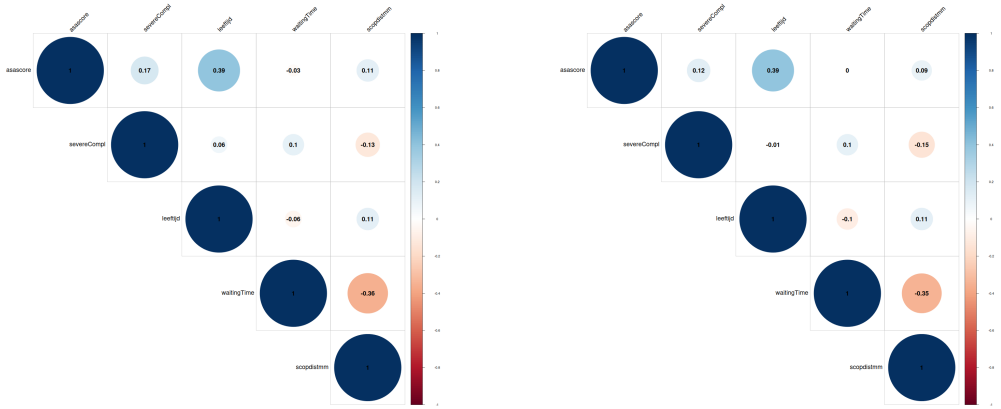
6.1.3 Pairwise analysis

Before we can retrieve meaningful results with the random forest classifier, we need to know whether we have marginal and pairwise stability of different years of different variables. We do already know per variable what the effect size is and whether it is small, medium or large. For small effect sizes we say that the marginal distribution of each variable is more similar across years. However, we do not yet know if we can also pool different variables. To find the answer to this, we will work with correlation matrices and their difference across consecutive years. For a continuous and a discrete variable, we use *polyserial correlation*. For two continuous variables, we use *Spearman's rho*. For two discrete variables, we use *polychoric correlation*. When the absolute value of this is around 0,1 it is classified as a small correlation, around 0,3 is a medium correlation and around 0,5 is a large correlation [52]. Note that this is a convention to use, not a theoretical cut-off.

To perform the pairwise analysis, we start with five variables to investigate the relations between these variables. This can be expanded in later research to a larger sample size and even include all of the variables in the dataset. The matrices used for the pairwise analysis in this section are also constructed for all 123 variables in Appendix H. This appendix shows the correlation matrices per year and corresponding difference matrices per two sequential years for all the 123 selected variables. These matrices can be used and investigated in future research. The correlation matrices in Section H.1 show that for some groups of variables, the correlation is very big. These are for example the comorbidities, the location of the tumour, and the complications. The difference matrices in Section H.2 show that it is not possible for all variables to pool them together, since their correlation measures can be classified as medium or even large. Further investigation into these matrices is outside the scope of this thesis.

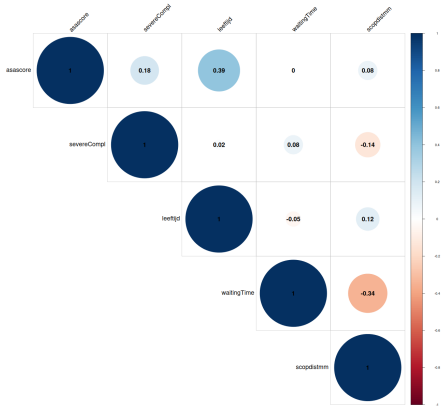
The variables that we selected to work with in this section are `asascore` (ASA score), `severeCompl` (severe complications), `leeftijd` (age), `waitingTime`, and `scopdistmm` (distance in mm from the lower edge of the first tumour to the anorectal junction on sagittal MRI (or CT, if applicable)). The variables `asascore` and `severeCompl` are discrete variables. Of these variables `asascore` is an ordinal variable, and `severeCompl` is a binary variable. The variables `leeftijd`, `waitingTime` and `scopdistmm` are continuous variables. We have constructed the correlation matrices per year. These are in Figure 6.3. Note that in these plots both positive and negative correlation is shown. When we want to classify the correlation, we consider the absolute values.

It is important to note that it is also possible to include nominal variables, but that one should be very careful when doing so. Because when looking at a nominal variable to estimate a correlation, you look at whether a higher value of the nominal variable is associated with for example more severe complications (value 1). However, for a nominal variable, the numbers represent categories and hence it is not always the case that the higher the number for the category, the more severe complications are met with it. Therefore, to solve this, you should either (i) use one-hot encoding for the nominal variable, or (ii) reorder the nominal variable by looking at the proportion of severe complications for each value of the nominal variable, and regive the values (1, 2, 3, etcetera) to each category.

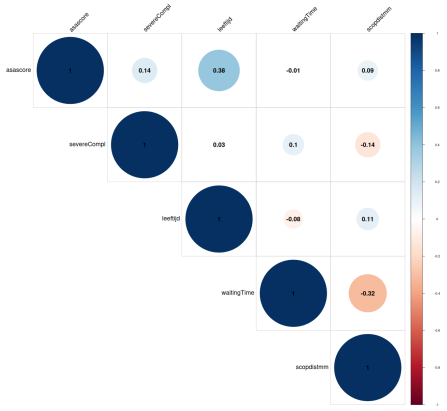


(a) 2018

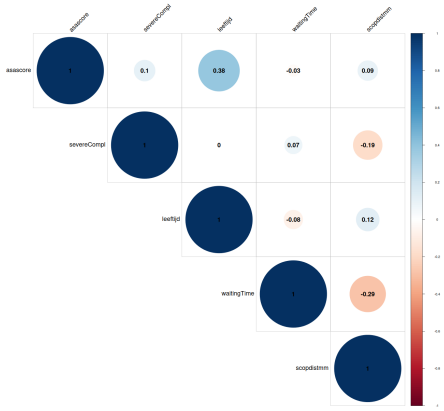
(b) 2019



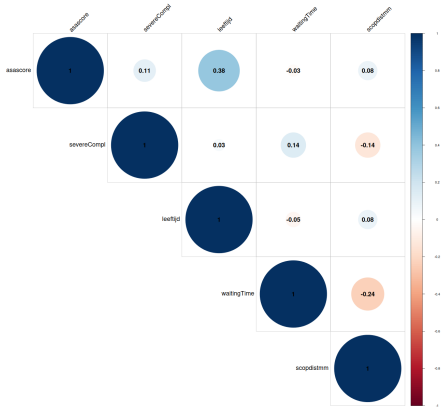
(c) 2020



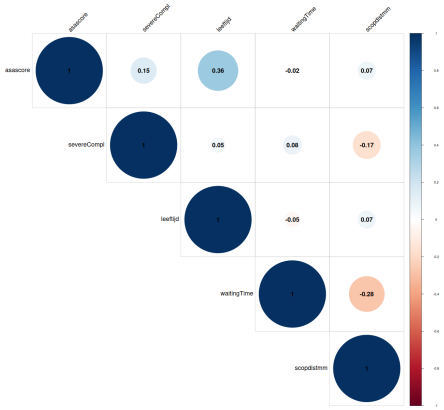
(d) 2021



(e) 2022



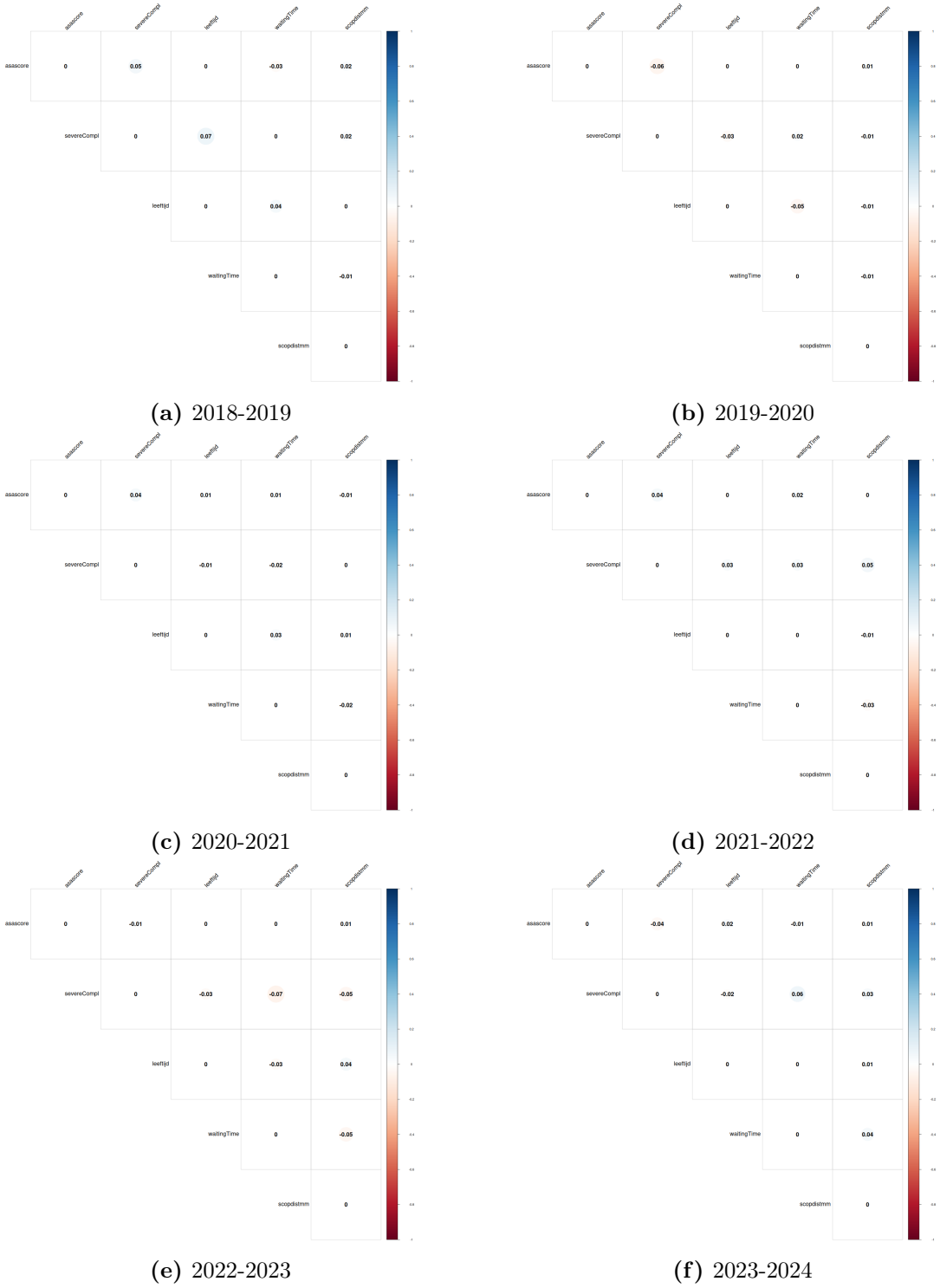
(f) 2023



(g) 2024

Figure 6.3: Correlation matrices for the years 2018 to 2024 for the variables `asascore`, `severeCompl`, `leeftijd`, `waitingTime`, and `scopdistmm`.

To find out whether there are shifts occurring over the years, we construct difference matrices from the correlation matrices. The difference matrices per two consecutive years are in Figure 6.4. A value in one of these matrices can be interpreted as the change in association of the variables over the years.



(a) 2018-2019 **(b)** 2019-2020

(c) 2020-2021 **(d)** 2021-2022

(e) 2022-2023 **(f)** 2023-2024

Figure 6.4: Difference matrices constructed from the correlation matrices for the years 2018 to 2024 for the variables `asascore`, `severeCompl`, `leeftijd`, `waitingTime`, and `scopdistmm`.

The conclusion that we can draw from looking at these difference matrices, is that there are

no changes in association for which the absolute value is bigger than 0,07. Hence, we conclude that for the selected variables we observed small differences in marginal summaries or pairwise associations for sequential years. It is interesting to note that there is not a trend in general visible that the values differ significantly around 2020, even though this is the COVID year. We do not compare for other than sequential years. That can be done in future research but is outside the scope of this thesis.

6.2 Random forest

Since we concluded that we have marginal and pairwise stability for the variables `asascore`, `severeCompl`, `leeftijd`, `waitingTime`, and `scopdistmm` for sequential years, we can continue with the two sample problem using the random forest classifier, which accounts for conditional relations in the data.

The two-sample problem is described in Section 2.2.4. For readability, we will write the notation for the two sample problem here as well.

Let $X_1, ..., X_{n_0}$ and $Y_1, ..., Y_{n_1}$ be a collection of $\mathbb{R}^d$-valued random vectors, such that $X_i \overset{iid}{\sim} P$ for $i = 1, 2, ..., n_0$ and $Y_i \overset{iid}{\sim} Q$ for $i = 1, 2, ..., n_1$, where P and Q are some Borel probability measure on $\mathbb{R}^d$. The goal is to test

$$H_0 : P = Q, \qquad H_A : P \neq Q.$$

Given these i.i.d. samples of vectors, we define labels $l_i = 1$ for each X_i and $l_i = 0$ for each Y_i to obtain the data (Z_j, l_j), $j = 1, ..., N$, for $N = n_0 + n_1$, and $Z_j = X_i$ or $Z_j = Y_i$. On this data we train a classifier $g : \mathbb{R}^d \to \{0, 1\}$. If g predicts the labels l better than expected under H_0, it is taken as evidence against H_0 [21].

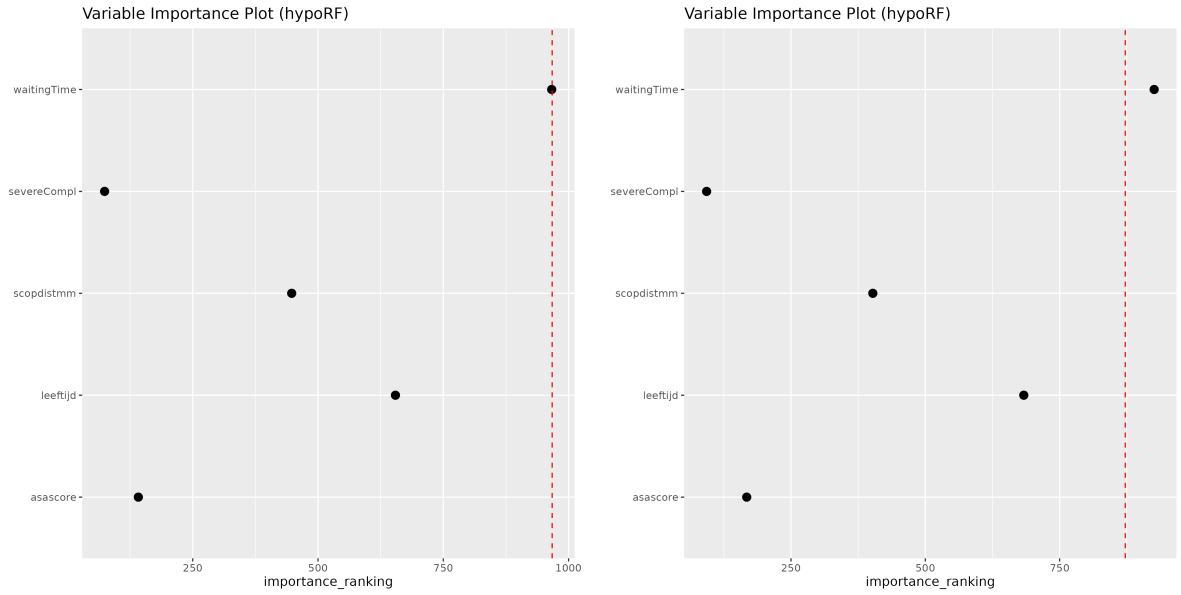
In our case with the DCRA dataset, g will be a (permuted) random forest classifier, see Chapter 2. There are d variables that are taken into account. We can split the data on a certain condition to investigate which variables could be possible predictors for the condition. The condition is that a variable should take a certain value and in doing so, it separates the data in two classes. A class that does satisfy the condition (condition a), and a class that does not satisfy the condition (condition b). The $\mathbb{R}^d$-valued random vectors $X_1, ..., X_{n_0}$ will be the observations of condition a. There are n_0 observations in the dataset of this condition. The $\mathbb{R}^d$-valued random vectors $Y_1, ..., Y_{n_1}$ will be the observations of condition b. There are n_1 observations in the dataset of this condition.

We split on two conditions, the first are the years, and the second are the severe complications. These are discussed in Section 6.2.1 and Section 6.2.2, respectively.

6.2.1 Years

We want to compare two sequential years. We do this for two pairs. The class label is the year. We look at the variable importance measure to decide which variables might have changed and whether it is consistent with the effect size analysis. If it is not consistent, then that could be because of conditional dependence of the variables. Note that the size of variable importance under the Gini index is between 0 and 1. However, since the R package ranger multiplies the Gini index by the number of observations in each node, we observe much bigger numbers for the variable importance measure. The compared years are in Figure 6.5. With these plots we examine which variables contribute most to discrimination between two year labels under the

chosen variable importance measure. We do not see big differences in these variable importance plots. For now, that is enough. In later research, it might be of interest to quantify the difference.



(a) 2018-2019 **(b)** 2019-2020

Figure 6.5: Plots of the variable importance measure for the selected variables with subsequent years as label. In plot 6.5a the years 2018-2019 are compared. In plot 6.5b the years 2019-2020 are compared. The vertical red dashed line represents the 95% quantile of the importance distribution.

The variable `leeftijd` (age) has a slightly bigger variable importance measure in 2019-2020 (Figure 6.5a) than in 2018-2019 (Figure 6.5b). The variable `scopdistmm` (distance in mm from the lower edge of the first tumour to the anorectal junction on sagittal MRI (or CT, if applicable)) on the other hand has a slightly smaller variable importance measure in 2019-2020 than in 2018-2019. The other variables have more or less the same variable importance measure (note the change in to where the x-axis goes). The 95% quantile of the importance distribution does shift however quite a bit from 2018-2019 to 2019-2020. The plots corresponding to the effect size analysis of the variables `leeftijd` and `scopdistmm` are in Figure G.15e and Figure G.15c, respectively. We see that for the variable `leeftijd` the absolute effect size increased slightly from 2018-2019 to 2019-2020. For the variable `scopdistmm` the absolute effect size decreased slightly from 2018-2019 to 2019-2020.

Comparing this to our variable importance measure plots retrieved with the random forest classifier, we see that they align. We check this also with the other variables and their effect size analyses since this is also relevant. The other variables are `waitingTime`, `severeCompl`, and `asascore`. The plots of their effect size analyses are in Figures 6.1a, G.14d, G.2h, respectively. Looking at these plots we see that for the years 2018-2019 and 2019-2020 the effect size analyses definitely did change. So we cannot say for `leeftijd` and `scopdistmm` their change in variable importance can be explained by their change in effect sizes. Hence, there should be another explanation for this. A possibility is the conditional dependence of the variables. All the selected variables contribute one way or another to the variable importance measure in determining which year we are dealing with.

6.2.2 Severe complications

Since we have marginal and pairwise stability for the five variables that we have selected for the sequential years between 2018 to 2024, we can use the random forest classifier on these years to account for conditional relations in data. To get more insights into the data, we run the random forest classifier per year for the selected variables with `severeCompl` as the label. To do so, we use the hypoRF package in R, provided by [21].

To investigate this, we need to split the data into two parts. Hence, we split it into group Y (yes) and group N (no). In group Y we have the observations for which `severeCompl` $= 1$, so severe complications is 'yes'. In group N we have the observations for which `severeCompl` $= 0$, so severe complications is 'no'. We then have the null hypothesis (H_0) and the alternative hypothesis (H_A), as follows:

$$H_0 : P = Q, \quad \text{you can not distinguish with these factors whether it is more likely that } \texttt{severeCompl} \text{ is 1 (yes) or 0 (no)}$$

$$H_A : P \neq Q, \quad \text{you can distinguish with these factors whether it is more likely that } \texttt{severeCompl} \text{ is 1 (yes) or 0 (no)}$$

The results are in Table 6.3. This table shows the variable importance measure[1] per variable per year and the corresponding 95% cut-off of that year. A variable importance plot for the year 2018 is in Figure 6.6. The vertical red dashed line represents the 95% quantile of the importance distribution. If a variable importance measure is above the 95% quantile of the importance distribution of the importance distribution[2], it is highlighted. This is only the case for the variable `waitingTime` in the years 2018 and 2023. In general, the table shows that the ranking of the distributions of the variable importance measure over the years are similar. An exact measure or classification for the difference over the years is outside the scope of this thesis and can be investigated in future research. Therefore, no rigorous conclusions can be drawn from these insights.

	Variable Importance Measure				95% quantile
	`waitingTime`	`scopdistmm`	`leeftijd`	`asascore`	
2018	495,34	215,74	364,23	69,48	491,59
2019	423,59	188,63	332,94	66,22	434,68
2020	380,38	160,10	315,04	63,22	394,36
2021	428,26	165,16	347,18	64,87	442,99
2022	359,49	144,88	286,91	56,32	369,49
2023	426,98	159,43	310,69	60,51	419,42
2024	396,38	147,64	306,18	56,45	414,83

Table 6.3: Variable importance measures per variable per year with `severeCompl` as the label with the corresponding 95% quantile of the importance distribution. Values higher than the empirical 95% importance quantile are highlighted.

[1] Named by 'importance ranking' in the hypoRF package.
[2] The 'cutoff' in the hypoRF package.

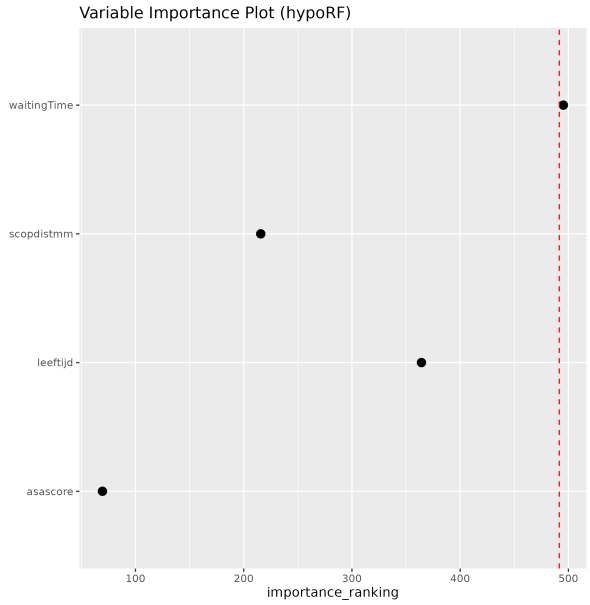


Figure 6.6: Variable importance plot for 2018 for the selected variables, split on `severeCompl`. The vertical red dashed line represents the 95% quantile of the importance distribution.

In Table 6.4 the variable importance measure per variable per two consecutive years and the corresponding 95% quantile of the importance distribution of those years are shown. In general, the table shows that the distributions of the variable importance measure over the consecutive years are similar in ranking. If a variable importance measure is above the 95% quantile of the importance distribution of the importance distribution, it is highlighted. For the years 2018-2019, 2021-2022, 2022-2023, and 2023-2024 the variable `waitingTime` is above the 95% quantile of the importance distribution.

	Variable Importance Measure				95% quantile
	`waitingTime`	`scopdistmm`	`leeftijd`	`asascore`	
2018-2019	820,34	387,85	624,61	114,36	808,95
2019-2020	723,14	323,01	579,33	109,23	730,30
2020-2021	711,12	299,91	571,50	108,27	738,06
2021-2022	715,10	289,70	555,56	102,28	709,55
2022-2023	712,39	285,80	523,08	98,19	687,28
2023-2024	757,01	289,03	534,23	97,34	744,38

Table 6.4: Variable importance measures per variable per year with `severeCompl` as the label with the corresponding 95% quantile of the importance distribution. Values higher than the empirical 95% importance quantile are highlighted.

The change in which years the variable importance measure of `waitingTime` is above the 95% quantile of the importance distribution when comparing Table 6.3 and Table 6.4, is remarkable. To cross validate this change, we have done a pairwise analysis for the two consecutive years 2018 and 2019. The corresponding correlation matrix is in Figure 6.7. Comparing this correlation matrix to the correlation matrix of 2018 (Figure 6.3a) we see that the correlation between `waitingTime` and `severeCompl` decreased slightly. For the correlation matrix of 2019 (Figure 6.3b) the correlation between `waitingTime` and `severeCompl` increased slightly compared to the correlation matrix of 2018 and 2019 together. The slight change in correlation could be just enough to arise above the 95% quantile of the importance distribution with the variable importance measure. This correlation matrix is thus consistent with the analysis from Table 6.3 and Table 6.4.

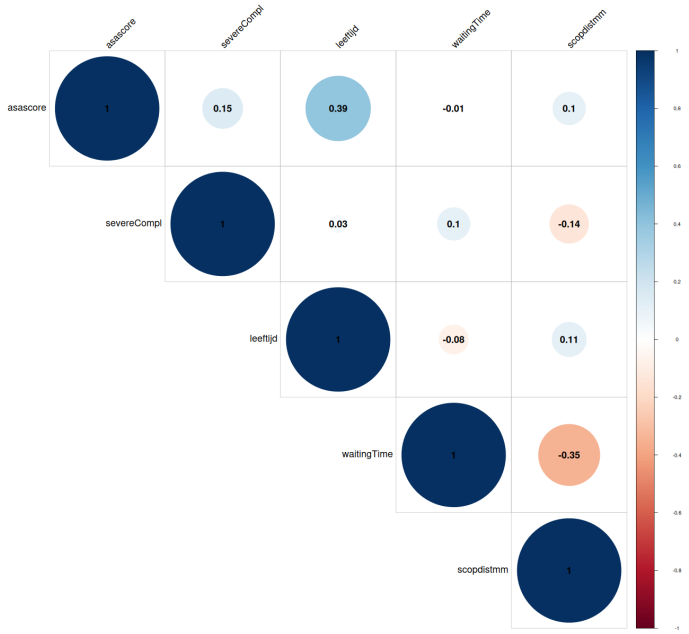


Figure 6.7: Correlation matrix for the years 2018 and 2019.

We thus conclude that in the setting used for individual years the variable `waitingTime` has a variable importance measure above the 95% quantile of the importance distribution for the years 2018 and for 2023. For consecutive years, the variable `waitingTime` has a variable importance measure above the 95% quantile of the importance distribution for the years 2018-2019, 2021-2022, 2022-2023, and 2023-2024.

7 | Conclusion and Discussion

In this chapter we discuss the conclusion in Section 7.1, the discussion in Section 7.2, and the possibilities for further research in Section 7.3.

7.1 Conclusion

The main question of this thesis is: Which years of the Dutch ColoRectal Audit (DCRA) registry can be pooled for statistical analysis without introducing practically relevant temporal heterogeneity?

In order to achieve the answer to this question, multiple steps were taken.

We started by harmonising the data using registry documentation and input from a medical expert. We have retrieved a complete dataset with 47242 rows and 123 variables including data from the years 2018 to 2024. To gain more insight into this final dataset, a univariate data analysis was done to see how the variables behave over the years.

To assess marginal temporal stability, we compared the distributions of individual variables across years with an effect size analysis. The effect sizes were classified as small and hence we concluded that the corresponding marginal distributions in the univariate setting were sufficiently similar to pool the data of one variable of consecutive years. Note that this conclusion is subject to the limitations of the selected measures and thresholds. There was thus not one year for which the differences in marginal distribution were classified as bigger than small. This is especially remarkable for the year 2020, the COVID year.

After marginal temporal stability, temporal stability in the dependence structure is investigated. This analysis examines whether variables may have similar marginal distributions while their joint structure changes over time. This is achieved with a pairwise correlation analysis and from that constructing pairwise correlation matrices per year. To quantify the change in association over the years, difference matrices were constructed from the pairwise correlation matrices per two sequential years. An element of a difference matrix could be interpreted as the change in association over the two years. The biggest change in correlation obtained was 0,07. Hence, for all the subsequent years and all the selected variables, the change in correlation is classified as small. We thus found evidence that the pairwise dependence structure among the selected variables remained stable between consecutive years. Combined with the analysis of the marginal distributions, we concluded that these results support pooling the selected variables across the compared consecutive years.

Having support that the data of the subsequent years of the selected variables can be pooled, a multivariate classification approach is used to inspect and support in pairs of years whether there are changes in variable importance amongst the selected variables. We do so by using the random forest classifier. With this information, we can decide which variables might have changed and whether it is consistent with our effect size analysis. When comparing the variable

importance measures for the selected variables for the years 2018-2019 and 2019-2020, we saw very small changes in some variables. This would imply that the analysis with random forests is consistent with the effect size analysis. Quantifying these changes can be investigated in further research.

The results obtained imply that the selected DCRA data for the years 2018-2024 for five selected variables (`waitingTime` (waiting time), `severeCompl` (severe complications), `scopdistmm` (distance in mm from the lower edge of the first tumour to the anorectal junction on sagittal MRI (or CT, if applicable)), `leeftijd` (age), and `asascore` (ASA score)) for sequential years can be pooled since temporal stability is concluded by a pairwise correlation analysis as well as the random forest classifier.

Additionally, by combining the evidence from the marginal, dependence, and multivariate analyses, we formulated recommendations on which years can be pooled. Continuing with the pooled consecutive years, to inspect whether the selected variables contain possible predictors for severe complications, we used the random forest classifier as well. If the variable importance measure of a variable is above the 95% quantile of the importance distribution we conclude that the variable could be a potentially relevant variable. This approach revealed the time between diagnosis and surgery as a potentially relevant variable for severe postoperative complications. The exact relation between the variables can be investigated in further research.

The results imply that we can group some of the variables together, not all (which can be seen by the difference matrices in Appendix H, discussed in Chapter 6). This implies that, if we would pool all the variables, there would be temporal heterogeneity that could introduce bias. This can have far-reaching consequences for results of medical related scientific research. In all of the papers discussed in the literature, no temporal stability analysis was conducted. Also the papers [18], [26], [30], [31] have used DCRA data both including data before and after 2018. The findings of this thesis do not imply that their conclusions are invalid since those studies used different variables, outcomes, inclusion criteria, and statistical methods. But, they demonstrate the value of including temporal stability analysis when data from evolving registries are pooled across years.

7.2 Discussion

During the retrieval of the final dataset, multiple choices on which data to include or exclude were made.
We have chosen to include the data from 2018 onwards. Of course there are convincing arguments in favour of this decision, which are discussed in Section 5.4. However, it is important to note that due to this decision, quite some observations are left out which decrease our sample size.
Another choice we have made is to do include some of the 'unknown' values. Values with unknown are only included when it was originally also possible for that variable to have the value unknown assigned to it, with the exception of the comorbidities due to their possible importance for predicting. This is discussed in detail in Section 5.6. In further research the included unknown values should be carefully handled and there should not be given meaning to these unknown values, in the sense that an unknown value becomes a predictor.
We have also decided to remove rows that had missing values after certain steps were taken. Even though we checked the change in distribution of severe complications to make sure that by our selection no bias occurred, similarity in this single outcome does not rule out selection bias. Complete case selection may change other marginal distributions or the dependence structure among variables. Future research should compare the original and final samples more systematically

Furthermore, we have used splits in data to assess temporal stability. We have decided to split

the data using years (12 months) with the date of the surgery as reference.
Our decision to split the data in blocks of 12 months influences the results of the temporal stability analysis. It could be possible for example that changes are occurring within these 12 months but that we do not notice them since we group it all as one. The significance of these changes is up for discussion. It could also be possible that if we would group more data as one, for example 24 months, it would lead to us being able to include more data.
Using the date of surgery as reference for splitting the data, also influences the temporal stability analysis. It could be possible that by choosing a different date as reference, for example the date of diagnosis, the outcome of the temporal stability analysis would be different.

The classification of small, medium, and large for effect size and correlation is not a hard theoretical cut-off but more a convention to use [52]. We used these classifications in the effect size analysis to decide whether the changes in marginal distribution were small enough. We also used it to decide whether changes in association were small enough in the pairwise analyses. Another classification might have led to different conclusions.

With the random forest classifier, variables are investigated which could be possible predictors for severe complications. For the definition of severe complications we have decided to be consistent with [29]. However, some aspects of this definition can be questioned. Readmission (`reintreq`) is for example not always severe, and for the mortality (`status90` $= 1$) it is the question whether this is really due to the surgery or because of other factors.

Attention must also be paid to the practical feasibility of the results. For some variables a certain value might be associated with a certain value for the outcome variable. However, it is not always practically feasible to act on this information, since the values of these variables can be inevitable. For example, for the variable `conversien` (Did a conversion (a change of approach during the surgery) took place) we identified that a conversion is associated with more severe complications. Nevertheless, we cannot give any advice on this matter. The conversion in the approach of the surgery takes place because the surgeon decides that it is necessary, not because it was planned.

7.3 Further research

The final dataset retrieved and worked with in this thesis underwent many transformations compared to the original dataset. In future research, the change in the dependence structure of the original and the final dataset can be investigated with for example correlation matrices, which we used for pairwise analyses in Chapter 6.

The results gained in this thesis can possibly be extended by selecting more variables and by checking whether data can be pooled for 2018-2024 in total, instead of only sequential years.

The methods used in this thesis can be expanded to investigate the temporal stability in other (medical) datasets. Note that these datasets should have a similar structure as the DCRA dataset, in the sense that we have one measurement per variable per observation, and no more than that. If there would be more measurements on the same variable per observation (patient), also types of temporal stability on an individual level should be investigated.

In general, there are many analyses possible with this dataset as is, since there are many variables and the sample size is large as well. For now, we provide two specific possibilities to investigate in further research in which the data can be split when using the random forest classifier. Obviously there are many other possibilities to split on. First, it would be a possibility to select only the data with severe complications and split on the waiting time above and below 30 days. With the random forest classifier, information can be gathered on which variables are associated with a waiting time of less or more than 30 days for patients with severe complications.

Second, it could be interesting to look at the `typok` variable that describes the type of surgery. In this thesis, only the elective surgeries were included, since it is already known that urgent surgeries are a risk factor for severe complications. The elective surgeries also have different types. The types that we included are 'elective' (1), 'elective after stent' (2), 'elective after stent or stoma' (5), and 'unknown' (9). For 2 and 5 it means that the stent or the stoma was placed in order to not have to do an urgent surgery. According to our medical expert, 2 and 5 do imply that the tumour was already quite large, since it means that it was necessary to place this tumour or stent because the bowel was blocked. This means that it could also be interesting to focus on the specific value of `typok` and on what the different chances of developing severe complications or other complications are for these patients.

The DCRA dataset contains quite some variables that are important but do not occur often, for example mortality. This phenomenon is known as class imbalance. In [53] probabilistic classifiers and systematic variable selection are used instead of models that ignore class imbalance. Class imbalance leads to poorer calibration and additional costs for data collection. They recommend to validate this externally. It would be possible to validate their models with DCRA data. This can be done in further research with the dataset that is constructed in this thesis.

Even though the data that is collected in DCRA is extensive, there are some recommendations to include more variables in this dataset. It would for example be interesting to take additional biological information into account, such as blood tests, medication use, lifestyle, etcetera. It is certainly a limiting factor of the DCRA data that this is not included. This is also emphasised in other papers, amongst which [27].

References

[1] Dutch Institute for Clinical Auditing. *Darmkanker – DCRA*. Accessed: 5-6-2026. 2025. URL: https://dica.nl/registratie/darmkanker-dcra/.

[2] Integraal Kankercentrum Nederland. "Trendrapport darmkanker; dikkedarm- en endeldarmkanker in Nederland". In: (Mar. 2024). URL: https://iknl.nl/getmedia/80578c3d-ca2d-49fd-87a5-0426e320fb4b/trendrapport_darmkanker_def.pdf.

[3] Integraal Kankercentrum Nederland. *Kankersoorten*. Accessed: 8-6-2026. 2026. URL: https://iknl.nl/kankersoorten.

[4] Integraal Kankercentrum Nederland. "Overlevingscijfers van darmkanker". In: (Feb. 2026). Accessed: 8-6-2026. URL: https://www.kanker.nl/kankersoorten/darmkanker-dikkedarmkanker/algemeen/overlevingscijfers-van-darmkanker.

[5] Gidius van de Kamp. "Patient level predictions in bowel surgery comparing variable selections for rare outcome modelling on real surgery data". MA thesis. Delft University of Technology, 2024.

[6] Stichting Darmkanker. *Darmkanker onderzoeken*. Accessed: 9-6-2026. 2026. URL: https://www.darmkanker.nl/over-darmkanker/darmkanker-onderzoeken/.

[7] J. H. Park et al. "The American Society of Anesthesiologists score influences on postoperative complications and total hospital charges after laparoscopic colorectal cancer surgery." In: *Medicine* 97 (2018), pp. 631–632. DOI: 10.1097/MD.0000000000010653.

[8] Medilex. *Colon Anatomy and Colonoscopy*. Accessed: 12-6-2026. Apr. 2023. URL: https://medilexinc.com/a-spoonful-of-medicine-blog/colon-anatomy-and-colonoscopy.

[9] Cleveland Clinic. *Stoma*. Accessed: 12-6-2026. June 2025. URL: https://my.clevelandclinic.org/health/articles/stoma.

[10] Cleveland Clinic. *Cerebrovascular Disease*. Accessed: 12-6-2026. Sept. 2022. URL: https://my.clevelandclinic.org/health/diseases/24205-cerebrovascular-disease.

[11] Stichting Darmkanker. *Operatie bij darmkanker*. Accessed: 9-6-2026. 2026. URL: https://www.darmkanker.nl/over-darmkanker/behandeling/operatie-bij-darmkanker/.

[12] Integraal Kankercentrum Nederland. *Cijfers over darmkanker*. Mar. 2024. URL: https://iknl.nl/Kankersoorten/Darmkanker/Cijfers. (accessed: 06-01-2026).

[13] Fred Hutch Cancer Center. *Stages of Colon Cancer*. Accessed: 5-3-2026. Dec. 2025. URL: https://www.fredhutch.org/en/diseases/colon-cancer/stages.html.

[14] N.J. Van Leersum et al. "The Dutch Surgical Colorectal Audit". In: *European Journal of Surgical Oncology* 39 (2013), pp. 1063–1070. URL: https://doi.org/10.1016/j.ejso.2013.05.008.

[15] J A Govaert et al. "Costs of complications after colorectal cancer surgery in the Netherlands: Building the business case for hospitals". In: *European Journal of Surgical Oncology* 41 (2015), pp. 1059–1067. URL: 10.1016/j.ejso.2015.03.236.

[16] D Henneman et al. "Hospital variation in failure to rescue after colorectal cancer surgery: results of the Dutch Surgical Colorectal Audit". In: *Annals of Surgical Oncology* 20 (2013), pp. 2117–2123. URL: https://doi.org/10.1245/s10434-013-2896-7.

[17] N E Kolfschoten et al. "Variation in case-mix between hospitals treating colorectal cancer patients in the Netherlands". In: *European Journal of Surgical Oncology* 37 (2011), pp. 956–963. URL: 10.1016/j.ejso.2011.08.137.

[18] V.T. Hoek et al. "A preoperative prediction model for anastomotic leakage after rectal cancer resection based on 13.175 patients". In: *European Journal of Surgical Oncology* 48 (2022), pp. 2495–2501. URL: https://doi.org/10.1016/j.ejso.2022.06.016.

[19] J.A. Bondy and U.S.R. Murty. *Graph Theory with Applications*. 5th ed. North-Holland, 1982. ISBN: 0.:444-19451-7.

[20] K.T. Huber, V. Moulton, and G.E. Scholz. "Forest-Based Networks". In: *Bulletin of Mathematical Biology* 84.1 (2022), p. 119. DOI: 10.1007/s11538-022-01081-9. URL: https://doi.org/10.1007/s11538-022-01081-9.

[21] Simon Hediger, Loris Michel, and Jeffrey Naf. "On the use of random forest for two-sample testing". In: *Computational Statistics & Data Analysis* 170 (2022), p. 107435. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2022.107435. URL: https://www.sciencedirect.com/science/article/pii/S0167947322000159.

[22] Haiyan Cai, Bryan Goggin, and Qingtang Jiang. "Two-sample test based on classification probability". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 13 (2019). URL: https://doi.org/10.1002/sam.11438.

[23] S. Hediger, L. Michel, and J. Naef. *Package 'hypoRF'*. Accessed: 4-3-2026. July 2025. URL: https://cran.r-project.org/web/packages/hypoRF/hypoRF.pdf.

[24] PerfusionTech. *Colorectal Surgery*. Accessed: 5-3-2026. 2023. URL: https://www.perfusiontech.com/colorectal-surgery.

[25] W.A.A. Borstlap et al. "Anastomotic Leakage and Chronic Presacral Sinus Formation After Low Anterior Resection: Results From a Large Cross-sectional Study". In: *Annals of Surgery* 5 (2017), pp. 870–877. DOI: 10.1097/SLA.0000000000002429. URL: https://pubmed.ncbi.nlm.nih.gov/28746154/.

[26] E.W. Ingwersen et al. "One Decade of Declining Use of Defunctioning Stomas After Rectal Cancer Surgery in the Netherlands: Are We on the Right Track?" In: *Diseases of the Colon & Rectum* 66 (2023), pp. 1003–1011. URL: https://journals.lww.com/dcrjournal/fulltext/2023/07000/one_decade_of_declining_use_of_defunctioning.20.aspx.

[27] Tom van den Bosch et al. "Predictors of 30-Day Mortality Among Dutch Patients Undergoing Colorectal Cancer Surgery, 2011-2016". In: *JAMA Network Open* 4 (2021). URL: 10.1001/jamanetworkopen.2021.7737.

[28] Lindsey C F de Nes et al. "Postoperative mortality risk assessment in colorectal cancer: development and validation of a clinical prediction model using data from the Dutch ColoRectal Audit". In: *BJS Open* 6 (2022). URL: 10.1093/bjsopen/zrac014.

[29] E.T.D. Souwer et al. "A Prediction Model for Severe Complications after Elective Colorectal Cancer Surgery in Patients of 70 Years and Older". In: *Cancers (Basel)* 13 (2021), p. 3110. URL: https://www.mdpi.com/2072-6694/13/13/3110.

[30] D.J. Sikkenk et al. "Nationwide practice in CT-based preoperative staging of colon cancer and concordance with definitive pathology". In: *European Journal of Surgical Oncology* 49 (2023), p. 106941. URL: https://www.sciencedirect.com/science/article/pii/S0748798323005103?via%3Dihub.

[31] Anne C M Cuijpers et al. "Development and external validation of preoperative clinical prediction models for postoperative outcomes including preoperative aerobic fitness in patients approaching elective colorectal cancer surgery". In: *European Journal of Surgical Oncology* 50 (2024). URL: 10.1016/j.ejso.2024.108338.

[32] Regina Beets-Tan. "Kwaliteitsevaluatie van beeldvormende diagnostiek bij oncologische patiënten". In: *Dutch Institute for Clinical Auditing* (2024). URL: https://dica.nl/wp-content/uploads/2024/06/eindrapport_skms_project_evaluatie_beeldvorming_dcra_2024-002.pdf.

[33] MRDM. *DICA*. Accessed: 19-6-2026. 2026. URL: https://support.mrdm.com/en/dica/.

[34] Dutch Institute for Clinical Auditing. "Handleiding DCRA dataset". In: (2024). Not publicly available.

[35] Slingeland Ziekenhuis. *Afwachtend beleid (wait-and see) bij endeldarmkanker*. Accessed: 12-6-2026. Feb. 2025. URL: https://folders.slingeland.nl/folders/2308-afwachtend_beleid_bij_darmkanker.

[36] MRDM. *Datadictionary*. Accessed: 9-3-2026. URL: https://support.mrdm.com/nl/batch/handleiding/ondersteunende-documentatie/datadictionary/.

[37] N. Ohja and A.S. Dhamoon. *Treasure Island*. StatPearls Publishing, 2026. URL: https://www.ncbi.nlm.nih.gov/books/NBK537076/.

[38] University of Pittsburgh Medical Center. *Peripheral Aneurysm*. Accessed: 12-6-2026. Feb. 2025. URL: https://www.upmc.com/conditions/p/peripheral-aneurysm.

