Universitätsklinikum Hamburg-Eppendorf

Klinik und Poliklinik für Interdisziplinäre Endoskopie

Direktor: Prof. Dr. med. Thomas Rösch

Middle- and Long-Term Effects of Per-Oral Endoscopic Myotomy in Treatment-Naïve and Previously Treated Achalasia

Dissertation

zur Erlangung des Grades eines Doktors der Medizin
an der Medizinischen Fakultät der Universität Hamburg

vorgelegt von:

Arne Kowalewski, B.Sc.

Hamburg, 2021

Angenommen von der Medizinischen Fakultät der Universität Hamburg am: **25. Januar 2021.**

Veröffentlicht mit Genehmigung der Medizinischen Fakultät der Universität Hamburg.

Prüfungsausschuss, der Vorsitzende: **Prof. Dr. Henning Wege**

Prüfungsausschuss, zweiter Gutachter: **Prof. Dr. Thomas Rösch**


DATA
LES
MODEL
YEARS
IRP
REGRESSION
DYSPHAGIA
MISSING DATA
ANALYSIS
TABLE
IMPUTED
PATIENT
AGE
LHM
FIT
STATISTICAL
TREATMENT-NAÏVE
PRE-TREATED
REFLUX
FOLLOW-UP
PATIENTS
ACHALASIA
TREATMENT SUCCESS
HRM
TREATMENT FAILURE
MMHG
COX
POEM
PRIOR TREATMENTS
ECKARDT
RESULTS
LOGISTIC REGRESSION
MULTIPLE IMPUTATION
ENDOSCOPIC
REVIEW
REFLUX DISEASE
RISK
GROUPS
STUDY
LA GRADE
DISTRIBUTION
SPASTIC
SIGMOID SHAPE
ACHALASIA TYPE II
OUTCOME
WEIGHT LOSS
SPASMS
PERORAL ENDOSCOPIC MYOTOMY
JACKHAMMER
NUMBER
IMPUTATION
RETROSTERNAL PAIN
PROPORTIONAL HAZARDS
COX REGRESSION
OBSERVATIONS
SHAPIRO-WILK
CLINICAL
SUCCESS
HAZARD
NULL
REPORTED
HAZARD RATIO
PRIOR TREATMENT
ESOPHAGEAL
REFLUX DISEASE
NUTCRACKER
PREDICTORS
LOG-RANK
ENDOSCOPY
STATISTICAL
MODELS
THESIS
PRESSURE
MYOTOMY
MONTHS
REPORTED
PAN-ESOPHAGEAL
FUNDOPLICATION
COMPARED
ESOPHAGRAM
TREATMENT FAILURE
TREATMENT SUCCESS
WEIGHT LOSS
AGE
MNAR
SIGNIFICANT
MCAR
R
PMM
ODDS
PPI
SURVIVAL
INFORMATION
STUDIES
MNAR
PAIRS
GERD
MULTIVARIATE
REGURGITATIONS
MANOMETRY
MOTILITY DISORDER
PREDICTORS
IRP
POEM
AMOTILITY
LOG-RANK
ACHALASIA
MMHG
COX
MONTHS
AMOTILITY
GERD
GROUPS
RECORDS
OUTCOME
THESIS
ESOPHAGUS
NOTES
PRIOR TREATMENTS
ECKARDT
INTEGRATED RELAXATION PRESSURE
EVIDENCE
EQUATION
STRAIGHT SHAPE
BOTULINUMTOXIN INJECTION
CASE
FACTOR
NUTCRACKER
STUDIES
CHICAGO CLASSIFICATION
OBSERVED
PREDICTIVE MEAN MATCHING
AGE
PAIRS
TREATMENT FAILURE
HRM
MOTILITY DISORDER
UPPER ENDOSCOPY
BALLOON DILATATION
ACHALASIA
PATIENT
ESOPHAGRAM
ESTIMATES
RESULTS
NORMALITY
LOWER ESOPHAGEAL SPHINCTER
PRE-TREATED
HELLER MYOTOMY
YEARS
REGRESSION
IMPUTATION
PMM
DCI
NA
STUDY
PATIENTS
GROUP
DATA
YEARS
SURVIVAL
AIC
REVIEW
BALLOON DILATATION
PAIRS
ENDOSCOPIC
SIGNIFICANT
DYSPHAGIA
MYOTOMY
PRESSURIZATION
PREDICTOR
OUTCOME
CASE
FACTOR
SPASMS
NUTCRACKER
SCHOENFELD
OBSERVED
CHICAGO CLASSIFICATION
STUDIES
NA
PREDICTIVE MEAN MATCHING
AGE
PAIRS
TREATMENT FAILURE
HRM
MOTILITY DISORDER
UPPER ENDOSCOPY
BALLOON DILATATION
ACHALASIA
PATIENT
ESOPHAGRAM
ESTIMATES
RESULTS
NORMALITY
LOWER ESOPHAGEAL SPHINCTER
PRE-TREATED
HELLER MYOTOMY
YEARS
REGRESSION
IMPUTATION
PMM
DCI
NA
STUDY
PATIENTS
GROUP
DATA
YEARS
SURVIVAL
AIC
REVIEW
BALLOON DILATATION
PAIRS
ENDOSCOPIC
SIGNIFICANT
DYSPHAGIA
MYOTOMY
PRESSURIZATION
PREDICTOR
OUTCOME
CASE
FACTOR
SPASMS
NUTCRACKER
SCHOENFELD
OBSERVED
CHICAGO CLASSIFICATION
STUDIES
NA
BOTULINUMTOXIN INJECTION
STRAIGHT SHAPE
EQUATION
EVIDENCE
INTEGRATED RELAXATION PRESSURE
ECKARDT
PRIOR TREATMENTS
NOTES
ESOPHAGUS
THESIS
GERD
GROUPS
RECORDS
OUTCOME
MMHG
COX
MONTHS
AMOTILITY
POEM
PPI
PMM
ODDS
IMPUTED
SURVIVAL
INFORMATION
STUDIES
MNAR
PAIRS
FOLLOW-UP
REFLUX
PRE-TREATED
TREATMENT-NAÏVE
STATISTICAL
FIT
PATIENTS
ACHALASIA
TREATMENT SUCCESS
HRM
TREATMENT FAILURE
MMHG
COX
POEM
PRIOR TREATMENTS
ECKARDT
RESULTS
LOGISTIC REGRESSION
MULTIPLE IMPUTATION
ENDOSCOPIC
REVIEW
REFLUX DISEASE
RISK
GROUPS
STUDY
LA GRADE
DISTRIBUTION
SPASTIC
SIGMOID SHAPE
ACHALASIA TYPE II
OUTCOME
WEIGHT LOSS
SPASMS
PERORAL ENDOSCOPIC MYOTOMY
JACKHAMMER
NUMBER
IMPUTATION
RETROSTERNAL PAIN
PROPORTIONAL HAZARDS
COX REGRESSION
OBSERVATIONS
SHAPIRO-WILK
CLINICAL
SUCCESS
HAZARD
NULL
REPORTED
HAZARD RATIO
PRIOR TREATMENT
ESOPHAGEAL
REFLUX DISEASE
NUTCRACKER
PREDICTORS
LOG-RANK
ENDOSCOPY
STATISTICAL
MODELS
THESIS
PRESSURE
MYOTOMY
MONTHS
DATA
YEARS
MODEL
LES
TIME
IRP
REGRESSION
DYSPHAGIA
MISSING DATA
ANALYSIS
TABLE
IMPUTED
PATIENT
AGE
LHM
FIT
STATISTICAL
TREATMENT-NAÏVE
PRE-TREATED
REFLUX
FOLLOW-UP
PATIENTS
ACHALASIA
TREATMENT SUCCESS
HRM
TREATMENT FAILURE
MMHG
COX
POEM
PRIOR TREATMENTS
ECKARDT
RESULTS

Table of Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

AIC ..... Akaike Information Criterion
CC ..... Chicago Classification / Complete Case
CI ..... Confidence Interval
DCI ..... Distal Contractile Integral
df ..... Degrees of Freedom
DfIE ..... Department for Interdisciplinary Endoscopy
EGJ ..... Esophagogastric Junction
FU ..... Follow-Up
HRM ..... High-Resolution Manometry
IRP ..... Integrated Relaxation Pressure
LA ..... Los Angeles
LES ..... Lower Esophageal Sphincter
LL ..... Log-Likelihood
MAR ..... Missing at Random
MCAR ..... Missing Completely at Random
MNAR ..... Missing Not at Random
POEM ..... Per-Oral Endoscopic Myotomy
p value ..... Probability Value
SD ..... Standard Deviation
SE ..... Standard Error
UE ..... Upper Endoscopy
UES ..... Upper Esophageal Sphincter
UKE ..... University Medical Center Hamburg-Eppendorf

1 Introduction

1.1 Achalasia

Achalasia is a rare yet well-known motility disorder of the esophagus with an estimated incidence of about 0.3 to 1.6 per 100,000 people per year (Sadowski et al. 2010, Schlottmann et al. 2018, Tebaibia et al. 2016). Its observed prevalence ranges from about 2 to 15 per 100,000 people (Arber et al. 1993, Ho et al. 1999, Sadowski et al. 2010, Samo et al. 2017). While some studies suggest a diagnostic peak at 30 years (Arber et al. 1993, Ho et al. 1999), others depict achalasia primarily as a disease of the elderly beyond 60 years of age (Farrukh et al. 2008, Gennaro et al. 2011, Mayberry and Atkinson 1985). Yet, the disease can generally occur at any age, and even in children (see Liu et al. 2020, Nabi et al. 2019). Major clinical symptoms typically reported by patients are dysphagia, recurrent episodes of chest pain, food regurgitation, weight loss, and, eventually, pulmonary aspiration (Pressman and Behar 2017). In rare end-stage achalasia, the esophagus tends to dilate and bend, thus losing its straight form in favor of a characteristic sigmoidal shape (Herbella and Patti 2015).

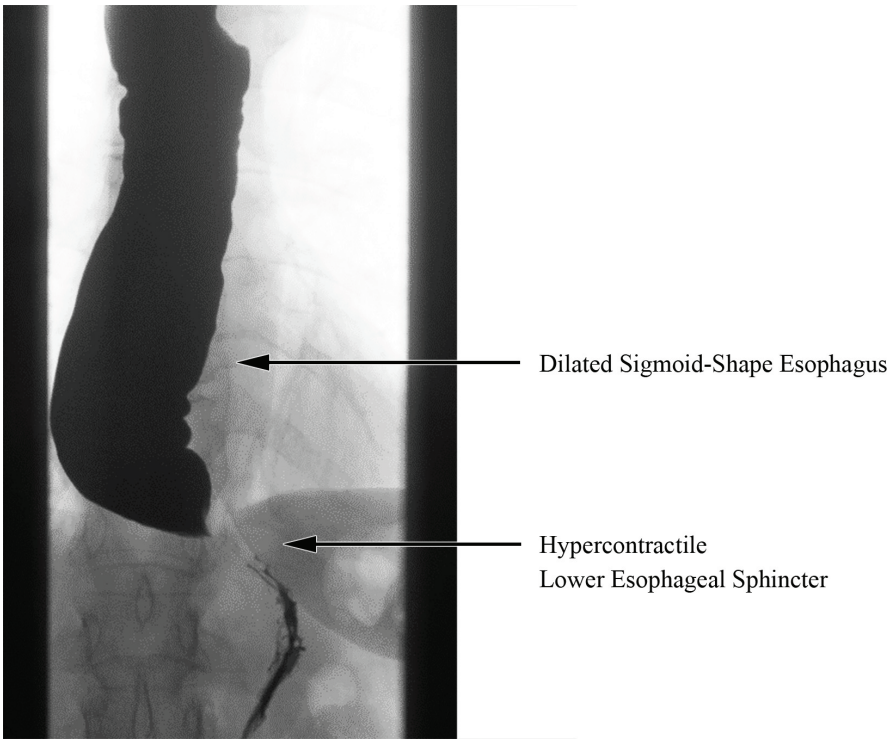


Figure 1: Barium Swallow Esophagram of End-Stage Sigmoidal Achalasia. Picture modified from unpublished records of the Department for Interdisciplinary Endoscopy (DfIE), University Medical Center Hamburg-Eppendorf (UKE), Germany.

Despite intensive research, the causes of primary achalasia remain suspects of speculation. A degeneration or total absence of the ganglia cells of the esophageal Auerbach's plexuses has been observed, accompanied by T cell, mast cell, and plasma cell infiltration (Pressman and Behar 2017). Schlottmann et al. (2018) state that the lower esophageal sphincter (LES) usually preserves a myogenic tone to prevent the reflux of gastric fluids, while at the same time providing the ability to relax during deglutition. They argue that this relaxation is regulated by the Auerbach's plexuses and consequentially malfunctions in the wake of their absence.

Etiologic causes for the demise of the Auerbach's plexuses that are currently investigated by researchers range from genetic predispositions over chronic inflammation to viral infections as potential trigger factors for auto-immune mediated processes (Pressman and Behar 2017).

Genetic predispositions can be assumed based on a strong association of achalasia with specific amino acid polymorphisms found in major histocompatibility complex signal molecules that are involved in the immune response, especially in HLA-DQ (Gockel et al. 2014). Besides, Zárate et al. (1999) found achalasia to develop significantly more frequently in patients with Down syndrome, a disease of known genetic origin. They argue that such a link between two rare diseases seems unlikely to be a coincidence and may therefore hint at the development of achalasia being affected by genetic predispositions. Ultimately, associations with other autoimmune diseases such as type I diabetes, hyperthyroidism, Sjögren syndrome, and systemic lupus erythematosus have been described as well (Booy et al. 2012).

LES muscle biopsies taken from achalasia patients show significantly elevated levels of the T cell subtypes T_h1, T_h2, T_h17, and T_h22 (Furuzawa-Carballeda et al. 2015). All these cell lines play vital roles in the regulation of the immune system and in a multitude of tissue inflammation processes (Akdis et al. 2012). Furthermore, Furuzawa-Carballeda et al. (2015) found both anti-myenteric autoantibodies and herpes simplex virus type 1 DNA in every single achalasia patient included in their study. By contrast, they found neither of them in any patient of the healthy control group. The development of achalasia may therefore, at least to a certain extent, be facilitated by herpes virus infections.

Unlike primary achalasia, whose etiology is still unclear, secondary achalasia usually occurs as a complication of Chagas disease (Pressman and Behar 2017, Schlottmann et al. 2018). Secondary achalasia is not subject of this thesis, though.

In their clinical practice guidelines for per-oral endoscopic myotomy, Inoue et al. (2018) cite the triad of upper endoscopy (UE), timed barium swallow, and high-resolution manometry (HRM) as the recommended diagnostic procedures for patients with suspected achalasia.

1.2 High-Resolution Manometry

HRM provides an intuitive graphical visualization of the intra-esophageal pressure between the pharynx and the stomach during swallowing (Schlottmann et al. 2017). It also offers advanced metrics for the analysis of the esophageal pressure topography (Kahrilas et al. 2015).

One of the most essential HRM metrics is the *integrated relaxation pressure* (IRP). It is a measure of the relaxation capability of the lower esophageal sphincter. As such, it represents the main criterion for the diagnosis of achalasia. The *distal contractile integral* (DCI) is another vital metric, which quantifies peristaltic contraction vigor. As such, it can identify and characterize esophageal spasms, just as it can differentiate between normal, weak, and failed peristalsis.

Figures 2 and 3 demonstrate a typical HRM setup. The manometry probe is calibrated before the measurement. It is then applied like a gastric tube. First, the probe is coated in a gel containing local anesthetics. Decongestant nasal drops may be applied beforehand to allow for easier passage through the nasal cavity. They should also ease some of the discomforts the patient may experience during the procedure. The probe is inserted through the patient's nose and pushed forward down the pharynx. Upon arrival at the larynx, the patient is given a cup of water and instructed to swallow. As soon as the epiglottis occludes the entrance to the trachea during deglutition, the probe is pushed forward into the esophagus and ultimately down through the LES into the stomach. Its position can be monitored in real-time based on the pressure topography that is continuously being measured by the probe.

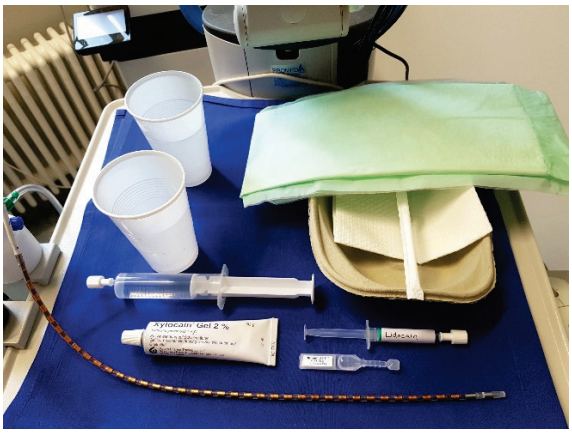


Figure 2: High-Resolution Manometry Setup.

Figure 3: The Manometry Probe.

Figure 4 illustrates the HRM of a healthy person's normal swallow. Usually, ten swallows are measured. The upper esophageal sphincter (UES) opens during deglutition to allow for food and fluids to pass from the pharynx into the esophagus. A peristaltic wave then unfolds that travels down the esophagus toward the stomach. The LES relaxes early on during the swallow and thus allows for peristalsis to push the swallowed substances through the esophagogastric junction (EGJ).

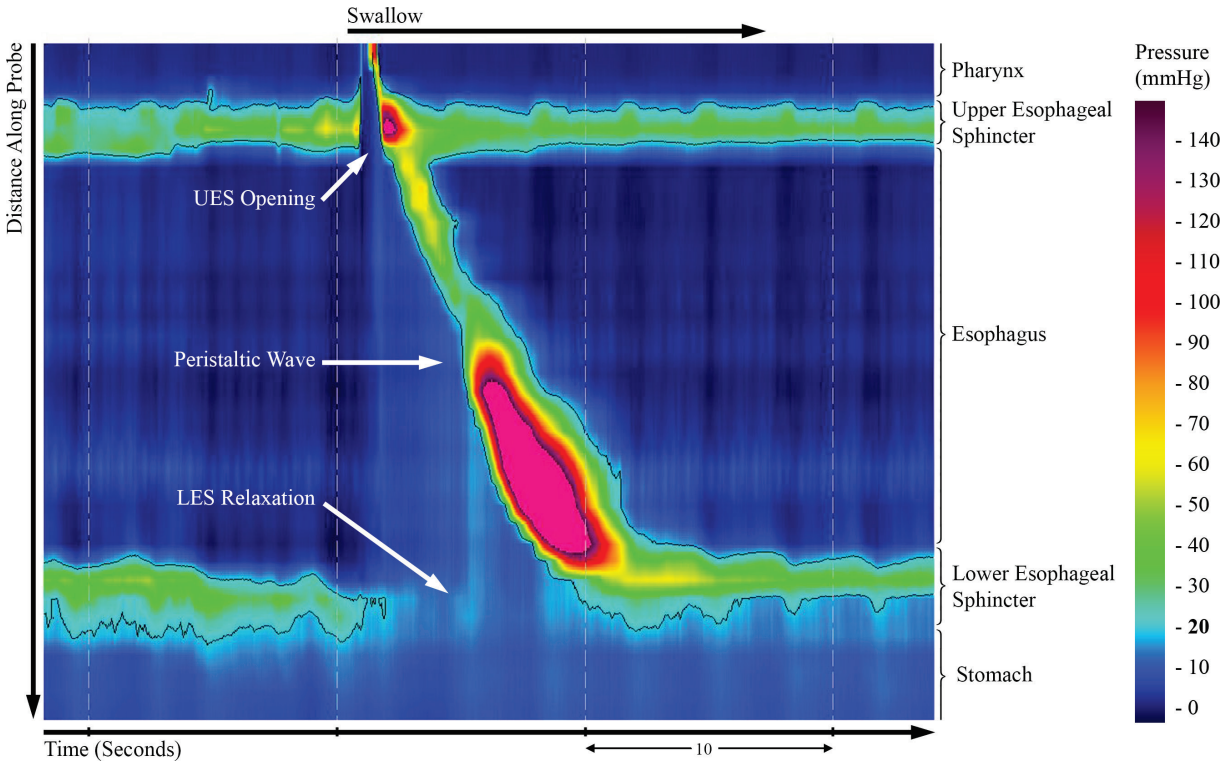


Figure 4: High-Resolution Manometry of a Healthy Person. The y-axis represents the distance along the probe. The x-axis represents time. The coloring indicates the pressure measured at a specific position and time. Black outlines visualize the pressure transition across a manually adjustable threshold, in this example set to 20 mmHg. Dashed white lines delimit consecutive time intervals, in this example 10 seconds each. Picture modified from unpublished records of the DfIE, UKE, Germany.

1.2.1 Integrated Relaxation Pressure

The IRP is a measure of the deglutitive relaxation of the LES (Ghosh et al. 2007). It was initially defined as the mean minimal pressure through the esophagogastric junction for a specific cumulative period of time during a 10 seconds long relaxation window that begins at the very moment the UES starts to relax during deglutition (Kahrilas et al. 2015).

Choosing a cumulative time span of 4 seconds for the IRP estimation, the so-called *4 s IRP*, in combination with an upper cut-off value of 15 mmHg was found to deliver the best distinction between normal

peristalsis of healthy individuals and impaired EGJ relaxation as it is typically found in achalasia patients (Pandolfino et al. 2009). The mean 4 s IRP has been reported to provide a sensitivity of up to 98 % and a specificity of up to 96 % in the detection of achalasia (Ghosh et al. 2007).

The latest revision of the Chicago Classification, which will be introduced in the next chapter, replaced the mean IRP with the median IRP to make it less vulnerable to outlying pressure measurements (Kahrilas et al. 2015). If not specified otherwise, the IRP nowadays usually refers to the *median 4s IRP*.