[39] Spinalcord.com Team. *Types of Paralysis: Monoplegia, Hemiplegia, Paraplegia and Quadriplegia*. Accessed: 12-6-2026. Dec. 2025. URL: https://www.spinalcord.com/types-of-paralysis.

[40] Federatie Medisch Specialisten. *Colorectaal carcinoom (CRC): Locoregionale stadiëring rectumcarcinoom*. Accessed: 15-6-2026. Sept. 2025. URL: https://richtlijnendatabase.nl/richtlijn/colorectaal_carcinoom_crc/diagnostiek_bij_crc/locoregionale_stadi_ring_rectumcarcinoom.html.

[41] Federatie Medisch Specialisten. *Colorectaal carcinoom (CRC): MMR/ MSI-onderzoek bij CRC*. Accessed: 15-6-2026. Dec. 2020. URL: https://richtlijnendatabase.nl/richtlijn/colorectaal_carcinoom_crc/pathologie_bij_crc/mmr_msi-onderzoek_bij_crc.html.

[42] Antoni van Leeuwenhoek. *'Verlaag leeftijdsgrens bevolkingsonderzoek darmkanker naar 50 jaar'*. Dec. 2019. URL: https://www.avl.nl/nieuwsberichten/2019/verlaag-leeftijdsgrens-bevolkingsonderzoek-darmkanker-naar-50-jaar/. (accessed: 06-01-2026).

[43] Rijksinstituut voor Volksgezondheid en Milieu. *Vragen en antwoorden tijdelijke stop bevolkingsonderzoeken naar kanker*. 2020. URL: https://www.rivm.nl/nieuws/tijdelijke-stop-bevolkingsonderzoeken/vragen-en-antwoorden. (accessed: 06-01-2026).

[44] Integraal Kankercentrum Nederland. *Minder darmkankerdiagnoses in 2020, maar op niveau tijdens tweede coronagolf*. Mar. 2021. URL: https://iknl.nl/nieuws/2021/minder-darmkankerdiagnoses-in-2020. (accessed: 06-01-2026).

[45] Bevolkigsonderzoek Nederland. *Achterstand bevolkingsonderzoeken in 2021 ingehaald*. Oct. 2022. URL: https://www.bevolkingsonderzoeknederland.nl/nieuws/achterstand-bevolkingsonderzoeken-in-2021-ingehaald/. (accessed: 06-01-2026).

[46] Rijksinstituut voor Volksgezondheid en Milieu. *Achterstand bevolkingsonderzoeken in 2021 ingehaald*. Oct. 2022. URL: https://www.rivm.nl/nieuws/achterstand-bevolkingsonderzoeken-in-2021-ingehaald. (accessed: 06-01-2026).

[47] Integraal Kankercentrum Nederland. *Bevolkingsonderzoeken 2021: meer uitnodigingen na COVID-19 pandemie*. Oct. 2022. URL: https://iknl.nl/nieuws/2022/achterstand-bevolkingsonderzoeken-in-2021-ingehaal. (accessed: 06-01-2026).

[48] C. Ialongo. "Understanding the effect size and its measures". In: *Biochemia Medica* 26 (2016), pp. 150–163. URL: https://pubmed.ncbi.nlm.nih.gov/27346958/.

[49] D. Lakens. "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs". In: *Frontiers in Psychology* 4 (2013). URL: https://doi.org/10.3389/fpsyg.2013.00863.

[50] Larry Hedges and Ingram Olkin. "Statistical Methods in Meta-Analysis". In: vol. 20. Jan. 1985. DOI: 10.2307/1164953.

[51] Geoff Cumming. "Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis". In: 2012. DOI: https://doi.org/10.4324/9780203807002.

[52] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Lawrence Erlbaum Associates, 1988. ISBN: 0-8058-0283-5.

[53] Özge Şahin et al. "Bowel surgery risk prediction: class-imbalance-aware models with reduced data-collection burden". In: (2025).

A | Calculations of Entropy and Information Gain

In Section A.1 the calculations for the entropy and information gain if node a of Figure 2.6 are shown. In Section A.2 the calculations for the entropy and information gain if node b of Figure 2.6 are shown.

A.1 Entropy and information gain node a

First, we calculate the entropy per relevant node per row of Table 2.6.

Let us consider option (1). This gives us entropy $E_b = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) + -\frac{3}{6}\log_2\left(\frac{3}{6}\right) = -\log_2\left(\frac{1}{2}\right) = 1$, and entropy $E_c = -0 \cdot \log_2(0) + -0\log_2(0) = 0$.

Let us consider option (2). This gives us entropy $E_b = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) + -\frac{3}{5}\log_2\left(\frac{3}{5}\right) \approx 0,971$, and entropy $E_c = -\frac{1}{1}\log_2\left(\frac{1}{1}\right) + -0\log_2(0) = 0$.

Let us consider option (3). This gives us entropy $E_b = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) + -\frac{2}{5}\log_2\left(\frac{2}{5}\right) \approx 0,971$, and entropy $E_c = -0\log_2(0) + -\frac{1}{1}\log_2\left(\frac{1}{1}\right) = 0$.

Let us consider option (4). This gives us entropy $E_b = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) + -\frac{1}{4}\log_2\left(\frac{1}{4}\right) \approx 0,811$, and entropy $E_c = -0\log_2(0) + -\frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0$.

Let us consider option (5). This gives us entropy $E_b = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) + -\frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$, and entropy $E_c = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) + -\frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$.

Let us consider option (6). This gives us entropy $E_b = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) + -\frac{3}{4}\log_2\left(\frac{3}{4}\right) \approx 0,811$, and entropy $E_c = -\frac{2}{2}\log_2\left(\frac{2}{2}\right) + -0\log_2(0) = 0$.

Let us consider option (7). This gives us entropy $E_b = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) + -\frac{1}{3}\log_2\left(\frac{1}{3}\right) \approx 0,918$, and entropy $E_c = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) + -\frac{2}{3}\log_2\left(\frac{2}{3}\right) \approx 0,918$.

Let us consider option (8). This gives us entropy $E_b = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) + -\frac{2}{3}\log_2\left(\frac{2}{3}\right) \approx 0,918$, and entropy $E_c = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) + -\frac{1}{3}\log_2\left(\frac{1}{3}\right) \approx 0,918$.

Let us consider option (9). This gives us entropy $E_b = -\frac{3}{3}\log_2\left(\frac{3}{3}\right) + -\frac{0}{3}\log_2\left(\frac{0}{3}\right) = 0$, and entropy $E_c = -\frac{0}{3}\log_2\left(\frac{0}{3}\right) + -\frac{3}{3}\log_2\left(\frac{3}{3}\right) = 0$.

Let us consider option (10). This gives us entropy $E_b = -\frac{0}{3}\log_2\left(\frac{0}{3}\right) + -\frac{3}{3}\log_2\left(\frac{3}{3}\right) = 0$, and entropy $E_c = -\frac{3}{3}\log_2\left(\frac{3}{3}\right) + -\frac{0}{3}\log_2\left(\frac{0}{3}\right) = 0$.

Now we calculate the information gain per row, corresponding to Table 2.6.

Let us consider option (1). This gives us information gain $IG = E(parent) - \sum w_i E(child_i) = E_a - w_b E_b - w_c E_c = 1 - \frac{6}{6} \cdot 1 - \frac{0}{6} \cdot 0 = 1 - 1 = 0$.

Let us consider option (2) and (3). This gives us information gain $IG = E(parent) - \sum w_i E(child_i) = E_a - w_b E_b - w_c E_c = 1 - \frac{5}{6} \cdot 0,971 - \frac{1}{6} \cdot 0 \approx 0,191$.

Let us consider option (4) and (6). This gives us information gain $IG = E(parent) - \sum w_i E(child_i) = E_a - w_b E_b - w_c E_c = 1 - \frac{4}{6} \cdot 0,811 - \frac{2}{6} \cdot 0 \approx 0,459$.

Let us consider option (5). This gives us information gain $IG = E(parent) - \sum w_i E(child_i) = E_a - w_b E_b - w_c E_c = 1 - \frac{4}{6} \cdot 1 - \frac{2}{6} \cdot 1 = 0$.

Let us consider option (7) and (8). This gives us information gain $IG = E(parent) - \sum w_i E(child_i) = E_a - w_b E_b - w_c E_c = 1 - \frac{3}{6} \cdot 0,918 - \frac{3}{6} \cdot 0,918 = 0,082$.

Let us consider option (9) and (10). This gives us information gain $IG = E(parent) - \sum w_i E(child_i) = E_a - w_b E_b - w_c E_c = 1 - \frac{3}{6} \cdot 0 - \frac{3}{6} \cdot 0 = 0$.

A.2 Entropy and information gain node b

First, we calculate the entropy per relevant node per row of Table 2.7.

Let us consider option (1). This gives us entropy $E_d = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) + -\frac{1}{4}\log_2\left(\frac{1}{4}\right) \approx 0,811$, and entropy $E_e = 0$.

Let us consider option (2). This gives us entropy $E_d = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) + -\frac{1}{3}\log_2\left(\frac{1}{3}\right) \approx 0,918$, and entropy $E_e = -\frac{1}{1}\log_2\left(\frac{1}{1}\right) + -0\log_2(0) = 0$.

Let us consider option (3). This gives us entropy $E_d = -\frac{3}{3}\log_2\left(\frac{3}{3}\right) + -\frac{0}{3}\log_2\left(\frac{0}{3}\right) = 0$, and entropy $E_e = -0\log_2(0) + -\frac{1}{1}\log_2\left(\frac{1}{1}\right) = 0$.

Now we calculate the information gain per row of Table 2.7.

Let us consider option (1). This gives us information gain $IG = E(parent) - \sum w_i E(child_i) = E_b - w_d E_d - w_e E_e = 0,81 - \frac{4}{4} \cdot 0,81 - \frac{0}{4} \cdot 0 = 0$.

Let us consider option (2). This gives us information gain $IG = E(parent) - \sum w_i E(child_i) = E_b - w_d E_d - w_e E_e = 0,81 - \frac{3}{4} \cdot 0,918 - \frac{1}{4} \cdot 0 = 0,12$.

Let us consider option (3). This gives us information gain $IG = E(parent) - \sum w_i E(child_i) = E_b - w_d E_d - w_e E_e = 0,81 - \frac{3}{4} \cdot 0 - \frac{1}{4} \cdot 0 = 0$.

B | Variables with Options

In this appendix, all the variables are given with all of their options, which can be found in the datadictionary. In Table B.1 the variables with their options for subdataset *patient* are given. In Table B.2 the variables with their options for subdataset *comorbidities* are given. In Table B.3 the variables with their options for subdataset *presentation* are given. In Table B.4 the variables with their options for subdataset *diagnostics* are given. In Table B.5 the variables with their options for subdataset *surgery* are given. In Table B.6 the variables with their options for subdataset *pathology* are given. In Table B.7 the variables with their options for subdataset *radiotherapy* are given. In Table B.8 the variables with their options for subdataset *systemic therapy* are given. When an option is highlighted in the table, this means I have constructed and added this option, so the option was not in the original dataset.

Variable	Options	Meaning of Option
patient_uri		
id		
gebdat		
geslacht		
	1	Man
	2	Woman
	7	Undifferentiated
	9	Unknown
datovl		

Table B.1: The variables with their options, belonging to the subdataset *patient*.

Variable	Options	Meaning of Option
datom		
commyo		
	0	No
	1	Yes
	9	Unknown
comhartfaal		
	0	No
	1	Yes
	9	Unknown
comperif		
	0	No
	1	Yes
	9	Unknown
comcva		
	0	No
	1	Yes
	9	Unknown

comdem		
	0	No
	1	Yes
	9	Unknown
comlong		
	0	No
	1	Yes
	9	Unknown
combind		
	0	No
	1	Yes
	9	Unknown
comgiulcus		
	0	No
	1	Yes
	9	Unknown
comlever		
	0	No
	1	Mild
	2	Moderate
	3	Severe
	9	Unknown
comdiam		
	0	No
	1	Yes, without organ damage
	2	Yes, with organ damage (such as retinopathy, nephropathy and neuropathy)
	3	Without chronic complications
	4	With chronic complications
	9	Unknown
comparalys		
	0	No
	1	Yes
	9	Unknown
comnier		
	0	No
	1	Yes
	9	Unknown
commalig		
	0	No
	1	No metastasis
	2	Metastatic
	9	Unknown
comhiv		
	0	No
	1	Yes
	9	Unknown

Table B.2: The variables with their options, belonging to the subdataset *comorbidities*.

Variable	Options	Meaning of Option
tumor_id		
idaaba		
lengte		
	999	Unknown
gewicht		

	999	Unknown
asascore		
	1	I - Normal healthy patient
	2	II - Mild systemic disease
	3	III - Severe systemic disease
	4	IV - Chronic life-threatening systemic disease
	5	V - Moribund
	9	Unknown
	10	Not mandatory before 2018, so not completed
resdarm		
	0	No
	1	Yes
	9	Unknown
	10	Not mandatory before 2018, so not completed
stomavg		
	0	No
	1	Ileostomy
	2	Colostomy
	3	Yes, type unknown
	9	Unknown
datpa1		
presentatie_uri		

Table B.3: The variables with their options, belonging to the subdataset *presentation*.

Variable	Options	Meaning of Option
primcrc		
	1	First primary tumour
	2	Second primary tumour
	9	Unknown
scopnumb		
	1	1
	2	2
	3	3 or more
diagn		
	1	Symptoms (e.g., anaemia)
	2	Screening
	3	Surveillance
	4	Coincidence
	7	Other
	9	Unknown
locprim		
	1	Blind
	2	Appendix
	3	Ascending colon
	4	Hepatic flexure
	5	Transverse colon
	6	Splenic flexure
	7	Descending colon
	8	Sigmoid colon
	9	Rectum
histd		
	1	Adenoma
	2	Adenocarcinoma
	8	Other
	9	Unknown

	10	Not mandatory before 2018, so not completed
`complpre`		
	0	No
	1	Yes
	9	Unknown
	10	Not mandatory before 2018, so not completed
`preanemie`		
	0	No
	1	Yes
	9	Unknown
`preobstruct`		
	0	No
	1	Yes
	9	Unknown
`preabces`		
	0	No
	1	Yes
	9	Unknown
`scorect`		
	1	cT1
	2	cT2
	3	cT3
	4	cT4
	5	cT4a
	6	cT4b
	9	cTX / Unknown
`scorecn`		
	0	cN0
	1	cN1
	2	cN2
	9	cNX / Unknown
`locprim2`		
	0	No second colorectal carcinoma
	1	Blind
	2	Appendix
	3	Ascending colon
	4	Hepatic flexure
	5	Transverse colon
	6	Splenic flexure
	7	Descending colon
	8	Sigmoid colon
	9	Rectum
	999	Unknown
`scorect2`		
	0	No second colorectal carcinoma
	1	cT1
	2	cT2
	3	cT3
	4	cT4
	5	cT4a
	6	cT4b
	9	cTX / Unknown
`scorecn2`		
	0	cN0
	1	cN1
	2	cN2
	9	cNX / Unknown
	10	No second colorectal carcinoma

`scorecm`

0	cM0
1	cM1
9	cMX / Unknown
92	Unknown, imaging not performed
93	Unknown, imaging inconclusive

`prelocmri`

0	No, does not satisfy condition
1	Yes
2	No, due to claustrophobia
3	No, due to metal implants (e.g., pacemaker)
7	No, other reason
8	No, reason unknown
9	Unknown
10	Not mandatory before 2018, so not completed

`prelocmri2`

0	No, does not satisfy condition
1	Yes
2	No, due to claustrophobia
3	No, due to metal implants (e.g., pacemaker)
7	No, other reason
8	No, reason unknown
9	Unknown
10	Not mandatory before 2018, so not completed

`scopdistmm`

999	Unknown
888	Irrelevant

`scopdist2mm`

999	Unknown
888	Irrelevant

`afstmrf`

111	MRF-free
222	MRF not free
888	Irrelevant
999	Unknown

`afstmrf2`

111	MRF-free
222	MRF not free
888	Irrelevant
999	Unknown

`prelocstad`

0	No
1	Endo-ultrasound
2	CT scan of the pelvis
3	Pelvic MRI
4	Endoscopic ultrasound and MRI

`emimri`

2	None
0	<=5mm
1	>5mm
9	Undetermined
10	Does not satisfy condition

`emimri2`

2	None
0	<=5mm
1	>5mm
9	Undetermined
10	Does not satisfy condition

genan		
	0	MSS (Microsatellite Stable)
	1	MSI (Microsatellite Instability)
	9	Not performed / not described
immuun		
	0	No
	1	Yes
	9	Unknown
mlh1		
	0	No
	1	Yes
	9	Unknown
pms2		
	0	No
	1	Yes
	9	Unknown
hyperm		
	0	No
	1	Yes
	9	Unknown
msh2		
	0	No
	1	Yes
	9	Unknown
msh6		
	0	No
	1	Yes
	9	Unknown
diagnostiek_uri		

Table B.4: The variables with their options, belonging to the subdataset *diagnostics*.

Variable	Options	Meaning of Option
datbez1		
verwezen		
	0	No
	1	Yes
	9	Unknown
darmperf		
	0	No
	1	Yes
	9	Unknown
ileusblow		
	0	No
	1	Yes
	9	Unknown
	888	Irrelevant
perfperi		
	0	No
	1	Yes
	9	Unknown
	888	Irrelevant
perfscopie		
	0	No
	1	Yes
	9	Unknown

888	Irrelevant

`prechirstomanew`

0	No
1	Diverting colostomy
2	Diverting ileostomy
3	Yes, a different type of stoma
8	Unknown (after conversion)
9	Unknown
10	Not mandatory before 2019, so not completed

`datstoma`

09-09-1909	Unknown

`prechirstent`

0	No
1	Yes
9	Unknown
10	Not mandatory before 2019, so not completed
888	Irrelevant

`datstent`

08-08-1808	No stent was placed
09-09-1809	Stent placed, date unknown

`prechirmetanew`

0	No
1	Liver resection
2	Local ablation of the liver
3	Other local treatment
4	Yes, but the type of treatment is unknown
9	Unknown
10	Not mandatory before 2019, so not completed
888	Irrelevant

`datprechirmetaeen`

08-08-1808	Irrelevant
09-09-1809	Unknown

`datprechirmetatwee`

08-08-1808	Irrelevant
09-09-1809	Unknown

`prechirappendix`

0	No
1	Yes
9	Unknown
10	Not mandatory before 2019, so not completed
888	Irrelevant

`datappendix`

08-08-1808	Did not have an appendectomy
09-09-1809	Appendix removed, date unknown

`prechircoloscex`

0	No
1	Yes, lysis-excision
2	Yes, EMR
3	Yes, ESD
4	Yes, full-thickness excision
5	Yes, laparoscopically assisted endoscopic polypectomy
6	Yes, EID

`datcoloscex`

08-08-1808	Irrelevant

`chirenkel`

0	No
1	Yes
9	Unknown

`datok`

08-08-1808	Did not have surgery as part of a watchful waiting approach
09-09-1809	Died without undergoing surgery
11-11-1811	No surgery took place

`typok`

1	Elective
2	Elective after stent
3	Urgent
4	Urgent, immediate
5	Elective following decompression stenting or stoma creation
9	Unknown

`procok`

0	Ileocecal resection
1	(extended) right hemicolectomy
2	Transversectomy
3	(extended) left hemicolectomy
21	(Low) anterior resection/sigmoid resection
5	Subtotal colectomy (from the cecum to the rectum)
15	Partial Mesorectal Excision (PME)
16	Total Mesorectal Excision (TME)
22	Abdomino-perineal resection (APR)
23	Proctocolectomy (from the cecum to the rectum)
8	Local excision
26	Local excision of the colon
9	Local excision + TME
10	Local excision + APR
13	Wig resection
25	Sigmoid resection
20	Transanal local excision
77	Other
24	Resection similar to the procedure for the first tumour
999	Unknown

`ingreep`

1	Transabdominal open
2	Transabdominal endoscopic
4	Transanal open local excision
5	Transanal minimally invasive local excision (local excision according to TAMIS/TEM)
6	Transanal and transabdominal endoscopic TME (TaTME or TAMIS TME)
7	Robot-assisted laparoscopic
8	Endoscopically assisted transabdominal laparoscopy
9	Unknown

`conversien`

0	No
1	Strategic
2	Reactive
9	Nature unknown
888	Irrelevant

`aprtyp`

888	Irrelevant
9	Unknown

`typprocok`

888	Irrelevant
9	Unknown

`datok2`

08-08-1808	Irrelevant
09-09-1809	Unknown

10-10-1810	No second surgery was performed

`procok2`

0	Ileocecal resection
1	(extended) right hemicolectomy
2	Transversectomy
3	(extended) left hemicolectomy
21	(Low) anterior resection/sigmoid resection
5	Subtotal colectomy (from the cecum to the rectum)
15	Partial Mesorectal Excision (PME)
16	Total Mesorectal Excision (TME)
22	Abdomino-perineal resection (APR)
23	Proctocolectomy (from the cecum to the rectum)
8	Local excision
26	Local excision of the colon
9	Local excision + TME
10	Local excision + APR
13	Wig resection
25	Sigmoid resection
20	Transanal local excision
77	Other
24	Resection similar to the procedure for the first tumour
888	Irrelevant
999	Unknown

`procok3`

0	Ileocecal resection
1	(extended) right hemicolectomy
2	Transversectomy
3	(extended) left hemicolectomy
21	(Low) anterior resection/sigmoid resection
5	Subtotal colectomy (from the cecum to the rectum)
15	Partial Mesorectal Excision (PME)
16	Total Mesorectal Excision (TME)
22	Abdomino-perineal resection (APR)
23	Proctocolectomy (from the cecum to the rectum)
8	Local excision
26	Local excision of the colon
9	Local excision + TME
10	Local excision + APR
13	Wig resection
25	Sigmoid resection
20	Transanal local excision
77	Other
24	Resection similar to the procedure for the first tumour
888	Irrelevant

`ingreep2`

1	Transabdominal open
2	Transabdominal endoscopic
4	Transanal open local excision
5	Transanal minimally invasive local excision (local excision according to TAMIS/TEM)
6	Transanal and transabdominal endoscopic TME (TaTME or TAMIS TME)
7	Robot-assisted laparoscopic
8	Endoscopically assisted transabdominal laparoscopy
9	Unknown
888	Irrelevant

`conversien2`

0	No

1	Strategic
2	Reactive
9	Nature unknown
888	Irrelevant

`aprtyp2`

1	Limited
2	Extensive
9	Unknown
888	Irrelevant

`resadd`

0	No
1	Extensive
2	Limited
888	Irrelevant

`locaddmaag`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locaddmilt`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locaddduod`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locaddddoverig`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locaddoment`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locaddbuikw`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locaddvag`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locadduter`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locaddovarl`

0	No
1	Strategic
9	Unknown

888	Irrelevant

`locaddovarr`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locadduretl`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locadduretr`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locaddblaas`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locaddprost`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locaddvesl`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locaddvesr`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locaddvesb`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locadddund`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locaddsacr`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`locaddoverig`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`resadm`

0	No

1	Strategic
9	Unknown
888	Irrelevant

`resomm`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`reslem`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`respem`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`reslmm`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`resbum`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`perrtxchemo`

0	No
1	IORT
2	HIPEC
888	Irrelevant

`anastom1`

0	No
1	Strategic
9	Unknown
888	Irrelevant

`stoma`

0	No
1	Ileostomy drainage
2	Terminal ileostomy
3	Colostomy drainage
4	Terminal colostomy
5	Stoma, type unknown
9	Unknown
888	Irrelevant

`stomahef`

0	No
1	Ileostomy reversed
2	Colostomy reversed
3	Yes, unknown type of stoma reversed
9	Unknown
888	Irrelevant

`compl90`

0	No
1	Yes
9	Unknown

`compchir`

0	No
1	Yes
888	Irrelevant

`datcompchir`

08-08-1808	Irrelevant

`chirnaad`

0	No
1	Yes
9	Unknown
888	Irrelevant

`chirnaadoccult`

0	No
1	Yes
9	Unknown
888	Irrelevant

`chirabces`

0	No
1	Yes
9	Unknown
888	Irrelevant

`chirbloeding`

0	No
1	Yes
9	Unknown
888	Irrelevant

`chirileus`

0	No
1	Yes
9	Unknown
888	Irrelevant

`chirgastro`

0	No
1	Yes
9	Unknown
888	Irrelevant

`chirfascie`

0	No
1	Yes
9	Unknown
888	Irrelevant

`chirdarmperf`

0	No
1	Yes
9	Unknown
888	Irrelevant

`chirlekkage`

0	No
1	Yes
9	Unknown
888	Irrelevant

`chirwondinf`

0	No
1	Yes
9	Unknown
888	Irrelevant

`datcompl`

08-08-1808	Irrelevant

09-09-1809	Unknown

`comppul`

0	No
1	Yes
9	Unknown
888	Irrelevant

`compcar`

0	No
1	Yes
9	Unknown
888	Irrelevant

`compthrom`

0	No
1	Yes
9	Unknown
888	Irrelevant

`compinf`

0	No
1	Yes
9	Unknown
888	Irrelevant

`compneu`

0	No
1	Yes
9	Unknown
888	Irrelevant

`compx`

0	No
1	Yes
9	Unknown
888	Irrelevant

`letsel`

0	No
1	Yes
9	Unknown
888	Irrelevant

`icdg`

999	Admitted to the ICU, but the number of days is unknown
888	No admission to ICU
777	Irrelevant

`icreden`

1	Single-organ failure
2	Multi-organ failure
3	Admission to the ICU/MC for other reasons
9	Unknown
888	Irrelevant

`transf`

0	No
1	Yes
9	Unknown
888	Irrelevant

`heropn90`

0	No
1	Yes
9	Unknown
888	Irrelevant

`datheropn`

08-08-1808	No readmission

	09-09-1809	Readmitted, date unknown
	09-09-1809	Unknown
reintreq		
	0	No
	1	Yes
	888	Irrelevant
datreint		
	08-08-1808	Irrelevant
reintaard		
	1	Radiological intervention
	2	Endoscopy performed by a gastroenterologist
	3	Laparoscopy
	4	Laparotomy
	7	Other
	888	Irrelevant
reintsts		
	0	No
	1	Ileostomy drainage
	2	Terminal ileostomy
	3	Colostomy drainage
	4	Terminal colostomy
	5	Stoma, type unknown
	9	Unknown
	888	Irrelevant
datont		
	08-08-1808	Was not discharged because of watchful waiting
	09-09-1809	Discharged, date unknown
	11-11-1811	No discharge occurred
status90		
	0	Alive at discharge and 90 days after resection
	1	Died during hospitalisation or within 90 days of resection
	2	Unknown
doodoorz		
	1	Death following surgery
	2	Death caused by the tumour in question
	3	Death from other causes
	4	Death from unknown causes
	888	Irrelevant
klaar		
	0	Nee
	1	Ja
chirurgie_uri		

Table B.5: The variables with their options, belonging to the subdataset *surgery*.

Variable	Options	Meaning of Option
tumorrest		
	0	Yes
	1	No, in cases following polypectomy
	2	No, in the case of status after neoadjuvant therapy (ypT0)
	3	No, in the case of status following an appendectomy
	10	Does not satisfy condition
tumorrest2		
	0	Yes
	1	No, in cases following polypectomy
	2	No, in the case of status after neoadjuvant therapy (ypT0)

	3	No, in the case of status following an appendectomy
	4	No, in cases where the status is following transanal minimally invasive local excision (TAMIS/TEM)
	5	No, in cases where the status is following other resections/dissections
	10	Does not satisfy condition
`histtype`		
	1	Adenocarcinoma
	2	Mucinous adenocarcinoma
	3	Signet ring carcinoma
	77	Other
`stadpt`		
	0	Does not satisfy condition
	5	pTX
	7	pTis
	6	pT0
	1	pT1
	2	pT2
	3	pT3
	4	pT4
	11	pT4a
	12	pT4b
	13	yT0
	9	Unknown
`stadpnnew`		
	0	pN0
	1	pN1a
	2	pN1b
	3	pN1c
	4	pN2a
	5	pN2b
	6	pNx
	7	pN1
	8	pN2
	10	Does not satisfy condition
`stadpn`		
	0	pN0
	1	pN1
	2	pN2
	5	pNx, gland status cannot be assessed
	9	Unknown
	10	Does not satisfy condition
`histtype2`		
	0	Irrelevant
	1	Adenocarcinoma
	2	Mucinous adenocarcinoma
	3	Signet ring carcinoma
	77	Unknown
`stadpt2`		
	0	Irrelevant
	5	pTX
	7	pTis
	6	pT0
	1	pT1
	2	pT2
	3	pT3
	4	pT4
	11	pT4a
	12	pT4b

13	yT0
9	Unknown

`stadpn2`

0	pN0
1	pN1a
2	pN1b
3	pN1c
4	pN2a
5	pN2b
6	pNx
10	Irrelevant

`stadpm`

0	pM0
1	pM1
5	pMX
2	pM-
10	Irrelevant

`resectien`

0	R0 local excision
1	R1 local excision
2	R2 local excision
3	Microscopically radical (margin $>$ 1 mm)
4	Microscopically narrow radical with a margin of between 0 and 1 mm
5	Microscopic irradical with tumour in the stained resection margin
6	Macroscopic irradical
9	Unknown
10	Irrelevant

`cirmarge`

444	$>$20 mm
666	$>$10 mm
888	Irrelevant
999	Unknown

`dmarge`

999	Unknown
888	Irrelevant

`resectie2n`

0	R0 local excision
1	R1 local excision
2	R2 local excision
3	Microscopically radical (margin $>$ 1 mm)
4	Microscopically narrow radical with a margin of between 0 and 1 mm
5	Microscopic irradical with tumour in the stained resection margin
6	Macroscopic irradical
9	Unknown
10	Irrelevant

`cirmarge2`

444	$>$20 mm
666	$>$10 mm
888	Irrelevant
999	Unknown

`dmarge2`

999	Unknown
888	Irrelevant

`cirmargelok`

444	$>$20 mm
666	$>$10 mm
888	Irrelevant

	999	Unknown
`cirmargebasaal`		
	444	>20 mm
	666	>10 mm
	888	Irrelevant
	999	Unknown
`macrotme`		
	0	On the muscularis propria
	1	In the mesorectal fat
	2	On the mesorectal fascia
	8	Irrelevant, no cover flap
	10	Irrelevant
`numlym`		
	888	Irrelevant
`numlympos`		
	999	Unknown
	888	Irrelevant
`cirmargeklier`		
	444	>20 mm
	666	>10 mm
	888	Irrelevant
	999	Unknown
`cirmargeklier2`		
	444	>20 mm
	666	>10 mm
	888	Irrelevant
	999	Unknown
`tumdepositnew`		
	888	Irrelevant
`anginvnew_0`		
	0	No
	1	Yes
	888	Irrelevant
`anginv_1`		
	0	No
	1	Yes
	888	Irrelevant
`anginv_3`		
	0	No
	1	Yes
	888	Irrelevant
`diffgraad`		
	1	Good/moderate
	2	Bad
	9	Unknown
	888	Irrelevant
`paperfdarm`		
	0	No
	1	Yes
	9	Unknown
	888	Irrelevant
`regress`		
	0	Cannot be assessed
	1	Full regression
	2	Minimal flames
	3	Moderate regression
	4	Minor regression
	5	No signs of tumour regression

	8	Irrelevant
	9	Unknown
anginvnew2_0		
	0	No
	1	Yes
	888	Irrelevant
anginv2_1		
	0	No
	1	Yes
	888	Irrelevant
anginv2_3		
	0	No
	1	Yes
	888	Irrelevant
diffgraad2		
	1	Good/moderate
	2	Bad
	9	Unknown
	888	Irrelevant
paperfdarm2		
	1	No
	2	Yes
	9	Unknown
	888	Irrelevant
regress2		
	0	Cannot be assessed
	1	Full regression
	2	Minimal flames
	3	Moderate regression
	4	Minor regression
	5	No signs of tumour regression
	8	Irrelevant
	9	Unknown
	888	Irrelevant
braf		
	0	No
	1	Yes
	9	Unknown
ras		
	0	Mutation absent
	1	Mutation present
	9	Not specified/not described

Table B.6: The variables with their options, belonging to the subdataset *pathology*.

Variable	Options	Meaning of Option
radtiming		
	0	No
	1	Yes, 5x5 Gy prior to treatment
	2	Yes, neoadjuvant chemoradiotherapy
	3	Yes, post-operative following resection
	4	Yes, without any subsequent resection ('watch and wait' strategy)
	7	Radiotherapy
	8	Resection (after conversion)
	9	Unknown
datrad1		

`datprertxstop`		
	08-08-1808	Irrelevant
`rtxrestad`		
	0	No
	1	Yes
	9	Unknown
	888	Irrelevant
`rtxct`		
	0	ycT0
	1	ycT1
	2	ycT2
	3	ycT3ab
	4	ycT3cd
	5	ycT4
	888	Irrelevant
`rtxcn`		
	0	ycN0
	1	ycN1
	2	ycN2
	888	Irrelevant

Table B.7: The variables with their options, belonging to the subdataset *radiotherapy*.

Variable	Options	Meaning of Option
`datctx1`		
	08-08-1808	Irrelevant
`prectx`		
	0	No
	1	Yes
	9	Unknown
	10	Not mandatory before 2019, so not completed
	888	Irrelevant
`prectxaard`		
	1	As part of chemoradiotherapy
	2	For tumour reduction/staging of the primary tumour
	3	For the treatment of metastases
	7	Other
	10	Not mandatory before 2018, so not completed
	888	Irrelevant
`datctx1stop`		
	08-08-1808	Irrelevant
`datchemo`		
	08-08-1808	Irrelevant
	09-09-1909	Unknown
`chemopost`		
	0	No
	1	Yes
	9	Unknown
	888	Irrelevant
`chemo`		
	0	None
	1	HIPEC
	2	Adjuvant chemotherapy
	3	Palliative systemic therapy
	4	Other
	5	HIPEC followed by adjuvant

10	Not mandatory before 2018, so not completed
888	Irrelevant

`chemo1`

2	FOLFOX schedule
3	CAPOX schedule
4	Capecitabine
5	Uracil and Tegafur
7	Other
888	Irrelevant

`datchemostop`

08-08-1808	Irrelevant

`chemokuur`

888	Irrelevant

`chemokuuraanpas`

0	No
1	Dose reduction
2	Reduction in the number of treatment courses / treatment courses stopped earlier
3	Interrupted and resumed / postponed
4	One of the treatments has been discontinued
5	A combination of dose reduction and interruption and resumption / deferral
6	Dose reduction and one of the drugs has been discontinued
7	Pause and resume combination, and one of the devices has stopped
8	Other
9	Unknown
10	One of the treatments was not started at the beginning of the course
888	Irrelevant

`chemokuurreden`

0	No adjustment
11	Neutropenic fever / febrile neutropenia
12	Infection (confirmed presence of bacteria, viruses, fungi, etc.)
18	Haematological toxicity, other
21	Diarrhoea
22	Mucositis
23	Nausea/vomiting
24	Change in patient's weight
25	Constipation
28	Gastrointestinal, other
31	Neurotoxicity / polyneuropathy (PNP)
41	Hand-foot syndrome (HFS)
42	Papules and pustules
48	Skin toxicity, other
51	Cardiac
52	Pulmonary
53	Thrombosis, DVT or pulmonary embolism
54	Osteonecrosis of the jaw (ONJ)
55	Renal insufficiency
56	Liver insufficiency / liver failure
61	Progressive disease (tumour growth despite treatment)
62	No response (no change in tumour size or disease progression since the start of treatment)
71	Patient's condition
72	At the request of the patient/family
88	Other
99	Unknown
888	Irrelevant

`chemokuurgrad`

	0	No adverse event, so not applicable
	1	Disorders of the blood and lymphatic system
	2	Heart conditions
	3	Congenital, familial and genetic disorders
	4	Disorders of the ear and inner ear
	5	Endocrine disorders
	6	Eye conditions
	7	Gastrointestinal disorders
	8	General disorders and administration site conditions
	9	Hepatobiliary disorders
	10	Disorders of the immune system
	11	Infections and parasitic diseases
	12	Injuries, poisoning and procedural complications
	13	Research
	14	Nutritional and metabolic disorders
	15	Muscle and connective tissue disorders
	16	Benign, malignant and unspecified tumours (including cysts and polyps)
	17	Disorders of the nervous system
	18	Pregnancy, the postnatal and perinatal periods
	19	Psychiatric disorders
	20	Kidney and urinary tract disorders
	21	Genital and breast conditions
	22	Disorders of the respiratory tract, chest and mediastinum
	23	Disorders of the skin and subcutaneous tissue
	24	Social circumstances
	25	Surgical and medical procedures
	26	Vascular disorders
	30	Fatigue
	31	Pain
	32	Dehydration
	33	Diarrhoea
	34	Nausea
	35	Vomiting
	36	Anorexia - loss of appetite
	37	Oral mucositis
	38	Intestinal perforation
	39	Dysphagia
	40	Peripheral neuropathy
	41	Palmar-plantar erythrodysesthesia syndrome - hand-foot syndrome
	42	Febrile neutropenia - neutropenic fever
	43	Thromboembolic event - thromboembolism
	44	Decrease in left ventricular ejection fraction (LVEF) - heart failure
	45	High blood pressure
	46	Other
	88	Other
	99	Unknown
	888	Irrelevant
`systgeen`		
	1	Comorbidity
	2	Performance status / functional status / patient condition
	3	Social context / home situation
	4	Age
	5	Short life expectancy
	6	Refusal / patient's wishes / family's wishes
	7	Tumour volume / advanced disease
	77	Other
	99	Unknown

	888	Irrelevant
tumorgergeen		
	1	Comorbidity
	2	Performance status / functional status / patient condition
	3	Social context / home situation
	4	Age
	5	Short life expectancy
	6	Refusal / patient's wishes / family's wishes
	7	Tumour volume / advanced disease
	77	Other
	99	Unknown
	888	Irrelevant

Table B.8: The variables with their options, belonging to the subdataset *systemic therapy*.

C | Original Percentage Missing Values

In this appendix an overview is given of the original percentage of missing values (Not Available (NA) values) of all variables, so without any imputations or selections done. This is in Tables C.1, C.2, and C.3.

Variable	Original NA percentage
Patient	
id	0,00
gebdat	0,01
geslacht	0,01
datovl	96,82
Comorbidities	
datcom	36,16
commyo	58,98
comhartfaal	61,19
comperif	62,42
comcva	58,16
comdem	61,66
comlong	56,23
combind	60,85
comgiulcus	60,65
comlever	61,83
comdiam	53,71
comparalys	65,44
comnier	60,92
commalig	54,78
comhiv	62,40
Presentation	
uri	0,00
patient_uri	0,00
⋮	⋮
⋮	⋮
tumor_id	0,00
idaaba	0,98
lengte	1,03
gewicht	1,00
asascore	0,06
resdarm	0,86
stomavg	49,18
datpa1	1,59
presentatie_uri	0,00
updated	100,00
Diagnostics	
primcrc	0,00
scopnumb	0,73
diagn	55,07
locprim	0,00
histd	62,54
complpre	0,07
preanemie	69,56
preobstruct	69,92
preabces	70,44
scorect	22,76
scorecn	23,14
locprim2	97,00
scorect2	97,94
scorecn2	98,96
scorecm	22,52
⋮	⋮

Table C.1: Original percentage of NA values per variable. Table 1 of 3.

Variable	Original NA percentage
Diagnostics	
⋮	⋮
prelocmri	53,15
prelocmri2	99,89
scopdistmm	73,76
scopdist2mm	99,67
afstmrf	81,11
afstmrf2	99,91
prelocstad	83,10
emimri	91,32
emimri2	99,95
genan	57,88
imuun	76,07
mlh1	86,32
pms2	86,33
hyperm	98,19
msh2	86,36
msh6	86,38
diagnostiek_uri	0,00
Surgery	
datbez1	3,87
verwezen	36,91
darmperf	45,40
ileusblow	99,18
perfperi	99,16
perfscopie	99,18
prehirstomanew	0,00
datstoma	98,11
prechirstent	61,18
datstent	95,57
prechirmetanew	93,66
datprechirmetaeen	99,75
datprechirmetatwee	99,68
prechirappendix	61,27
datcoloscex	97,48
chirenkel	61,05
datok	0,00
typok	0,12
procok	0,05
ingreep	1,16
conversien	38,16
aprtyp	97,23
typprocok	99,55
datok2	99,84
⋮	⋮
procok2	97,50
procok3	99,70
ingreep2	99,63
conversien2	99,73
aprtyp2	99,94
resadd	2,89
locaddmaag	96,96
locaddmilt	96,96
locaddduod	96,95
locadddoverig	96,95
locaddoment	97,50
locaddbuikw	97,49
locaddvag	97,50
locadduter	96,95
locaddovarl	97,50
locaddovarr	97,50
locadduretl	96,95
locadduretr	96,95
locaddblaas	96,94
locaddprost	96,95
locaddvesl	96,25
locaddvesr	96,25
locaddvesb	96,25
locadddund	96,95
locaddsacr	96,95
locaddoverig	96,23
resadm	1,30
resomm	96,90
reslem	96,85
respem	96,91
reslmm	96,94
resbum	96,93
perrtxchemo	32,16
anastom1	9,48
stoma	2,20
stomahef	96,46
compl90	40,96
compchir	72,94
datcompchir	94,56
chirnaad	90,44
chirnaadoccult	95,13
chirabces	93,39
chirbloeding	93,48
chirileus	93,27
chirgastro	98,16
chirfascie	93,54
chirdarmperf	93,58
chirlekkage	93,58
chirwondinf	93,41
⋮	⋮

Table C.2: Original percentage of NA values per variable. Table 2 of 3.

Variable	Original NA percentage
Surgery	
⋮	⋮
datcompl	93,44
comppul	75,86
compcar	75,96
compthrom	76,29
compinf	75,84
compneu	76,21
compx	75,70
letsel	82,85
icdg	17,73
icreden	93,94
transf	6,84
heropn90	62,37
datheropn	94,86
reintreq	82,38
datreint	91,68
reintaard	90,88
reintst	90,53
datont	0,90
status90	58,32
doodoorz	99,78
klaar	1,10
chirurgie_uri	0,00
Pathology	
tumorrest	91,52
tumorrest2	99,76
histtype	2,98
stadpt	1,53
stadpnnew	3,29
stadpn	40,97
histtype2	97,13
stadpt2	97,16
stadpn2	98,92
stadpm	2,40
resectien	2,61
cirmarge	75,71
dmarge	97,43
resectie2n	97,55
cirmarge2	99,61
dmarge2	99,97
cirmargelok	99,16
cirmargebasaal	98,63
⋮	⋮
⋮	⋮
macrotme	92,05
numlym	3,21
numlympos	3,29
cirmargeklier	98,36
cirmargeklier2	99,96
tumdepositnew	74,15
anginvnew_0	9,07
anginv_1	50,00
anginv_3	56,71
diffgraad	18,27
paperfdarm	62,04
regress	64,02
anginvnew2_0	97,99
anginv2_1	98,87
anginv2_3	98,95
diffgraad2	99,00
paperfdarm2	99,02
regress2	99,01
braf	63,52
ras	63,87
Radiotherapy	
radtiming	4,73
datrad1	83,62
datprertxstop	94,91
rtxrestad	95,35
rtxct	96,78
rtxcn	96,66
Systemic Therapy	
datctx1	91,41
prectx	46,70
prectxaard	95,62
datctx1stop	96,27
datchemo	92,36
chemopost	93,42
chemo	33,43
chemo1	83,70
datchemostop	95,56
chemokuur	98,81
chemokuuraanpas	98,96
chemokuurreden	99,02
chemokuurgrad	99,20
systgeen	100,00
tumorgergeen	100,00
Other	
charlgrp	0,00
type_ziekenhuis	1,39

Table C.3: Original percentage of NA values per variable. Table 3 of 3.

D | Variables with Conditions from Datadictionary

The variables with their corresponding conditions from the datadictionary are in Table D.1 for subdataset *diagnostics*, in Table D.2 for subdataset *surgery*, in Table D.3 for subdataset *pathology*, in Table D.4 for subdataset *radiotherapy*, and in Table D.5 for subdataset *systemic therapy*.

Variable	Condition from datadictionary
preanemie	[complpre] == 1
preobstruct	[complpre] == 1
preabces	[complpre] == 1
locprim2	[scopnumb]==2 || [scopnumb]==3
scorect2	[scopnumb]==2 || [scopnumb]==3
scorecn2	[scopnumb]==2 || [scopnumb]==3
prelocmri	[locprim]==9 || ([locprim]==8 && totime([../datpa1])<totime('2020-01-01') && [../datpa1]!==null)
prelocmri2	([scopnumb]==2 || [scopnumb]==3) && ([locprim2]==9 || ([locprim2]==8 && totime([../datpa1])<totime('2020-01-01') && [../datpa1]!==null))
scopdistmm	[locprim]==9 || ([locprim]==8 && totime([../datpa1])<totime('2020-01-01') && [../datpa1]!==null)
scopdist2mm	([scopnumb]==2 || [scopnumb]==3) && ([locprim2]==9 || ([locprim2]==8 && totime([../datpa1])<totime('2020-01-01') && [../datpa1]!==null))
afstmrf	(([locprim]==9 && totime([../datpa1])<totime('2024-01-01')) || ([locprim] == 9 && ([scorect]==2 || [scorect]==3) && totime([../datpa1])>=totime('2020-01-01')) || ([locprim]==8 && totime([../datpa1])<totime('2024-01-01') && [../datpa1]!==null))
afstmrf2	(([scopnumb]==2 || [scopnumb]==3) && [locprim2]==9 && totime([../datpa1])<totime('2024-01-01')) || (([scopnumb]==2 || [scopnumb]==3) && [locprim2]==9 && ([scorect2] == 2 || [scorect2] == 3) && totime([../datpa1])>=totime('2020-01-01')) || (([scopnumb]==2 || [scopnumb]==3) && (([locprim2]==8 && totime([../datpa1])<totime('2024-01-01') && [../datpa1]!==null)))
prelocstad	[locprim]==9 && totime([../datpa1]) < totime('2022-01-01')
emimri	(([prelocmri]==1 || [prelocstad]==3 || [prelocstad]==4) && totime([../datpa1])<totime('2024-01-01')) || ([prelocmri]==1 && [scorect]==3 && totime([../datpa1])>=totime('2020-01-01'))
emimri2	(([scopnumb]==2 || [scopnumb]==3) && [prelocmri2]==1 && totime([../datpa1])<totime('2024-01-01')) || (([scopnumb]==2 || [scopnumb]==3) && [prelocmri2]==1 && [scorect2]==3 && totime([../datpa1])>=totime('2020-01-01'))

mlh1	[imuun]==1
pms2	[imuun]==1
hyperm	[mlh1]==1 && [pms2]==1
msh2	[imuun]==1
msh6	[imuun]==1

Table D.1: *Diagnostics.* Variables with corresponding conditions from datadictionary.

Variable	Condition from datadictionary
ileusblow	[darmperf]==1
perfperi	[darmperf]==1
perfscopie	[darmperf]==1
datstoma	[prechirstomanew]==1 || [prechirstomanew]==2 || [prechirstomanew]==3
prechirstent	totime([../../datpa1]) < totime('01-01-2024')
prechirmetanew	[../scorecm] == 1
datprechirmetaeen	[prechirmetanew] ==1 || [prechirmetanew] ==2 || [prechirmetanew] ==3
datprechirmetatwee	[prechirmetanew] ==1 || [prechirmetanew] ==2 || [prechirmetanew] ==3
prechirappendix	totime([../../datpa1]) < totime('01-01-2024')
datcoloscex	[prechircoloscex] >= 1
conversien	[ingreep]==2 || [ingreep]==6 || [ingreep]==7
aprtyp	(totime([datok])<totime('2019-01-01') && [procok]==10) || [procok]==22
typprocok	[procok]==77
datok2	(totime([datok]) < totime('2018-01-01') && [../scopnumb]==2 || [../scopnumb]==3) || [chirenkel]==='0'
procok2	[../scopnumb]==2 || [../scopnumb]==3
procok3	[../scopnumb]==1 && [chirenkel]==='0'
ingreep2	(totime([datok]) < totime('2018-01-01') && ([../scopnumb]==2 || [../scopnumb]==3)) || [chirenkel]==='0'
conversien2	((totime([datok]) < totime('2018-01-01') &&([../scopnumb]==2 || [../scopnumb]==3)) || [chirenkel]==='0') && ([ingreep2]==2 || [ingreep2]==6 || [ingreep2]==7)
aprtyp2	((totime([datok]) < totime('2018-01-01') &&([../scopnumb]==2 || [../scopnumb]==3)) || [chirenkel]==='0') && ([procok2]==22 || [procok3]==22)
resadd	[procok]!=8 && [procok]!=13 && (totime([datok]) < totime('2019-01-01') || ([ingreep]!=3 && [ingreep]!=4 && [ingreep2]!=3 && [ingreep2]!=4)) && (([procok]!=20 && [procok]!=26) || (totime([datok])>totime('2022-01-01') && (([procok2]!==null && [procok2]!=20 && [procok2]!=26) || ([procok3]!==null && [procok3]!=20 && [procok3]!=26))))
locaddmaag	[resadd]==1 || (totime([datok]) < totime('2022-01-01') && [resadd]==2)
locaddmilt	[resadd]==1 || (totime([datok]) < totime('2022-01-01') && [resadd]==2)

locaddduod	[resadd]==1 || (totime([datok]) < totime('2022-01-01') && [resadd]==2)
locadddoverig	[resadd]==1 || (totime([datok]) < totime('2022-01-01') && [resadd]==2)
locaddoment	[resadd]==1 || [resadd]==2
locaddbuikw	[resadd]==1 || [resadd]==2
locaddvag	[resadd]==1 || [resadd]==2
locadduter	[resadd]==1 || (totime([datok]) < totime('2022-01-01') && [resadd]==2)
locaddovarl	[resadd]==1 || [resadd]==2
locaddovarr	[resadd]==1 || [resadd]==2
locadduretl	[resadd]==1 || (totime([datok]) < totime('2022-01-01') && [resadd]==2)
locadduretr	[resadd]==1 || (totime([datok]) < totime('2022-01-01') && [resadd]==2)
locaddblaas	[resadd]==1 || (totime([datok]) < totime('2022-01-01') && [resadd]==2)
locaddprost	[resadd]==1 || (totime([datok]) < totime('2022-01-01') && [resadd]==2)
locaddvesl	[resadd]==1 || [resadd]==2
locaddvesr	[resadd]==1 || [resadd]==2
locaddvesb	[resadd]==1 || [resadd]==2
locadddund	[resadd]==1 || (totime([datok]) < totime('2022-01-01') && [resadd]==2)
locaddsacr	[resadd]==1 || (totime([datok]) < totime('2022-01-01') && [resadd]==2)
locaddoverig	[resadd]==1 || [resadd]==2
resadm	[procok]!=8 && ([procok]!=13 && ([procok]!=20 && [procok]!=26) || totime([datok]) >= totime('2021-01-01'))
resomm	[resadm]==1
reslem	[resadm]==1
respem	[resadm]==1
reslmm	[resadm]==1
resbum	[resadm]==1
perrtxchemo	[procok]!==null && [procok]!=20 && [procok]!=26
anastom1	[procok]!==null && (totime([datok])>totime('2019-01-01') || ([procok]!=8 && [procok]!=13)) && [procok]!=20 && [procok]!=22 && [procok]!=26
stoma	[procok]!==null && (totime([datok])>totime('2019-01-01') || ([procok]!=8 && [procok]!=13)) && [procok]!=20 && [procok]!=26
stomahef	([../../stomavg]==1 || [../../stomavg]==2 || [../../stomavg]==3 || [../../stomavg]==9 || [prechirstomanew]==1 || [prechirstomanew]==2 || [prechirstomanew]==3 || [prechirstomanew]==8 || [prechirstomanew]==9) && ([procok]!==null && [procok]!=8 && [procok]!=13 && [procok]!=20 && [procok]!=26)
compchir	[compl90]==1
datcompchir	[compchir]==1
chirnaad	[compchir]==1
chirnaadoccult	[compchir]==1
chirabces	[compchir]==1

chirbloeding	[compchir]==1
chirileus	[compchir]==1
chirgastro	[compchir]==1
chirfascie	[compchir]==1
chirdarmperf	[compchir]==1
chirlekkage	[compchir]==1
chirwondinf	[compchir]==1
datcompl	[compl90]==1
comppul	[compl90]==1
compcar	[compl90]==1
compthrom	[compl90]==1
compinf	[compl90]==1
compneu	[compl90]==1
compx	[compl90]==1
letsel	[compl90]==1
icdg	(totime([datok])<totime('2021-01-01') && [procok]!=8) || (totime([datok])>=totime('2021-01-01') && [procok]!=20 && [procok]!=26)
icreden	[icdg] >= 1 && ([icdg] != 999 && [icdg] != 888)
transf	(totime([datok])<totime('2021-01-01') && [procok]!=8) || (totime([datok])>=totime('2021-01-01') && [procok]!=20 && [procok]!=26)
heropn90	totime([../../datpa1]) < totime('01-01-2024')
reintreq	[compchir]==1
datreint	[reintreq] == 1
reintaard	[reintreq] == 1
reintst	[reintreq] == 1
doodoorz	[status90]==1 || [../../../datovl]!==null

Table D.2: *Surgery.* Variables with corresponding conditions from datadictionary.

Variable	Condition from datadictionary
tumorrest	[../procok]!==null
tumorrest2	(([../../scopnumb]==2 || [../../scopnumb]==3) && [../procok2]!==null) || [../chirenkel]==='0'
stadpt	[../procok]!==null
stadpnnew	[../procok]!==null
stadpn	[../procok]!==null && [../procok]!=8 && [../procok]!=13 && totime([../datok]) < totime('2018-01-01')
histtype2	[../../scopnumb]==2 || [../../scopnumb]==3
stadpt2	[../../scopnumb]==2 || [../../scopnumb]==3
stadpn2	[../../scopnumb]==2 || [../../scopnumb]==3
stadpm	[../procok]!==null && (totime([../datok])>totime('2021-01-01') || [../procok]!=8 && [../procok]!=13)
resectien	[../procok]!==null && [../procok]!=8 && [../procok]!=13
cirmarge	[../procok]!==null && [../../locprim]==9 && [../procok]!=8 && (totime([../datok])<totime('2021-01-01') || [../procok]!=20 || ([../procok3]!==null && [../procok3] != 20))
dmarge	([../procok]==15 || [../../locprim]==9) && [../ingreep]!=5
resectie2n	(([../../scopnumb]==2 || [../../scopnumb]==3) && ([../procok2]!==null && [../procok2]!=8 && [../procok2]!=13)) || [../chirenkel]==='0'

cirmarge2	(([../../scopnumb]==2 || [../../scopnumb]==3) && ([../../locprim2]==9 && [../procok2]!=20 && [../procok2]!=26)) || ([../../scopnumb]==1 &&[../../locprim]==9 && [../chirenkel]===’0’ && [../procok3]!=20 && [../procok3]!=26 && [../procok3]!==null)
dmarge2	([../procok2]==15 || [../../locprim2]==9) && [../ingreep2]!=5
cirmargelok	(totime([../datok])<totime(’2021-01-01’) && ([../procok]==8 || [../procok]==9 || [../procok]==10 || [../procok]==13 || [../procok]==20 || [../procok]==26 || [../procok2]==8 || [../procok2]==9 || [../procok2]==10 || [../procok2]==13 || [../procok2]==20 || [../procok2]==26 || [../procok3]==8 || [../procok3]==9 || [../procok3]==10 || [../procok3]==13 || [../procok3]==20 || [../procok3]==26)) || (totime([../datok])>=totime(’2021-01-01’) && ([../procok]==20 || [../procok2]==20 || [../procok3]==20))
cirmargebasaal	(totime([../datok])<totime(’2019-01-01’) && ([../procok]==8 || [../procok]==9 || [../procok]==10 || [../procok]==13 || [../procok2]==8 || [../procok2]==9 || [../procok2]==10 || [../procok2]==13 || [../procok3]==8 || [../procok3]==9 || [../procok3]==10 || [../procok3]==13)) || (([../procok]==20 || [../procok]==26 || [../procok2]==20 || [../procok2]==26 || [../procok3]==20 || [../procok3]==26) && (totime([../datok])<totime(’2021-01-01’) || [../ingreep]!=6))
macrotme	(totime([../datok])<totime(’2019-01-01’) && ([../procok]==9 || [../procok]==10 || [../procok2]==9 || [../procok2]==10 || [../procok3]==9 || [../procok3]==10)) || [../procok]==16 || [../procok]==22 || [../procok2]==16 || [../procok2]==22 || [../procok3]==16 || [../procok3]==22
numlym	([../procok]!==null && [../procok]!=8 && [../procok]!=13 && [../procok]!=20 && [../procok]!=26) || (totime([../datok])>=totime(’2020-01-01’) &&(([../procok2]!==null || [../procok3]!==null) && ([../procok2]!=20 && [../procok2]!=26 && [../procok3]!=20 && [../procok3]!=26))) || (totime([../datok])>=totime(’2021-01-01’) &&([../procok]==20 || [../procok]==26 || [../procok2]==20 || [../procok2]==26 || [../procok3]==20 || [../procok3]==26) && ([stadpnnew]!=6 || ([stadpn2]!==null && [stadpn2]!=6)))
numlympos	([../procok]!==null && [../procok]!=8 && [../procok]!=13 && [../procok]!=20 && [../procok]!=26) || (totime([../datok])>=totime(’2020-01-01’) &&(([../procok2]!==null || [../procok3]!==null) && ([../procok2]!=20 && [../procok2]!=26 && [../procok3]!=20 && [../procok3]!=26))) || (totime([../datok])>=totime(’2021-01-01’) && ([stadpnnew]!=6 || ([stadpn2]!=6 && [stadpn2]!==null)))
cirmargeklier	[../../locprim] == 9 && ([stadpnnew] !== ’0’ && [../procok] != 20)
cirmargeklier 2	([../../locprim2] == 9 && ([stadpn2] !== ’0’ && [../procok2] != 20)) || ([../../scopnumb]==1 && [../../locprim]==9 && [../chirenkel]===’0’ && [../procok3]!=20 && [../procok3]!==null && [stadpnnew] !== ’0’)

tumdepositnew	([../procok]!==null && [../procok]!=8 && [../procok]!=13 && [../procok]!=20 && [../procok]!=26) && (totime([../datok])<totime('2020-01-01') || ([../procok2]!==null &&[../procok3]===null) || ([../procok2]!=8 && [../procok2]!=13 && [../procok2]!=26 && [../procok3]!=8 && [../procok3]!=13 && [../procok3]!=26))
anginvnew-0	[../procok]!==null
anginv-1	[../procok]!==null
anginv-3	[../procok]!==null
diffgraad	[../procok]!==null
paperfdarm	[../procok]!==null && [../procok]!=20
anginvnew2-0	[../../scopnumb]==2 || [../../scopnumb]==3
anginv2-1	[../../scopnumb]==2 || [../../scopnumb]==3
anginv2-3	[../../scopnumb]==2 || [../../scopnumb]==3
diffgraad2	[../../scopnumb]==2 || [../../scopnumb]==3
paperfdarm2	([../../scopnumb]==2 || [../../scopnumb]==3) && [../procok2] != 20
regress2	[../../scopnumb]==2 || [../../scopnumb]==3

Table D.3: *Pathology.* Variables with corresponding conditions from datadictionary.

Variable	Condition from datadictionary
datprertxstop	[radtiming]!==null && [radtiming]!=='0'
rtxrestad	[radtiming]!==null && [radtiming]!=='0'
rtxct	[rtxrestad]==1
rtxcn	[rtxrestad]==1

Table D.4: *Radiotherapy.* Variables with corresponding conditions from datadictionary.

Variable	Condition from datadictionary
prectx	[../locprim] !== null && totime([datctx1]) != totime('08-08-1808')
prectxaard	[prectx]==1 && totime([datctx1]) != totime('08-08-1808')
datctx1stop	[prectx]==1 && totime([datctx1]) != totime('08-08-1808')
chemopost	[datchemo] !=null && totime([datchemo]) != totime('08-08-1808')
chemo	[chemopost]==1 || totime([../../datpa1]) < totime('2018-01-01')
chemo1	[chemo]!==null && [chemo]!=='0' && [chemo]!=='1'
datchemostop	[chemo]!==null && [chemo]!=='0' && [chemo]!=1 && [chemo]!=5
chemokuur	[chemofilter]==1 && ([chemo]==2)
chemokuuraanpas	[chemofilter]==1 && ([chemo]==2)
chemokuurreden	[chemofilter]==1 && ([chemo]==2)
chemokuurgrad	[chemofilter]==1 && ([chemo]!==null && [chemo]!=='0' && [chemo]!=1 && [chemo]!=5)
systgeen	[chemofilter]==1 && [chemo]==='0'
tumorgergeen	[chemofilter]==1 && [chemo]==='0'

Table D.5: *Systemic Therapy.* Variables with corresponding conditions from datadictionary.

E | Missing Values Before and After Imputations from Datadictionary

This appendix shows for the dataset with elective surgeries only the percentage of missing values (Not Available (NA) values) before and after the imputations done with the data dictionary. This is in Tables E.1, E.2, and E.3. The values are marked yellow if they are above 5%.

Variable	Before imputing	After imputing
Diagnostics		
preanemie	75,44	0,46
preobstruct	76,09	1,12
preabces	76,36	1,39
locprim2	96,93	0,06
scorect2	97,87	1,01
scorecn2	98,92	2,05
prelocmri	51,38	0,16
prelocmri2	99,87	1,17
scopdistmm	71,15	19,87
scopdist2mm	99,65	1,02
afstmrf	79,16	32,96
afstmrf2	99,90	1,19
prelocstad	81,47	6,84
emimri	90,38	17,21
emimri2	99,94	0,04
mlh1	85,85	0,17
pms2	85,86	0,18
hyperm	98,11	0,10
msh2	85,88	0,20
msh6	85,90	0,22
Surgery		
ileusblow	99,72	0,13
perfperi	99,70	0,12
⋮	⋮	⋮
⋮	⋮	⋮
perfscopie	99,72	0,13
datstoma	97,98	0,05
prechirstent	59,96	54,57
prechirmetanew	94,23	0,00
datprechirmetaeen	99,73	0,43
datprechirmetatwee	99,65	0,36
prechirappendix	60,05	4,43
datcoloscex	98,93	1,68
conversien	33,58	7,64
aprtyp	96,92	4,99
typprocok	99,59	0,52
datok2	99,84	2,28
procok2	97,44	0,58
procok3	99,70	96,07
ingreep2	99,63	2,03
conversien2	99,72	0,04
aprtyp2	99,93	0,08
resadd	3,04	1,54
locaddmaag	97,12	4,27
locaddmilt	97,12	4,27
locaddduod	97,12	4,27
locadddoverig	97,11	4,26
locaddoment	97,58	5,44
locaddbuikw	97,11	5,43
locaddvag	97,58	5,44
⋮	⋮	⋮

Table E.1: Percentage of NA values before and after imputation with conditions from datadictionary. Table 1 of 3.

Variable	Before imputing	After imputing
Surgery		
⋮	⋮	⋮
locadduter	97,11	4,26
locaddovarl	97,58	5,44
locaddovarr	97,58	5,44
locadduretl	97,11	4,26
locadduretr	97,11	4,26
locaddblaas	97,10	4,26
locaddprost	97,11	4,26
locaddvesl	96,43	4,29
locaddvesr	96,43	4,29
locaddvesb	96,43	4,29
locadddund	97,12	4,27
locaddsacr	97,11	4,26
locaddoverig	96,41	4,27
resadm	1,43	0,13
resomm	96,99	0,18
reslem	96,92	0,11
respem	96,99	0,18
reslmm	97,02	0,21
resbum	97,01	0,20
perrtxchemo	31,34	30,11
anastom1	10,42	0,44
stoma	2,39	0,40
stomahef	96,24	0,89
compchir	74,03	3,17
datcompchir	97,88	15,11
chirnaad	90,49	7,72
chirnaadoccult	95,12	12,34
chirabces	93,40	10,63
chirbloeding	93,48	10,71
chirileus	93,27	10,50
chirgastro	98,11	15,34
chirfascie	93,55	10,78
chirdarmperf	93,59	10,82
chirlekkage	93,59	10,81
chirwondinf	93,42	10,65
datcompl	94,74	23,88
comppul	76,92	6,07
compcar	77,02	6,16
compthrom	77,30	6,44
compinf	76,89	6,04
compneu	77,23	6,37
⋮	⋮	⋮
⋮	⋮	⋮
compx	76,77	5,91
letsel	83,25	12,40
icdg	17,50	16,06
icreden	94,63	6,00
transf	7,17	5,74
heropn90	61,42	55,93
reintreq	82,80	0,03
datreint	96,77	6,28
reintaard	91,05	0,56
reintsts	90,69	0,20
doodoorz	99,82	2,34
Pathology		
tumorrest	91,11	91,06
tumorrest2	99,75	2,70
stadpt	1,58	1,53
stadpnnew	3,36	3,31
stadpn	42,30	0,26
histtype2	97,07	0,20
stadpt2	97,10	0,23
stadpn2	98,88	2,01
stadpm	2,50	2,45
resectien	2,52	1,71
cirmarge	73,28	1,14
dmarge	97,17	25,75
resectie2n	97,49	0,41
cirmarge2	99,58	0,16
dmarge2	99,96	0,37
cirmargelok	99,06	1,25
cirmargebasaal	98,48	0,72
macrotme	91,20	6,76
numlym	3,38	2,18
numlympos	3,48	2,28
cirmargeklier	98,19	4,78
cirmargeklier2	99,96	0,32
tumdepositnew	73,36	71,32
anginvnew_0	9,19	9,14
anginv_1	49,89	49,84
anginv_3	55,57	55,52
diffgraad	17,47	17,42
paperfdarm	61,07	59,86
anginvnew2_0	97,94	1,07
anginv2_1	98,85	1,98
anginv2_3	98,92	2,05
diffgraad2	98,96	2,10
paperfdarm2	98,99	2,11
regress2	98,98	2,11

Table E.2: Percentage of NA values before and after imputation with conditions from datadictionary. Table 2 of 3.

Variable	Before imputing	After imputing
Radiotherapy		
datprertxstop	97,79	17,03
rtxrestad	94,85	14,08
rtxct	96,42	0,40
rtxcn	96,30	0,27
Systemic Therapy		
prectx	44,49	44,49
prectxaard	95,23	3,36
⋮	⋮	⋮
⋮	⋮	⋮
datctx1stop	98,45	6,57
chemopost	93,57	1,14
chemo	34,85	0,91
chemo1	84,81	38,56
datchemostop	95,61	49,31
chemokuur	98,82	15,79
chemokuuraanpas	98,97	15,94
chemokuurreden	99,03	16,00
chemokuurgrad	99,23	52,93
systgeen	100,00	44,77
tumorgergeen	100,00	44,77

Table E.3: Percentage of NA values before and after imputation with conditions from datadictionary. Table 3 of 3.

F | Graphs of Univariate Data Analysis

In this appendix are all the graphs that belong to the univariate data analysis, conducted in Section 5.10. In Section F.1 the 108 discrete variables are treated. In Section F.2 we move to the 7 continuous variables.

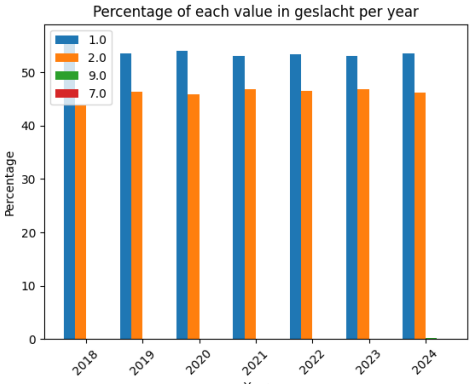
F.1 Discrete variables

In Section F.1.1 the percentage of each value per year for the discrete variables is shown. In Section F.1.2 the percentage change of each value per year for the discrete variables is shown.

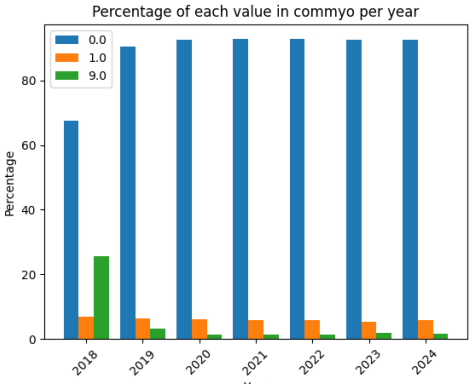
The 108 discrete variables are `geslacht`, `commyo`, `comhartfaal`, `comperif`, `comcva`, `comdem`, `comlong`, `combind`, `comgiulcus`, `comlever`, `comdiam`, `comparalys`, `comnier`, `commalig`, `comhiv`, `asascore`, `resdarm`, `stomavg`, `primcrc`, `scopnumb`, `diagn`, `locprim`, `histd`, `complpre`, `preanemie`, `preobstruct`, `preabces`, `scorect`, `scorecn`, `locprim2`, `scorect2`, `scorecn2`, `scorecm`, `prelocmri`, `prelocmri2`, `verwezen`, `darmperf`, `ileusblow`, `perfperi`, `perfscopie`, `prechirstomanew`, `prechirstent`, `prechirmetanew`, `prechirappendix`, `prechircoloscex`, `chirenkel`, `typok`, `procok`, `ingreep`, `conversien`, `aprtyp`, `typprocok`, `procok2`, `ingreep2`, `conversien2`, `aprtyp2`, `resadd`, `locaddmaag`, `locaddmilt`, `locaddduod`, `locadddoverig`, `locaddoment`, `locaddbuikw`, `locaddvag`, `locadduter`, `locaddovarl`, `locaddovarr`, `locadduretl`, `locadduretr`, `locaddblaas`, `locaddprost`, `locaddvesl`, `locaddvesr`, `locaddvesb`, `locadddund`, `locaddsacr`, `locaddoverig`, `resadm`, `resomm`, `reslem`, `respem`, `reslmm`, `resbum`, `perrtxchemo`, `anastom1`, `stoma`, `stomahef`, `radtiming`, `prectx`, `charlgrp`, `type_ziekenhuis`, `chirnaad`, `chirnaadoccult`, `chirabces`, `chirbloeding`, `chirileus`, `chirgastro`, `chirfascie`, `chirdarmperf`, `chirlekkage`, `chirwondinf`, `compppul`, `compcar`, `compthrom`, `compinf`, `compneu`, `compx`, and `severeCompl`.

F.1.1 Percentage of values per year

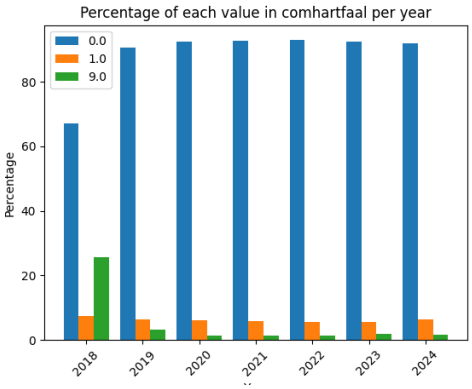
In this section, we show the plots of the discrete variables that show per year how much (in percentage) a value occurred for the variable.



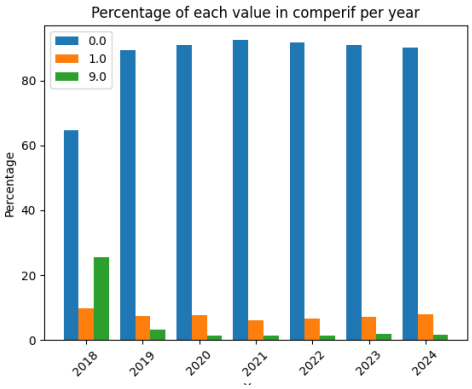
(a) geslacht



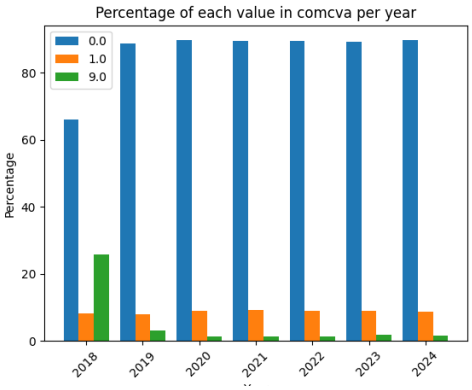
(b) commyo



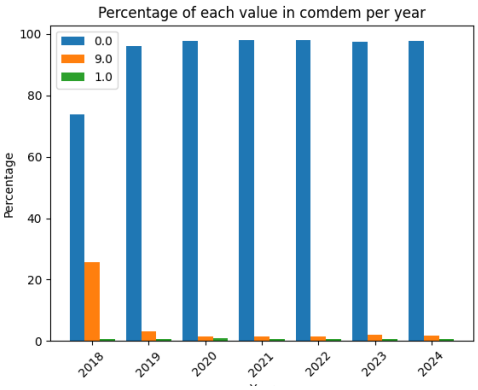
(c) comhartfaal



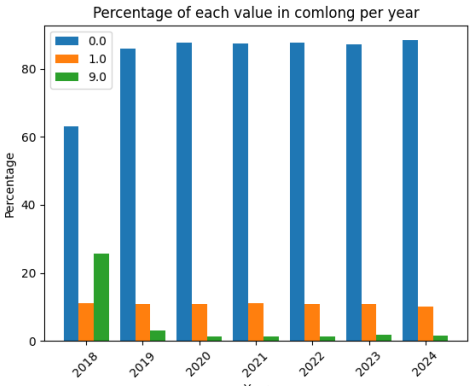
(d) comperif



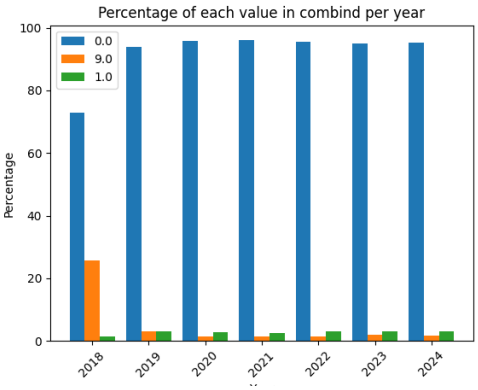
(e) comcva



(f) comdem

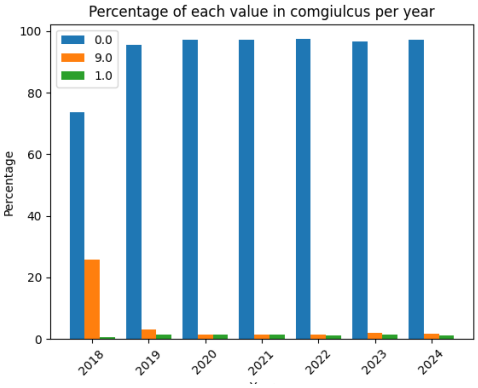


(g) comlong

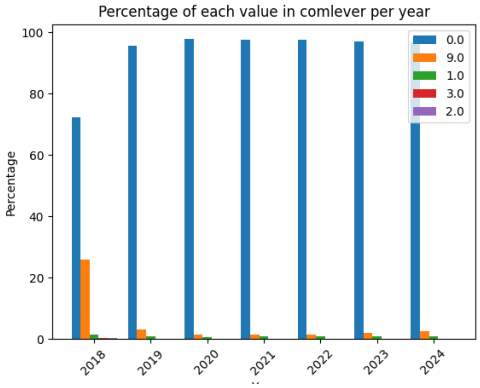


(h) combind

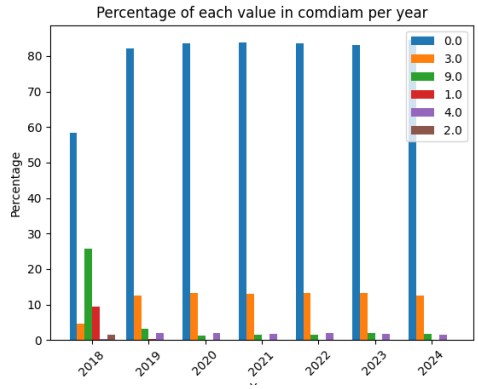
Figure F.1: Plots of the values of the discrete variables per year. Part 1/14.



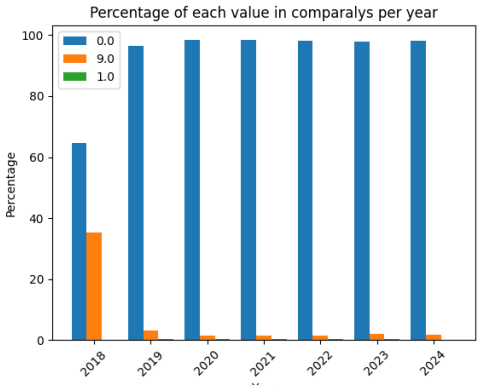
(a) comgiulcus



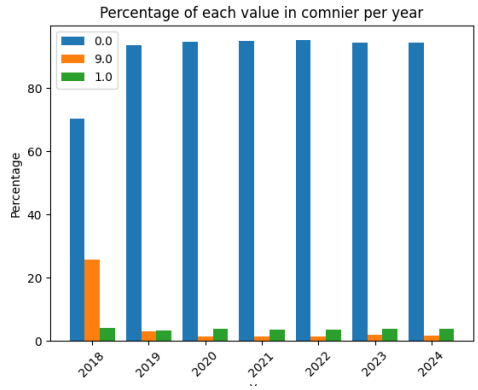
(b) comlever



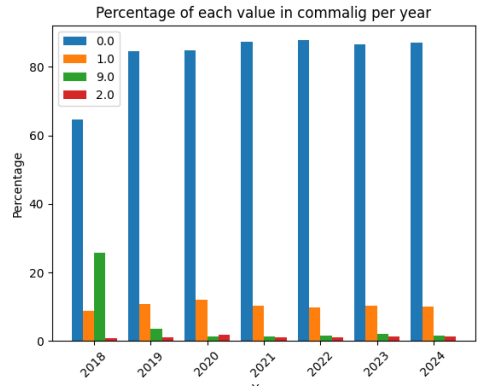
(c) comdiam



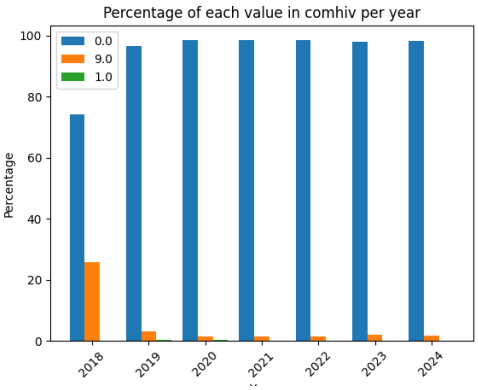
(d) comparalys



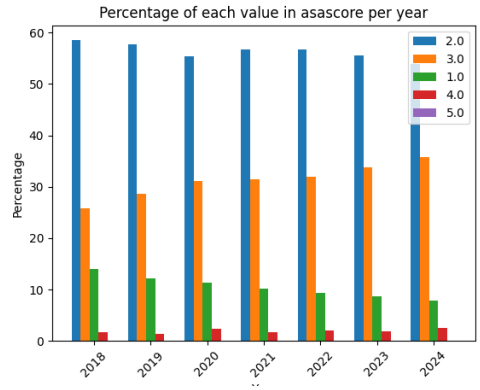
(e) comnier



(f) commalig

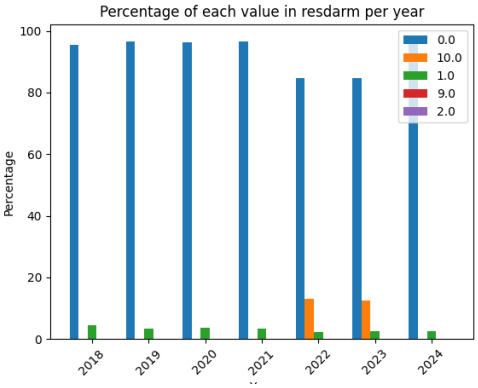


(g) comhiv

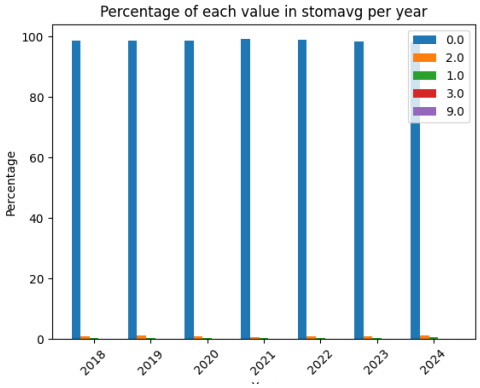


(h) asascore

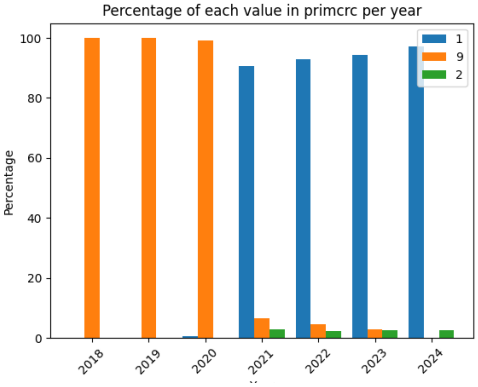
Figure F.2: Plots of the values of the discrete variables per year. Part 2/14.



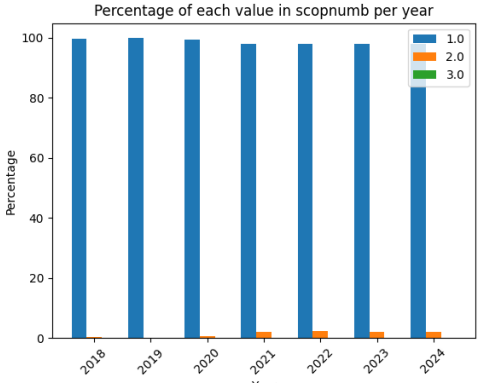
(a) `resdarm`



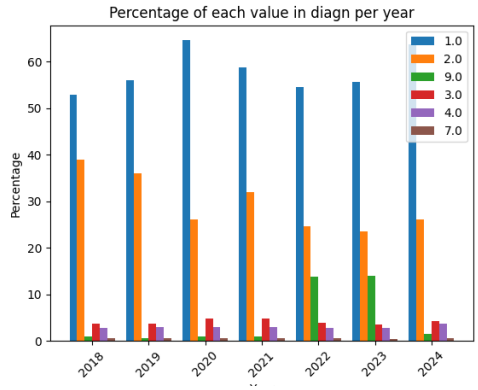
(b) `stomavg`



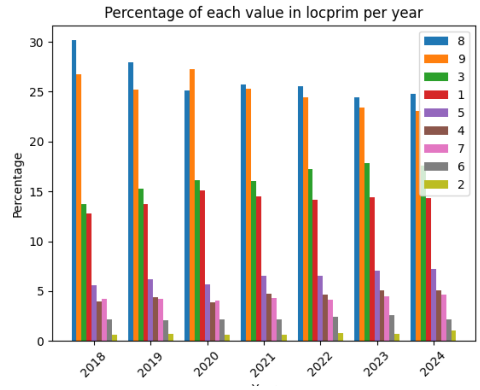
(c) `primcrc`



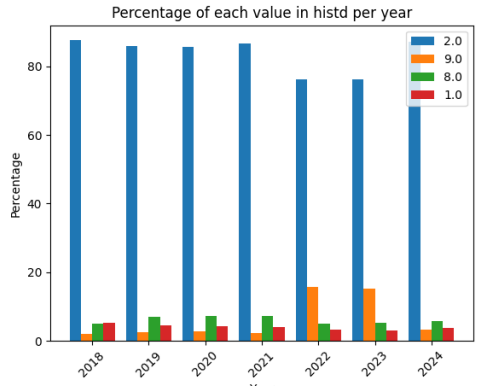
(d) `scopnumb`



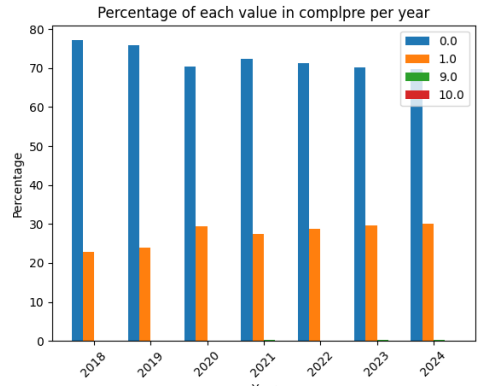
(e) `diagn`



(f) `locprim`

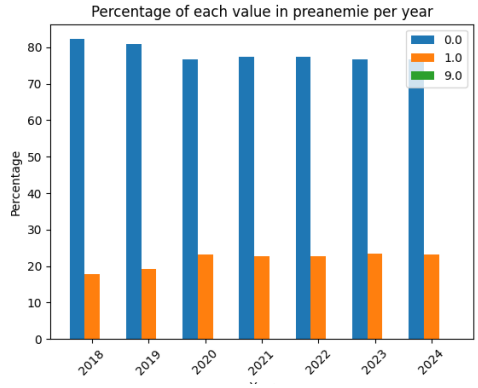


(g) `histd`

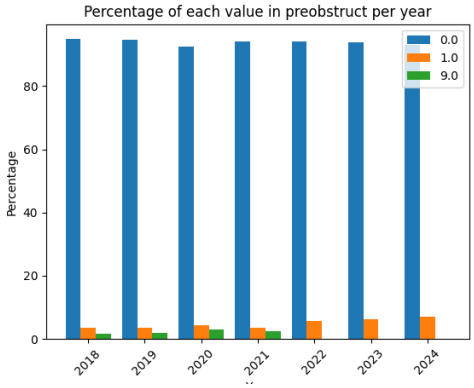


(h) `complpre`

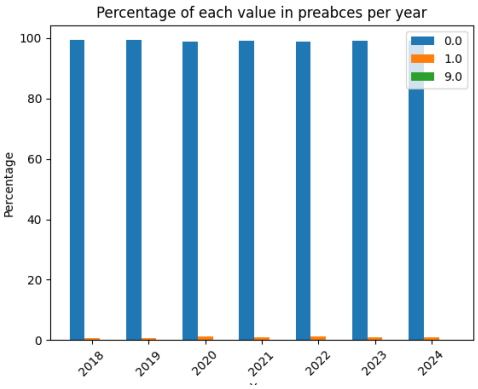
Figure F.3: Plots of the values of the discrete variables per year. Part 3/14.