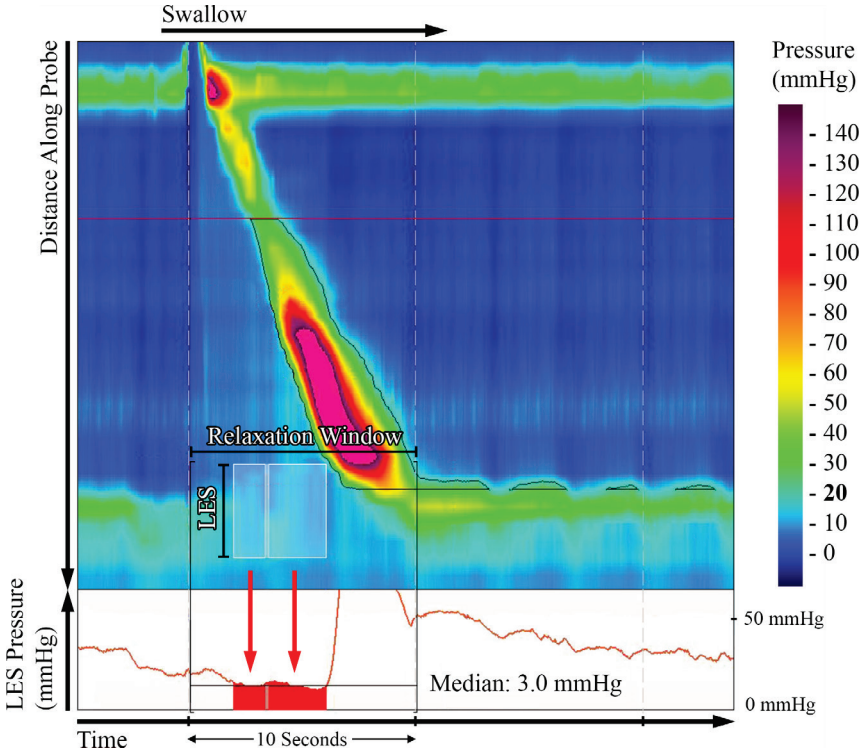


Figure 5: Computational Estimation of the Integrated Relaxation Pressure. The LES pressure is measured continuously during the relaxation window of 10 seconds. The red area under the curve represents four cumulative seconds of minimal pressure during said window. Their median is the IRP. Picture modified from unpublished records of the DfIE, UKE, Germany.

1.2.2 Distal Contractile Integral

The DCI is the product of amplitude, duration, and length of the distal esophageal contraction above 20 mmHg that reaches from the transition zone down to the proximal margin of the lower esophageal sphincter (Kahrilas et al. 2015). As such, it is a measure of esophageal contraction vigor. The transition zone is the anatomical area where the upper contraction wave of the proximal esophagus' striated muscle fibers segues into a lower contraction wave as it descends into the smooth muscle fibers of the distal esophagus (Ghosh et al. 2006).

Kahrilas et al. (2015) define a DCI between 450 and 8,000 mmHg·s·cm as the normal peristaltic pressure. They characterize weak peristalsis by a DCI between 100 and 450 mmHg·s·cm, and failed peristalsis by a DCI below 100 mmHg·s·cm. Ultimately, they define a DCI of 8,000 mmHg·s·cm or above as proof of hypercontractile peristalsis as it is found in jackhammer esophagus.

1.3 The Chicago Classification of Motility Disorders of the Esophagus

Motility disorders of the esophagus are commonly categorized according to the *Chicago Classification* (CC). Its latest revision by Kahrilas et al. (2015) is illustrated in figure 6. The most important differential diagnoses to achalasia are EGJ outflow obstruction and major peristaltic disorders. All of these will be introduced in chapter 1.5. The clinical relevance of minor peristaltic disorders is controversial (Kahrilas et al. 2015). They are not the focus of this thesis and will therefore not be discussed further.

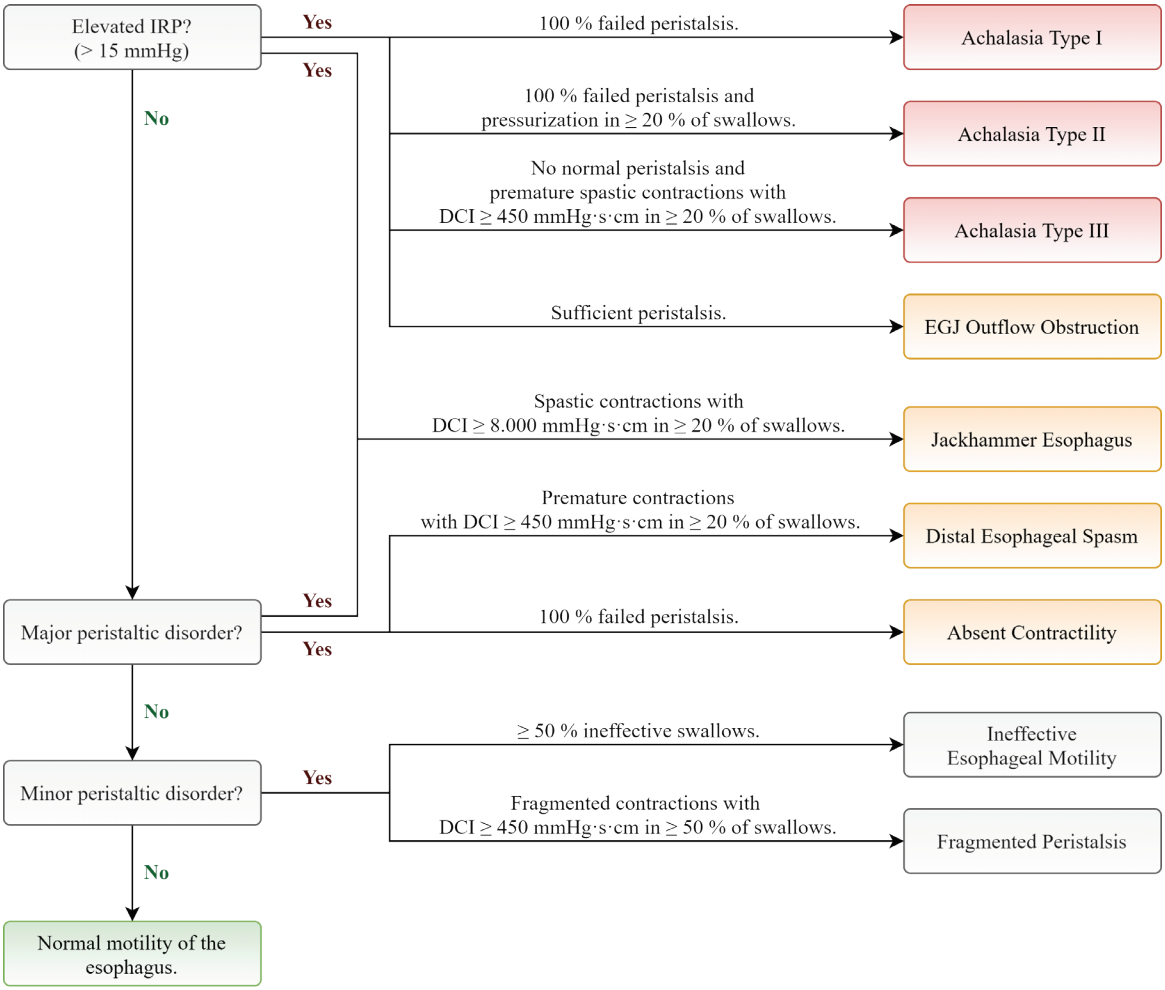


Figure 6: Diagnostic Criteria of Esophageal Motility Disorders. Simplified, based on Kahrilas et al. (2015).

1.4 Achalasia Type I, II, and III

As proposed by Kahrilas et al. (2015) and illustrated in figure 6, achalasia is characterized by an elevated IRP in conjunction with either failed peristalsis (type I), pan-esophageal pressurization (type II), or spasms (type III). They further state that in the case of incompletely expressed achalasia or mechanical obstruction, EGJ outflow obstruction must be diagnosed instead. Accordingly, if the IRP is normal, other rare motility disorders such as jackhammer esophagus should be considered. Figures 7 to 9 illustrate typical HRM findings in achalasia type I, II, and III.

Treatment success varies between the three achalasia types. In their meta-analysis, which included both endoscopic and surgical myotomy, Pandolfino and Gawron (2015) found type II achalasia to usually respond best to treatment with a success rate of about 96 %, followed by 86 % in type I, and both far ahead of 66 % in type III. Matching conclusions can be drawn from Podboy et al. (2020). In contrast, Greene et al. (2015) and Zheng et al. (2019) did not find significant differences in the post-interventional treatment responses between the achalasia types. The aforementioned meta-analysis undoubtedly carries the most scientific weight. This suggests that type II may in fact be the best treatable disease manifestation, and type III is the worst. However, the evidence on this subject remains somewhat conflicted.

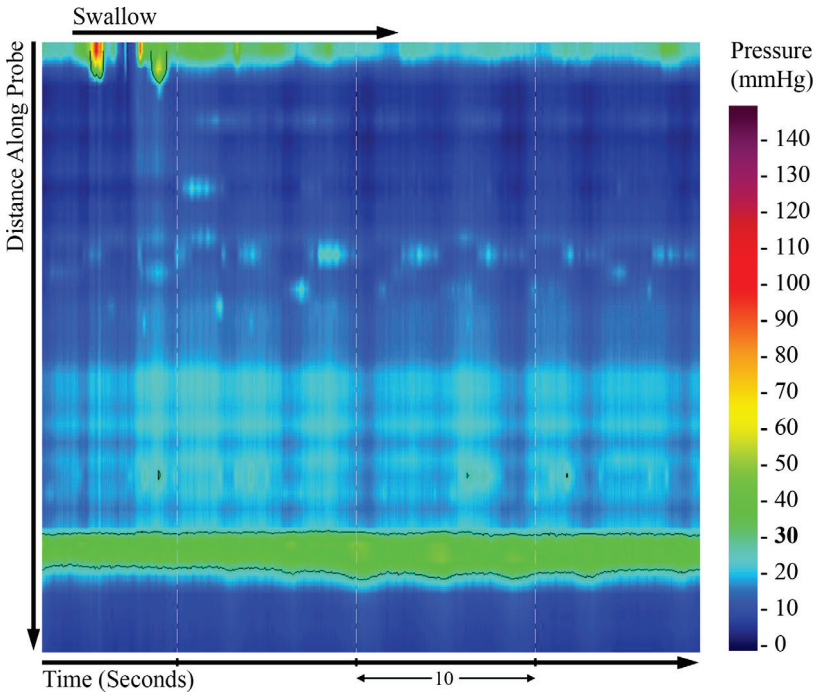


Figure 7: Typical High-Resolution Manometry in Type I Achalasia. After the opening of the UES during early deglutition, the utter absence of any peristalsis and LES relaxation paint the distinct picture of type I achalasia. Black outline: pressure transition from below to above 30 mmHg. Picture modified from unpublished records of the DfIE, UKE, Germany.

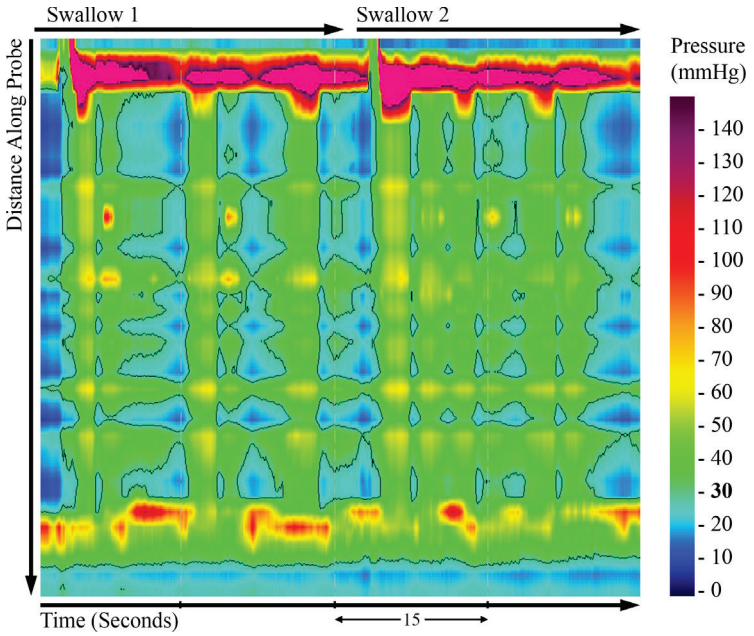


Figure 8: Typical High-Resolution Manometry in Type II Achalasia. As the swallow commences, panesophageal pressurization unaccompanied by any form of visible peristalsis or LES relaxation prevails. Black outline: pressure transition from below to above 30 mmHg. Picture modified from unpublished records of the DfIE, UKE, Germany.

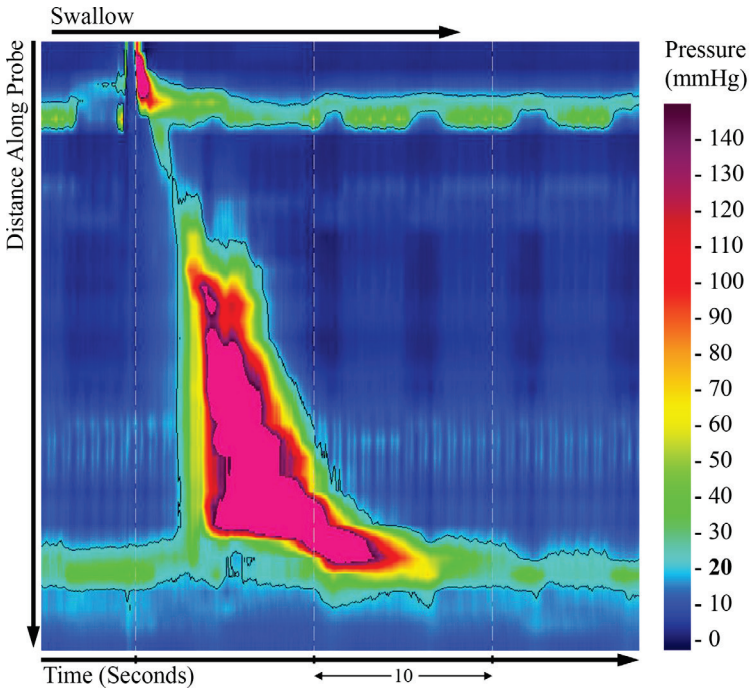


Figure 9: Typical High-Resolution Manometry in Type III Achalasia. Following early deglutition, immediate spastic contractions stretch across the esophagus, which shows no sign of LES relaxation. The DCI was measured between 450 and 8,000 mmHg·s·cm (not shown in the picture), thereby securing the diagnosis of type III achalasia. Black outline: pressure transition from below to above 20 mmHg. Picture modified from unpublished records of the DfIE, UKE, Germany.

1.5 Differential Diagnoses

As depicted in chapter 1.3, there are many differential diagnoses for achalasia among the motility disorders of the esophagus. Additionally, achalasia-like symptoms can also be induced or mimicked by other diseases and conditions that need not necessarily originate from the esophagus itself.

1.5.1 EGJ Outflow Obstruction and Pseudoachalasia

EGJ outflow obstruction is defined by the CC as a condition in which the median IRP is elevated, yet the criteria for achalasia are not met due to sufficient peristalsis (Kahrilas et al. 2015). Causes for EGJ outflow obstruction include a wide variety of underlying conditions. Examples for these are early or incomplete achalasia, eosinophilic esophagitis, mechanical processes such as strictures, varices, or tumors, fibrosis, extrinsic compression of the esophagus, obesity-induced intra-abdominal pressure, opiate abuse, and ultimately HRM measurement errors (Samo and Qayed 2019).

Figure 10 showcases exemplary HRM and upper endoscopy findings in a patient who presented with achalasia-like symptoms. They turned out to be caused by a submucosal tumor.

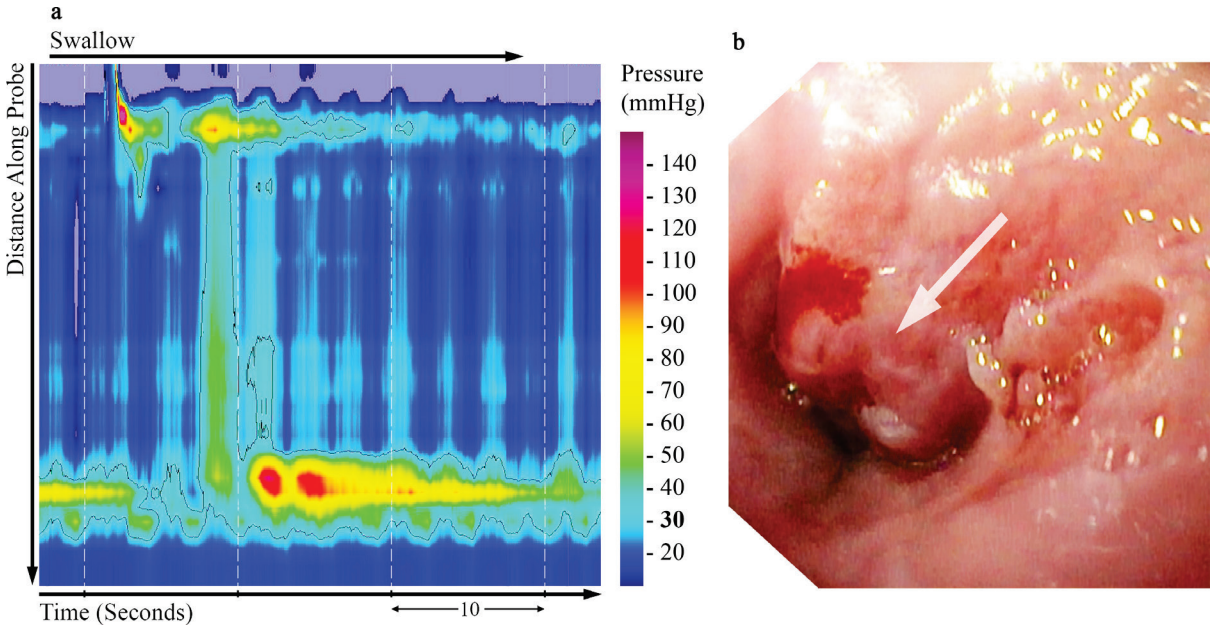


Figure 10: Pseudoachalasia Caused by a Submucosal Tumor. Pictures modified from unpublished records of the DfIE, UKE, Germany. **a: High-Resolution Manometry.** During the entire swallowing process, atypical motility patterns emerge throughout the entire esophagus, including the UES and, to a lesser extent, the LES. No peristalsis is visible. Black outline: pressure transition from below to above 30 mmHg. **b: Endoscopic Findings.** A submucosal tumor can be seen distal of the UES. It protrudes into the esophageal lumen, where it causes an obstruction.

1.5.2 Major Peristaltic Disorders

Distal esophageal spasms are defined by the CC as a condition in which a patient presents with a normal IRP in combination with premature contractions that show a DCI of more than 450 mmHg·s·cm in at least 20 % of the swallows (Kahrilas et al. 2015).

Hypercontractile esophagus, or *Jackhammer esophagus*, shows a pattern of extreme spastic contractions. It is defined by the CC as a condition in which at least 20 % of the observed swallows show a DCI above 8,000 mmHg·s·cm, independent of the IRP (Kahrilas et al. 2015). *Nutcracker esophagus* is a quite similar condition characterized by high-amplitude peristaltic contractions of at least 180 mmHg in the older conventional manometry that was used before HRM existed (Hong et al. 2016). Since HRM has become readily available, the diagnosis of nutcracker esophagus has gradually been abandoned in favor of jackhammer esophagus. It is still referred to in older reports, though.

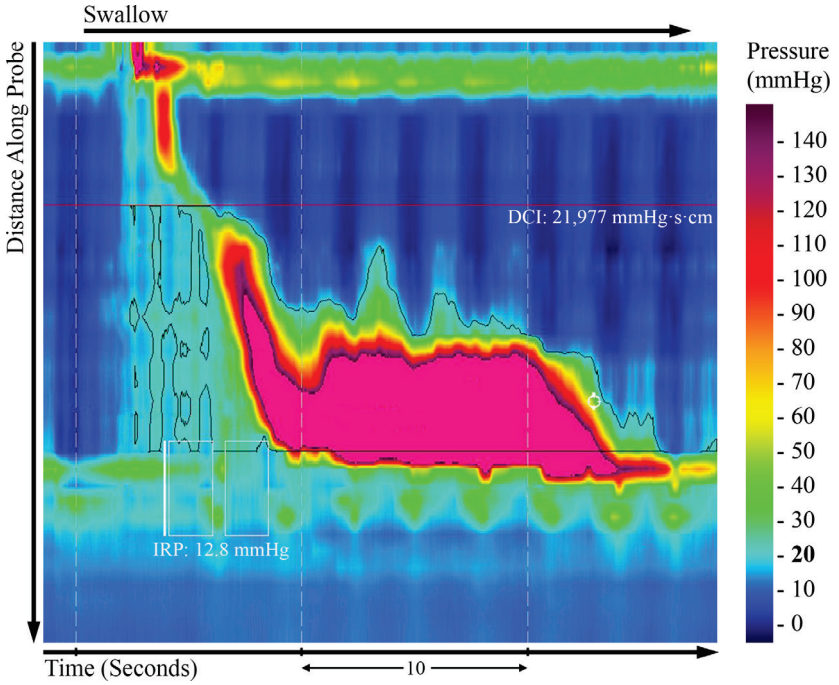


Figure 11: Typical High-Resolution Manometry in Jackhammer Esophagus. Following early deglutition and presumed initial peristalsis in the proximal esophagus, a massive spastic contraction emerges, which then persists in the distal esophagus for approximately 15 seconds. It builds up immense pressure: in this case, the DCI is about 22,000 mmHg·s·cm. In contrast, the IRP of 12.8 mmHg is normal. Black outline: pressure transition from below to above 20 mmHg, confined to the region below the red horizontal line and above the LES. Picture modified from unpublished records of the DfIE, UKE, Germany.

Ultimately, *absent contractility* is defined by the CC as the total failure of peristalsis in the presence of a normal IRP (Kahrilas et al. 2015). It is a rare condition mostly observed in patients with connective tissue disorders, such as systemic sclerosis (van Hoeij and Bredenoord 2016).