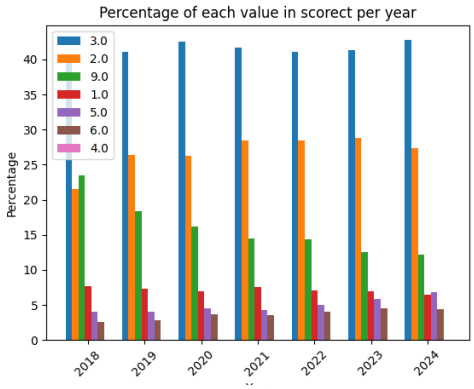
(a) `preanemie`



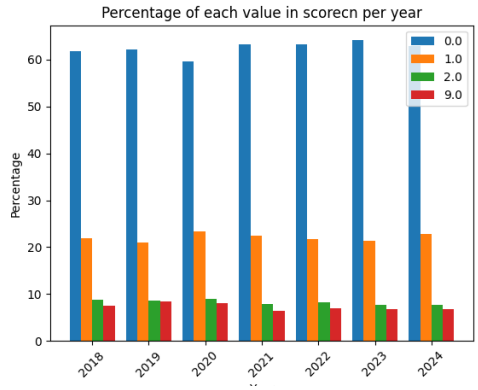
(b) `preobstruct`



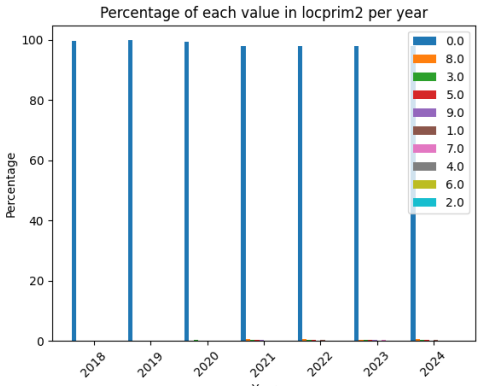
(c) `preabces`



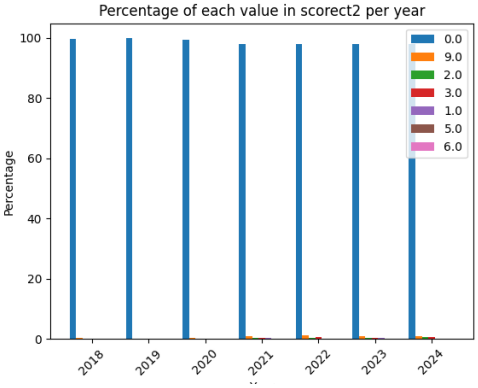
(d) `scorect`



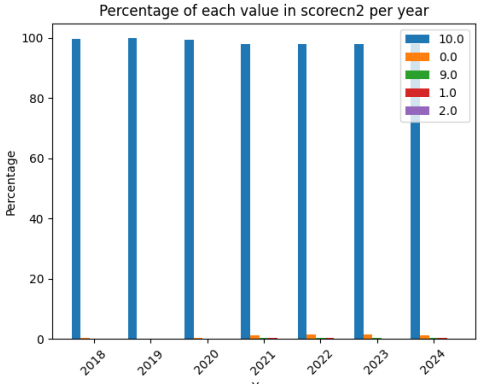
(e) `scorecn`



(f) `locprim2`

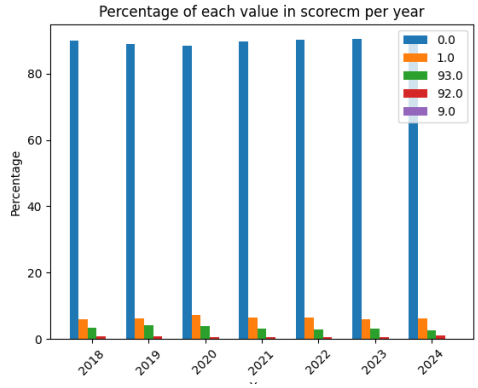


(g) `scorect2`

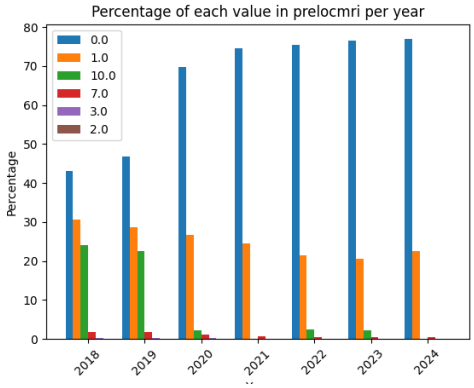


(h) `scorecn2`

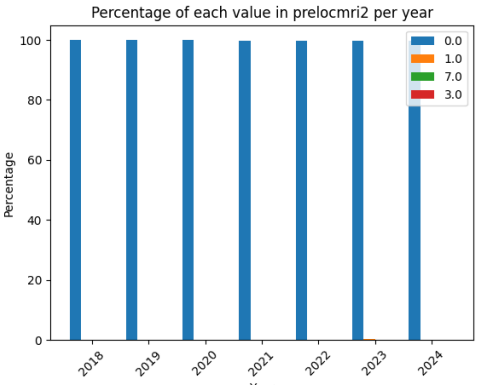
Figure F.4: Plots of the values of the discrete variables per year. Part 4/14.



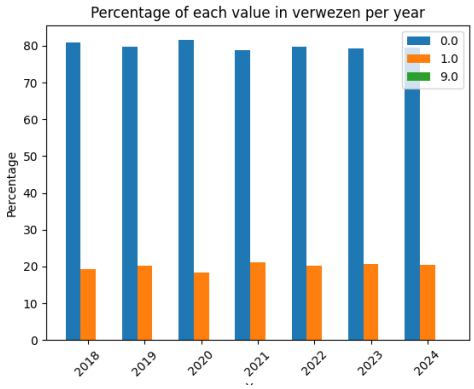
(a) `scorecm`



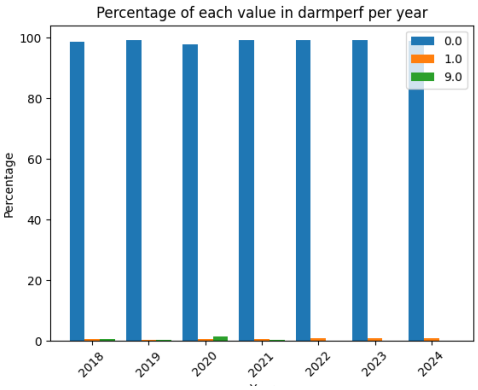
(b) `prelocmri`



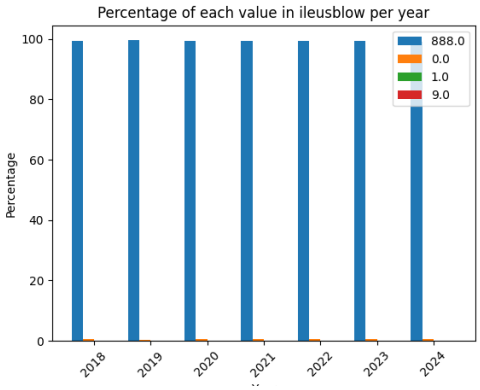
(c) `prelocmri2`



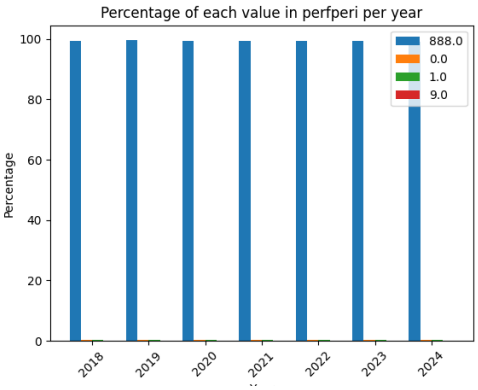
(d) `verwezen`



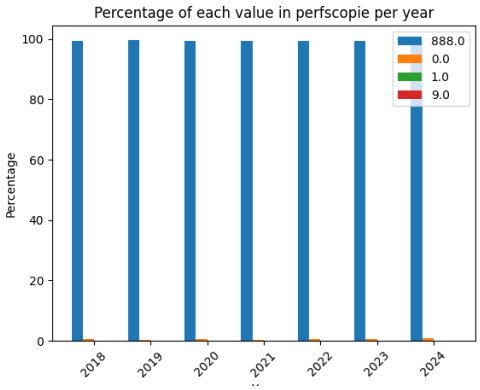
(e) `darmperf`



(f) `ileusblow`

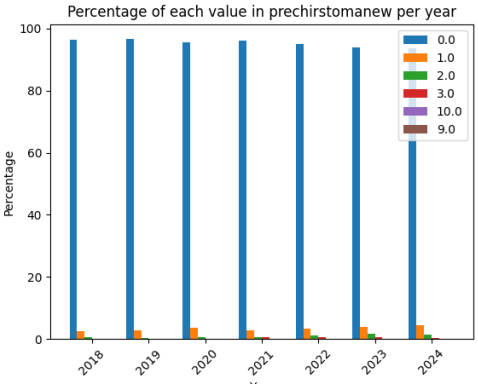


(g) `perfperi`

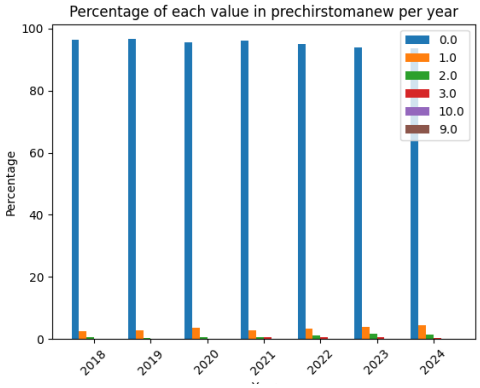


(h) `perfscopie`

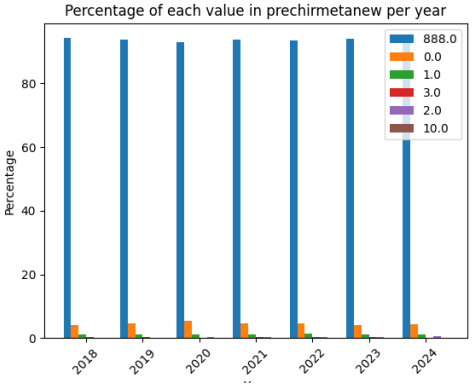
Figure F.5: Plots of the values of the discrete variables per year. Part 5/14.



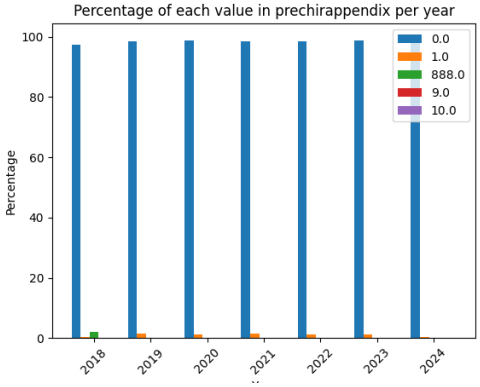
(a) `prechirstomanew`



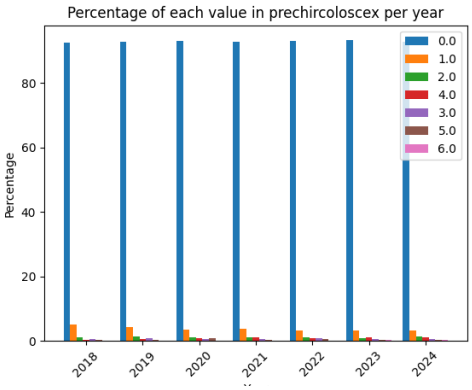
(b) `prechirstent`



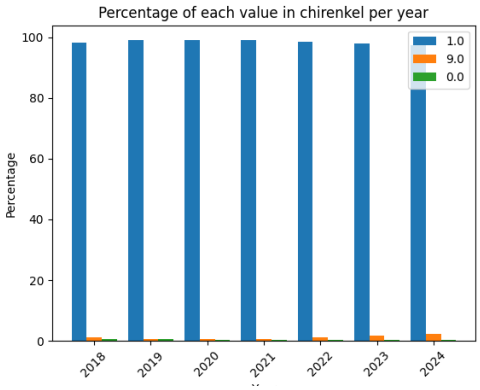
(c) `prechirmetanew`



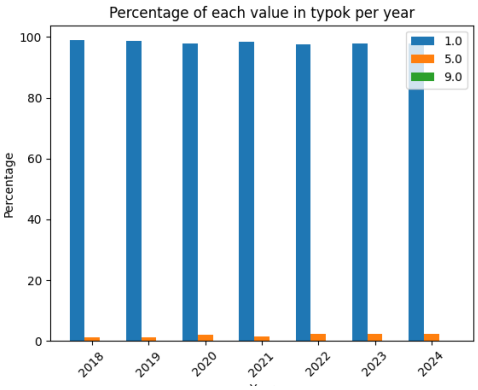
(d) `prechirappendix`



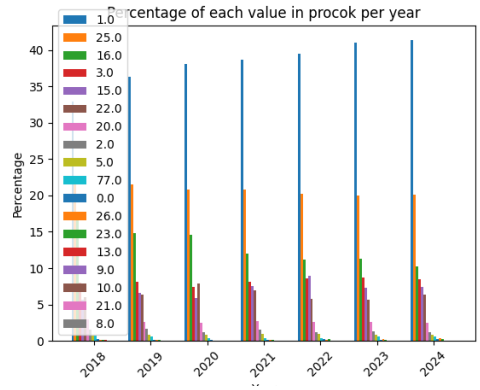
(e) `prechircoloscex`



(f) `chirenkel`

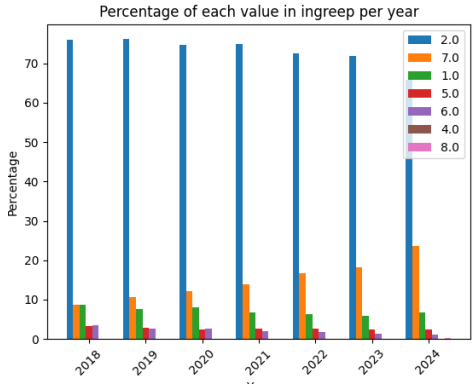


(g) `typok`

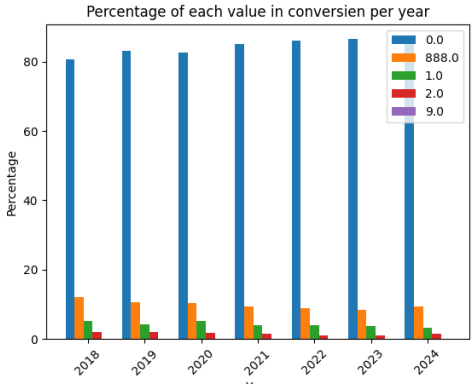


(h) `procok`

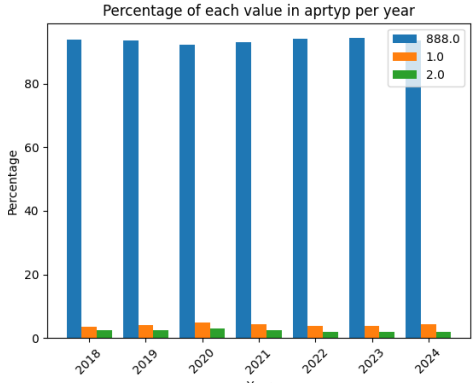
Figure F.6: Plots of the values of the discrete variables per year. Part 6/14.



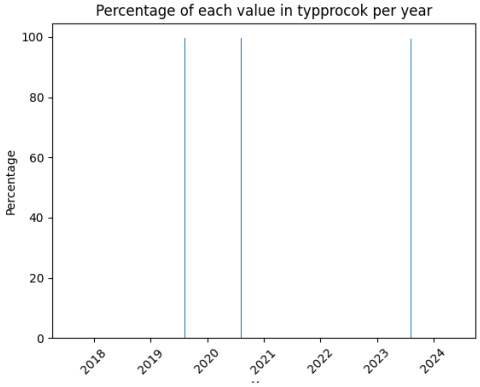
(a) `ingreep`



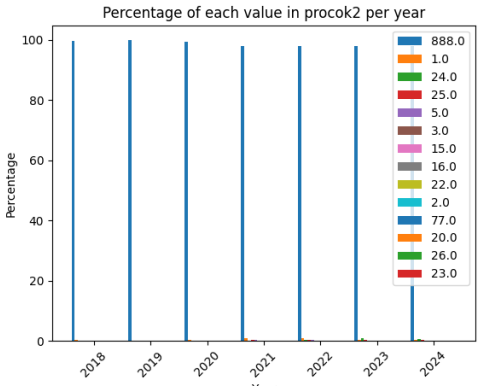
(b) `conversien`



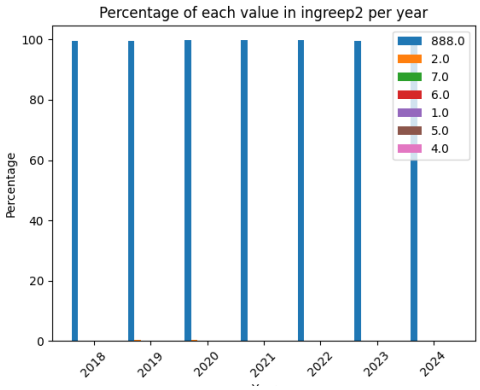
(c) `aprtyp`



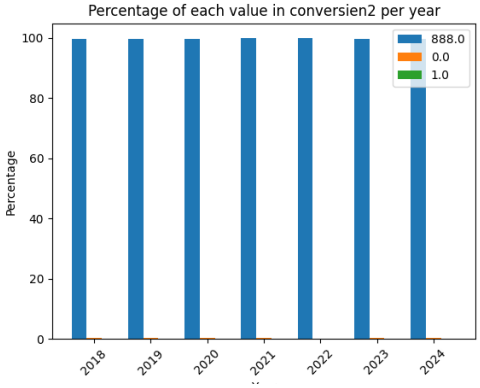
(d) `typprocok`



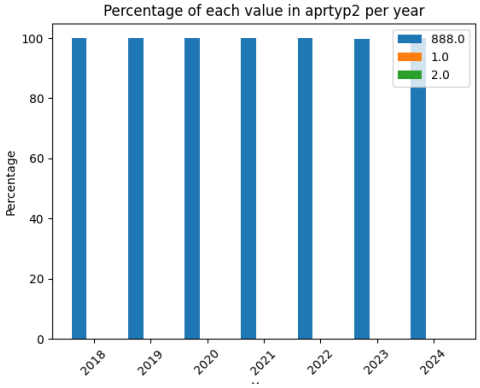
(e) `procok2`



(f) `ingreep2`

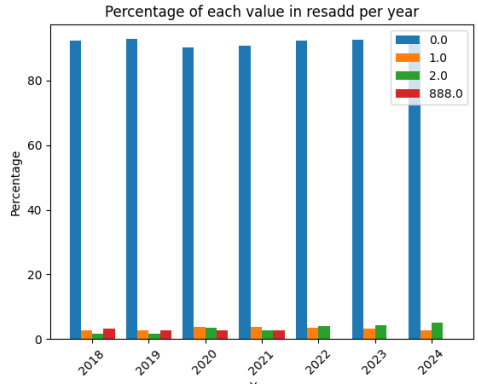


(g) `conversien2`

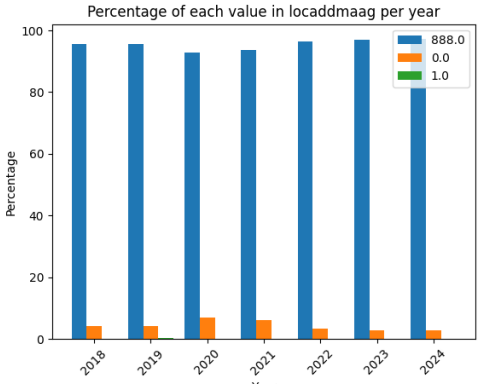


(h) `aprtyp2`

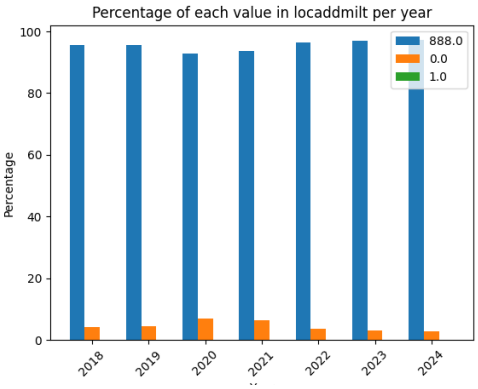
Figure F.7: Plots of the values of the discrete variables per year. Part 7/14.



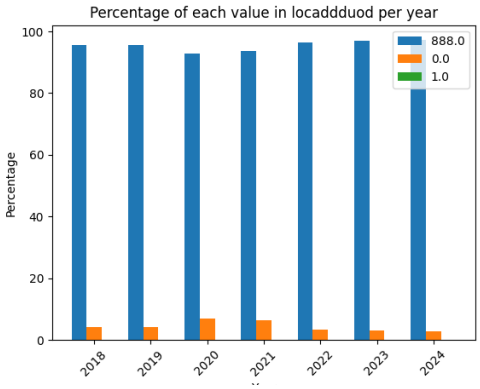
(a) `resadd`



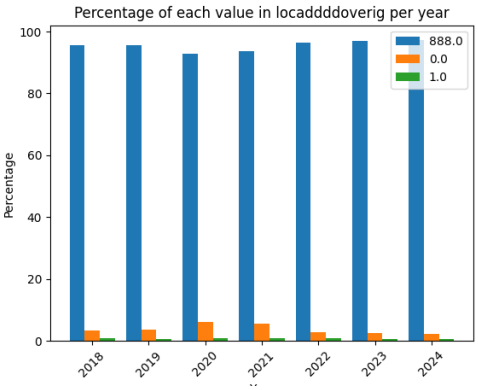
(b) `locaddmaag`



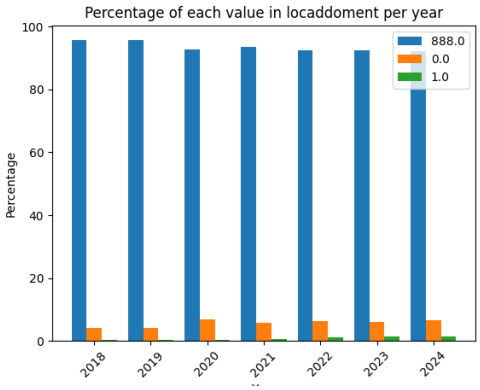
(c) `locaddmilt`



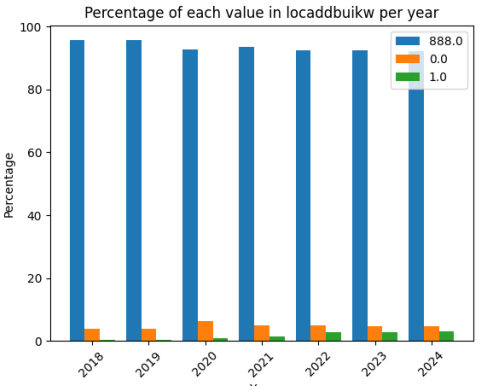
(d) `locaddduod`



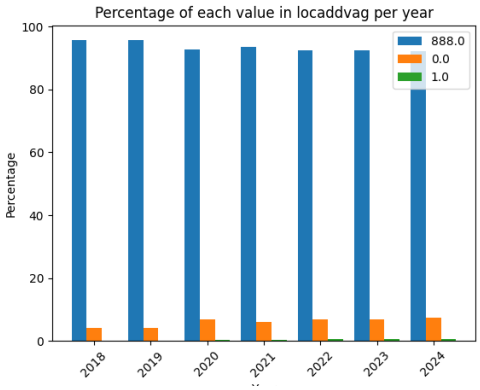
(e) `locadddoverig`



(f) `locaddoment`

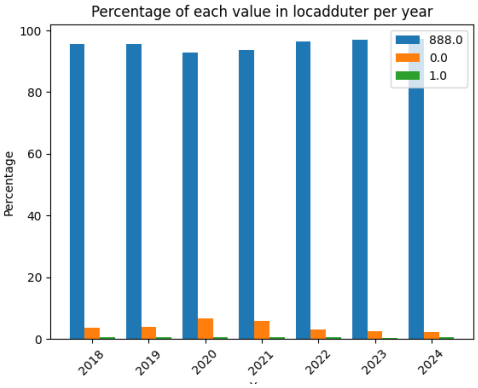


(g) `locaddbuikw`

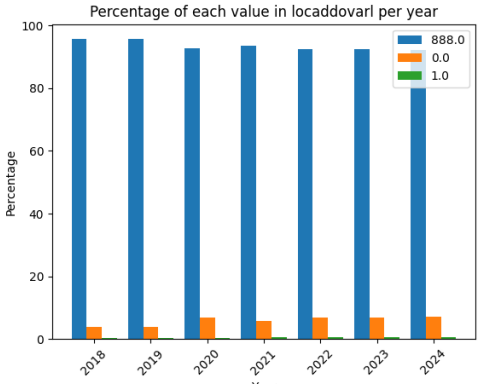


(h) `locaddvag`

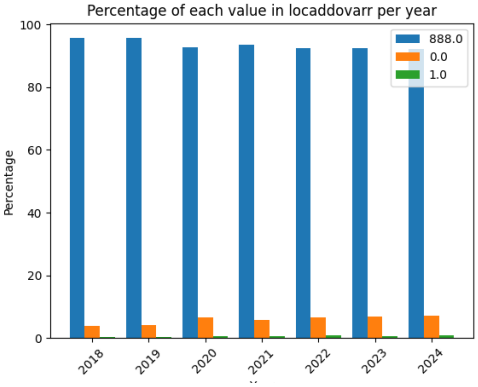
Figure F.8: Plots of the values of the discrete variables per year. Part 8/14.



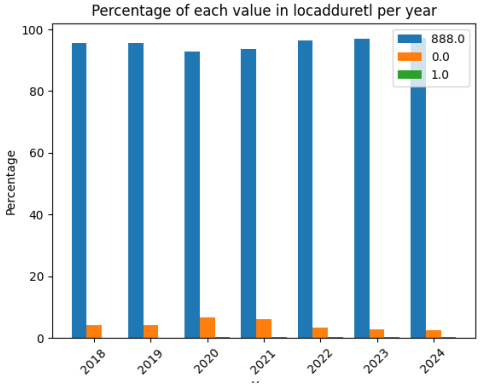
(a) `locadduter`



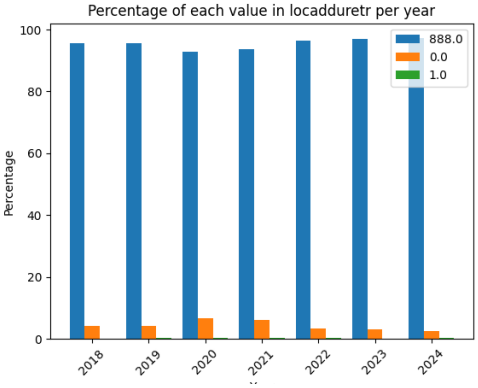
(b) `locaddovarl`



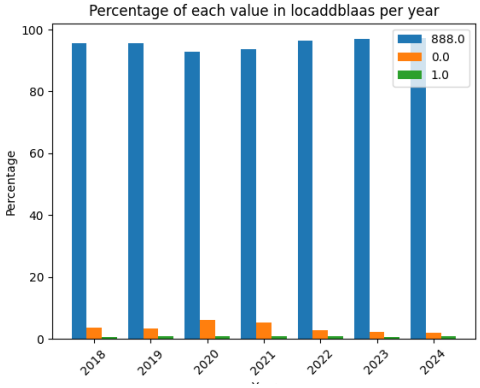
(c) `locaddovarr`



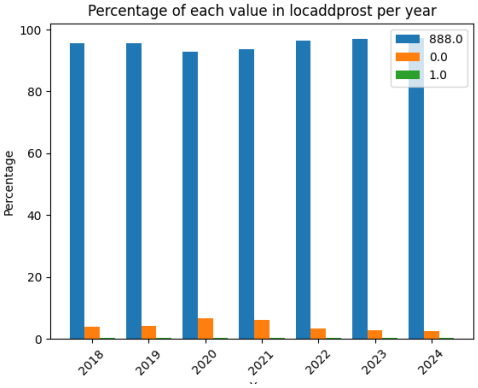
(d) `locadduretl`



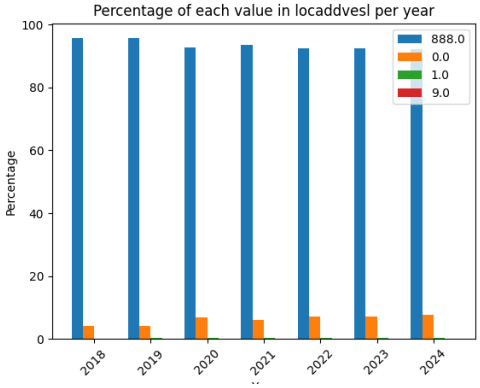
(e) `locadduretr`



(f) `locaddblaas`

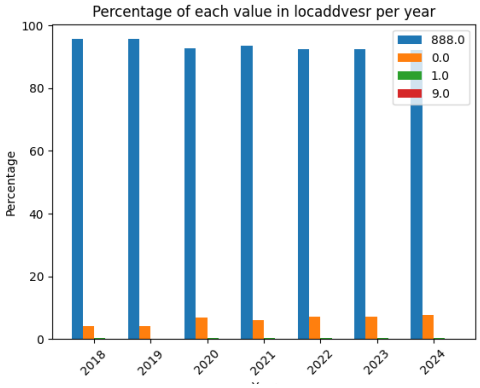


(g) `locaddprost`

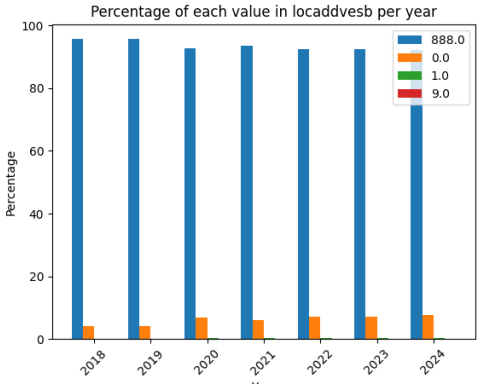


(h) `locaddvesl`

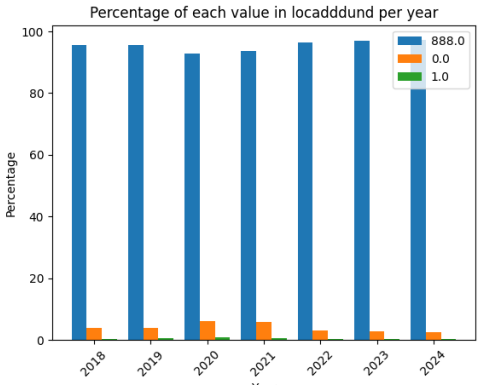
Figure F.9: Plots of the values of the discrete variables per year. Part 9/14.



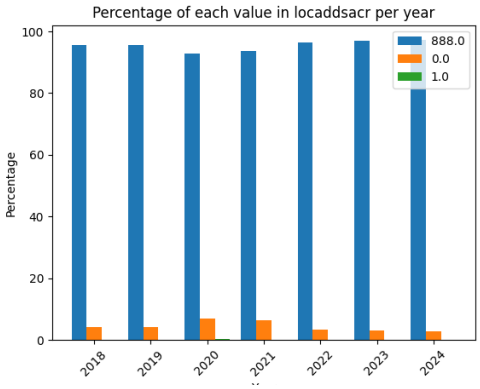
(a) `locaddvesr`



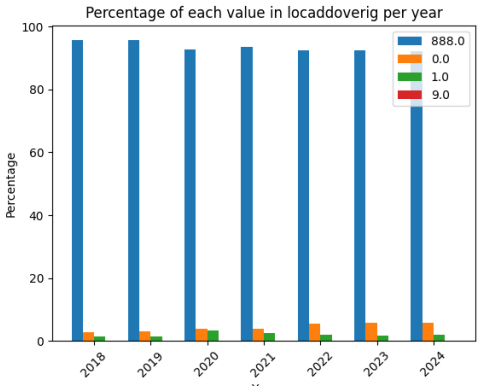
(b) `locaddvesb`



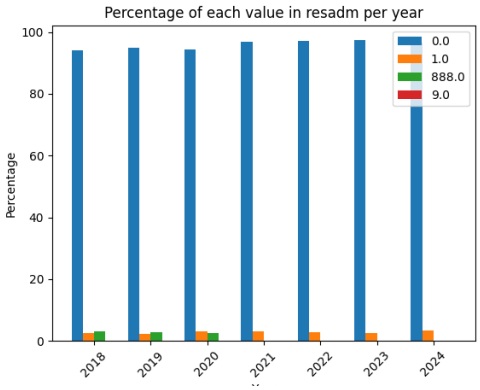
(c) `locadddund`



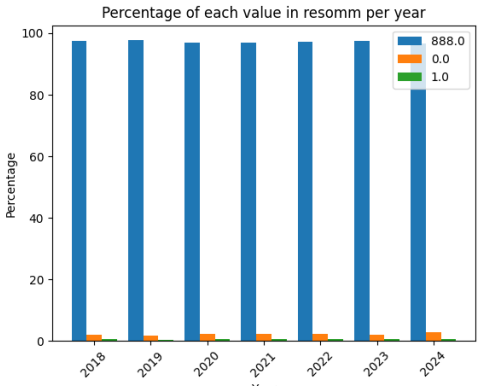
(d) `locaddsacr`



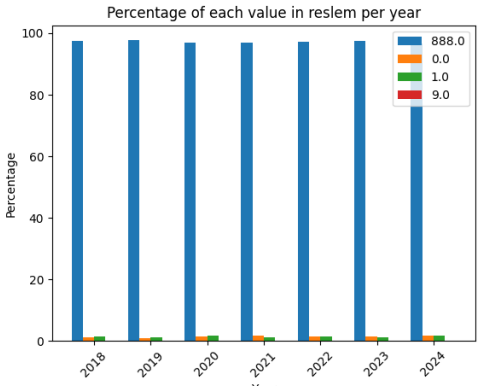
(e) `locaddoverig`



(f) `resadm`

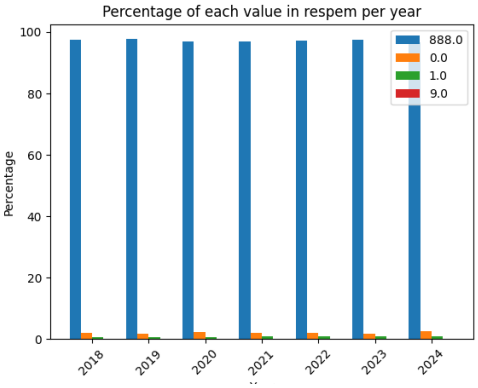


(g) `resomm`

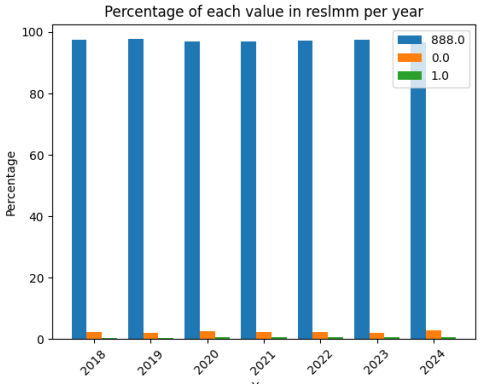


(h) `reslem`

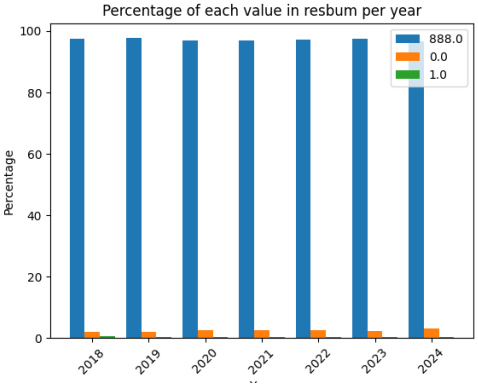
Figure F.10: Plots of the values of the discrete variables per year. Part 10/14.



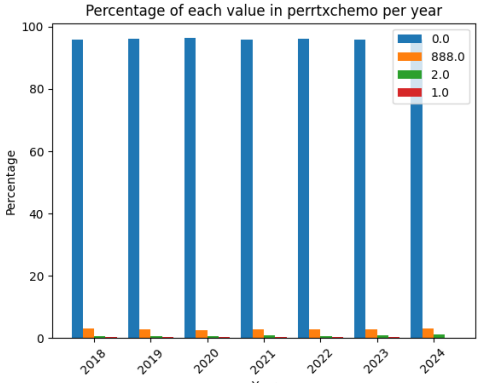
(a) `respem`



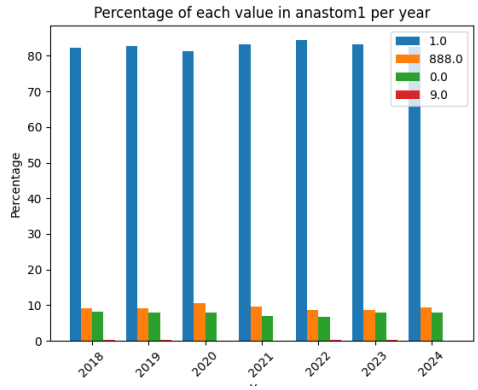
(b) `reslmm`



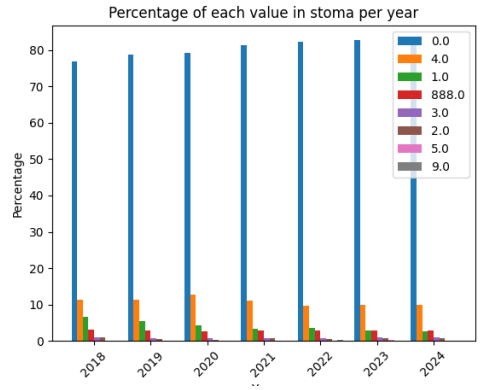
(c) `resbum`



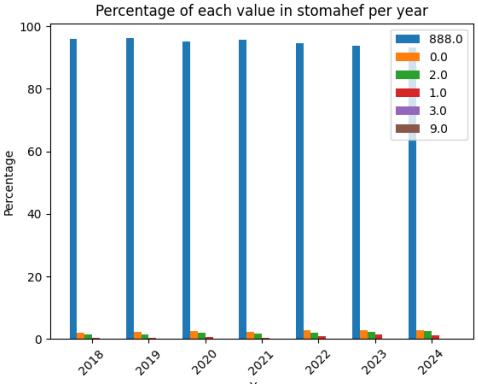
(d) `perrtxchemo`



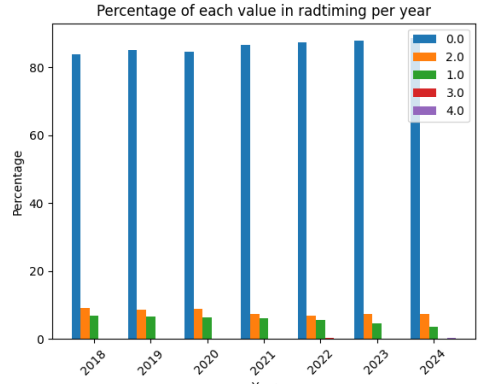
(e) `anastom1`



(f) `stoma`

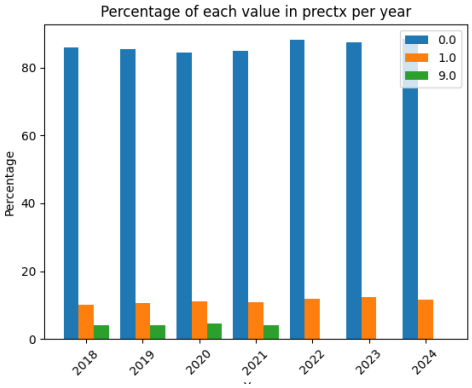


(g) `stomahef`

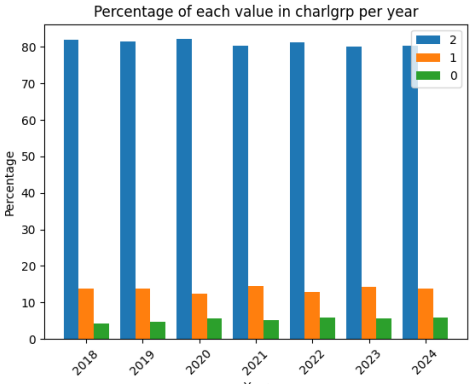


(h) `radtiming`

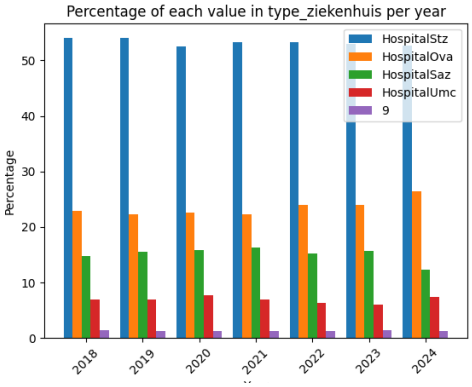
Figure F.11: Plots of the values of the discrete variables per year. Part 11/14.



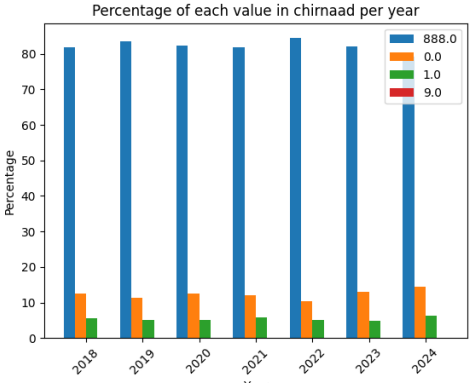
(a) `prectx`



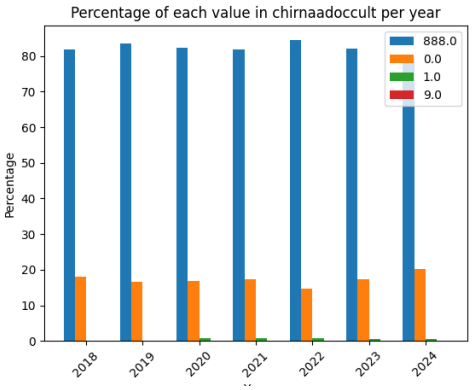
(b) `charlgrp`



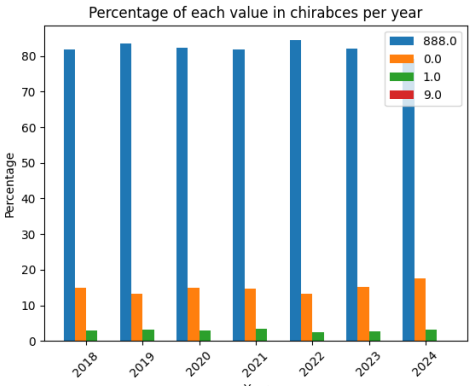
(c) `type_ziekenhuis`



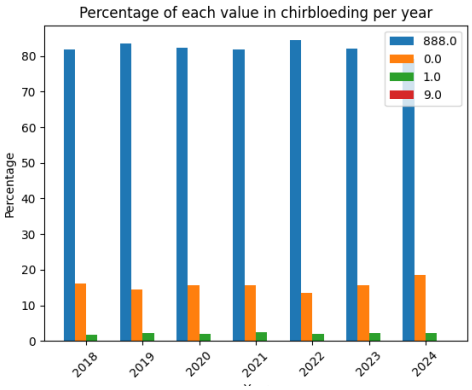
(d) `chirnaad`



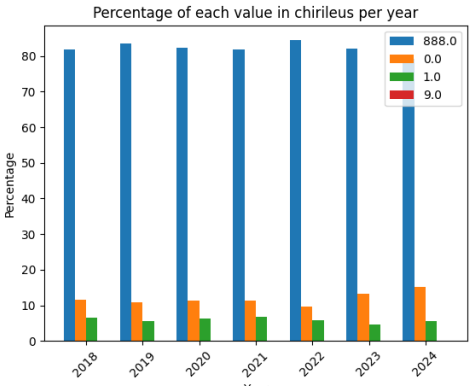
(e) `chirnaadoccult`



(f) `chirabces`

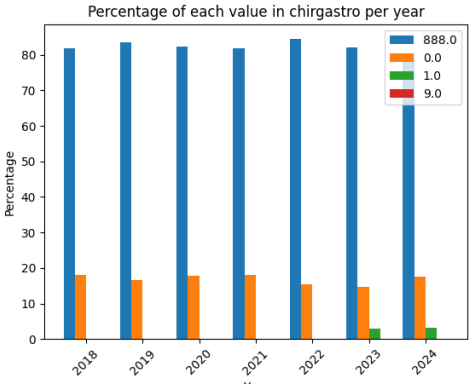


(g) `chirbloeding`

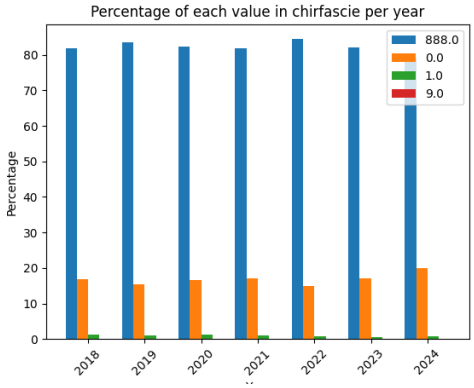


(h) `chirileus`

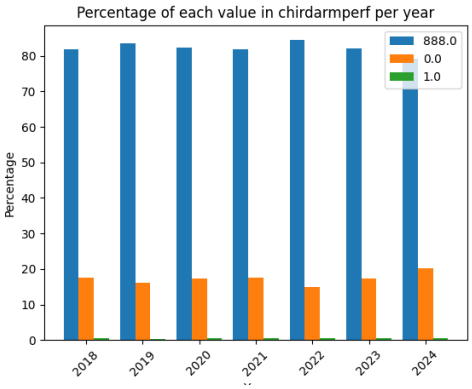
Figure F.12: Plots of the values of the discrete variables per year. Part 12/14.