1.5.3 Eosinophilic Esophagitis

Eosinophilic esophagitis is a chronic immune-mediated eosinophilic inflammation of the esophageal mucosa. It may occur at any age and present with unspecific symptoms, dysphagia, or even classic manifestations of gastro-esophageal reflux disease such as chest pain (Kumar et al. 2020).

Interestingly, eosinophilic esophagitis is also associated with abnormalities in esophageal motility. Spechler et al. (2018) claim that the accumulation of eosinophilic granulocytes in the esophageal muscularis propria may induce the release of toxic proteins that are well capable of destroying nearby nerve cells. Furthermore, they theorize that eosinophilic secretory products may disrupt peristalsis, restrict relaxation, and induce fibrosis. All this might cause motility abnormalities similar to those seen in achalasia. These symptoms may normalize after treating the eosinophilia found in the mucosa of affected patients (Spechler et al. 2018). Therefore, it is important to rule out or treat eosinophilic esophagitis before diagnosing or treating achalasia in a patient.

During upper endoscopy, mucosal biopsies are usually taken and later analyzed. Eosinophilic esophagitis is defined by the presence of 15 or more eosinophilic granulocytes per high-power field in the histopathologic examination (Kumar et al. 2020). In contrast, biopsies taken from achalasia patients are primarily characterized by ganglion cell loss (Sodikoff et al. 2016).

1.6 The Eckardt Score

The de-facto standard grading system to quantify the clinical severity of achalasia is the *Eckardt score* introduced by Eckardt et al. (1992):

Table 1: The Eckardt Score. Cited from Eckardt et al. (1992).

Score	Dysphagia	Regurgitations	Retrosternal Pain	Weight Loss
0	None	None	None	None
1	Occasional	Occasional	Occasional	< 5 kg
2	Daily	Daily	Daily	5 – 10 kg
3	Each meal	Each meal	Each meal	> 10 kg

The score is used to assess both the initial need for treatment, as well as post-interventional treatment response. Each of its four components – dysphagia, regurgitations, retrosternal pain, and weight loss – is assessed and given a score of either 0, 1, 2, or 3, based on the criteria depicted in table 1. Thus, the lowest possible Eckardt score is 0, and the highest possible score is 12. In recent literature, an Eckardt score above 3 has been established as the standard indicator of the need for treatment. Accordingly, a post-interventional Eckardt score above 3 is usually considered treatment failure (see Inoue et al. 2015, Minami et al. 2015, Shiwaku et al. 2016b).

1.7 The Los Angeles Classification System

Gastroesophageal reflux disease is one of the big concerns regarding the long-term outcome after many surgical and endoscopic achalasia treatments, as will be further explained in the next chapter. One of the most commonly used systems for the assessment of reflux severity is the *Los Angeles (LA)* classification introduced by Armstrong et al. (1996). It utilizes endoscopic assessments of the esophageal mucosa to divide cases of reflux disease into four grades:

Table 2: The Los Angeles Classification System. Based on Armstrong et al. (1996).

LA Grade	Criteria
A	≥ one lesion ≤ 5 mm in length, confined to the mucosal fold(s)
B	≥ one lesion > 5 mm in length, confined to the mucosal fold(s), not continuous between the tops of two folds
C	≥ one non-circumferential lesion that continues between the tops of at least two mucosal folds
D	Circumferential lesion

A major strength of the system is its high inter-observer agreement among endoscopists, as emphasized by Armstrong et al. (1996). They consciously refrained from including minimal mucosal changes such as erythema and edema into the gradings because they found that these were not consistently detectable. In Japan, a modified version of the LA system is somewhat prevalent, which re-introduces such minimal mucosal changes as criteria for an additional grade M (Miwa et al. 2008). However, the criteria for the shared grades A to D differ slightly from the original LA classification as well. This modified system is especially relevant since many studies on the matter of this thesis are conducted in Japan.

Table 3: The Modified Los Angeles Classification System. Cited from Miwa et al. (2008).

LA Grade	Criteria
N	Normal mucosa
M	Minimal mucosal changes such as erythema or whitish turbidity
A	Non-confluent mucosal breaks < 5 mm in length
B	Non-confluent mucosal breaks > 5 mm in length
C	Confluent mucosal breaks < 75 % circumferential
D	Confluent mucosal breaks > 75 % circumferential

1.8 Treatments

Because primary Achalasia cannot currently be cured, all known treatments aim to relieve the functional obstruction caused by the hypercontractile LES (Eckardt et al. 1992). Established treatments include botulinum toxin injections into the sphincter, as well as a multitude of interventional procedures that either stretch the muscle by force (pneumatic balloon dilatation) or cut through it (laparoscopic Heller myotomy and POEM).

1.8.1 Medication

For symptomatic relief that bridges the time gap until operative intervention, temporary relaxation of the LES may be achieved by the oral intake of nitrates (Gelfond et al. 1981), calcium channel blockers (Short and Thomas 1992), or phosphodiesterase type 5 inhibitors (Bortolotti et al. 2000). However, the effects of these drugs are usually short-lived (Kahrilas and Pandolfino 2017). Also, even though they may help to reduce the manometrically measured sphincter pressure, this does not necessarily translate well into actual clinical symptom relief (Short and Thomas 1992). Overall, there is little evidence for the treatment of achalasia with drugs (Kahrilas and Pandolfino 2017).

1.8.2 Botulinum Toxin Injection

As the name suggests, during *botulinum toxin injection*, a neurotoxin is injected endoscopically into the hypercontractile LES. Botulinum toxin inhibits the release of acetylcholine from nerve endings and thus effectively induces temporary flaccid paralysis in the esophageal sphincter (Cariati et al. 2019). While the procedure provides decent initial mitigation of dysphagia in many patients, most relapse within a year, and subsequent injections tend to be increasingly ineffective (Kahrilas and Pandolfino 2017).

1.8.3 Pneumatic Balloon Dilatation

Pneumatic balloon dilatation is performed by positioning a cylindrical balloon across the LES and then inflating it with a pre-defined pressure up to a specific diameter, effectively dilating the hypercontractile sphincter muscle (see Chuah et al. 2010). Naturally, too large diameters bear an increased risk of esophageal perforation, whereas too small diameters may diminish the therapeutic effect of the procedure. A balloon diameter of 30 to 40 mm is recommended (Chuah et al. 2010, Kahrilas and Pandolfino 2017, Mikaeli et al. 2004). The initial dilatation is typically performed with a 30 mm balloon, whereas larger diameters are reserved for subsequent re-dilatations.

1.8.4 Heller Myotomy and Fundoplication

Heller myotomy refers to the surgical cutting of the LES. Its roots reach back to the open cardiomyotomy originally introduced by Heller (1913). After the inception of minimally invasive surgery, it was quickly adapted to utilize laparoscopic access to the abdominal cavity (Hunter et al. 1997). Since then, the *laparoscopic Heller Myotomy* has become the most prevalent variant of the surgical procedure. These days, it may even be performed with the assistance of surgery robots (Huffmann et al. 2007).

To hinder the reflux of gastric fluids into the esophagus once the sphincter has been cut, Heller myotomy is usually followed by *fundoplication* (Bloomston et al. 2003). During the latter procedure, the fundus of the stomach is wrapped around the esophagus (Engstrom et al. 2007). This is either performed in the way of the classic *Nissen fundoplication*, which involves a full 360 degrees wrap, or by creating a partial wrap, which is characteristic of the variants named after *Toupet* or *Dor* (Bramhall and Mourad 2019).

1.8.5 Per-Oral Endoscopic Myotomy

Peroral endoscopic myotomy (POEM) was introduced by Inoue et al. (2010) as a minimally invasive endoscopic treatment approach for achalasia, and, as such, especially as an alternative for the older and more invasive Heller myotomy. As detailed by them, an endoscope is inserted into the patient's esophagus under general anesthesia. They proceed to cut the esophageal mucosa with an ESD knife or a comparable endoscopic tool from inside the esophageal lumen, just about 13 cm proximal of the esophagogastric sphincter. The endoscope is then used to dissect the mucosa from the underlying muscle layers while being pushed further down the esophagus to create a submucosal tunnel that reaches down to the sphincter, whose luminal circular muscle layer is then cut until at least 1 to 2 centimeters into the stomach (Inoue et al. 2018). Ultimately, the endoscope is pulled back into the esophageal lumen and the mucosal cut is closed with clips (Inoue et al. 2010). After the procedure, the patient is allowed to start with an oral diet as soon as leaks and other adverse events have been ruled out (Inoue et al. 2018).

The primary adverse events of POEM are mucosal injury, submucosal hematoma, and mucosal perforation, of which the last can lead to mediastinitis (Inoue et al. 2018). Besides these, post-interventional reflux has most notably been reported as a grave long-term side-effect (Rösch et al. 2017). POEM cannot be combined with fundoplication because it has no access to the abdominal cavity. Therefore, post-interventional reflux caused by the procedure is expected to be much worse than after Heller myotomy with subsequent fundoplication (Sanaka et al. 2019). Long-term exposure of the esophagus to gastric acid might entail late effects such as an increased risk for the development of esophageal carcinoma. Recently, this major disadvantage of POEM was addressed by the development of novel endoscopic

extensions of the procedure. Inoue et al. (2019) introduced an approach which they called *"POEM+F"*. They describe a technique that involves advancing the endoscope through the submucosal tunnel into the abdominal cavity and then wrapping the stomach around itself, thus basically aiming to imitate a fundoplication. Prior to that, Tyberg et al. (2018) had already introduced "transoral incisional fundoplication" (TIF), a procedure which aims to achieve a 270° wrap from inside the stomach. It has yet to be seen if such new approaches will become established and help reduce the reported reflux issues after POEM.

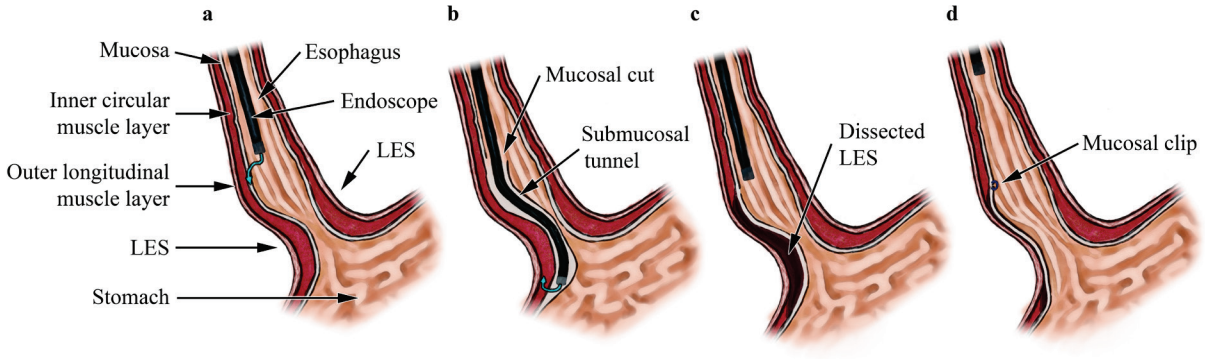


Figure 12: Per-Oral Endoscopic Myotomy. Inspired by Inoue et al. (2010). **a:** The endoscope is inserted into the esophagus. The mucosa is cut to create an opening for a submucosal tunnel. **b:** The endoscope is pushed forward down the esophagus until about 1 to 2 cm into the stomach while cutting the tissue between the mucosa and the inner circular muscle layer. This creates a submucosal tunnel. **c:** The inner circular muscle layer is cut from distal to proximal while retracting the endoscope back toward the tunnel entrance. The outer longitudinal muscle layer remains untouched to prevent esophageal perforation. **d:** The endoscope is pulled back into the esophageal lumen. The tunnel entrance is clipped.

1.8.6 Treatment After Treatment

Many studies have been conducted to assess which treatment approach grants the best outcomes, yet the results vary and are even partially conflicting (see Hanna et al. 2018, Kahrilas and Pandolfino 2017, Kumbhari et al. 2015, Martins et al. 2020, Moonen et al. 2016, Rohof et al. 2013). Balloon dilatation, Heller myotomy, and POEM are the de-facto only known long-term achalasia treatments capable of providing a lasting effect. However, achalasia remains incurable and tends to relapse. Therefore, many patients who consider undergoing POEM have already been treated with one or more of the older methods before. Since prior treatments typically induce inflammation and submucosal fibrosis in the esophagus, they may severely complicate the creation of the submucosal tunnel during POEM (Nabi et al. 2017, Richardson et al. 2003, Shiwaku et al. 2016a). This may increase the technical difficulty of the procedure (Liu et al. 2019). It is therefore reasonable to assume that previous treatments have an impact on the long-term outcome after POEM as well. This is what this study ultimately wants to clarify.

2 Material and Methods

2.1 Software

Table 4: List of Used Software.

Software and Packages	Version	Purpose
R	4.0.0	Core statistics; data aggregation, manipulation, and visualization.
BaylorEdPsych	0.5	Little's MCAR test.
generalhoslem	1.3.4	Hosmer-Lemeshow test.
ggplot2	3.3.0	Graph plotting.
metaviz	0.3.1	Rainforest plots.
mice	3.8.0	Multiple imputation.
MissMech	1.0.2	Hawkins' test and Jamshidian's test.
moments	0.14	Skew and kurtosis calculation.
rsm	5.1-4	Regression model validation.
survival	3.1-8	Cox proportional hazards regression analysis.
survminer	0.4.6	Survival curve plotting.
finalfit	1.0.0	Missing data analysis and graph plotting.
PHP	7.4.5	Data aggregation and manipulation.
MySQL	8.0.20	Data storage, aggregation, and manipulation.

2.2 Statistical Background

2.2.1 Null Hypothesis, Significance Level, and Probability Value

In statistics, scientific theories are usually formulated as testable hypotheses. The *alternative hypothesis* denotes the assumption that the theorized effect in question exists (Field et al. 2012). In contrast, the *null hypothesis* is the opposed assumption that it does *not* (Harrell 2016). Since most scientific hypotheses are formulated in a way such that it is impossible to prove their validity, the usual approach is to test the null hypothesis instead to see if there is sufficient evidence to reject it.

To achieve this, a measure is required that allows for the differentiation between whether a hypothesis test's result is more likely an expression of randomness than it is the result of a real underlying phenomenon, i.e. the presence of the effect in question. This standard is the *significance level* α, usually chosen to be 5 %. It is the probability of falsely rejecting the null hypothesis (Field et al. 2012). In other words, α is a threshold for how much data must deviate from the expectation for said deviation to be assumed not random. Once an α level has been defined, the statistical test of choice is performed. It results in a probability value, or p value, which is the probability of the test's result occurring by chance (see Field et al. 2012). Intuitively, if $p < \alpha$, there is strong evidence against the null hypothesis, which is then rejected. The alternative hypothesis is accepted instead, and the test result is called *significant*.

2.2.2 Fisher's Exact Test and Student's T Test

A common interest in clinical research is the analysis of *factorial models* that compare groups based on the observation counts of a factorial variable like treatment success and failure. *Metric models* on the other hand compare groups based on distinct observations of a continuous variable like age or pressure. *Fisher's exact test* (Fisher 1922) and *Student's t test* (Gosset 1908 under his pseudonym "Student") compare the differences between groups of a factorial or metric model, respectively. Either test reaching significance indicates strong evidence for real differences between the groups, i.e. differences that are unlikely to be mere artifacts of chance.

2.2.3 Normality and Deviation from Normality

2.2.3.1 Skew and Kurtosis

Many statistical procedures rely on the assumption of normality for the data they analyze, i.e. the assumption that the data's probability distribution does not differ significantly from a normal distribution. It is therefore of vital interest to test data for normality to avoid inaccurate or even wrong predictions.

Skew and *kurtosis* are descriptive metrics of a variable's probability distribution. Skew is a measure of symmetry around the mean: a value of 0 is considered symmetric, a positive skew indicates an accumulation of values below the mean, and a negative skew indicates an accumulation of values above the mean (Field et al. 2012). Kurtosis is a measure of the extent to which values accumulate in the tails, i.e. how frequent outliers appear at the higher and lower extremes of the distribution (Field et al. 2012). Pearson (1905) originally described the concept of kurtosis as a measure of deviance from normality. He coined the terms *mesokurtic* for a normal distribution, *platykurtic* for a kurtosis above that of a normal distribution, and *leptokurtic* for a kurtosis below that of a normal distribution.

The normal distribution's skew is 0 (Field et al. 2012). Its kurtosis is 3, though it is sometimes wrongly referred to in the literature as being 0, which is actually the *excess* kurtosis (DeCarlo 1997). Figures 13 to 17 show examples of a normal distribution and common deviations from normality that have been drawn based on randomly generated pseudo-data.

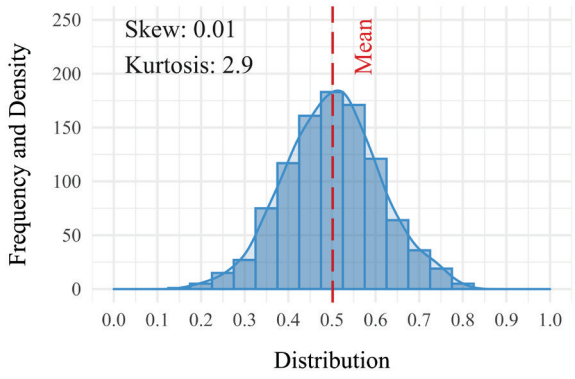


Figure 13: A Normal Distribution. It is symmetric around its mean (skew ≈ 0), and mesokurtic (kurtosis ≈ 3).

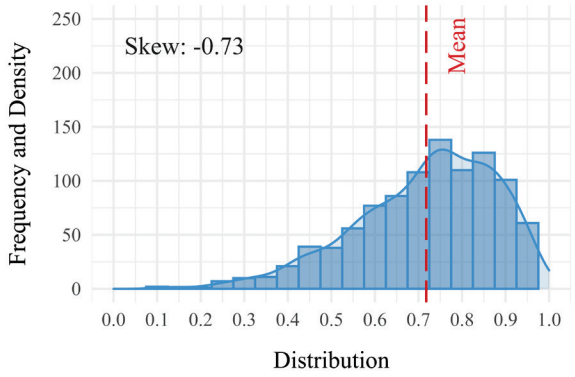


Figure 14: A Negatively Skewed Distribution. Its most frequent values accumulate above the mean.

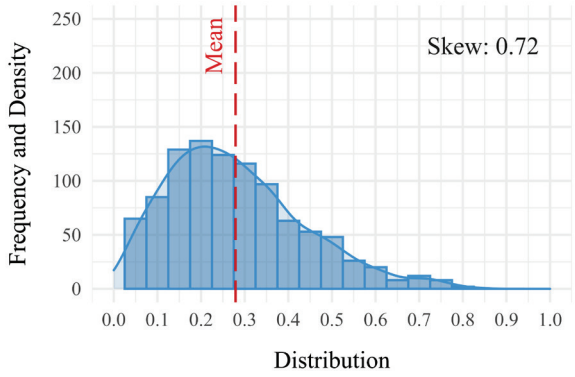


Figure 15: A Positively Skewed Distribution. Its most frequent values accumulate below the mean.

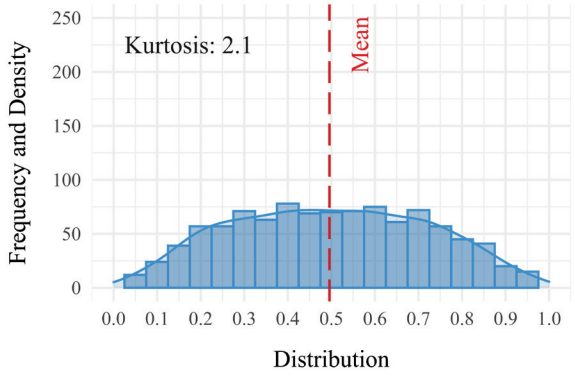


Figure 16: A Platycurtic Distribution. The tails of the curve show a reduced incidence of values compared to a normal distribution.

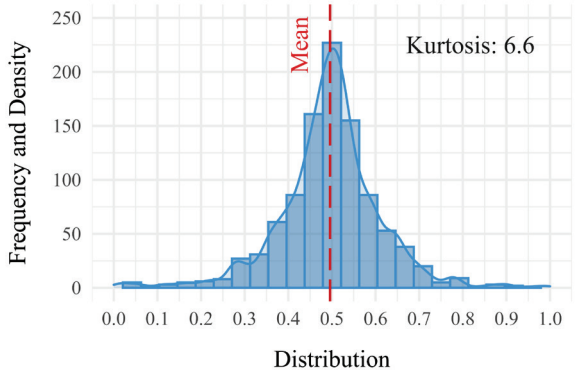


Figure 17: A Leptokurtic Distribution. The tails of the curve show an elevated incidence of values compared to a normal distribution.

2.2.3.2 The Shapiro-Wilk Test

Only looking at a histogram does not always suffice to distinguish a normal from a non-normal distribution. Shapiro and Wilk (1965) introduced a test for whether a set of complete observations deviates significantly from normality. In simplified terms, it compares the variance of the data, i.e. the squared deviations from the distribution's mean, with the variance expected to be found if the distribution was normal. If the test reaches significance, its null hypothesis of normality should be rejected. Else, evidence against normality is weak.

As emphasized by Field et al. (2012), even minuscule deviances from normality may lead to the *Shapiro-Wilk test* reaching significance if the distribution's sample size is large. However, they state that such deviances do not necessarily need to be influential enough to actually bias the results of subsequent statistical procedures. In other words, the *Shapiro-Wilk test* may reach significance in large data sets even though the distribution does not deviate enough from normality to invalidate statistical calculations based on the assumption of normality. Therefore, it is important to also visualize a distribution in addition to the test, and to assess the extent of deviation from normality manually. In this sense, basic histograms and measures like the skew and kurtosis are still very relevant.

2.2.4 Outlier Analysis

To attain reliable results from statistical procedures, not only must data satisfy distributional assumptions made by these procedures, but the data should also be plausible. A common way of detecting outliers amid observed continuous data is based upon their *interquartile range*. This method reaches back to a way of drawing box plots that was initially introduced by Tukey (1977).