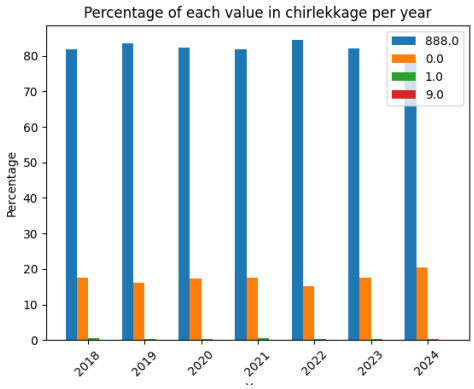
(a) `chirgastro`



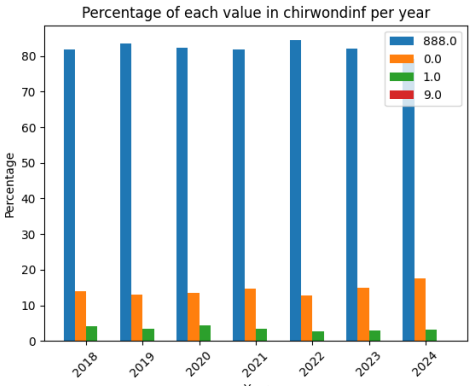
(b) `chirfascie`



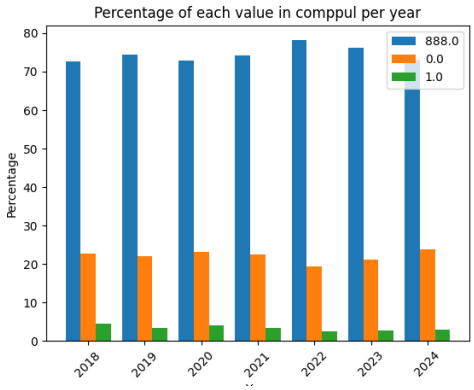
(c) `chirdarmperf`



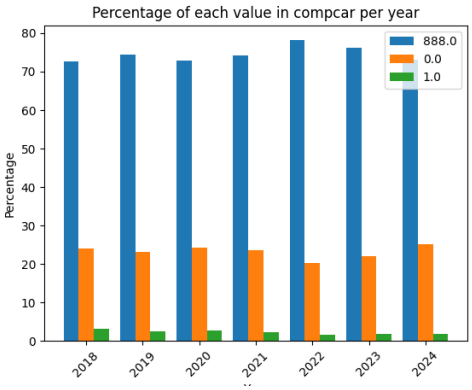
(d) `chirlekkage`



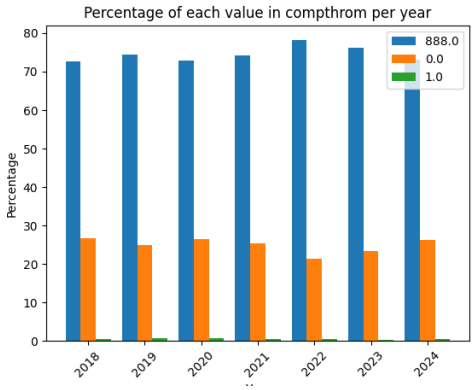
(e) `chirwondinf`



(f) `comppul`

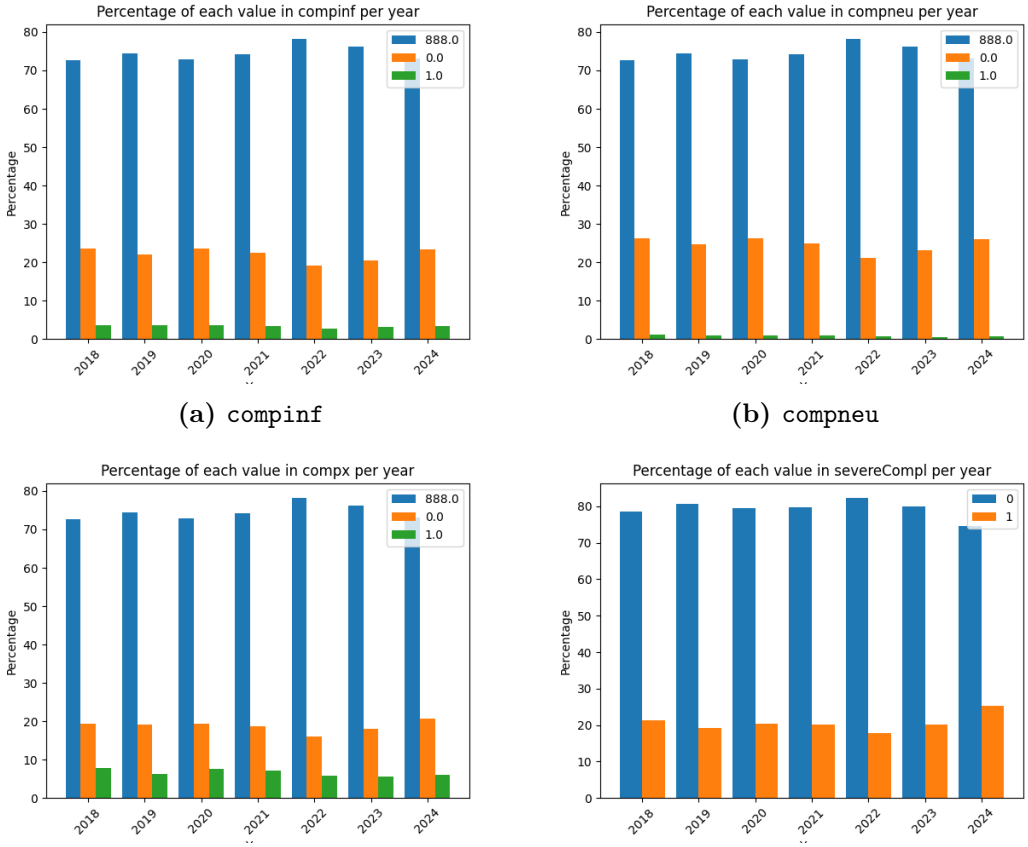


(g) `compcar`



(h) `compthrom`

Figure F.13: Plots of the values of the discrete variables per year. Part 13/14.



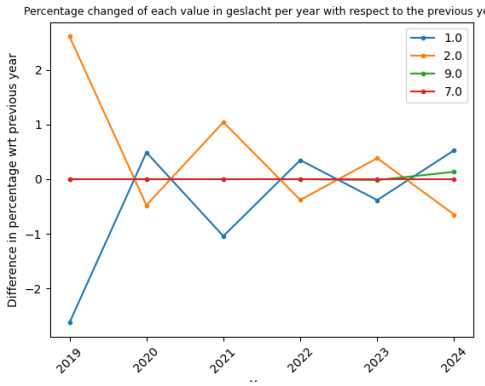
(a) `compinf` (b) `compneu`

(c) `compx` (d) `severeCompl`

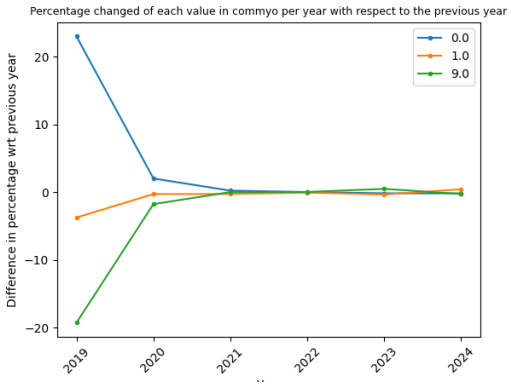
Figure F.14: Plots of the values of the discrete variables per year. Part 14/14.

F.1.2 Percentage change of values per year compared to previous year

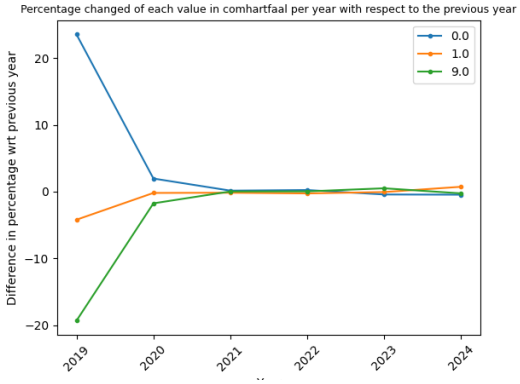
Here we show the plots of the discrete variables. The plots show per year how much (in percentage) a value changed compared to the previous year.



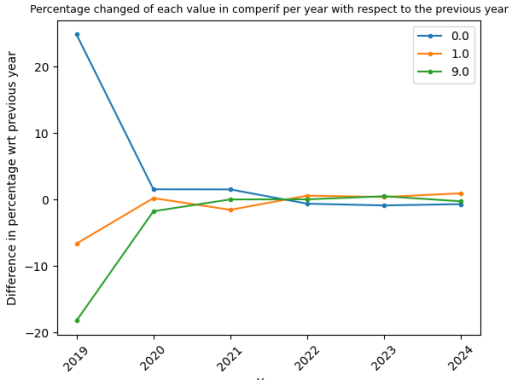
(a) `geslacht`



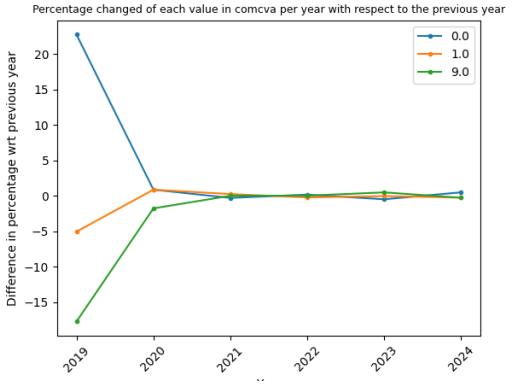
(b) `commyo`



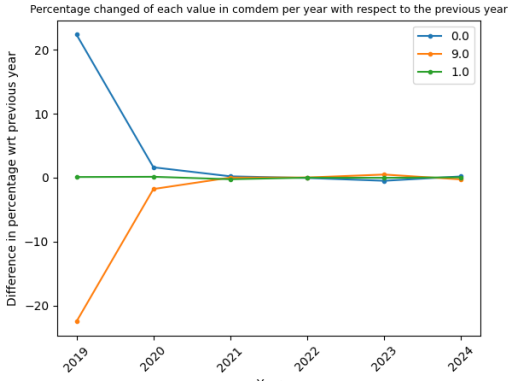
(c) `comhartfaal`



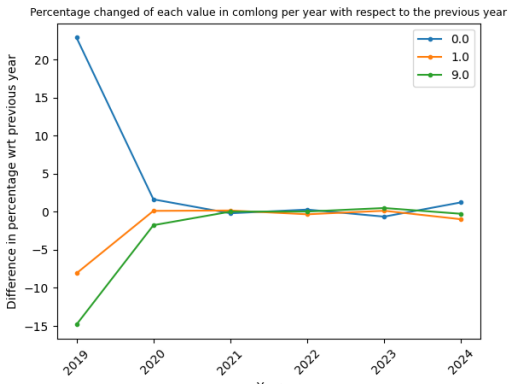
(d) `comperif`



(e) `comcva`



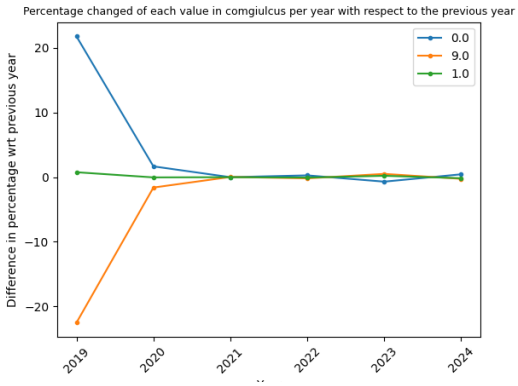
(f) `comdem`



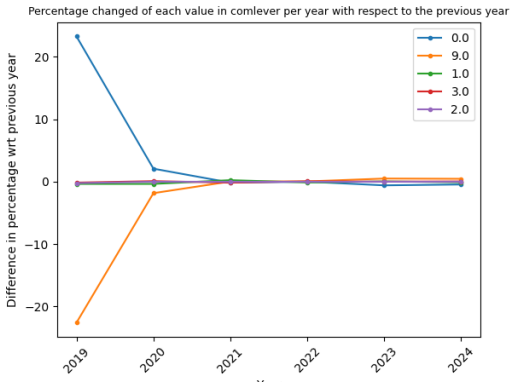
(g) `comlong`

(h) `combind`

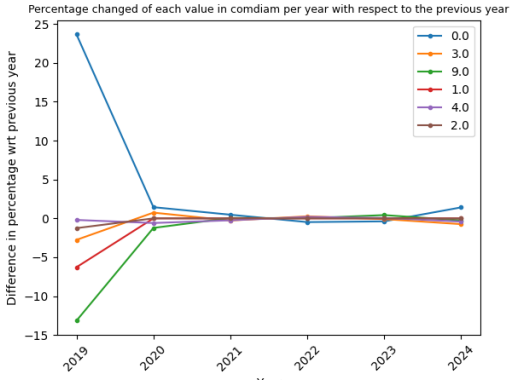
Figure F.15: Plots of the percentage change in values compared to the previous year of the discrete variables per year. Part 1/14.



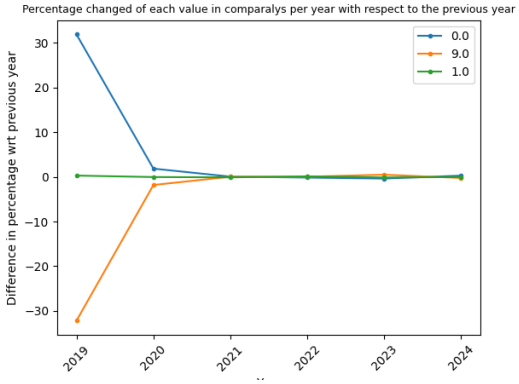
(a) comgiulcus



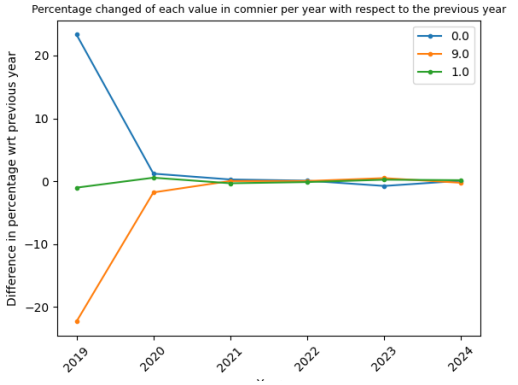
(b) comlever



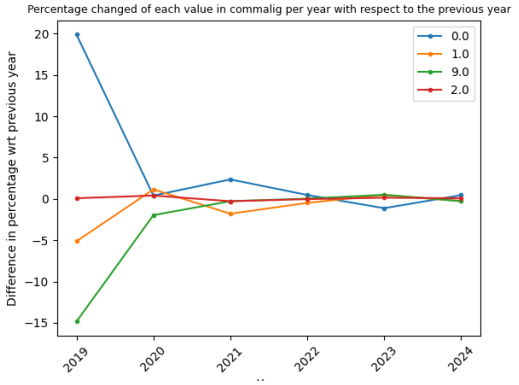
(c) comdiam



(d) comparalys

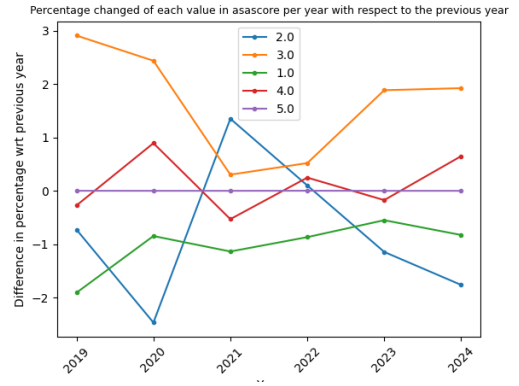


(e) comnier



(f) commalig

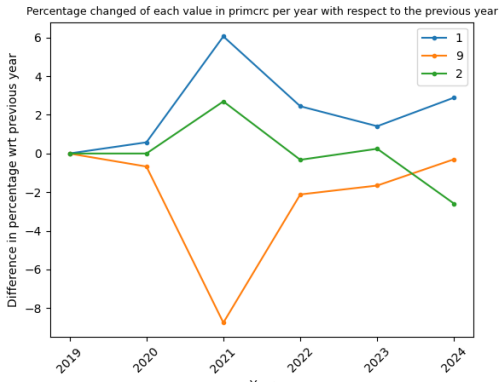
(g) comhiv



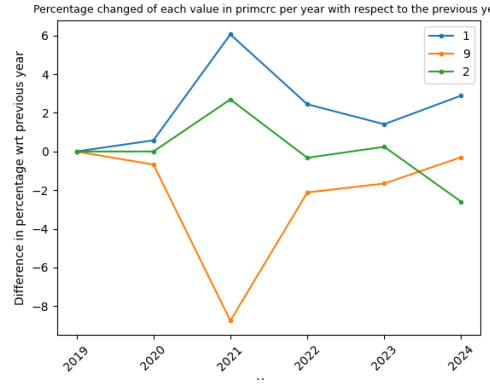
(h) asascore

Figure F.16: Plots of the percentage change in values compared to the previous year of the discrete variables per year. Part 2/14.

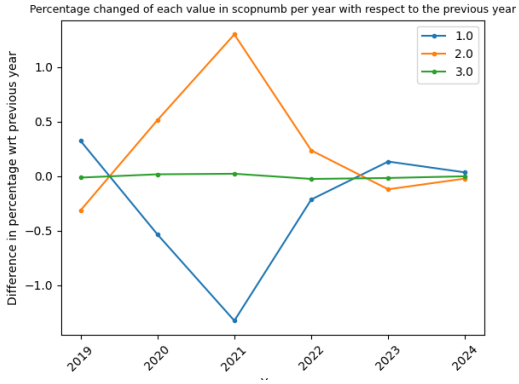
(a) `resdarm`



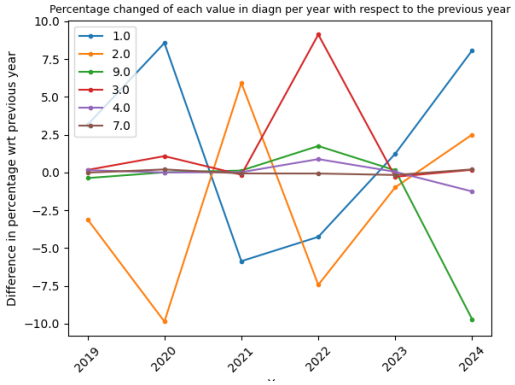
(b) `stomavg`



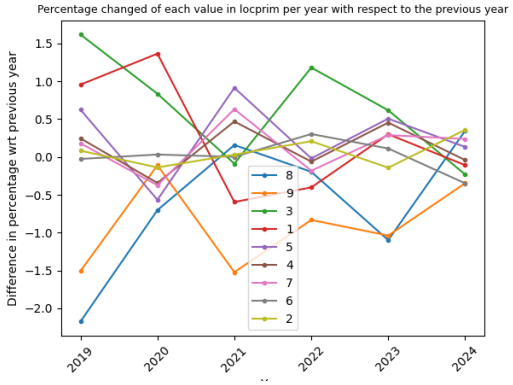
(c) `primcrc`



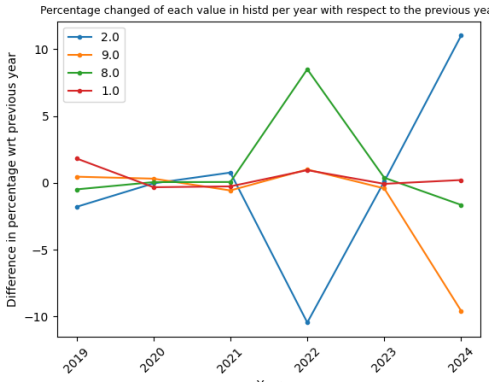
(d) `scopnumb`



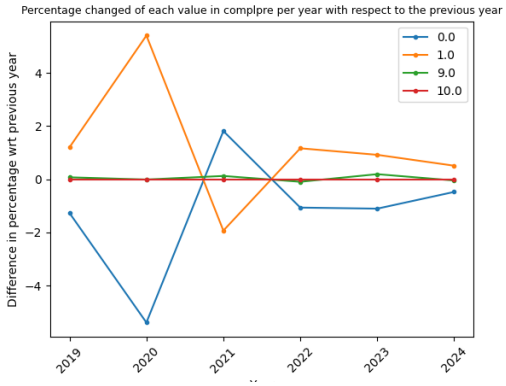
(e) `diagn`



(f) `locprim`

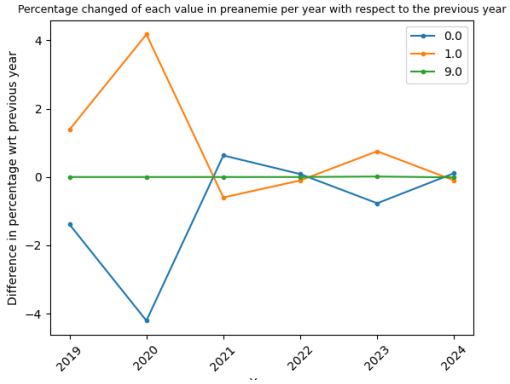


(g) `histd`

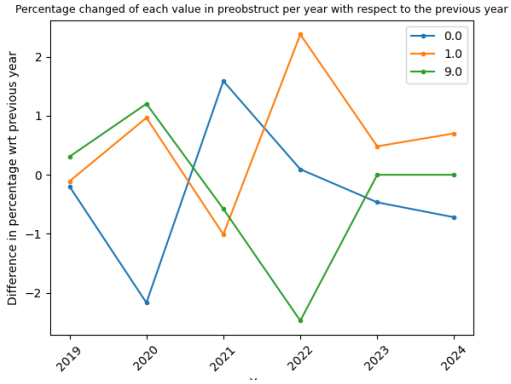


(h) `complpre`

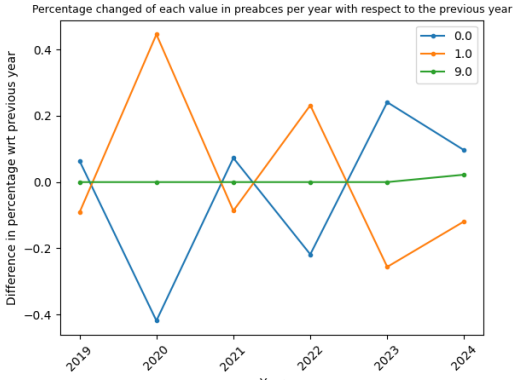
Figure F.17: Plots of the percentage change in values compared to the previous year of the discrete variables per year. Part 3/14.



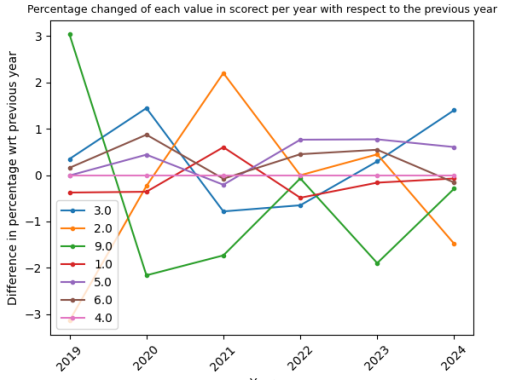
(a) `preanemie`



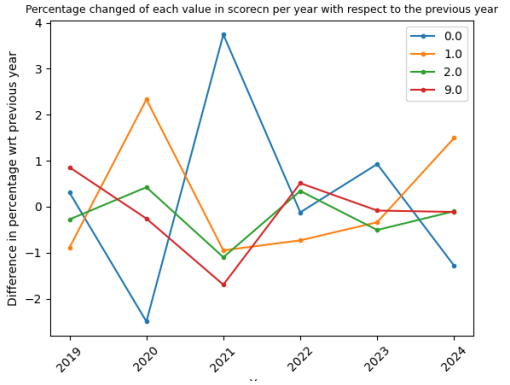
(b) `preobstruct`



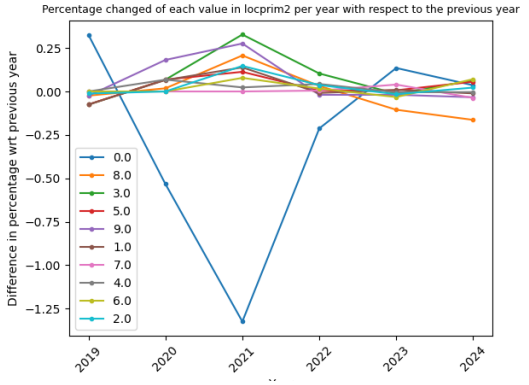
(c) `preabces`



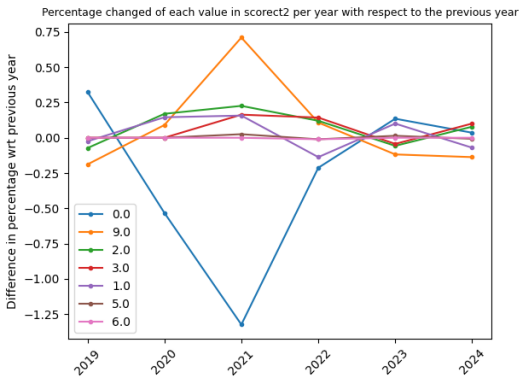
(d) `scorect`



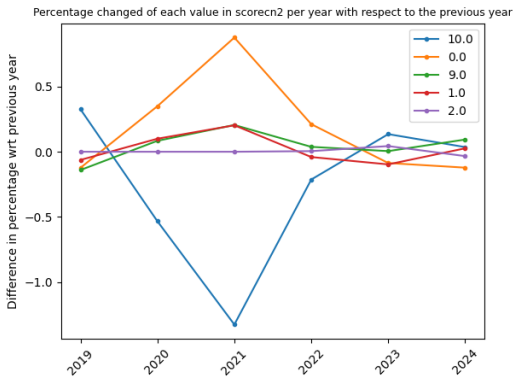
(e) `scorecn`



(f) `locprim2`

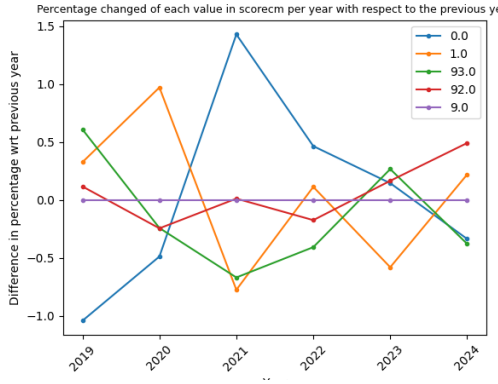


(g) `scorect2`

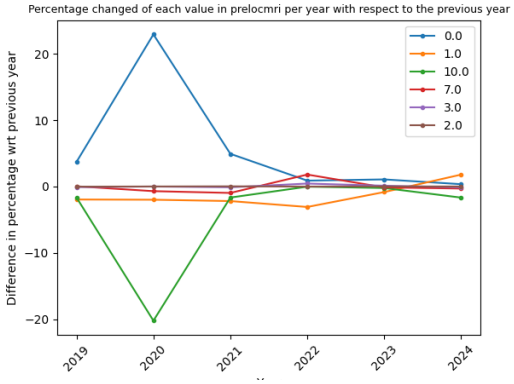


(h) `scorecn2`

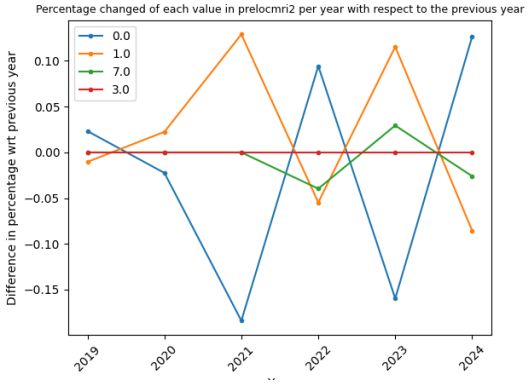
Figure F.18: Plots of the percentage change in values compared to the previous year of the discrete variables per year. Part 4/14.



(a) `scorecm`

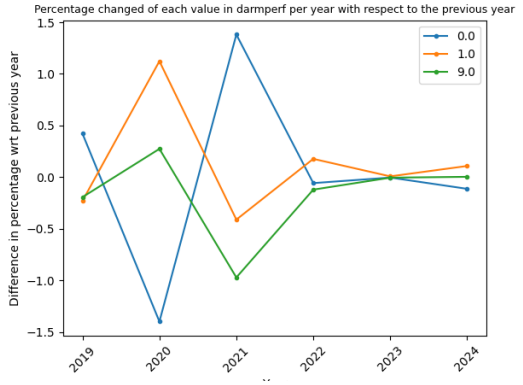


(b) `prelocmri`

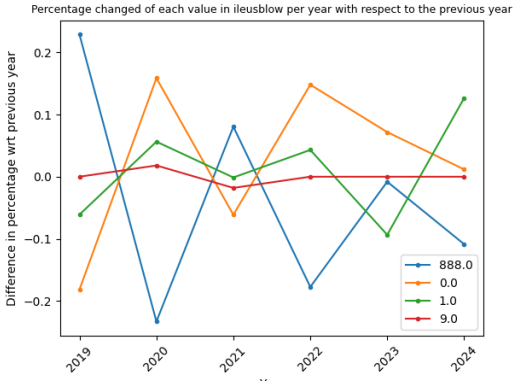


(c) `prelocmri2`

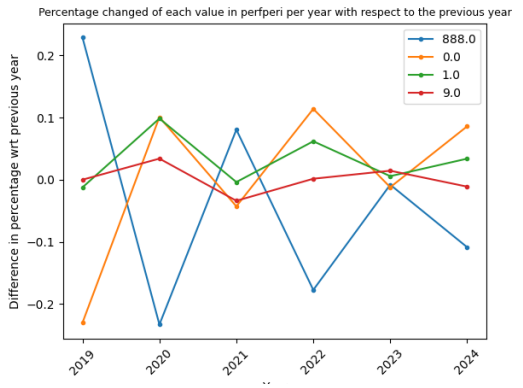
(d) `verwezen`



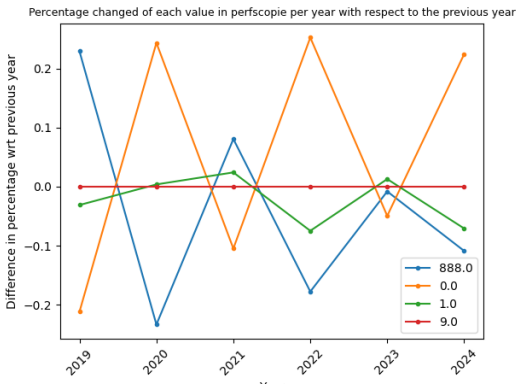
(e) `darmperf`



(f) `ileusblow`

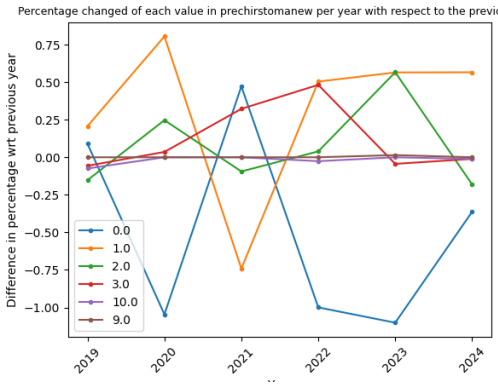


(g) `perfperi`

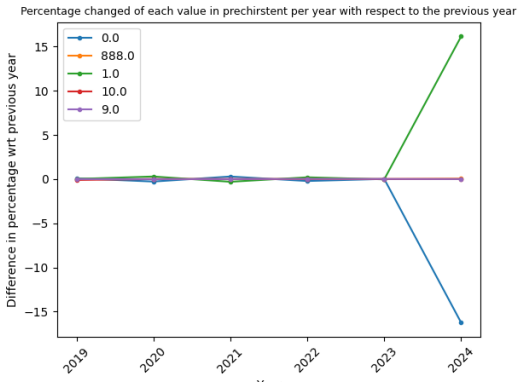


(h) `perfscopie`

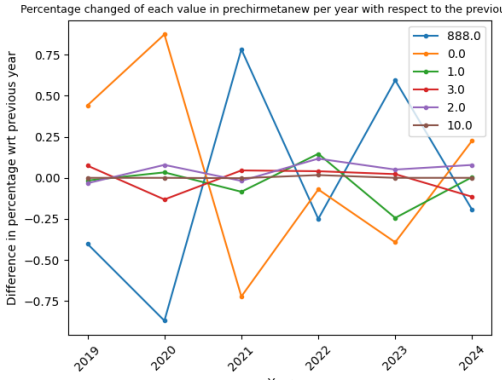
Figure F.19: Plots of the percentage change in values compared to the previous year of the discrete variables per year. Part 5/14.



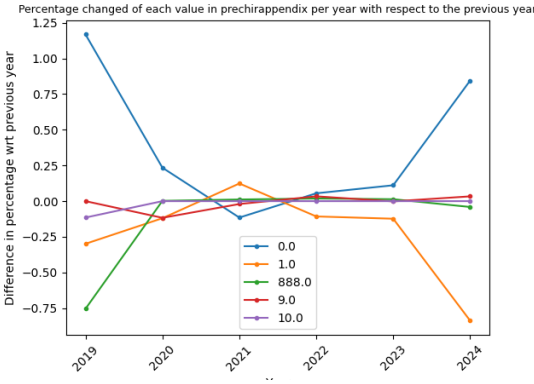
(a) `prechirstomanew`



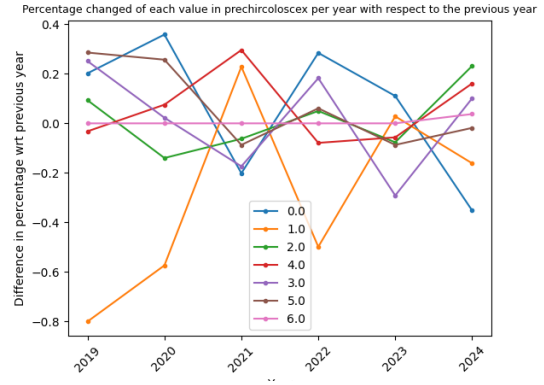
(b) `prechirstent`



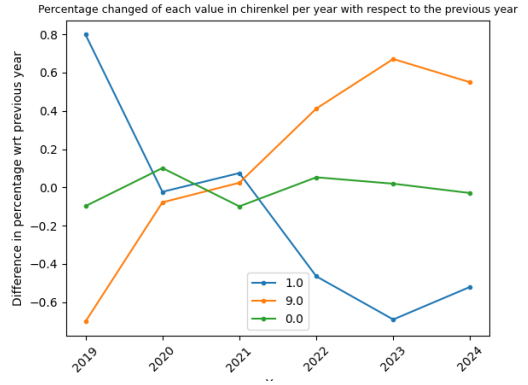
(c) `prechirmetanew`



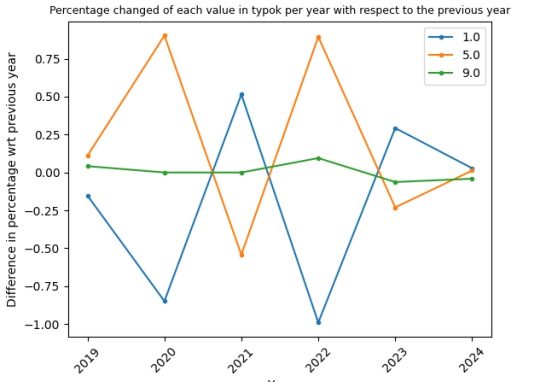
(d) `prechirappendix`



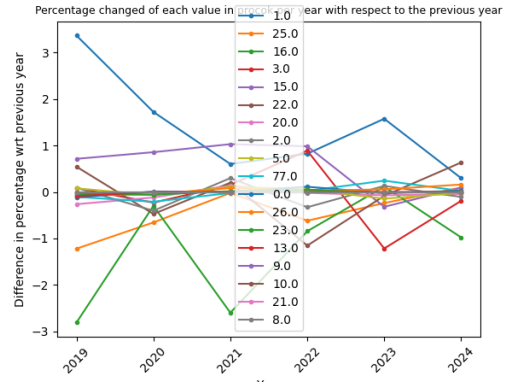
(e) `prechircoloscex`



(f) `chirenkel`

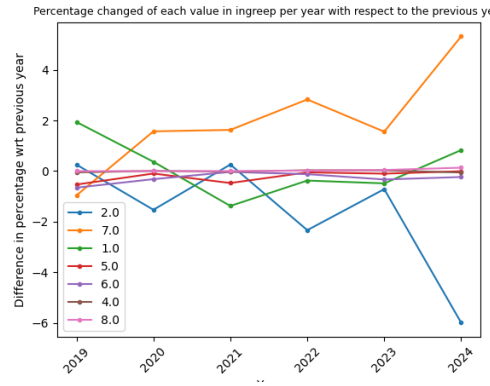


(g) `typok`

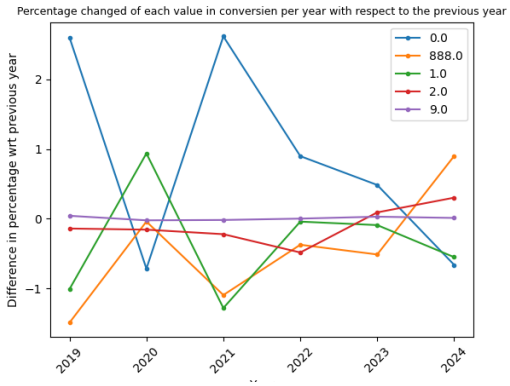


(h) `procok`

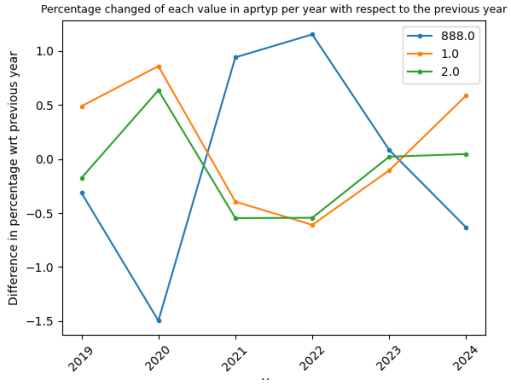
Figure F.20: Plots of the percentage change in values compared to the previous year of the discrete variables per year. Part 6/14.



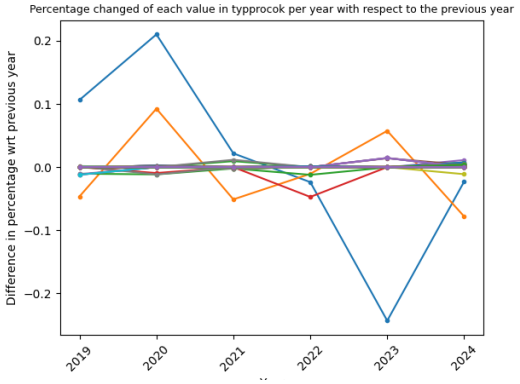
(a) `ingreep`



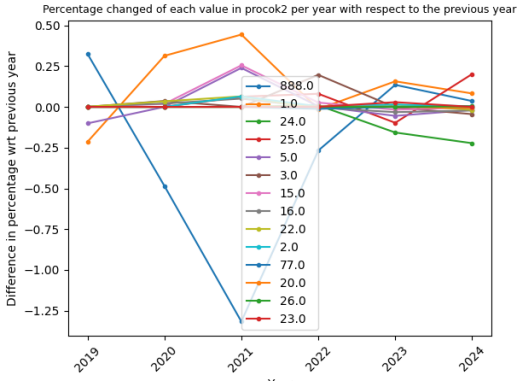
(b) `conversien`



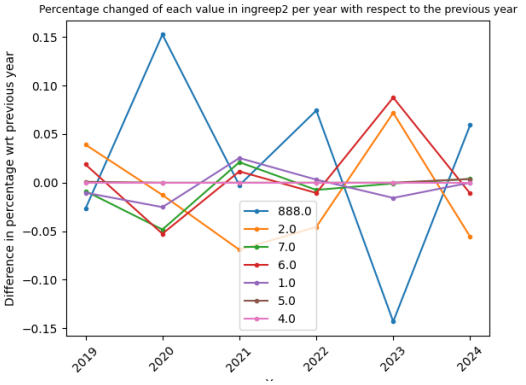
(c) `aprtyp`



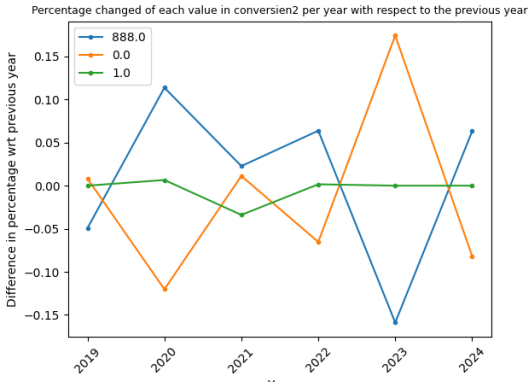
(d) `typprocok`



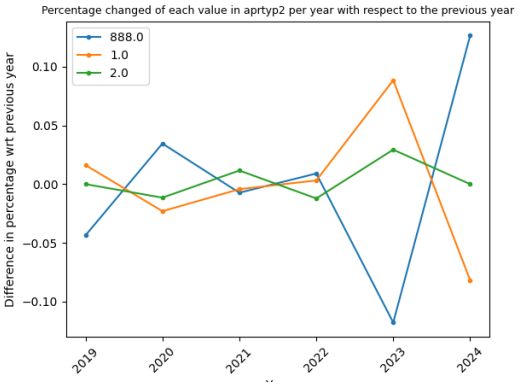
(e) `procok2`



(f) `ingreep2`

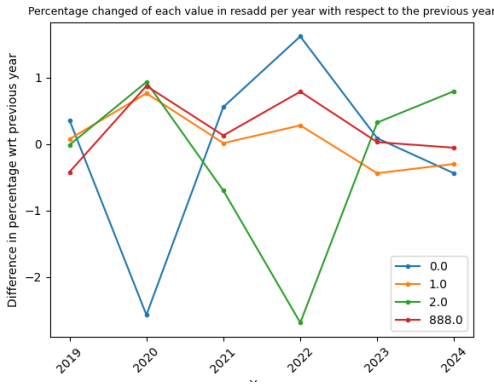


(g) `conversien2`

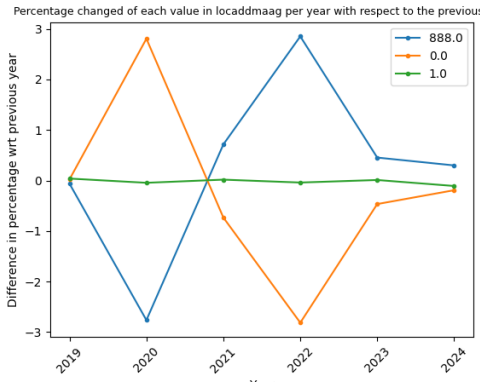


(h) `aprtyp2`

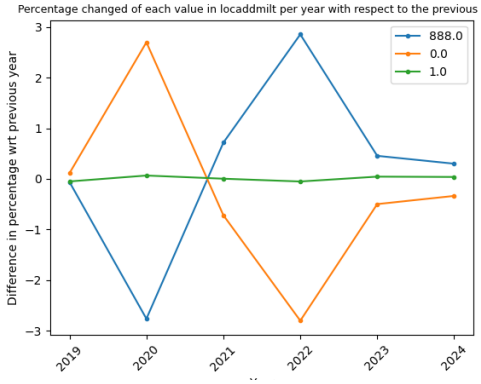
Figure F.21: Plots of the percentage change in values compared to the previous year of the discrete variables per year. Part 7/14.



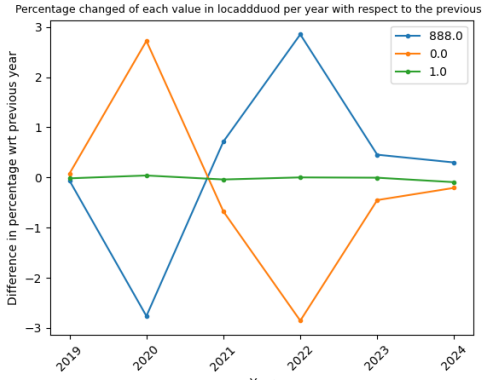
(a) `resadd`



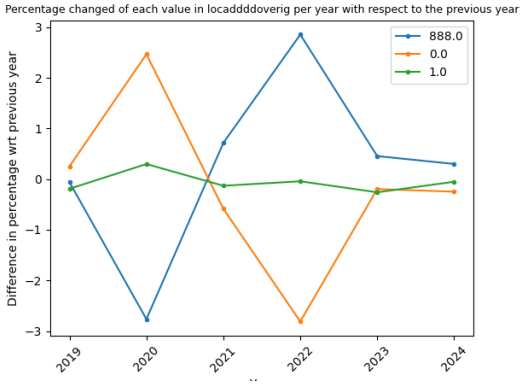
(b) `locaddmaag`



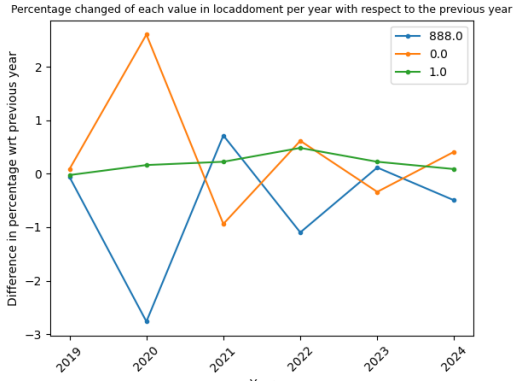
(c) `locaddmilt`



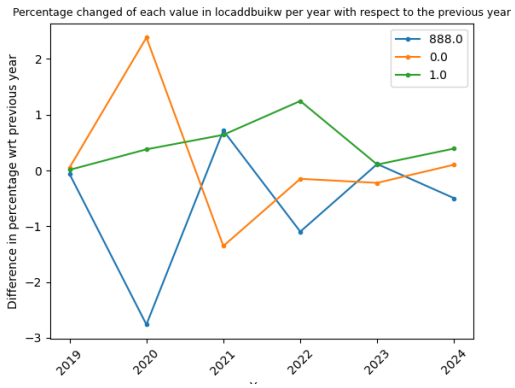
(d) `locaddduod`



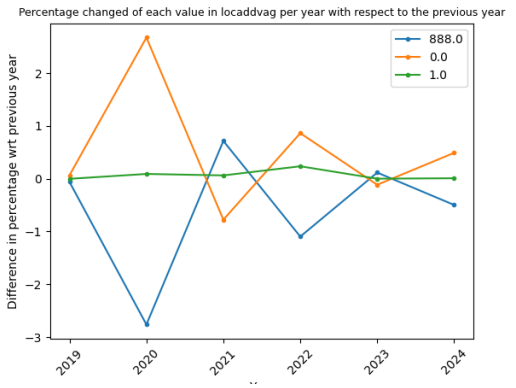
(e) `locadddoverig`



(f) `locaddoment`

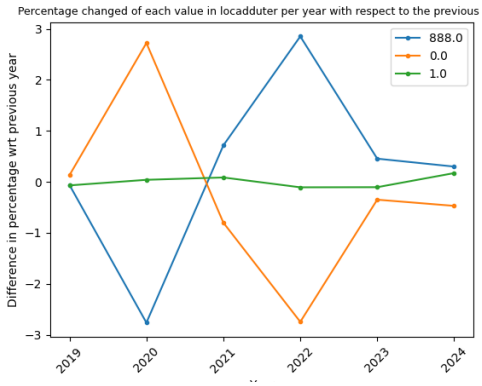


(g) `locaddbuikw`

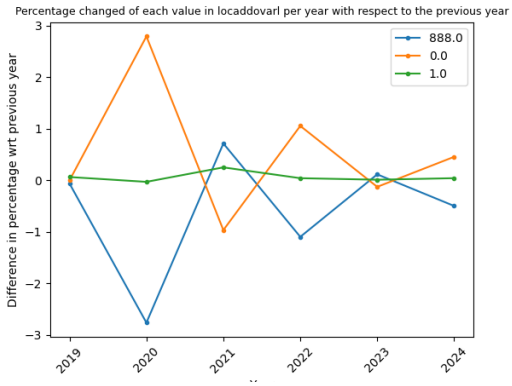


(h) `locaddvag`

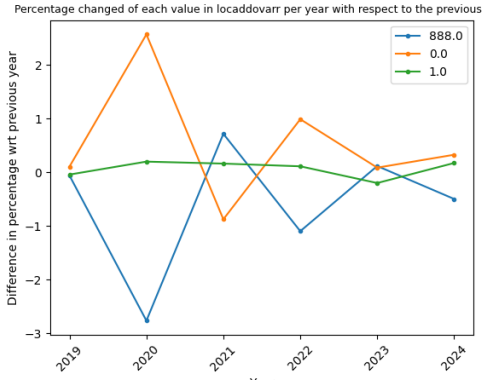
Figure F.22: Plots of the percentage change in values compared to the previous year of the discrete variables per year. Part 8/14.



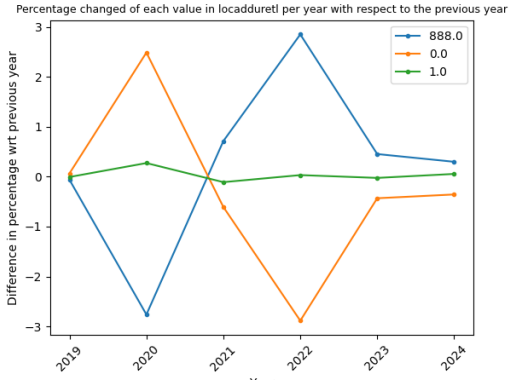
(a) `locadduter`



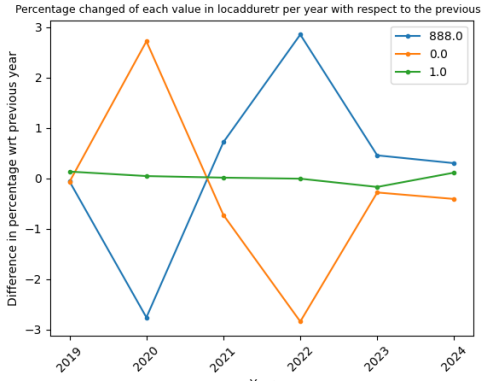
(b) `locaddovarl`



(c) `locaddovarr`

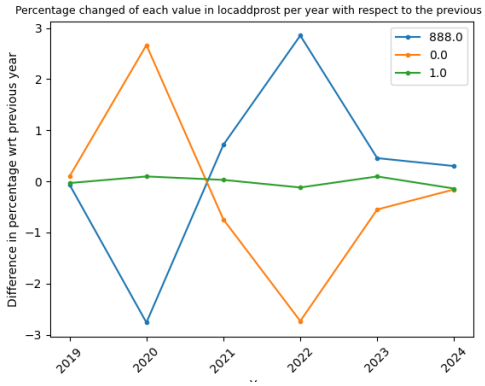


(d) `locadduretl`

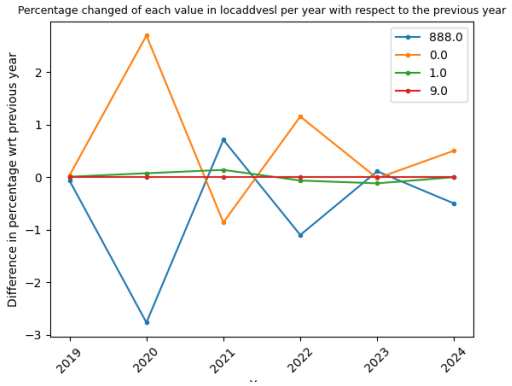


(e) `locadduretr`

(f) `locaddblaas`

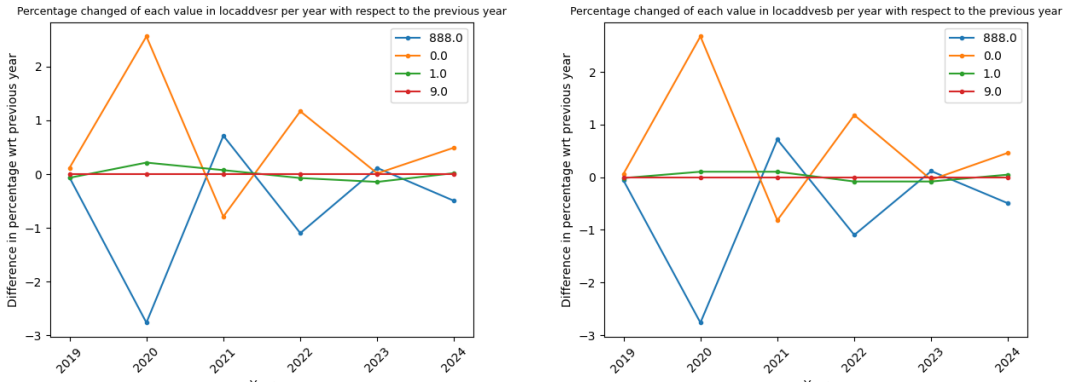


(g) `locaddprost`

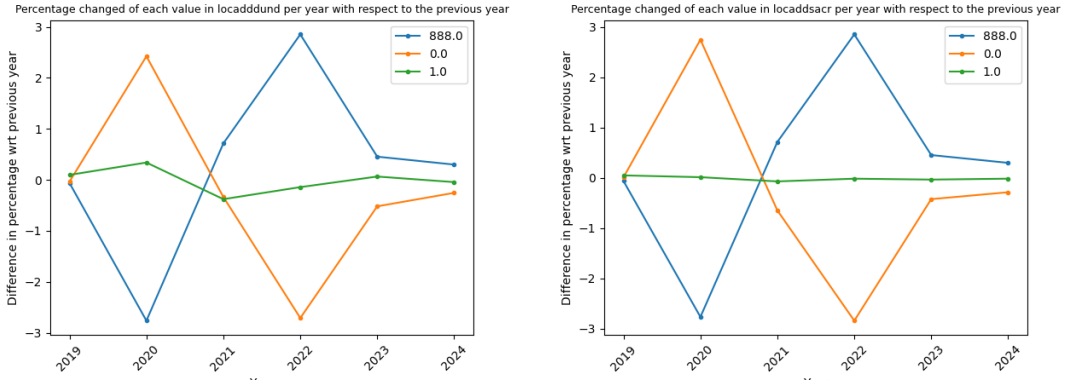


(h) `locaddvesl`

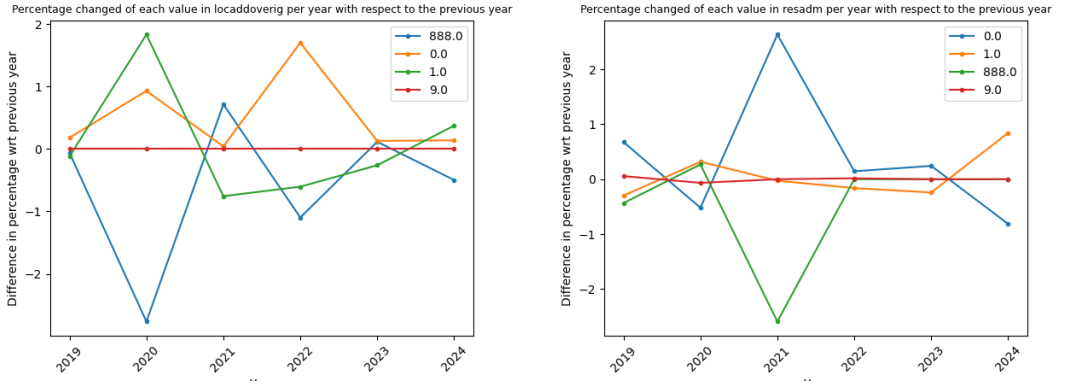
Figure F.23: Plots of the percentage change in values compared to the previous year of the discrete variables per year. Part 9/14.



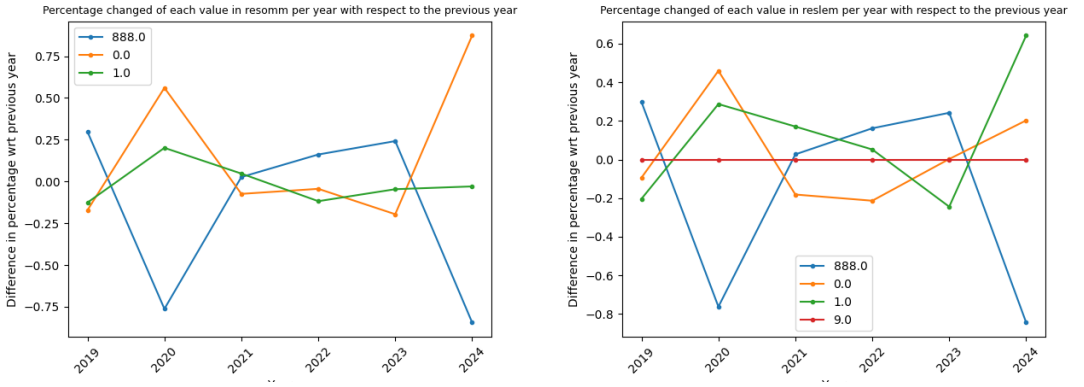
(a) `locaddvesr` **(b)** `locaddvesb`



(c) `locadddund` **(d)** `locaddsacr`

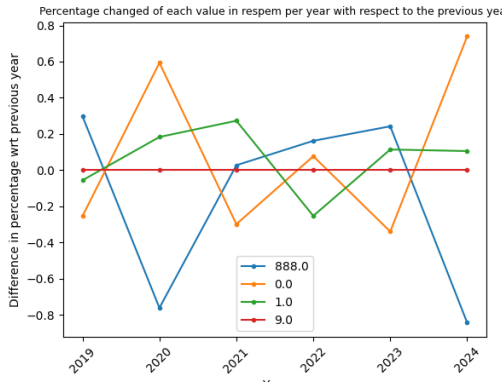


(e) `locaddoverig` **(f)** `resadm`

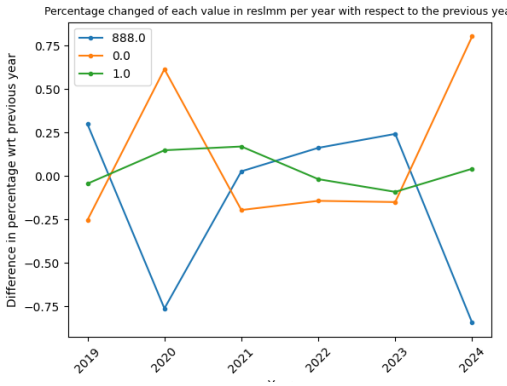


(g) `resomm` **(h)** `reslem`

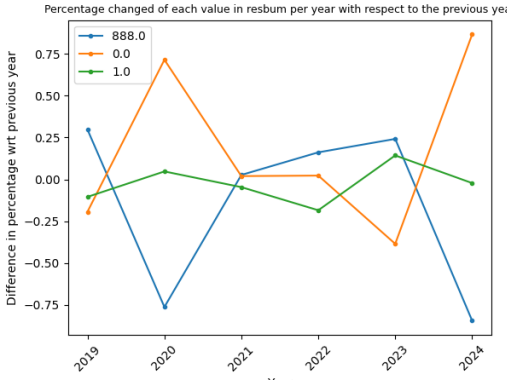
Figure F.24: Plots of the percentage change in values compared to the previous year of the discrete variables per year. Part 10/14.



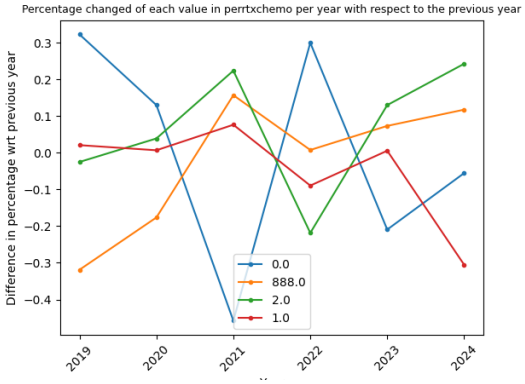
(a) `respem`



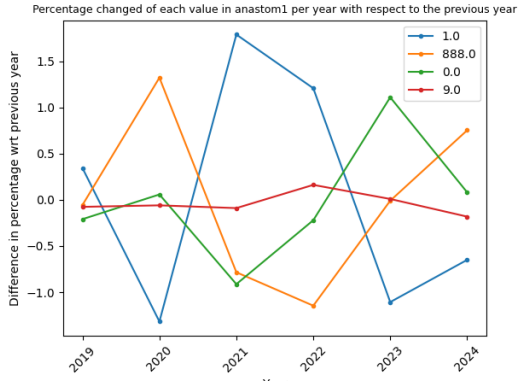
(b) `reslmm`



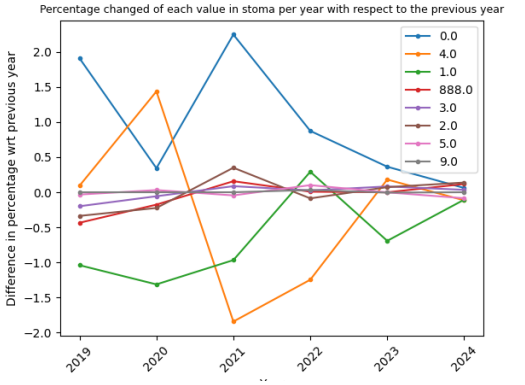
(c) `resbum`



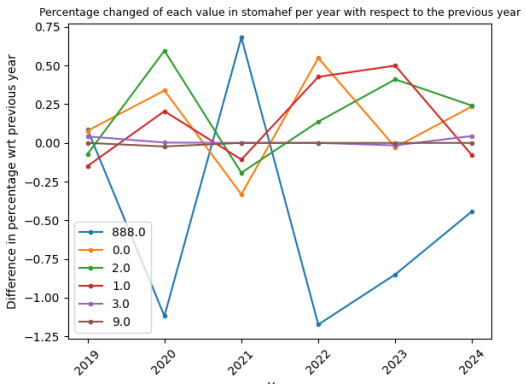
(d) `perrtxchemo`



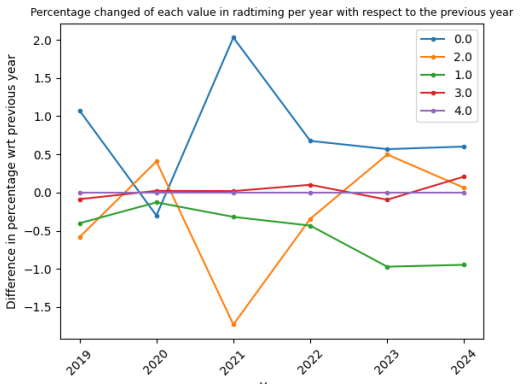
(e) `anastom1`



(f) `stoma`

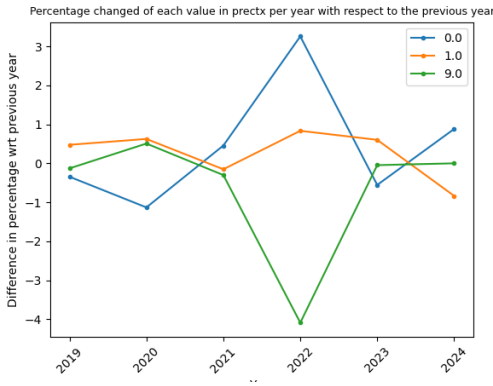


(g) `stomahef`

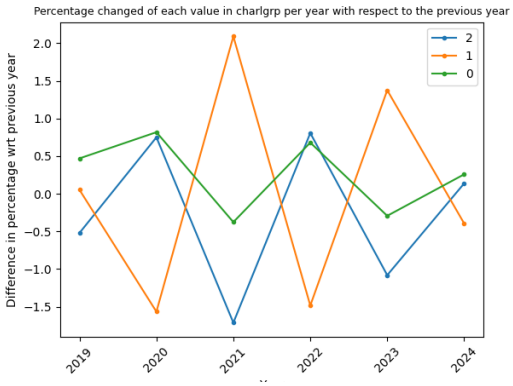


(h) `radtiming`

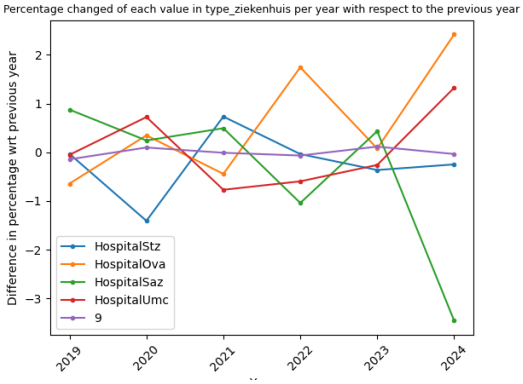
Figure F.25: Plots of the percentage change in values compared to the previous year of the discrete variables per year. Part 11/14.



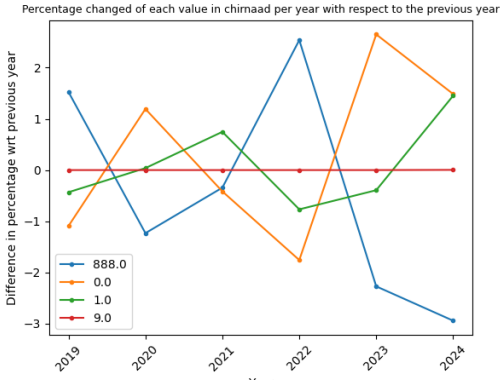
(a) `prectx`



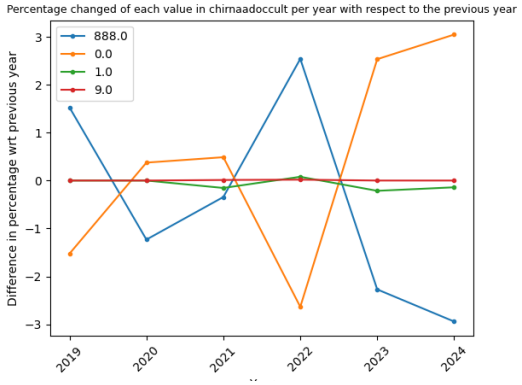
(b) `charlgrp`



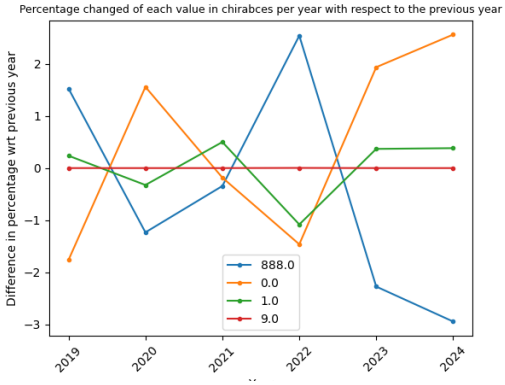
(c) `type_ziekenhuis`



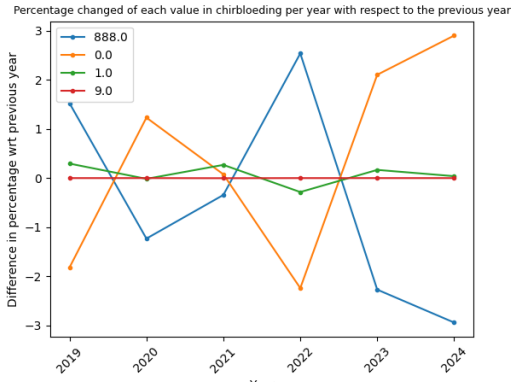
(d) `chirnaad`



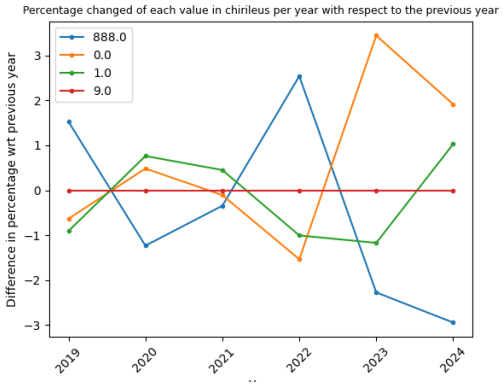
(e) `chirnaadoccult`



(f) `chirabces`

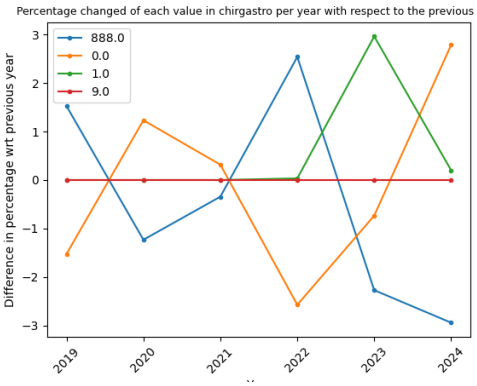


(g) `chirbloeding`

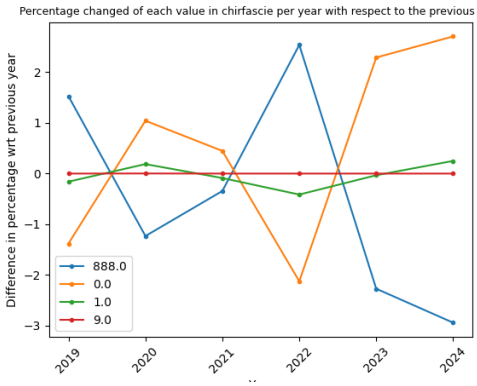


(h) `chirileus`

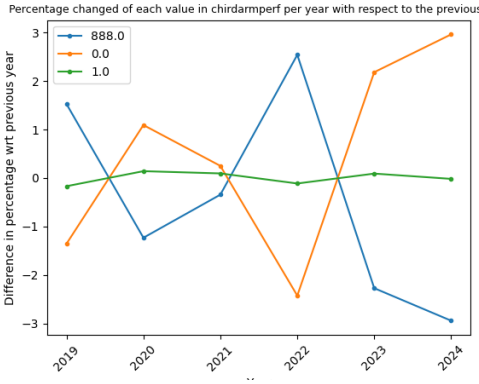
Figure F.26: Plots of the percentage change in values compared to the previous year of the discrete variables per year. Part 12/14.



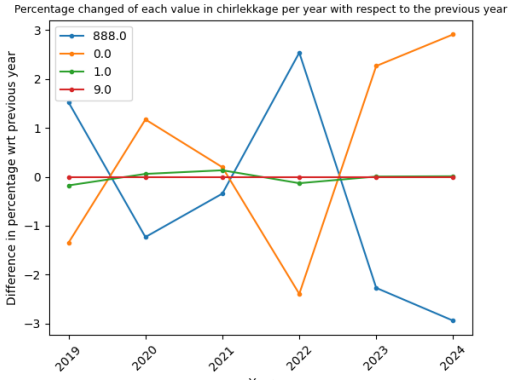
(a) `chirgastro`



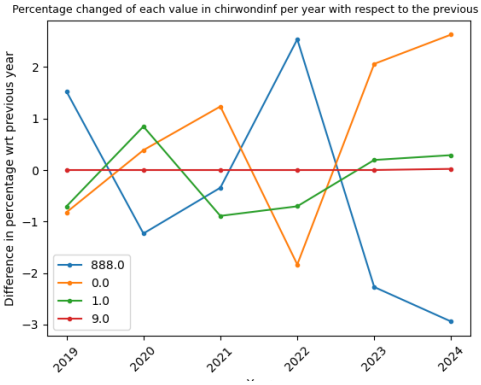
(b) `chirfascie`



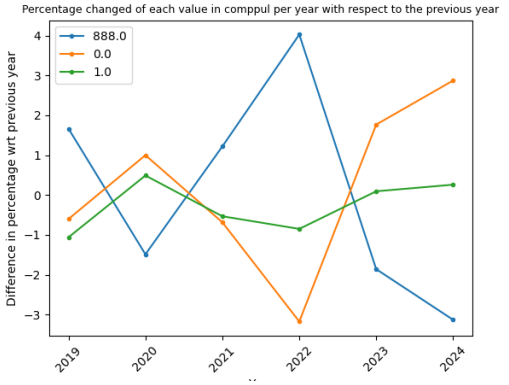
(c) `chirdarmperf`



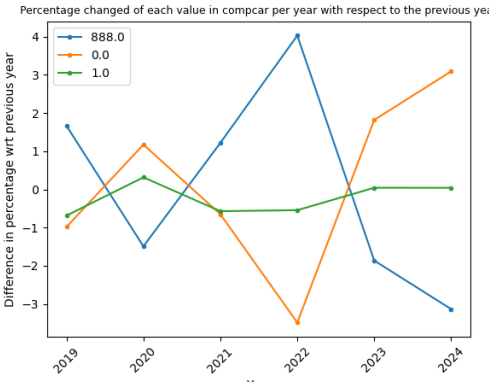
(d) `chirlekkage`



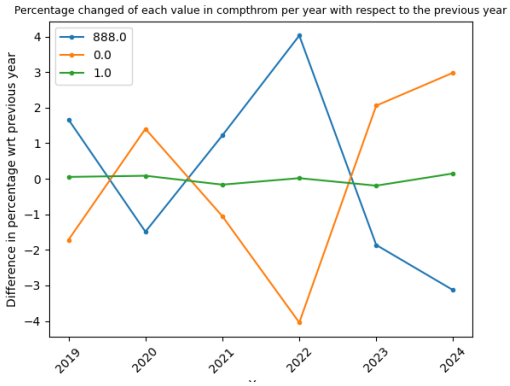
(e) `chirwondinf`



(f) `comppul`

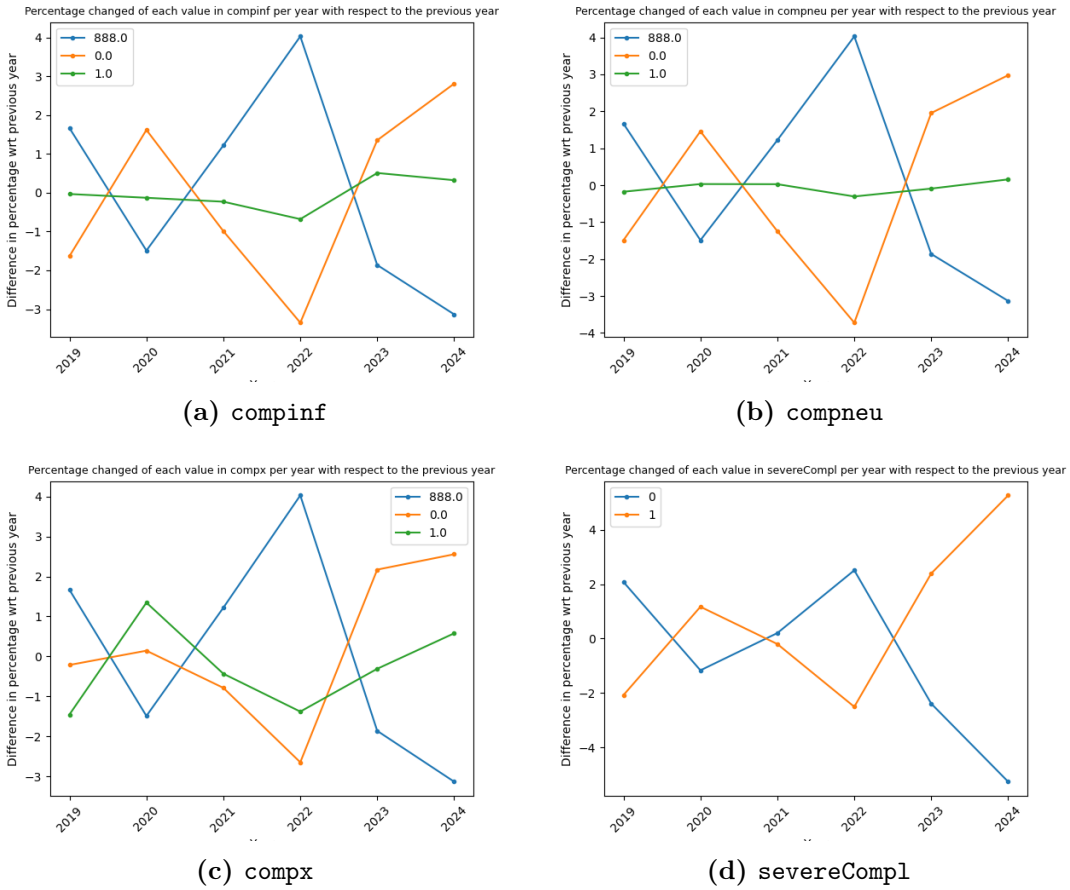


(g) `compcar`



(h) `compthrom`

Figure F.27: Plots of the percentage change in values compared to the previous year of the discrete variables per year. Part 13/14.



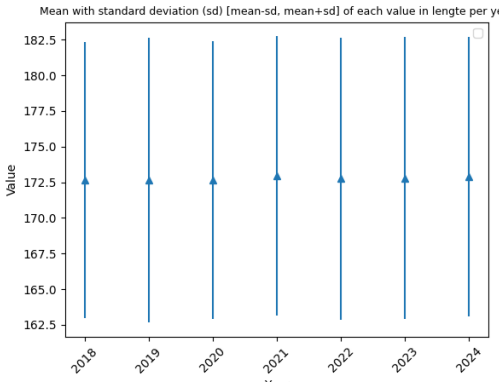
(a) `compinf` **(b)** `compneu`

(c) `compx` **(d)** `severeCompl`

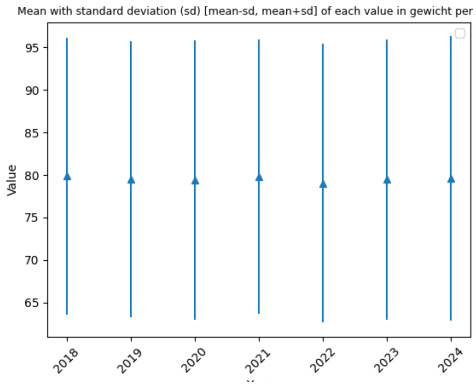
Figure F.28: Plots of the percentage change in values compared to the previous year of the discrete variables per year. Part 14/14.

F.2 Continuous variables

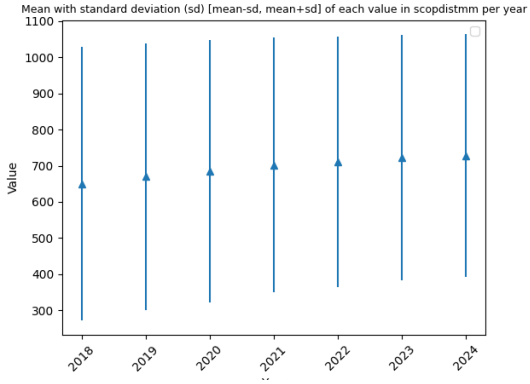
In this section we discuss the continuous variables and show their mean and standard deviation per year. This is shown in Figure F.29. The plots show the empirical mean (marked with an upward pointing arrow) and the empirical standard deviation (sd) (shown with the vertical lines above and below the mean). The 7 continuous variables are `lengte`, `gewicht`, `scopdistmm`, `scopdist2mm`, `leeftijd`, `waitingTime`, `lengthOfStay`.



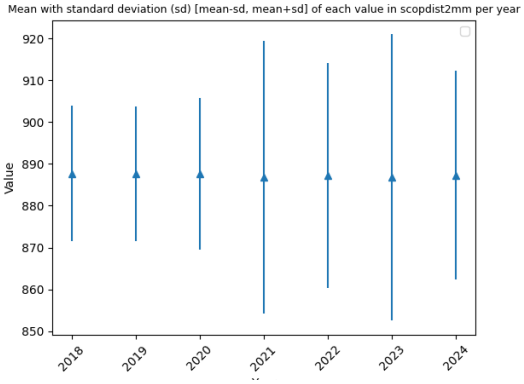
(a) `lengte`



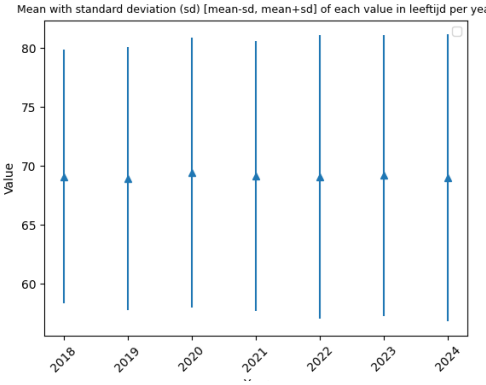
(b) `gewicht`



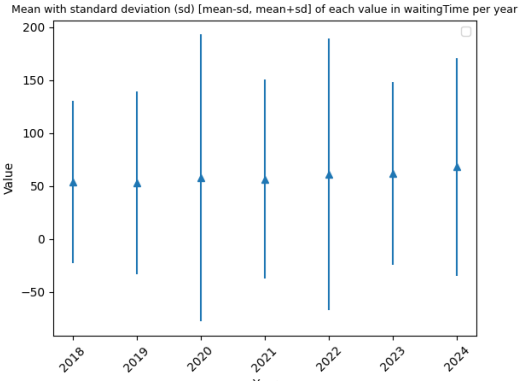
(c) `scopdistmm`



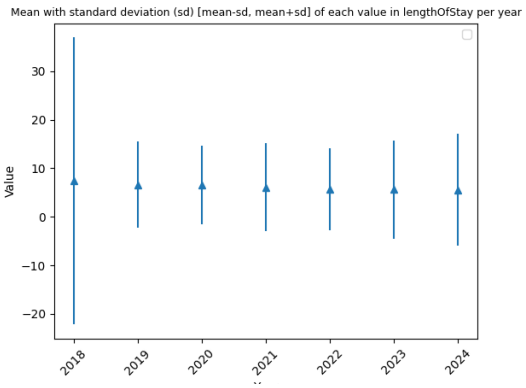
(d) `scopdist2mm`



(e) `leeftijd`



(f) `waitingTime`



(g) `lengthOfStay`

Figure F.29: Plots of the mean with standard deviation (sd) [mean-sd, mean+sd] of each value in the concerning variable per year.

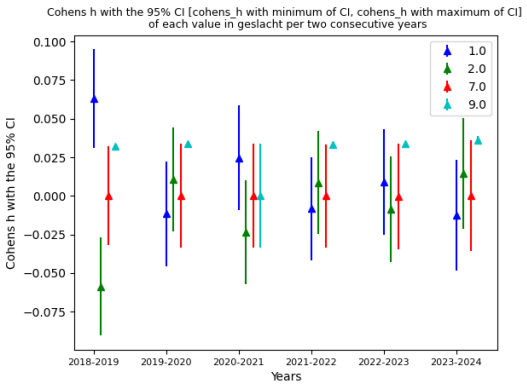
G | Graphs of Effect Size Analysis

In this appendix we show the graphs of the effect size analysis. In Section G.1 the 108 discrete variables are discussed. In Section G.2 the 7 continuous variables are treated.

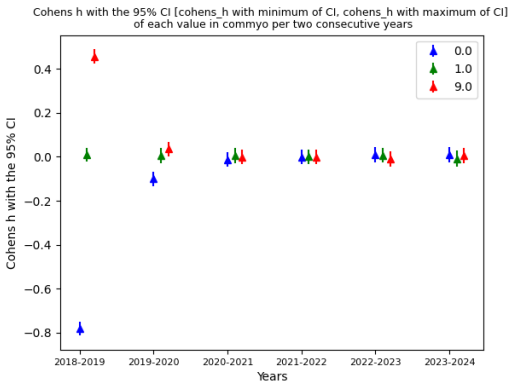
G.1 Discrete variables

In this section we show the plots of the effect size analysis of the discrete variables. The plots show the effect sizes per two consecutive years.

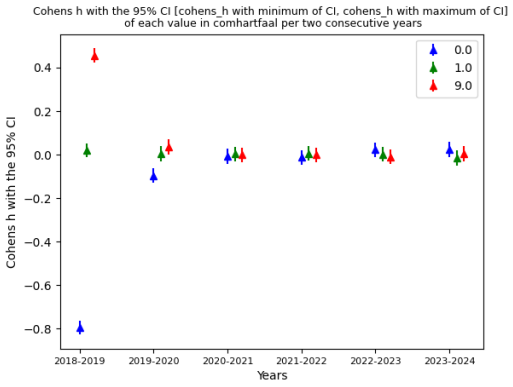
The 108 discrete variables are `geslacht`, `commyo`, `comhartfaal`, `comperif`, `comcva`, `comdem`, `comlong`, `combind`, `comgiulcus`, `comlever`, `comdiam`, `comparalys`, `comnier`, `commalig`, `comhiv`, `asascore`, `resdarm`, `stomavg`, `primcrc`, `scopnumb`, `diagn`, `locprim`, `histd`, `complpre`, `preanemie`, `preobstruct`, `preabces`, `scorect`, `scorecn`, `locprim2`, `scorect2`, `scorecn2`, `scorecm`, `prelocmri`, `prelocmri2`, `verwezen`, `darmperf`, `ileusblow`, `perfperi`, `perfscopie`, `prechirstomanew`, `prechirstent`, `prechirmetanew`, `prechirappendix`, `prechircoloscex`, `chirenkel`, `typok`, `procok`, `ingreep`, `conversien`, `aprtyp`, `typprocok`, `procok2`, `ingreep2`, `conversien2`, `aprtyp2`, `resadd`, `locaddmaag`, `locaddmilt`, `locaddduod`, `locadddoverig`, `locaddoment`, `locaddbuikw`, `locaddvag`, `locadduter`, `locaddovarl`, `locaddovarr`, `locadduretl`, `locadduretr`, `locaddblaas`, `locaddprost`, `locaddvesl`, `locaddvesr`, `locaddvesb`, `locadddund`, `locaddsacr`, `locaddoverig`, `resadm`, `resomm`, `reslem`, `respem`, `reslmm`, `resbum`, `perrtxchemo`, `anastom1`, `stoma`, `stomahef`, `radtiming`, `prectx`, `charlgrp`, `type_ziekenhuis`, `chirnaad`, `chirnaadoccult`, `chirabces`, `chirbloeding`, `chirileus`, `chirgastro`, `chirfascie`, `chirdarmperf`, `chirlekkage`, `chirwondinf`, `comppul`, `compcar`, `compthrom`, `compinf`, `compneu`, `compx`, and `severeCompl`.