He sorts the individual observations by their values in ascending order and then splits them into quartiles. The first quartile, Q_1, is the observation in the middle of the smallest observation and the median. The second quartile, Q_2, is the median. Consistently, the third quartile, Q_3, is the middle observation between the median and the highest observation. The interquartile range is the range of observed values between the first and third quartile: $IQR = Q_3 - Q_1$. Observations are considered potential outliers if their value is below $Q_1 - 1.5\ IQR$ or above $Q_3 + 1.5\ IQR$.

Although Tukey (1977) provides no statistical justification for having chosen these very thresholds, his approach and his box plots have been widely adopted among scientists since their introduction. However, his method does not identify outliers per se. It merely identifies extreme values that are suspicious of *potentially being* outliers. They always require further individual assessments to allow for a final verdict. Factorial variables cannot have outliers because each factor level is of course plausible by definition.

2.2.5 Regression Analysis

2.2.5.1 Dependent and Independent Variables

Regression models aim to predict the value of a *dependent outcome variable* as a function of one or more *independent predictor variables*. The contribution of each predictor to the outcome estimate is the respective predictor's *effect size* or *coefficient*. In general terms, regression tries to estimate these coefficients for a defined statistical model based on a given set of observed data. This process allows for the detection of general correlations and the derivation of predictions for specific predictor constellations.

2.2.5.2 Logistic Regression, Null Models, Odds, and the Logit Function

Let Y be the continuous dependent outcome variable of a statistical model with n independent predictor variables $x_1, \dots, x_n$, their respective coefficients $\beta_1, \dots, \beta_n$ and the residual intercept β_0. Assuming a linear correlation, this model can readily be described by the following linear regression equation:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$ **Equation 1: Linear Regression.**

Linear regression is limited to the prediction of continuous dependent outcome variables. In contrast, *logistic regression* is a generalization of the concept that allows for the prediction of binary dependent outcome variables. For the primary outcome of this thesis is *treatment failure*, logistic regression is its statistical model of choice. Instead of predicting the actual value of a variable Y as seen in equation 1, logistic regression aims to predict the probability P of Y occurring, i.e. $Y = 1$. As described by Field et al. (2012), this model is characterized by the logistic regression equation:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$ **Equation 2: Logistic Regression.**

As will explained in the next chapter, it is often interesting to assess whether the inclusion or exclusion of specific predictor variables in a regression model leads to increased predictive power. This may be achieved by comparing different models with varying sets of predictors. One model commonly used for such a basic comparison is the model which assumes the null hypothesis to be true. It is therefore called the *baseline model* (see Field et al. 2012), or the *null model*. As explained in chapter 2.2.1, the null hypothesis of a model is the assumption that observed deviations from the model's assumed distribution are random and not caused by real underlying effects described by the model's predictors. Hence it follows that the null model is the model for which $x_i = 0$. In other words, the null model is a constant expression of the probability of Y occurring, independent from any predictors. This model is described by equation 3.

$$P(Y = 1) = \frac{1}{1 + e^{-\beta_0}}$$

Equation 3: Logistic Regression (Null Model).

In addition to the probability of *Y* occurring, the *odds*, *o*, is another mathematical expression of interest. The odds are the probability of *Y* occurring divided by the probability of *Y* not occurring (Field et al. 2012). In mathematical terms, this is described by the following equation:

$$o = \frac{P(Y = 1)}{\neg \mathrm{P}(Y = 1)} = \frac{P(Y = 1)}{1 - \mathrm{P}(Y = 1)}$$

Equation 4: The Odds as a Function of Probability.

Substitution of equation 3 into equation 4 followed by transformation of the formula results in the equation for the odds of the logistic regression model:

$$o = \frac{\frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}}{1 - \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)} - 1} = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n}$$

Equation 5: The Odds in Logistic Regression.

The exponentiated coefficient e^{β} is an important standard expression commonly reported for logistic regression models in the literature. It allows for an intuitive interpretation of the model's coefficients. This becomes quite clear upon further transforming equation 5 using the basic mathematical power laws of $e^{a+b} = e^a e^b$ and $e^{ab} = (e^a)^b$:

$$o = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n} = e^{\beta_0} e^{\beta_1 x_1} \dots e^{\beta_n x_n} = e^{\beta_0} \left(e^{\beta_1}\right)^{x_1} \dots \left(e^{\beta_n}\right)^{x_n}$$

Equation 6: The Odds in Logistic Regression (Transformed).

It is easy to see that e^{β_i} corresponds to the relative change in the effect of its associated predictor x_i on the odds of *Y* occurring per 1-unit increase of x_i. Therefore, e^{β_i} is the *odds ratio* of the predictor x_i.

This is quite intuitive. For example, if *x* is *age* in years, then e^{β} can be interpreted as the odds ratio for each 1-year increase of x_i. If on the other hand *x* is *sex* with female ≔ 0 and male ≔ 1, then e^{β} is the odds ratio for males compared to females. As it is apparent in equation 6, these correlations are exponential. The effect of x_i changes depending on both its own value and the value of its coefficient β_i. For example, let *x* again be *age* in years and e^{β} be 1.5. This means that for each additional year of age, the odds for *Y* occurring increase by factor 1.5. If *x* is 2 years and the odds increase by factor 1.5 per year, the total change in odds accumulates to $(e^{\beta})^x = 1.5^2 = 2.25$, and so on.

Finally, equation 5 can be further transformed. Taking its logarithm results in equation 7, which is a concise way of describing a logistic regression model. The logarithm of the odds, or the *log-odds*, is known as the *logit* function (see Harrell 2016). Let $p = P(Y = 1)$, then:

$$logit(p) = ln(o) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Equation 7: The Logit Transformation in Logistic Regression.

2.2.5.3 Cox Proportional Hazards Regression, Hazard, and the Schoenfeld Test

The analysis of right-censored event times, also known as *survival analysis*, is commonly performed using a method first described by Cox (1972) and named after him as the *Cox proportional hazards model*. As explained by him, it estimates the *hazard* in regard to a specific event happening as a so-called *hazard function* of all independent predictor variables of the statistical model, each modulated by an unknown regression coefficient and multiplied by an unknown function of time. "Right-censored" means that throughout the observation period, patients may withdraw their participation and drop out of the study before having reported an event. It is a major benefit of survival analysis that the partial information gained from the event-free time until censoring can still be included in the regression model. An interesting metric of survival analysis is the *number at risk*. It is the number of patients at any given point in time who have neither been censored *before*, nor had an event *before or at*, that point in time.

Let *H* be the hazard function of time *t* with *n* independent predictor variables $x_1, \ldots, x_n$ and their respective coefficients $\beta_1, \ldots, \beta_n$. The function $H(t)$ describes the expected hazard at the time t. Let $\lambda(t)$ denote the underlying hazard function of time for a subject with the standard set of coefficients, i.e. $\sum_1^n \beta_n x_n = 0$ and thus $e^{\beta_1 x_1 + \cdots + \beta_n x_n} = 1$. This proportional hazards model is then described by the following equation (see Cox 1972, Harrell 2016):

$$H(t) = \lambda(t)\, e^{\beta_1 x_1 + \cdots + \beta_n x_n}$$

Equation 8: Proportional Hazards Regression.

Harrell (2016) explains that the so-called *underlying hazard function*, $\lambda(t)$, is usually unknown, yet of little interest anyway, whereas the so-called *relative hazard function*, $e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n}$, describes the effects of the predictor variables on the hazard. As such, it is the primary term of interest. Similar to the odds ratio in logistic regression, in a proportional hazards regression model, the *hazard ratio* for a predictor variable represents its influence on the overall hazard per 1-unit increase of said predictor. Let *T* be the time to event and let t again depict time. The underlying hazard function is then defined as follows (see Harrell 2016, notation amended for the sake of consistency):

$$\lambda(t) = \lim_{u \to 0} P(t < T \leq t + u \mid T > t)$$

Equation 9: The Hazard Function.

Thus, the *hazard* is the probability at any given time t of experiencing an event during the forthcoming infinitesimally short time span, given that no event has occurred before. In less technical terminology, it is the probability of a patient who did not experience an event yet to experience it the very next moment. As such, it needs to be noted that, despite the similar interpretation of the odds ratio and the hazard ratio, the odds and the hazard are fundamentally different measures.

The Cox model estimates the hazard as a function of time, depending on independent predictors, which are often derived from clinical experience or theoretical considerations. As can be seen in equation 8, this model assumes a proportional relationship between the predictors. In other words, it expects a linear term in the exponent that contains no interactions between any two predictors, and no interactions between any predictor and time. This is the *proportional hazards assumption*, which needs to hold true for the model to be appropriate (Cox 1972).

The proportional hazards assumption can be assessed by a test that was introduced by Schoenfeld (1980). It determines distinct measures, the so-called *Schoenfeld residuals*, for each predictor, each observation, and every time an event was observed in the data. His algorithm calculates these residuals from the predictor's observed and predicted values, for those observations that had an event occurring. Censored records are discarded. The test then assesses whether the *Schoenfeld residuals* are distributed in a random pattern over time, or if they show a trend as time progresses. The latter of course indicates a correlation between the predictor and time, and thus a breach of the proportional hazards assumption. If this *Schoenfeld test* reaches significance, a violation of proportional hazards should be assumed. The proportional hazards regression model may then not suffice for adequate coefficient estimation. If significance is not reached, evidence against proportional hazards is weak.

2.2.6 Model Selection and the Goodness of Fit

When constructing a model for regression analysis, the choice of which predictors to include is essential. Too few, too many, or a wrong selection of predictors may hurt the model's fit and reduce its predictive power. There is a multitude of statistical measures aiming to assert a good model fit. Some of them can be used for a quick comparison of the model's fit with the fit of another model, such as the null model, to assess whether the chosen predictors improved the model. Others are commonly used to determine the goodness of fit of a model as-is, without requiring a second model for comparison. This chapter explains the major measures for the goodness of fit used in this thesis.

For linear regression:

- the *coefficient of determination,* R^2.

For logistic regression:

- the *log-likelihood*, *deviance*, and *Akaike information criterion* (AIC),
- Nagelkerke's log-likelihood based *pseudo-*R^2,
- Kendall's τ_a, Goodman and Kruskal's γ, and Somers' D, and
- the *Hosmer-Lemeshow test*.

For proportional hazards regression:

- the *log-rank test* and Harrell's *concordance*.

As explained by Korn and Simon (1990), a model that has been fit well for a specific set of predictor variables will on average make correct predictions. However, they warned about mistaking the statistical significance of a predictor variable for the overall predictive power of the model, for even significant predictors may in fact contribute very little to the latter. Similarly, they state that predictive power should not be confused with the goodness of fit; although a well fit model will on average make correct predictions, these may still be of little precision for any individual patient.

2.2.6.1 The Coefficient of Determination: R^2

In linear regression models, the so-called *coefficient of determination*, R^2, is a measure of the spread of data points around their regression line (King 1986). In general terms, it is the fraction of variation shared between variables (Field et al. 2012). In this thesis, R^2 itself is not used since no linear regression is performed. Its basic knowledge is however beneficial for the understanding of the *pseudo-*R^2 measure that will be discussed in section 2.2.6.3, which is commonly reported for logistic regression models.

As a fraction, R^2 naturally yields values between 0 and 1. While 0 indicates that the model shares *none* of the variance found in the data, a value of 1 indicates that it shares *all* the variance. R^2 is often utilized as an estimate for a model's predictive power: higher values are considered indicative of a better fit.