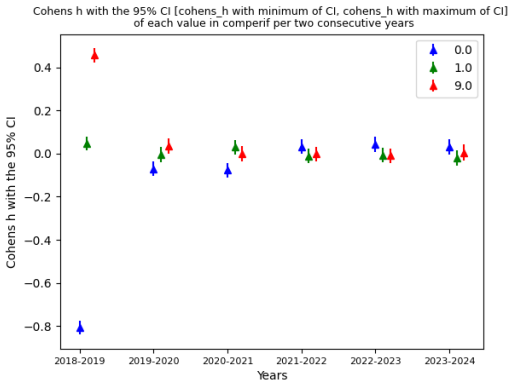
(a) `geslacht`



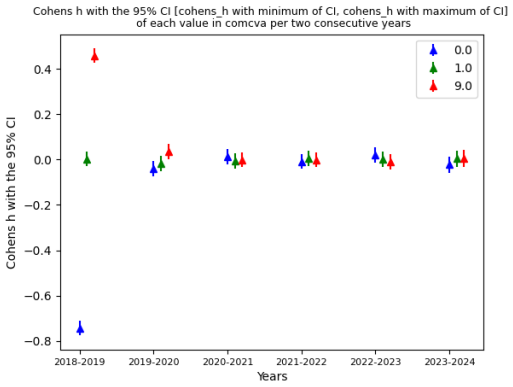
(b) `commyo`



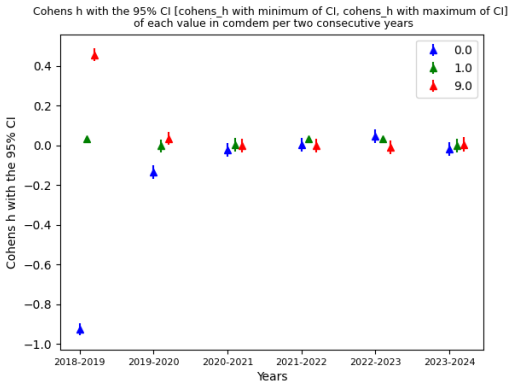
(c) `comhartfaal`



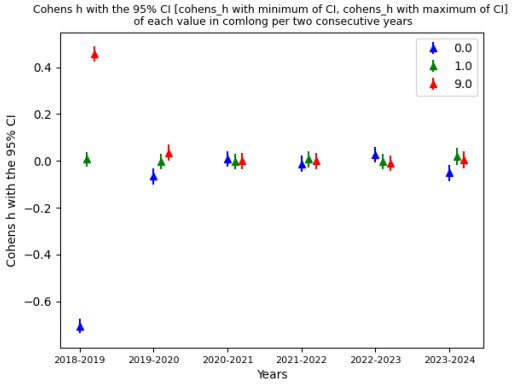
(d) `comperif`



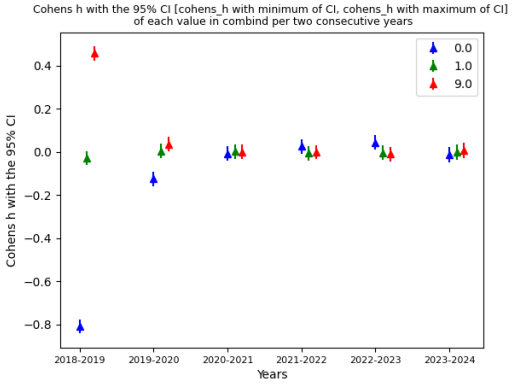
(e) `comcva`



(f) `comdem`

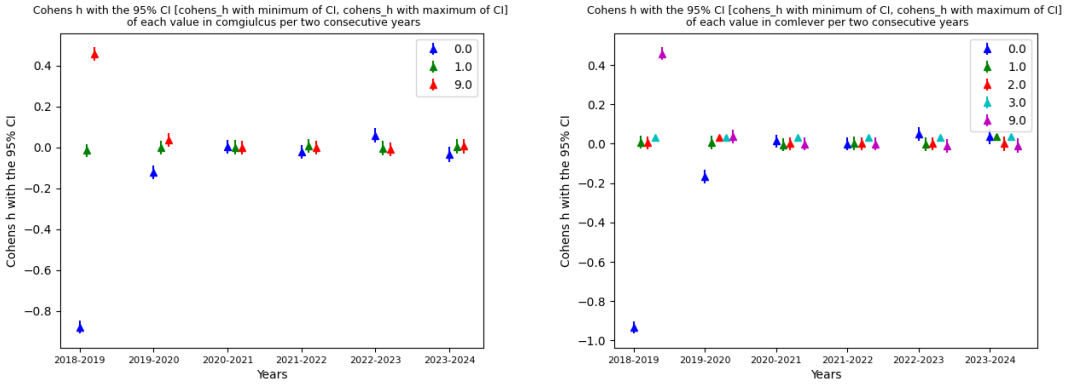


(g) `comlong`



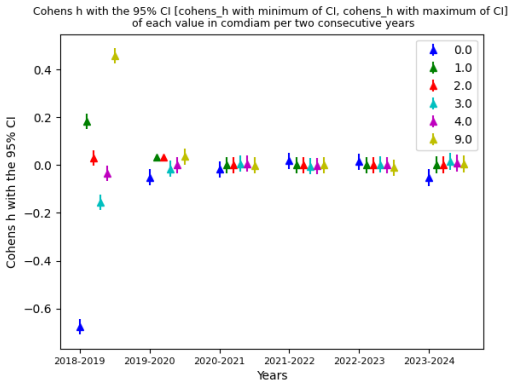
(h) `combind`

Figure G.1: Plots of the effect sizes of the discrete variables per two consecutive years. Part 1/14.

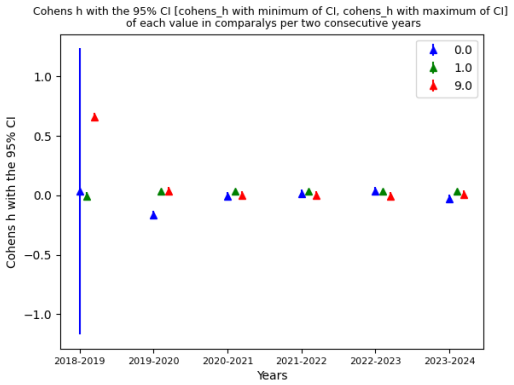


(a) `comgiulcus`

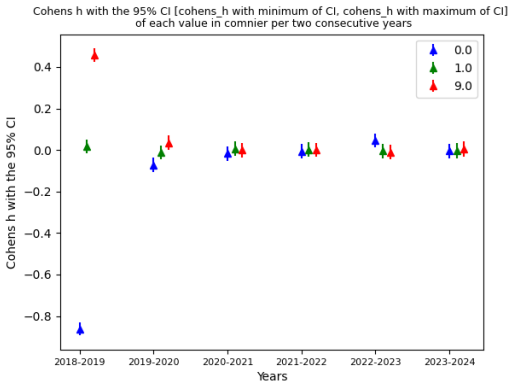
(b) `comlever`



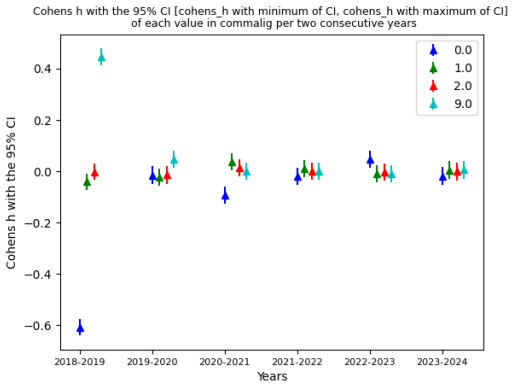
(c) `comdiam`



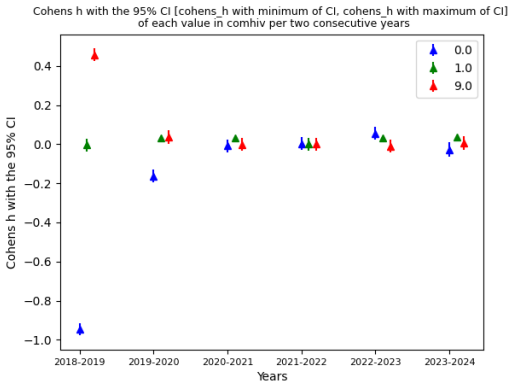
(d) `comparalys`



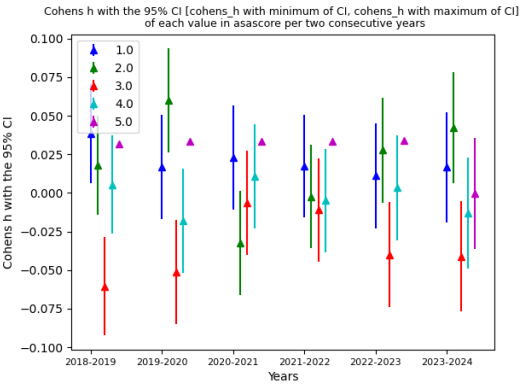
(e) `comnier`



(f) `commalig`

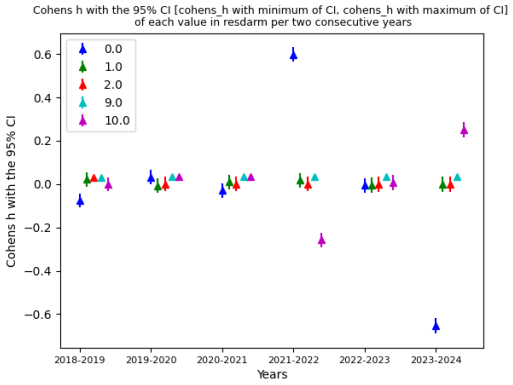


(g) `comhiv`

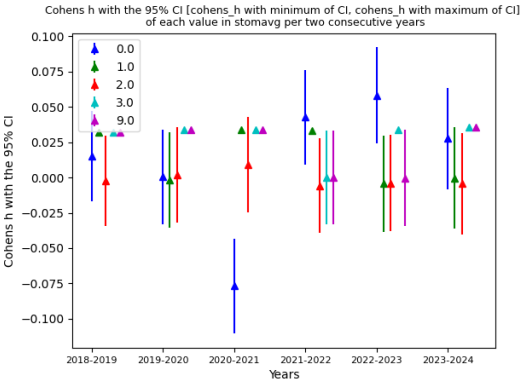


(h) `asascore`

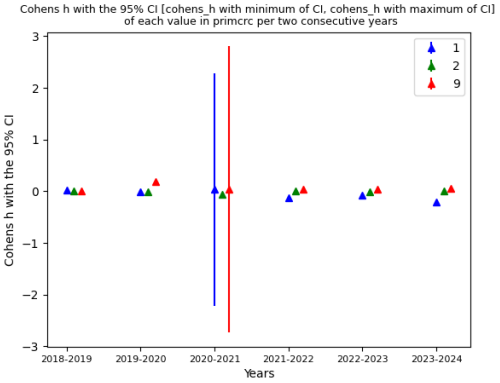
Figure G.2: Plots of the effect sizes of the discrete variables per two consecutive years. Part 2/14.



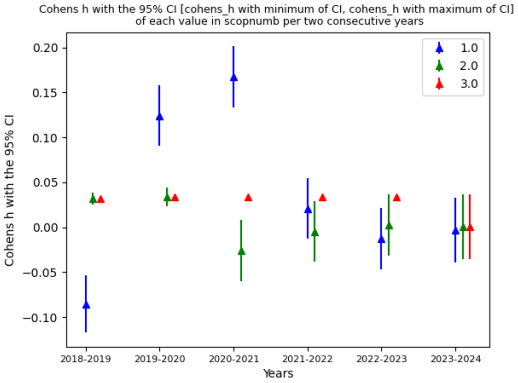
(a) `resdarm`



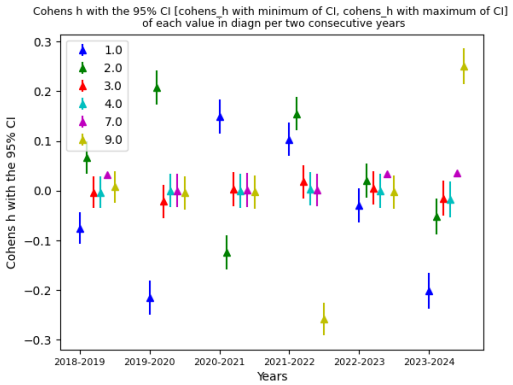
(b) `stomavg`



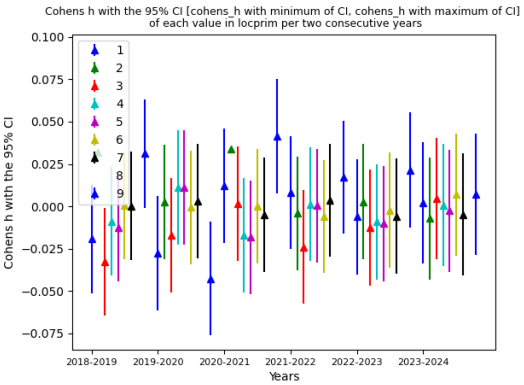
(c) `primcrc`



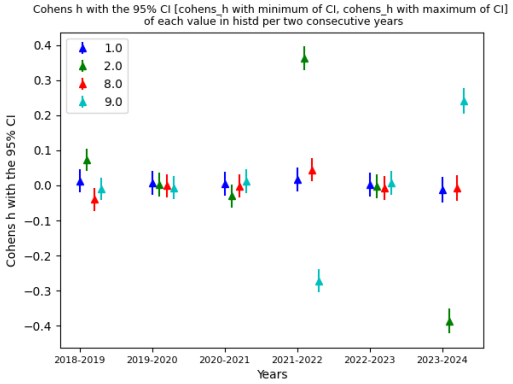
(d) `scopnumb`



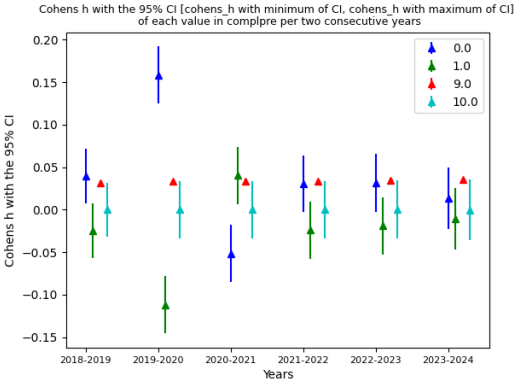
(e) `diagn`



(f) `locprim`

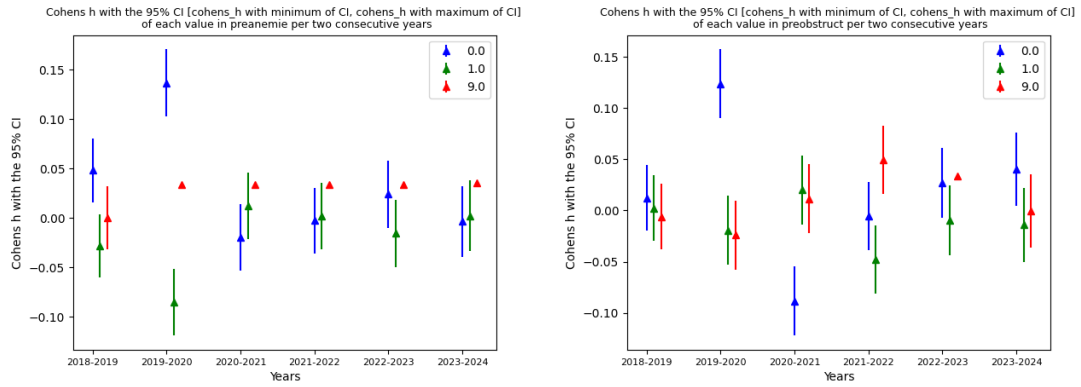


(g) `histd`



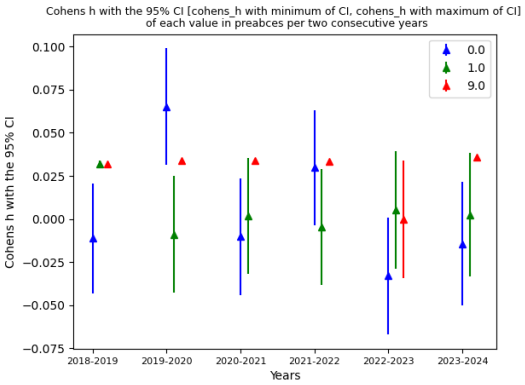
(h) `complpre`

Figure G.3: Plots of the effect sizes of the discrete variables per two consecutive years. Part 3/14.

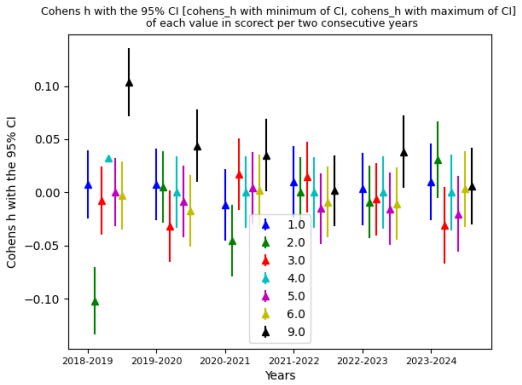


(a) `preanemie`

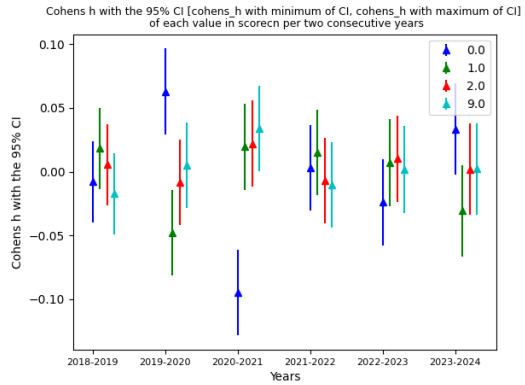
(b) `preobstruct`



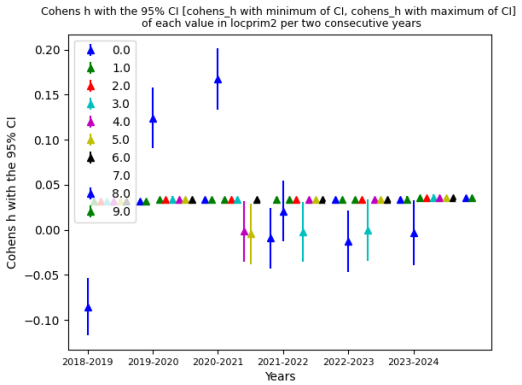
(c) `preabces`



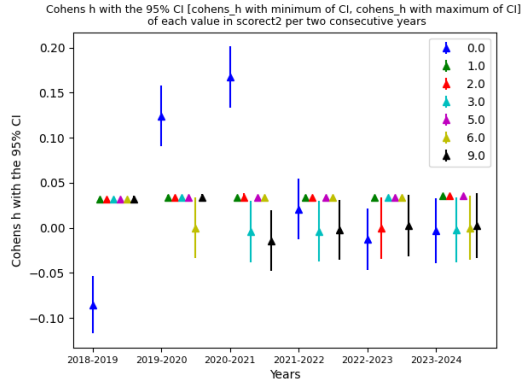
(d) `scorect`



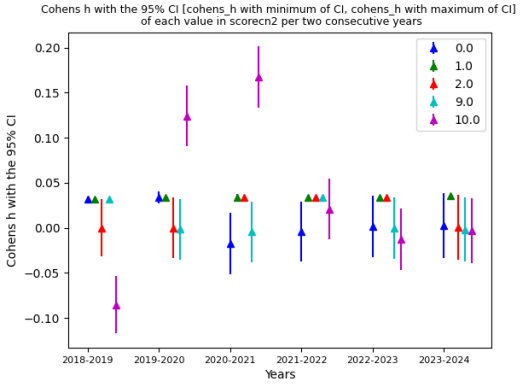
(e) `scorecn`



(f) `locprim2`

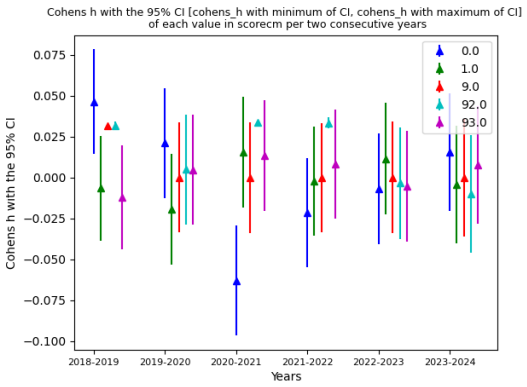


(g) `scorect2`

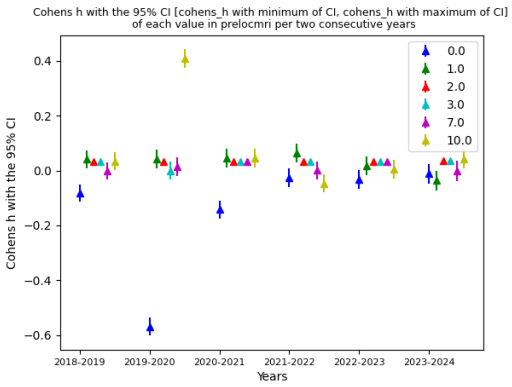


(h) `scorecn2`

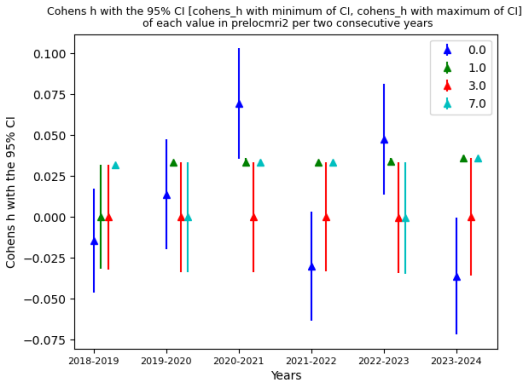
Figure G.4: Plots of the effect sizes of the discrete variables per two consecutive years. Part 4/14.



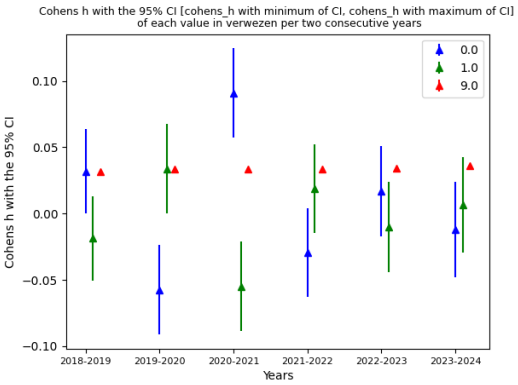
(a) `scorecm`



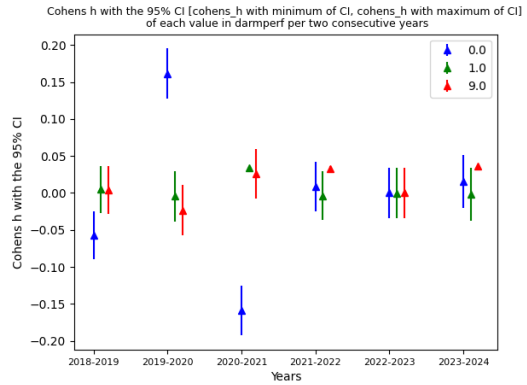
(b) `prelocmri`



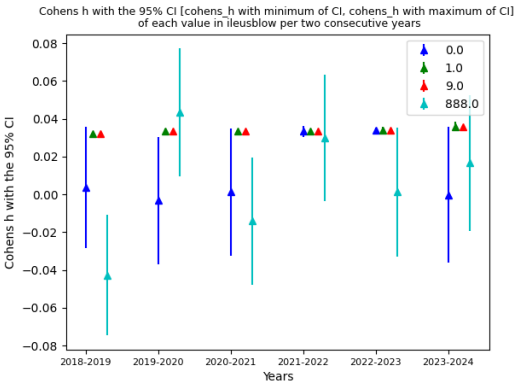
(c) `prelocmri2`



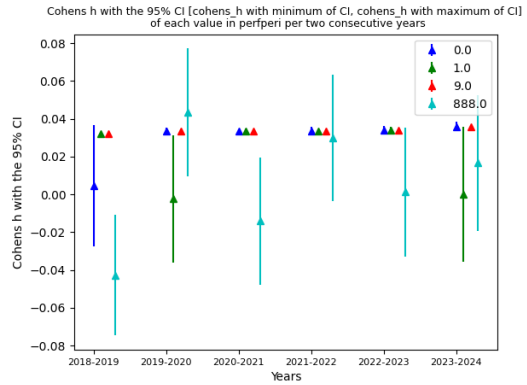
(d) `verwezen`



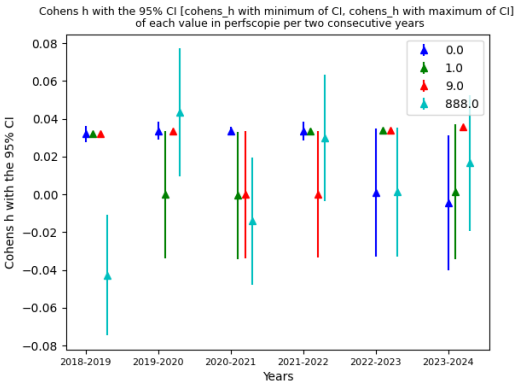
(e) `darmperf`



(f) `ileusblow`

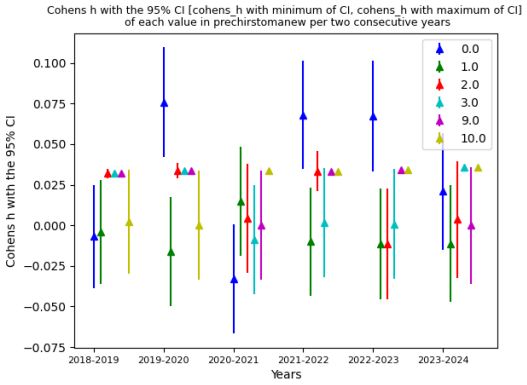


(g) `perfperi`

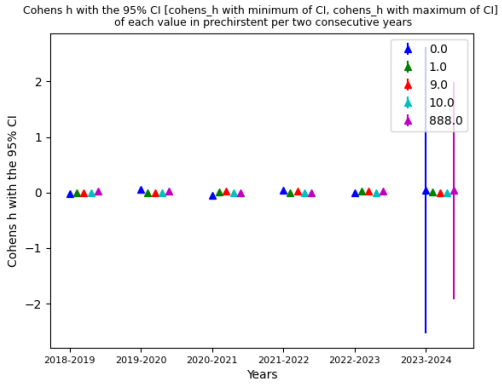


(h) `perfscopie`

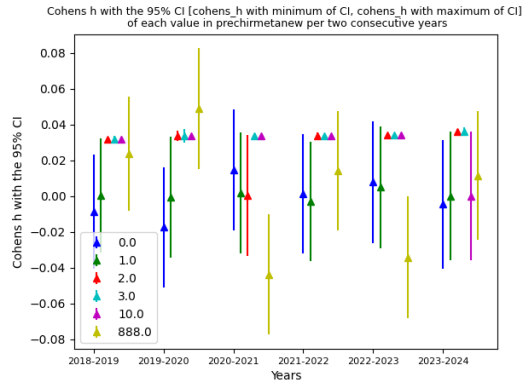
Figure G.5: Plots of the effect sizes of the discrete variables per two consecutive years. Part 5/14.



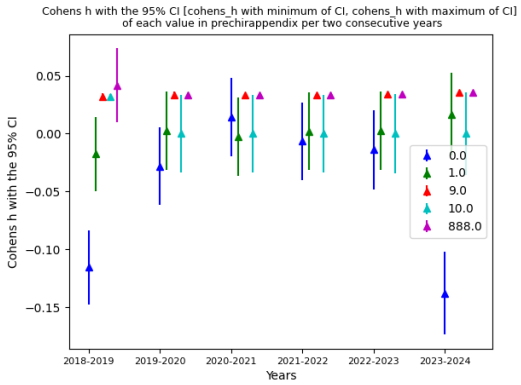
(a) `prechirstomanew`



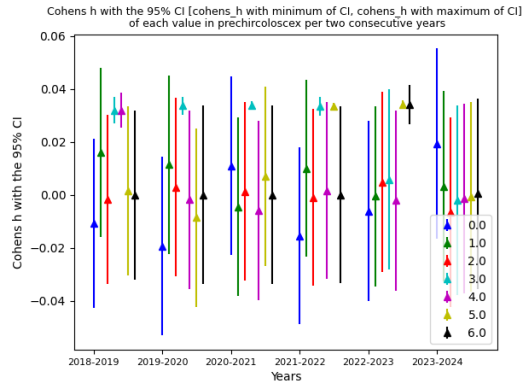
(b) `prechirstent`



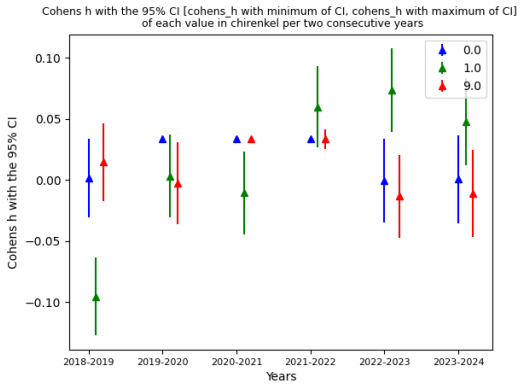
(c) `prechirmetanew`



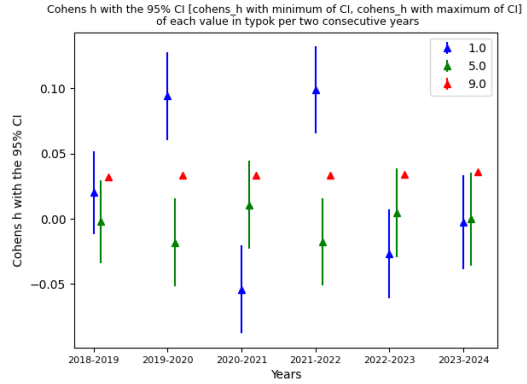
(d) `prechirappendix`



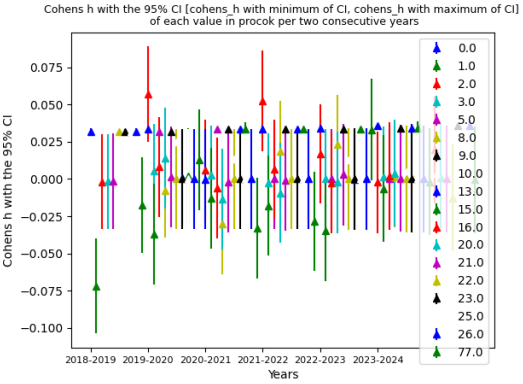
(e) `prechircoloscex`



(f) `chirenkel`

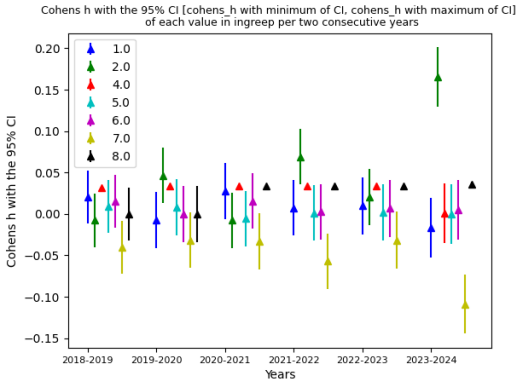


(g) `typok`

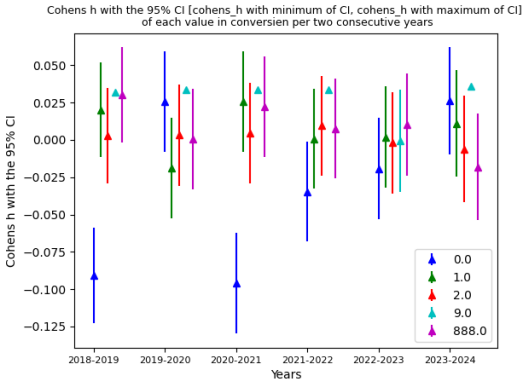


(h) `procok`

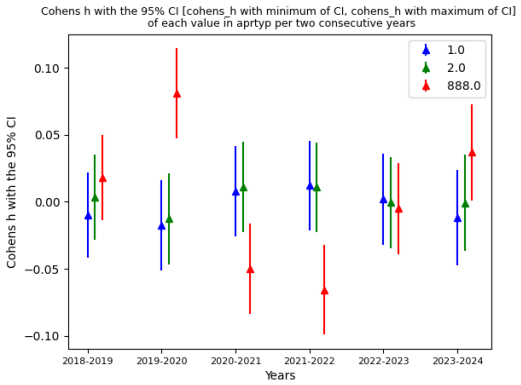
Figure G.6: Plots of the effect sizes of the discrete variables per two consecutive years. Part 6/14.



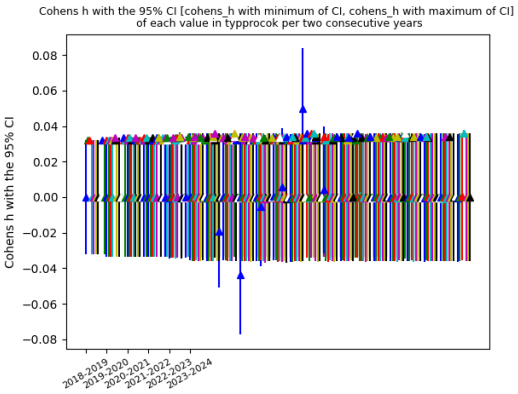
(a) `ingreep`



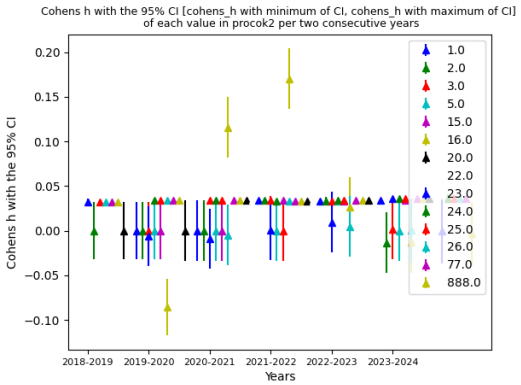
(b) `conversien`



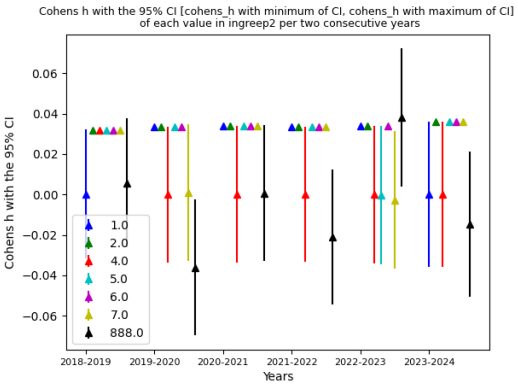
(c) `aprtyp`



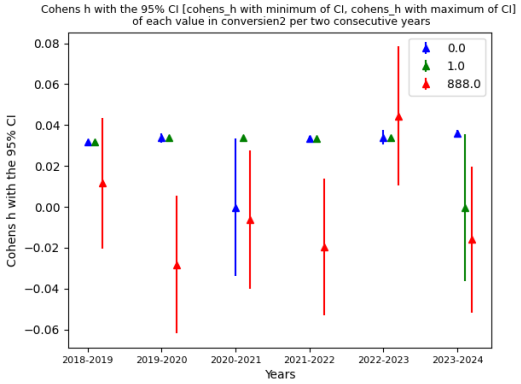
(d) `typprocok`



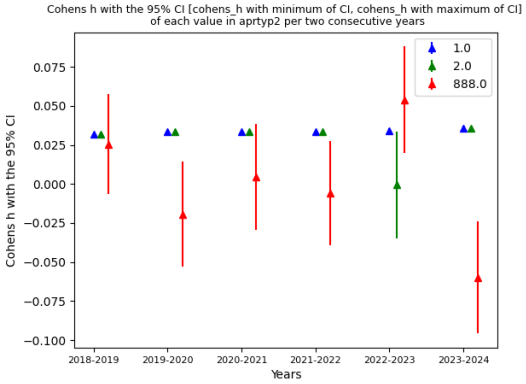
(e) `procok2`



(f) `ingreep2`

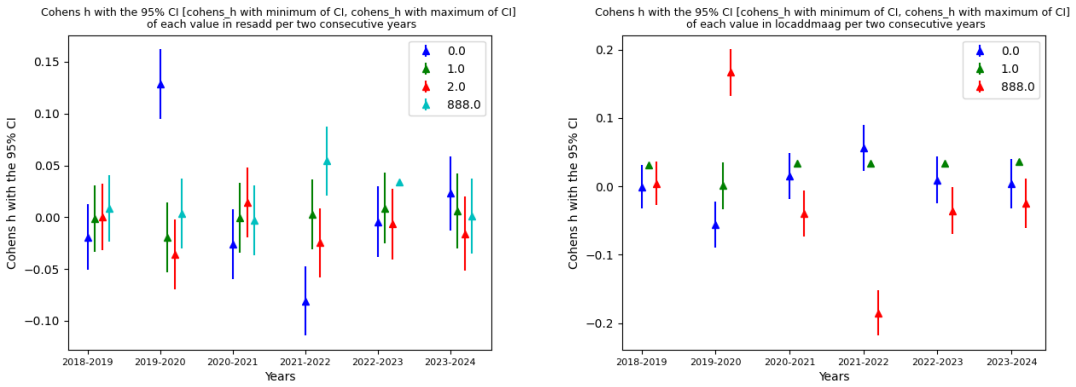


(g) `conversien2`



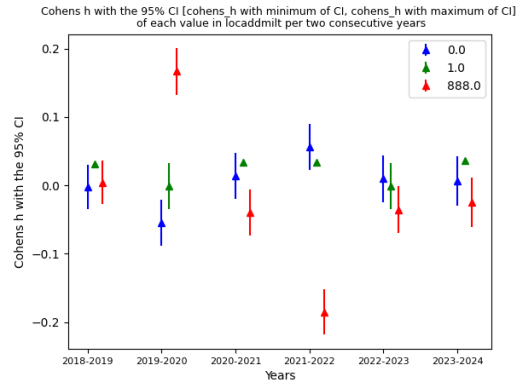
(h) `aprtyp2`

Figure G.7: Plots of the effect sizes of the discrete variables per two consecutive years. Part 7/14.

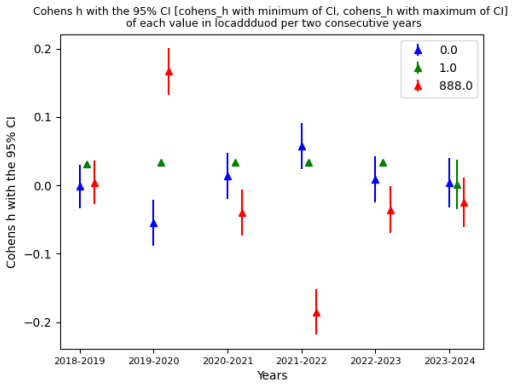


(a) `resadd`

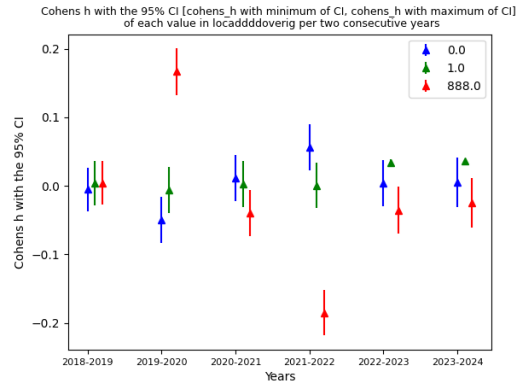
(b) `locaddmaag`



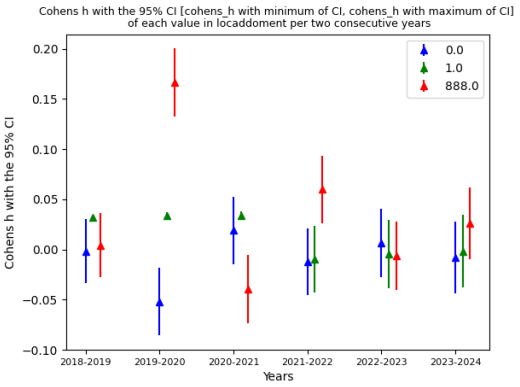
(c) `locaddmilt`



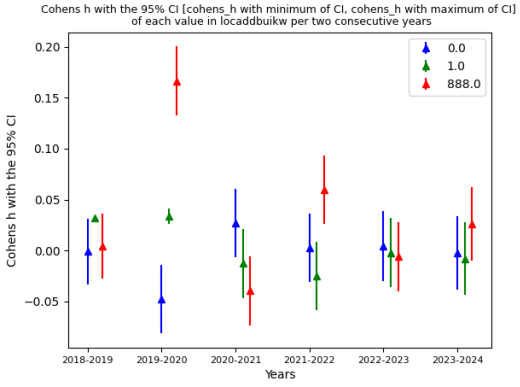
(d) `locaddduod`



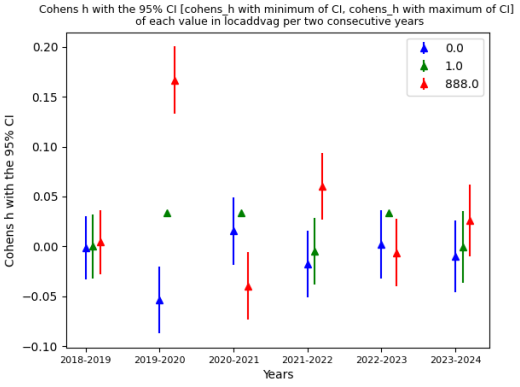
(e) `locadddoverig`



(f) `locaddoment`

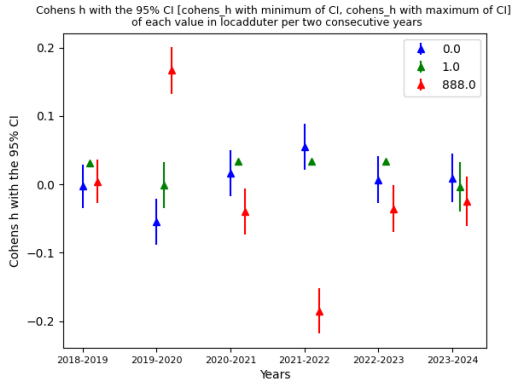


(g) `locaddbuikw`

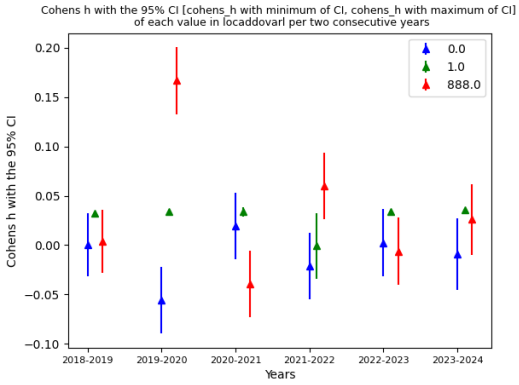


(h) `locaddvag`

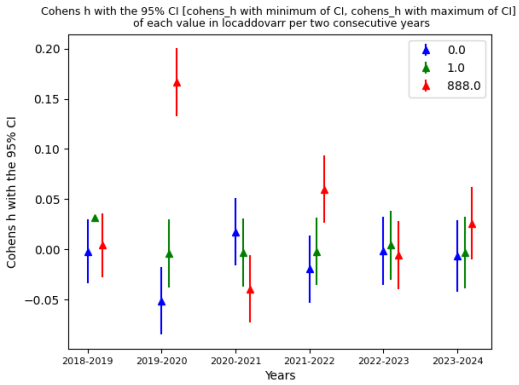
Figure G.8: Plots of the effect sizes of the discrete variables per two consecutive years. Part 8/14.



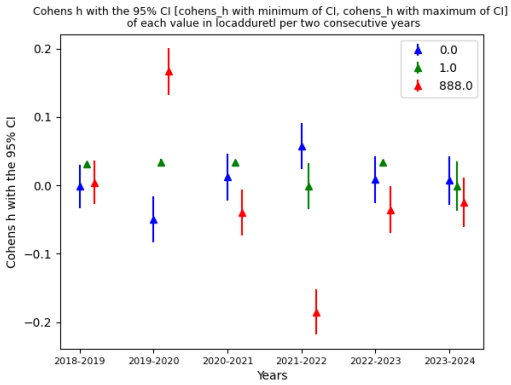
(a) `locadduter`



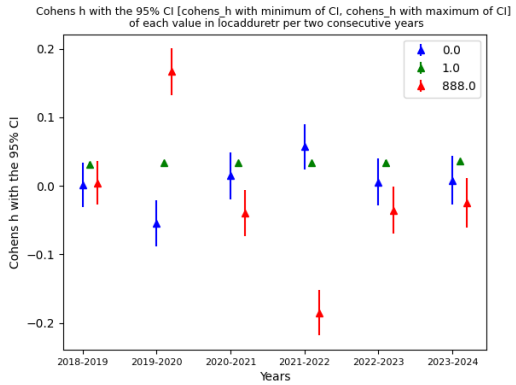
(b) `locaddovarl`



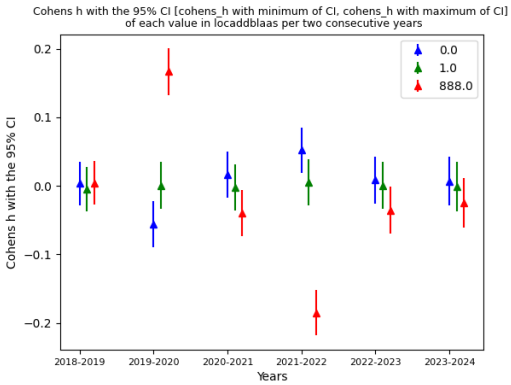
(c) `locaddovarr`



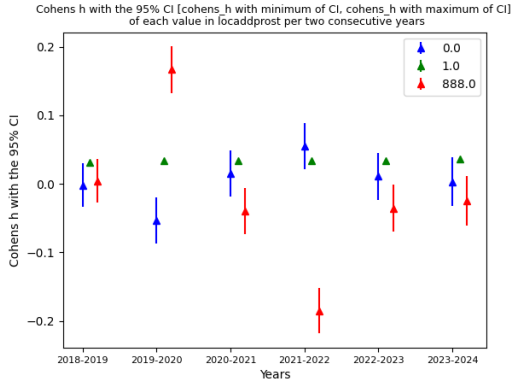
(d) `locadduretl`



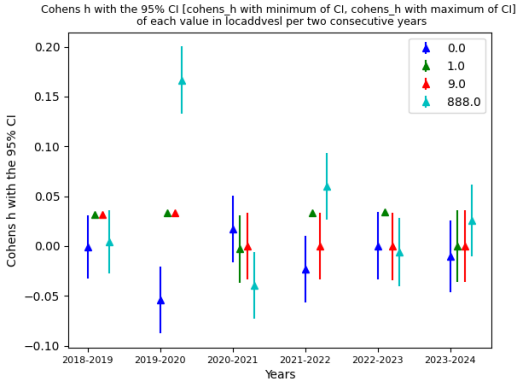
(e) `locadduretr`



(f) `locaddblaas`

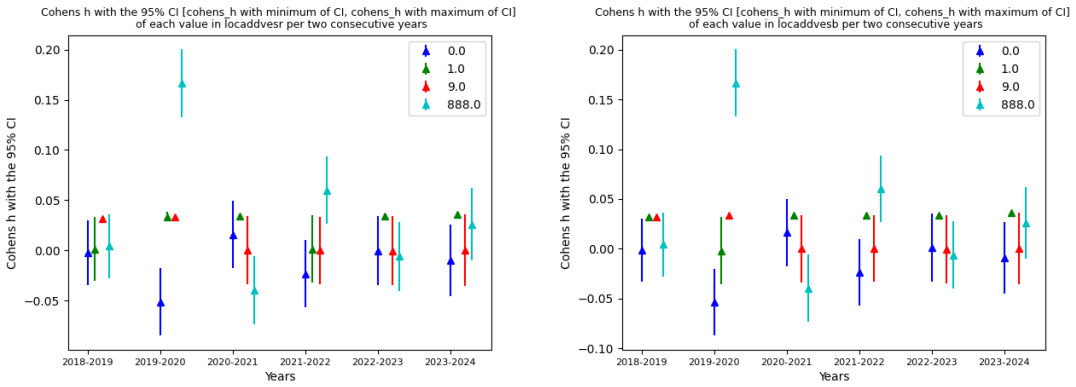


(g) `locaddprost`



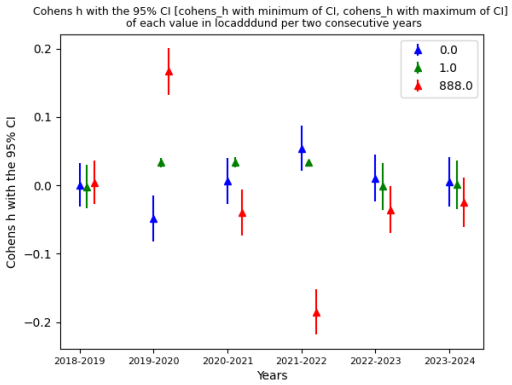
(h) `locaddvesl`

Figure G.9: Plots of the effect sizes of the discrete variables per two consecutive years. Part 9/14.

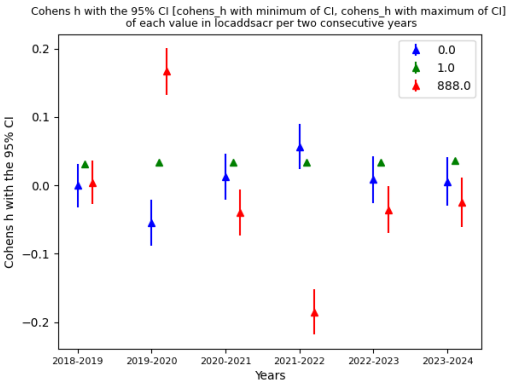


(a) `locaddvesr`

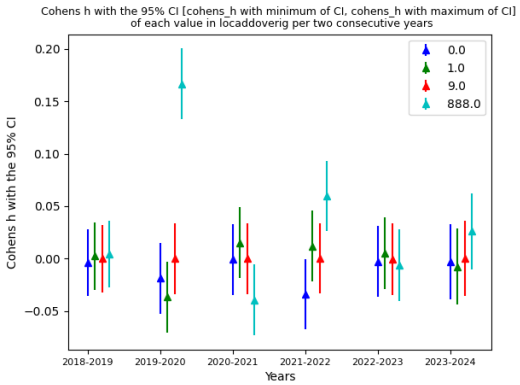
(b) `locaddvesb`



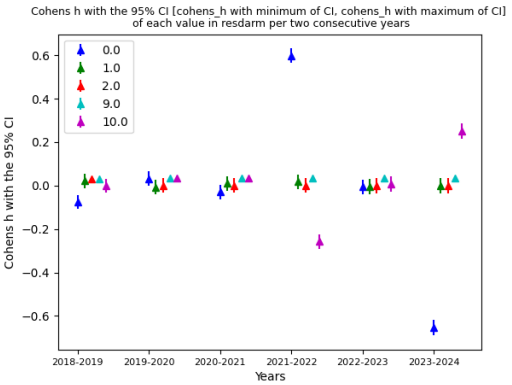
(c) `locadddund`



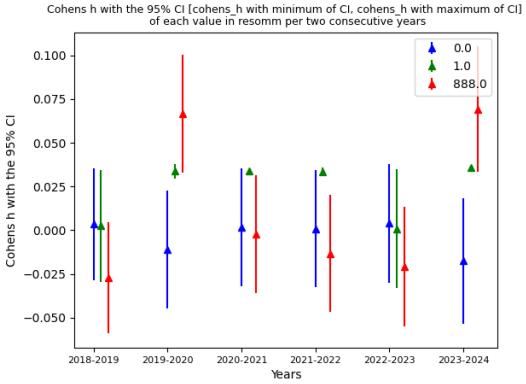
(d) `locaddsacr`



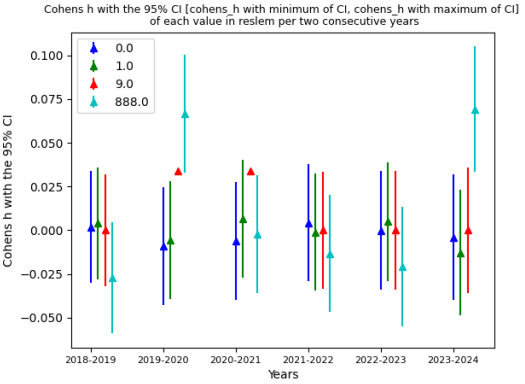
(e) `locaddoverig`



(f) `resadm`

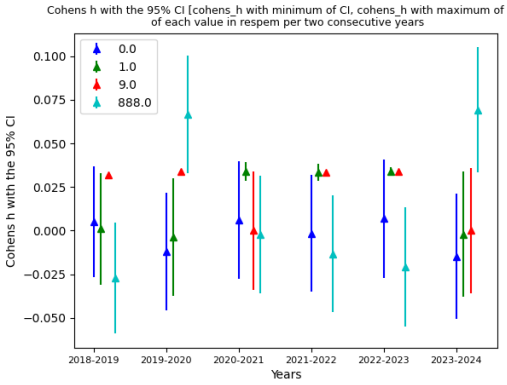


(g) `resomm`

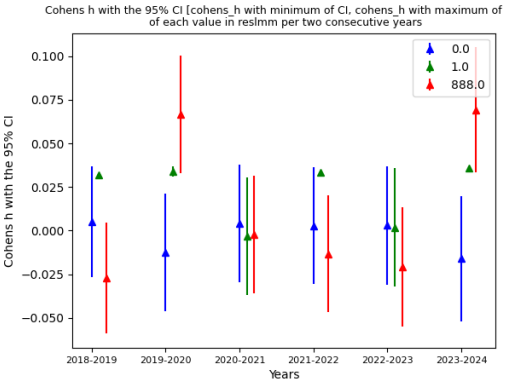


(h) `reslem`

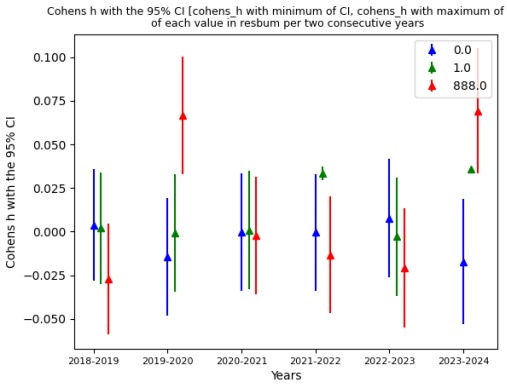
Figure G.10: Plots of the effect sizes of the discrete variables per two consecutive years. Part 10/14.



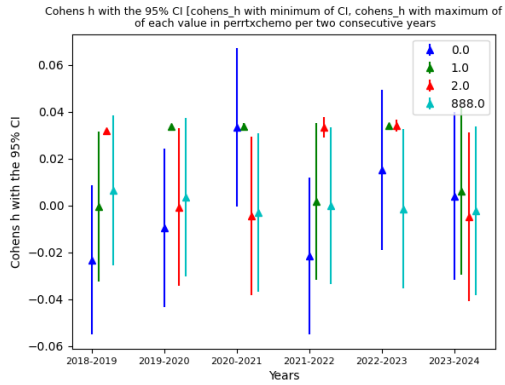
(a) `respem`



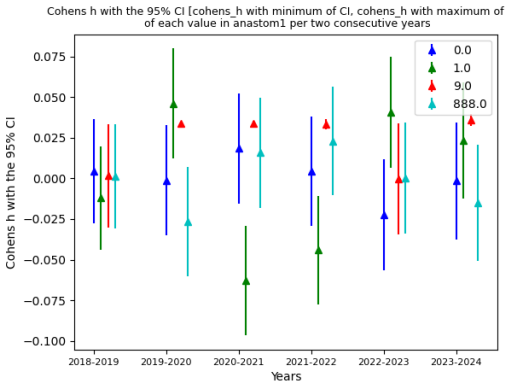
(b) `reslmm`



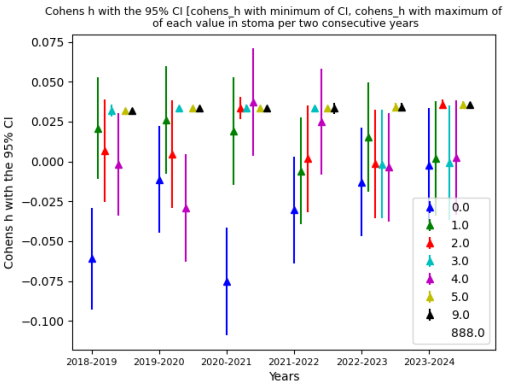
(c) `resbum`



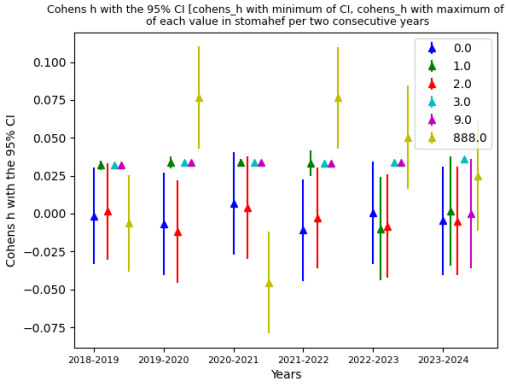
(d) `perrtxchemo`



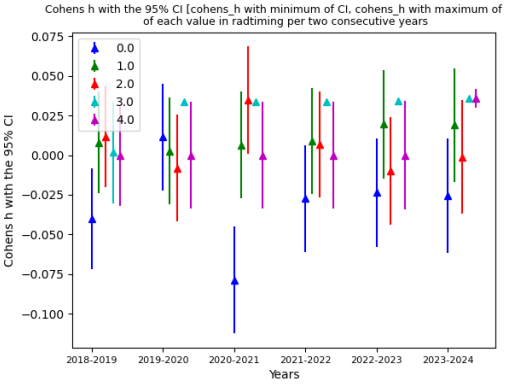
(e) `anastom1`



(f) `stoma`

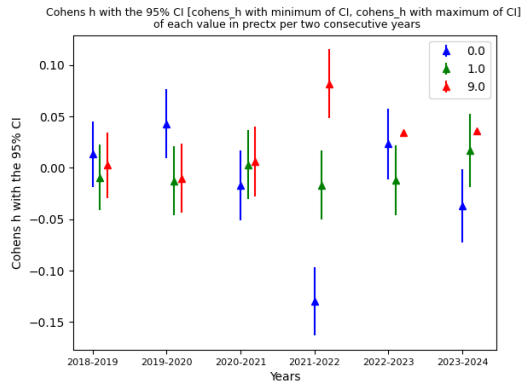


(g) `stomahef`

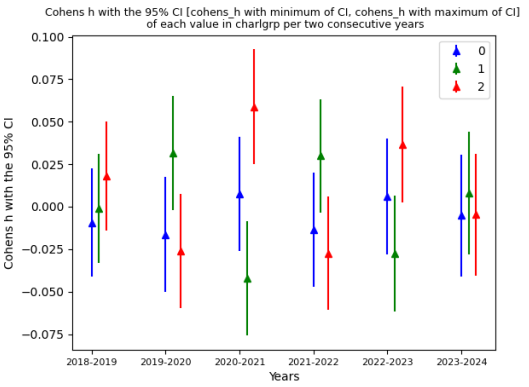


(h) `radtiming`

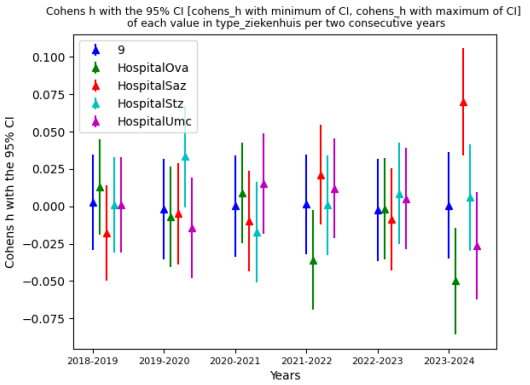
Figure G.11: Plots of the effect sizes of the discrete variables per two consecutive years. Part 11/14.



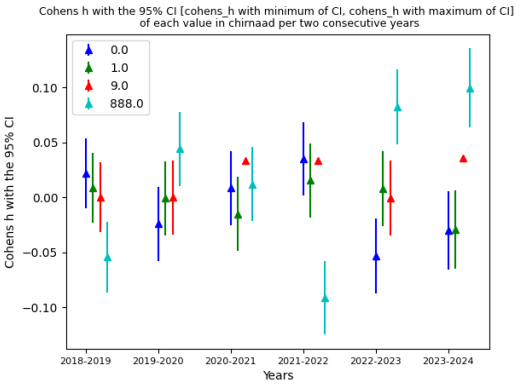
(a) `prectx`



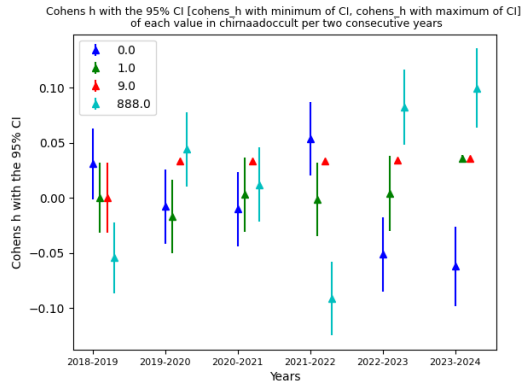
(b) `charlgrp`



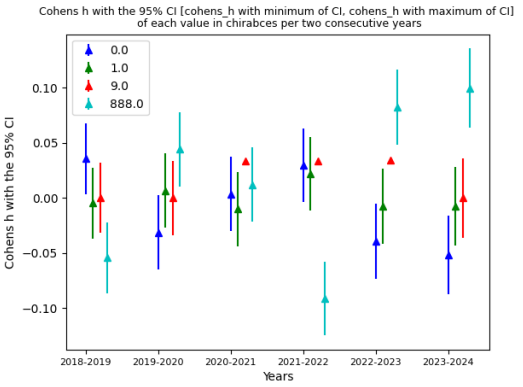
(c) `type_ziekenhuis`



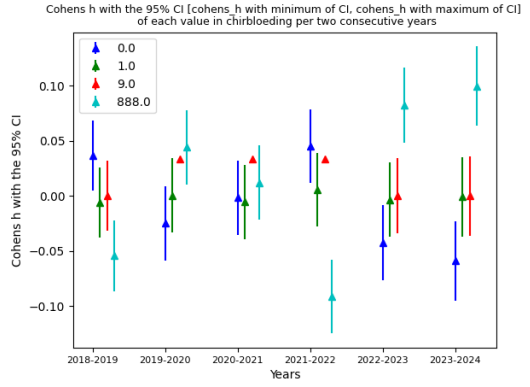
(d) `chirnaad`



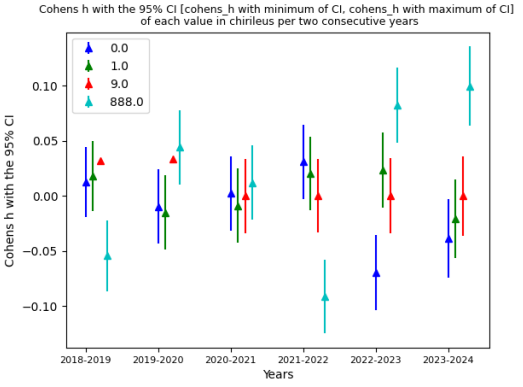
(e) `chirnaadoccult`



(f) `chirabces`

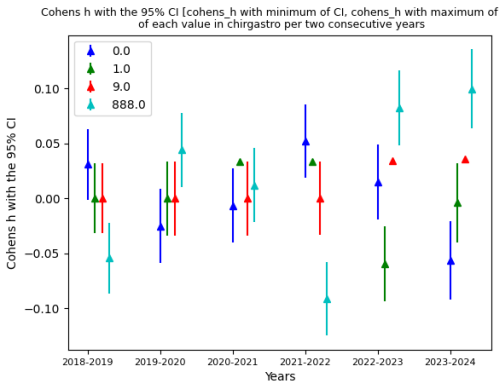


(g) `chirbloeding`

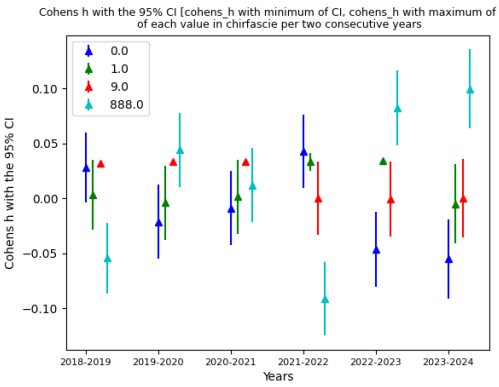


(h) `chirileus`

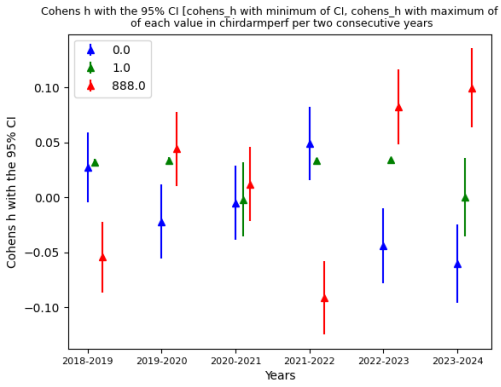
Figure G.12: Plots of the effect sizes of the discrete variables per two consecutive years. Part 12/14.



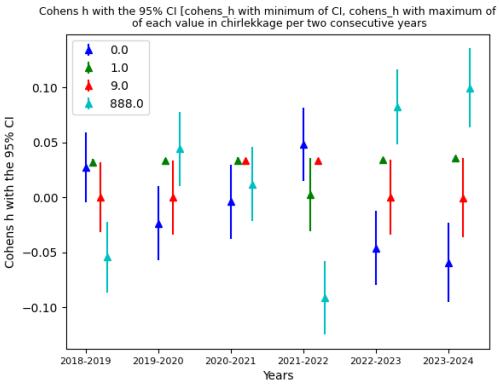
(a) `chirgastro`



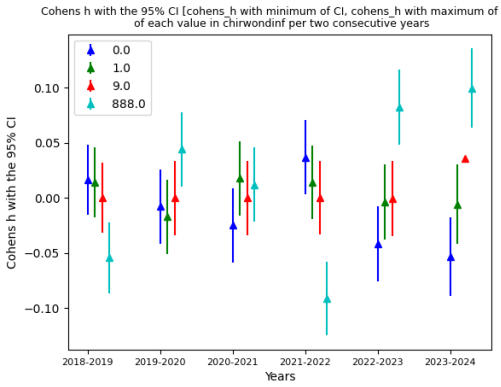
(b) `chirfascie`



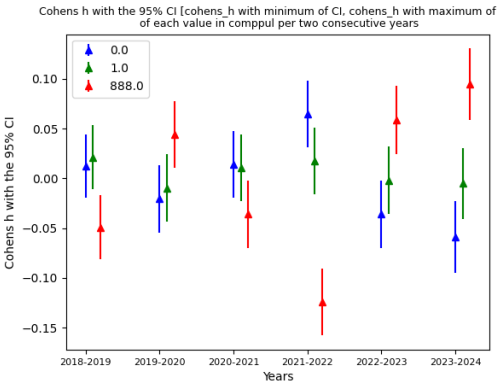
(c) `chirdarmperf`



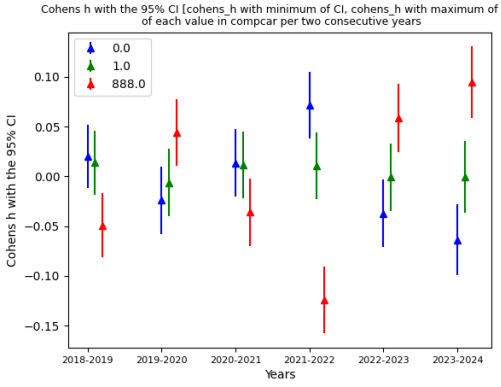
(d) `chirlekkage`



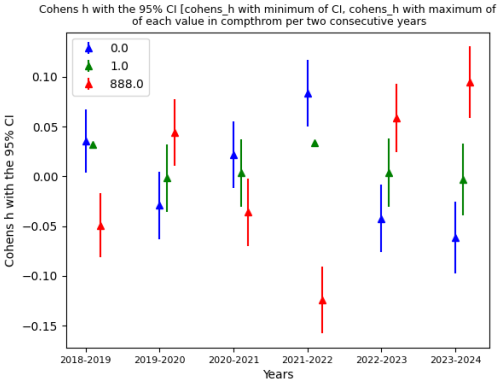
(e) `chirwondinf`



(f) `comppul`

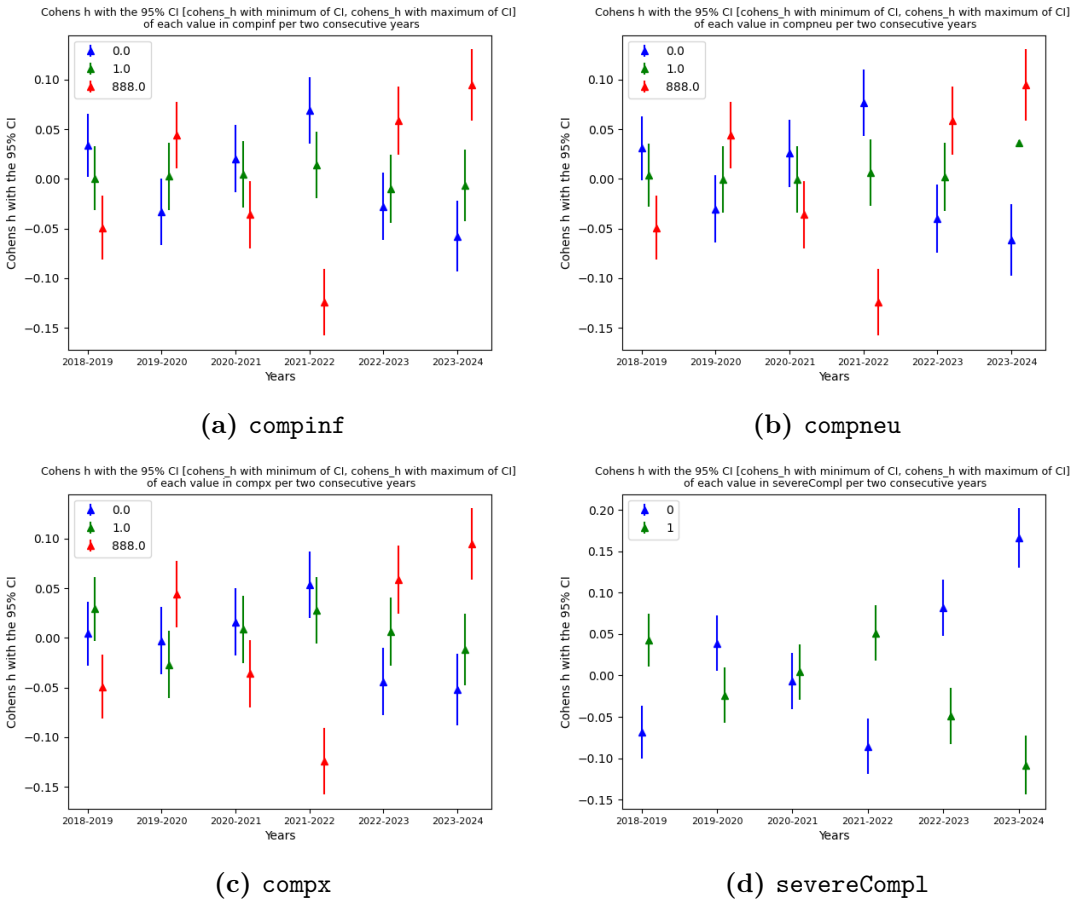


(g) `compcar`



(h) `compthrom`

Figure G.13: Plots of the effect sizes of the discrete variables per two consecutive years. Part 13/14.



(a) `compinf` **(b)** `compneu`

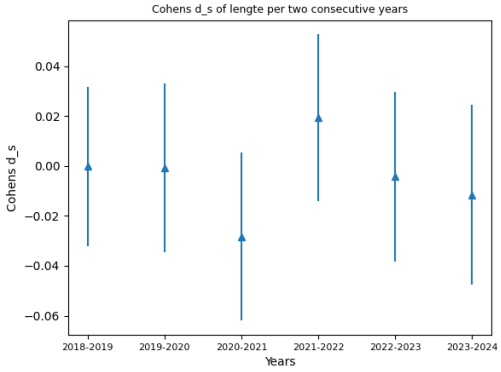
(c) `compx` **(d)** `severeCompl`

Figure G.14: Plots of the effect sizes of the discrete variables per two consecutive years. Part 14/14.

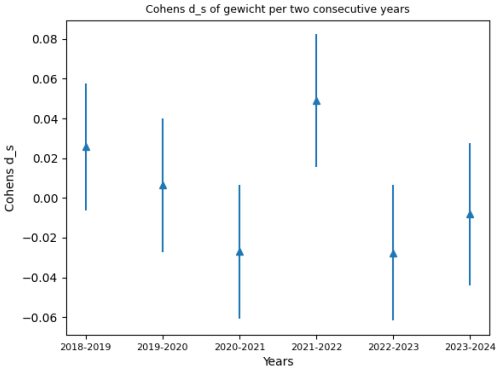
G.2 Continuous variables

In this section we show the plots of the effect size analysis of the continuous variables. The plots show the effect sizes per two consecutive years.

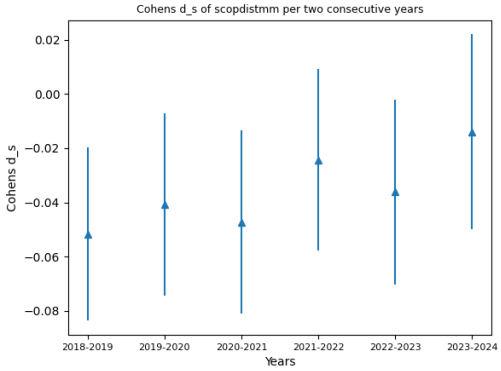
The 7 continuous variables are `lengte`, `gewicht`, `scopdistmm`, `scopdist2mm`, `leeftijd`, `waitingTime`, `lengthOfStay`.



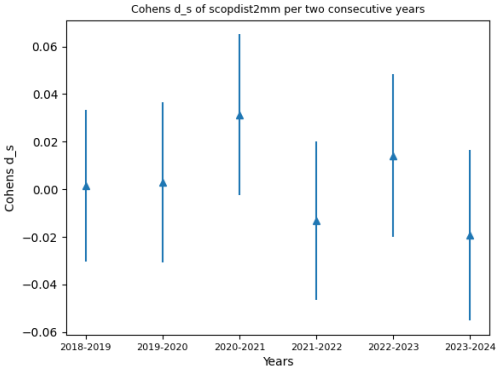
(a) `lengte`



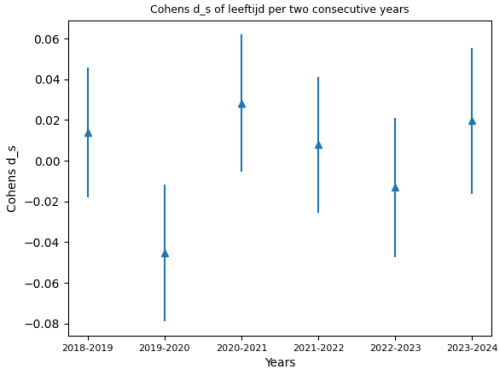
(b) `gewicht`



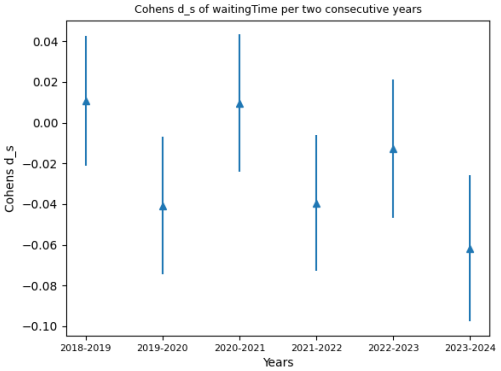
(c) `scopdistmm`



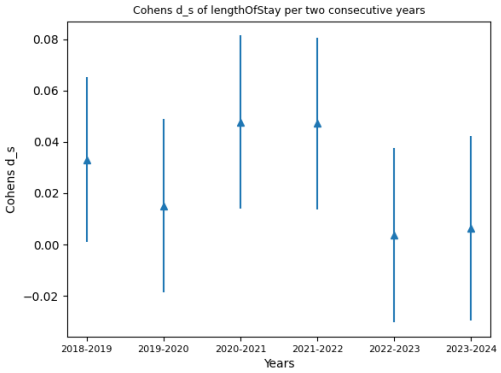
(d) `scopdist2mm`



(e) `leeftijd`



(f) `waitingTime`



(g) `lengthOfStay`

Figure G.15: Plots of the effect sizes of the continuous variables per two consecutive years.

H | Correlation and Difference Matrices for All Variables

In this appendix we show the correlation matrices per year (Section H.1) and the difference matrices per two sequential years (Section H.2) for all variables.

H.1 Correlation matrices

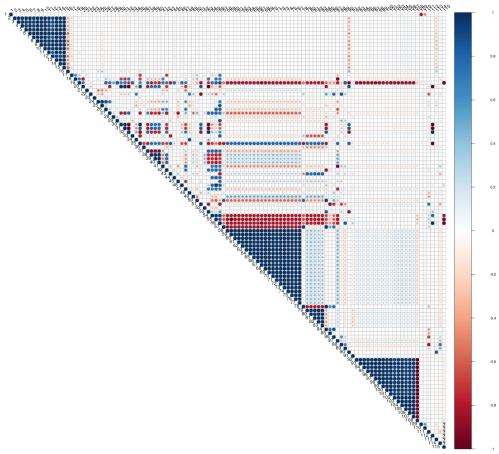


Figure H.1: Correlation matrix of all variables for 2018.

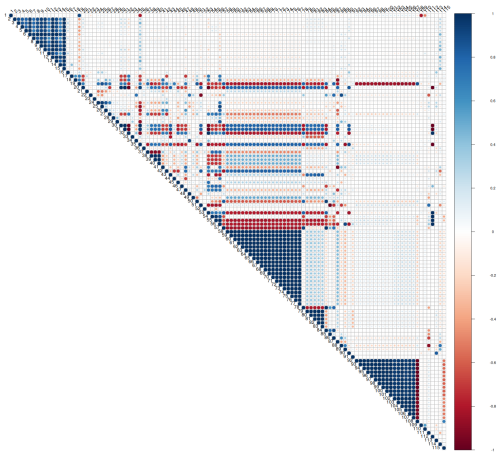


Figure H.2: Correlation matrix of all variables for 2019.

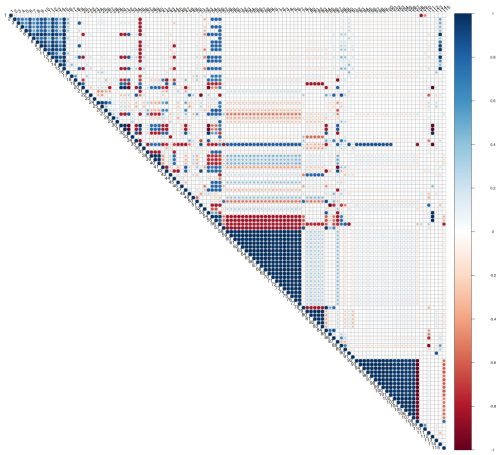


Figure H.3: Correlation matrix of all variables for 2020.

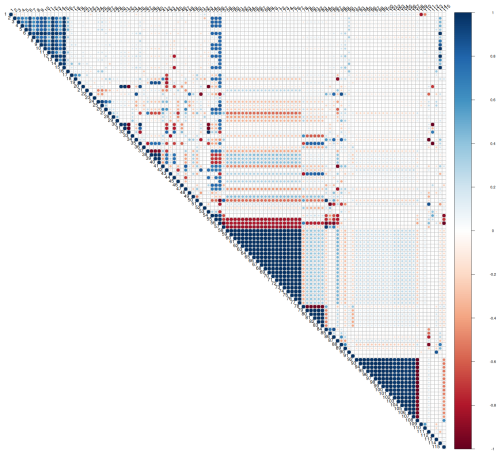


Figure H.4: Correlation matrix of all variables for 2021.

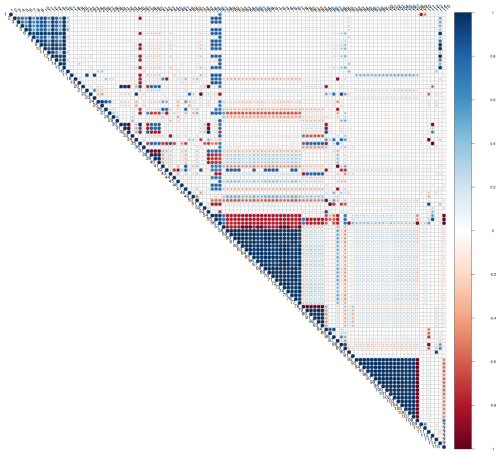


Figure H.5: Correlation matrix of all variables for 2022.

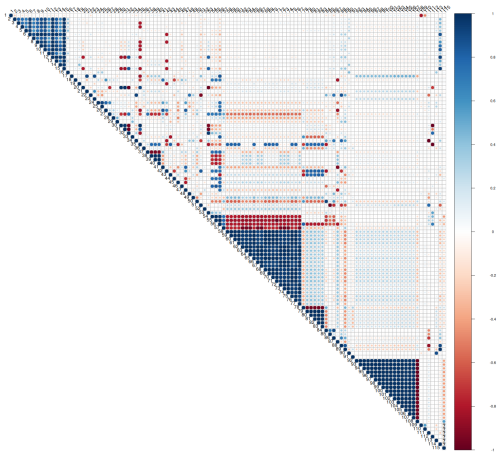


Figure H.6: Correlation matrix of all variables for 2023.

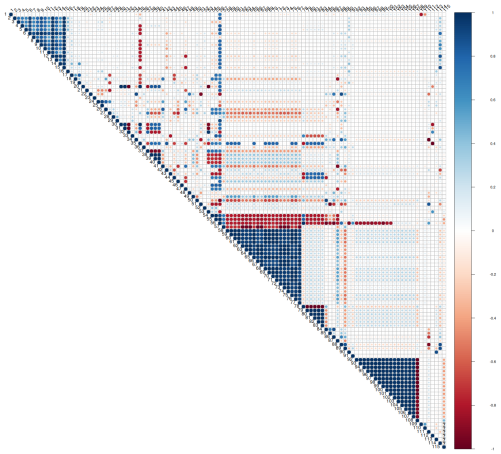


Figure H.7: Correlation matrix of all variables for 2024.

H.2 Difference matrices

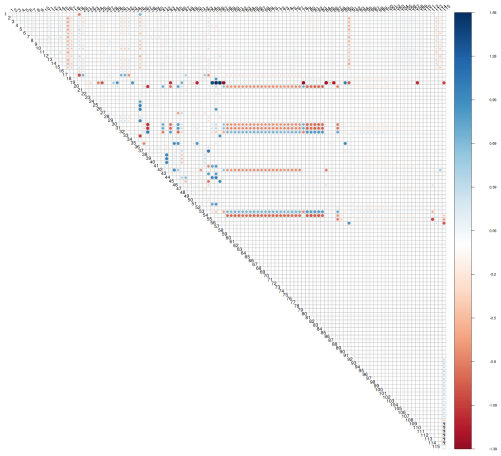


Figure H.8: Difference matrix of all variables for 2018-2019.

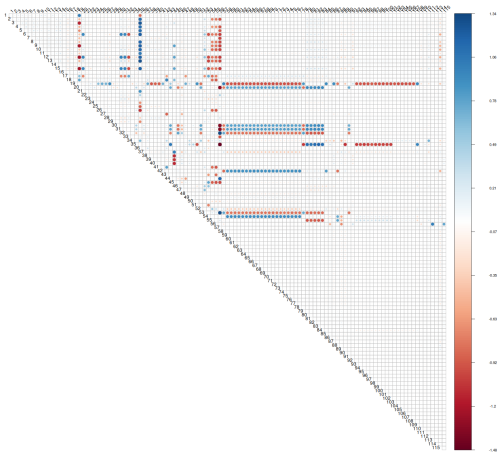


Figure H.9: Difference matrix of all variables for 2019-2020.

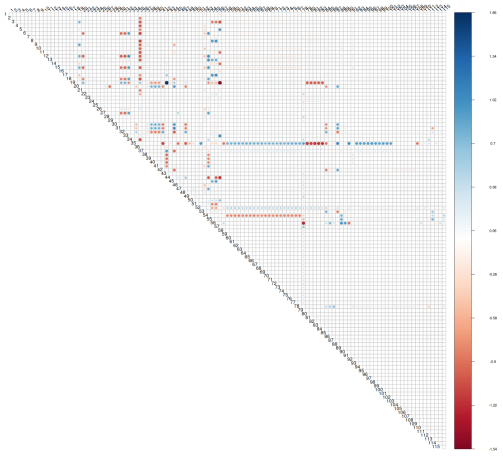


Figure H.10: Difference matrix of all variables for 2020-2021.

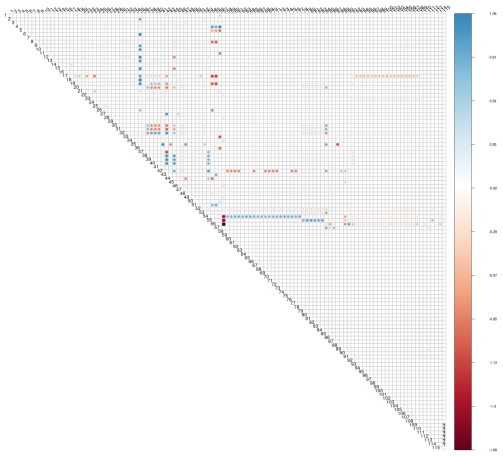


Figure H.11: Difference matrix of all variables for 2021-2022.

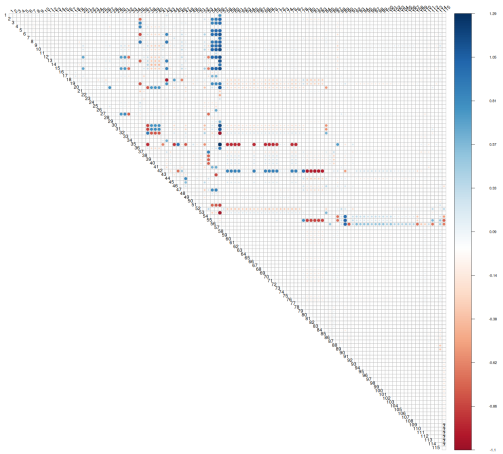


Figure H.12: Difference matrix of all variables for 2022-2023.

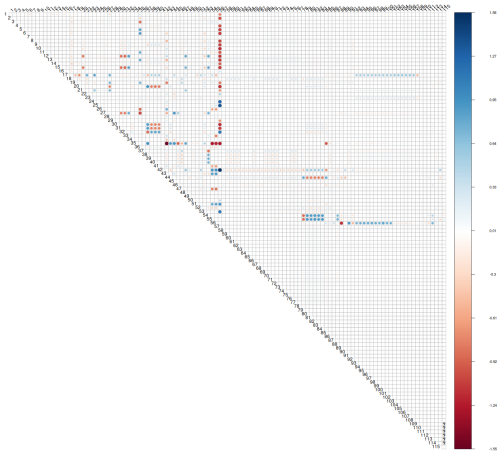


Figure H.13: Difference matrix of all variables for 2023-2024.