2.2.6.2 The Log-Likelihood and the Deviance

This section recalls two basic statistical measures as they are explained by Field et al. (2012). The mathematical notation has been slightly amended for the sake of consistency, with no changes to its meaning.

In logistic regression, the *log-likelihood*, *LL,* is a measure of how much observed information remains unexplained by a regression model after it has been fit. It compares the observed values to the values predicted by the model. Let Y be the observed outcome, let $\hat{Y}$ be the model's prediction for said Y, and let n denote the number of observations. The log-likelihood is then calculated as follows:

$$LL = \sum_{i=1}^{n} \left(Y_i \ln\left(\hat{Y}_i\right) + (1 - Y_i) \ln\left(1 - \hat{Y}_i\right) \right)$$ **Equation 10: The Log-Likelihood.**

The *deviance* is derived from the log-likelihood. It is commonly reported to compare the quality of multiple alternative logistic regression models that have been fit with different sets of predictors based on the same underlying data:

$$deviance = -2LL$$ **Equation 11: The Deviance.**

The more information in the data remains unexplained by the model, the larger the deviance. Therefore, a higher deviance indicates a worse model fit. For the log-likelihood, naturally, the opposite holds true.

2.2.6.3 The Akaike Information Criterion

Using likelihood-based estimates such as the deviance to assess the goodness of a model's fit is risky. Due to the way these estimates are calculated, they will generally favor the model with the highest dimension (Schwarz 1978). In other words, a model built upon a higher number of predictors will tend to yield a lower deviance – not because it is better fit, but simply because it incorporates more parameters. To address this issue, the *Akaike information criterion* introduces a proportional penalty for every additional model parameter k into the deviance (Akaike 1973):

$$AIC = -2LL + 2k$$ **Equation 12: The Akaike Information Criterion.**

In a logistic regression model, *k* equals the number of independent predictors, plus 1 for the residual intercept. Thus, the AIC penalizes increasing model complexity. Just like the deviance, the AIC has no inherent meaning on its own, but it can be calculated for different models based on the same data to assess which of them has a better fit (Field et al. 2012).

2.2.6.4 Nagelkerke's Pseudo-R^2

As pointed out earlier, the ordinary R^2 is only defined for linear regression models. In logistic regression, log-likelihood based *pseudo-*R^2 measures are used instead. They represent the improvement in model

likelihood after fitting, compared to the null model (Hemmert et al. 2016). As such, despite their similar names, R^2 and *pseudo-*R^2 measures are fundamentally different in both nature and interpretation.

Many different *pseudo-*R^2 have been described in the scientific literature over the years with little consensus over which one is best for what. In this thesis, the one proposed by Nagelkerke (1991) is used:

$$R_N^2 = \frac{1 - e^{\frac{2LL_{null} - 2LL_{fit}}{n}}}{1 - e^{\frac{2LL_{null}}{n}}}$$

Equation 13: Nagelkerke's Pseudo-R^2.

In this formula, LL_{null} is the log-likelihood of the null model, LL_{fit} is the log-likelihood of the fit regression model, and n is once again the number of observations. Again, the variables' names have been slightly altered from the source for the sake of achieving a consistent naming convention in this thesis.

2.2.6.5 Measures of Discrimination: τ_a, γ, and D

Many measures have been proposed in the literature that can assess the quality of a logistic regression model's fit. The most common ones are Kendall's τ_a, Goodman and Kruskal's γ, and Somers' D. All three are essentially ratios of ranked observation pairs, as will be explained in this section. While they share a common numerator, their denominators are composed of different terms that slightly change the interpretation of each measure. This chapter first introduces their general concepts and formulas. They will then be applied to logistic regression.

Let X be a ranked independent predictor variable and let Y be a ranked dependent outcome variable. Furthermore, let $(x_1, y_1), \dots, (x_n, y_n)$ be a data set of n observations. Any two observation pairs (x_a, y_a) and (x_b, y_b) are *concordant* if the observation with the larger x also has the larger y, they are *discordant* if the observation with the larger x has the smaller y, they are *tied on X* if $x_a = x_b$, and they are *tied on Y* if $y_a = y_b$ (see Kruskal 1958).

Out of all possible pairs of observations, let N be the sum of all pairs, let P be the sum of concordant pairs, let Q be the sum of discordant pairs, let Y_0 be the sum of pairs that are tied on Y, and let X_0 be the sum of pairs that are tied on X.

Kendall's τ_a is the sum of concordant and discordant pairs over the total number of pairs (Kendall 1938):

$$\tau_a = \frac{P - Q}{N} = \frac{P - Q}{\binom{n}{2}} = \frac{P - Q}{\frac{n(n-1)}{2}} = \frac{2(P - Q)}{n(n-1)}$$

Equation 14: Kendall's τ_a.

Goodman and Kruskal's γ is the sum of concordant pairs minus the sum of discordant pairs, divided by the sum of concordant and discordant pairs (Goodman and Kruskal 1954, 1959, 1963, 1972):

$$\gamma = \frac{P - Q}{P + Q}$$

Equation 15: Goodman and Kruskal's γ.

Somers' D is calculated like γ, but it also factors in the number of ties. Since a pair of observations can be tied on X or Y, *Somers' D* is an asymmetric measure with two possible formulas: D_{xy} is calculated from the sum of pairs that are tied on X, and used if X is the dependent outcome variable, whereas D_{yx} is calculated from the sum of pairs that are tied on Y, and used if Y is the dependent outcome variable (Somers 1962).

$$D_{yx} = \frac{P - Q}{P + Q + Y_0},\ D_{xy} = \frac{P - Q}{P + Q + X_0}$$

Equation 16: Somers' *D*.

As can be easily deduced from the formulas, the possible values of all three of these coefficients range between -1 and 1. As described by Harrell (2016) for D_{yx}, but obviously no less applicable to τ_A and γ, if a measure is 0, the model's predictions are pretty much random, whereas a value of 1 indicates perfect predictions. Consistently, a value of -1 depicts a perfect disagreement in the sense that all pairs are discordant. Such a model always predicts the exact opposite of the actual real-world observation.

The three measures of discrimination, τ_a, γ, and D, were initially introduced based on the aforementioned ideas. However, they can efficiently be utilized to assess the quality of a regression model's fit. For this application, they need to be slightly adjusted, though.

For a binary logistic regression model estimate, let Y once again be the ranked dependent outcome variable. Since the model is binary, Y can either be 0 (≔ event does not occur), or 1 (≔ event does occur). Let $\hat{Y}$ be the fit regression model's prediction for Y. As explained in chapter 2.2.5.2, a binary logistic regression model estimates the probability P of Y occurring, i.e. $P(Y = 1)$. Therefore, the possible values for $\hat{Y}$ range from 0 to 1. Let $(\hat{y}_1, y_1), \ldots, (\hat{y}_n, y_n)$ be a data set consisting of n observations y, each paired with their associated prediction $\hat{y}$. Since Y can only be 0 or 1, the pairing of every observation-prediction pair with each other would now lead to a very high number of ties on Y, for which a correct or wrong rank order cannot be determined. Because of this, instead, any two pairs $(\hat{y}_a, y_a)$ and $(\hat{y}_b, y_b)$ for which $y_a \neq y_b$ are now looked at (see Orth 2010). In other words, every observed event is paired with an observed non-event. A pair is considered *concordant* if $\hat{y}$ is higher for the observation y in which the event occurred, i.e where $y = 1$. In contrast, a pair is *discordant* if $\hat{y}$ is higher for the observation y in which the event did *not* occur, i.e where $y = 0$. If $\hat{y}$ is the same for both, the pair is tied (on $\hat{Y}$).

By applying these rules, τ_a, γ, and D_{xy} can be used as measures for the goodness of fit of a logistic regression model. To summarize, they are calculated on a data set that does not consist of pairings of X with Y, but of pairings of $\hat{Y}$ with Y instead (Harrell 2016).

Because ties are defined as pairs for which $\hat{y}_a = \hat{y}_b$, and $\hat{Y}$ takes the place of the original X described at the beginning of this section, D_{xy} is commonly used since it includes the pairs tied on $\hat{Y}$ in its formula. Incidentally, since the pairing criterion is for both observations y to have opposite values ($y_a \neq y_b$), there can be no pairs that are tied on Y, therefore $Y_0 = 0$ and $D_{yx} = \gamma$.

While τ_a includes the total number of pairs in its denominator, γ only incorporates the sum of concordant and discordant pairs. If the data contains a large number of ties, they may considerably outweigh the concordant and discordant pairs in the formula. This may lead to a very small τ_a that may be unfit to discriminate between the untied pairs. In such cases, γ may be the preferred measure, for it ignores the number of tied pairs. Ultimately, D_{xy} is especially adequate for logistic regression analysis since it includes the number of pairs that are tied on $\hat{Y}$, while ignoring the pairs that are tied on Y.

2.2.6.6 The Hosmer-Lemeshow Test

The *Hosmer-Lemeshow test* assesses the goodness of fit in a logistic regression model in quite a different way than τ_a, γ, and D. Rather than looking at the rank ratios, it tests for differences in event proportions. As introduced by Hosmer and Lemesbow (1980), the test first orders the data records based on their predicted probability $\hat{Y}$, before dividing them into a specific number g of equally sized groups. The number of groups is usually chosen to be g = 10, sometimes called the *deciles of risk* (Fagerland and Hosmer 2013). For each group, the number of observed events is then compared to the number of expected events, and the number of observed non-events is compared to the number of expected non-events. The results are ultimately summed up to a single χ^2 value. The test statistic H is calculated by the following equation (see Hosmer and Lemesbow 1980, notation amended for the sake of consistency):

$$H = \sum_{k=0}^{1} \sum_{i=1}^{g} \frac{\left(|Y_{ki}| - |\hat{Y}_{ki}|\right)^2}{|\hat{Y}_{ki}|}$$

Equation 17: The Hosmer-Lemeshow Test Statistic.

In this equation, k = 0 denotes non-events and k = 1 denotes events. Therefore, Y_{1i} are the observed events in the group i, Y_{0i} are the observed non-events in the group i, $\hat{Y}_{1i}$ are the predicted events in the group i, and $\hat{Y}_{0i}$ are the predicted non-events in the group i. Based on the test statistic, an associated p value is calculated. If the test reaches significance, there is strong evidence for a bad model fit. Else, evidence is weak.

2.2.6.7 The Log-Rank Test

The *log-rank test*, or *Mantel-Cox test*, is a rank test for distributive differences between survival tables (Mantel 1966, Peto and Peto 1972). Its most common application is to test for significant differences in the survival of specific subgroups within the observed data. As introduced by Mantel (1966), the test compares the groups based on their differences in the number of observed events at each point in time where an event was observed in either group, adjusted for the number of patients at risk at that same time. If the test reaches significance, there is strong evidence for differences in the observed survival between groups. Else, evidence is weak, and observed differences may well be mere artifacts of chance.

2.2.6.8 Harrell's Concordance

The *concordance c*, also called *Harrell's c*, is a ranked correlation coefficient quite like τ_a, γ, and D_{yx}, that has been modified for the use in survival analysis. It is the fraction of concordant pairs.

As described by Harrell et al. (1996), the concordance is calculated by aggregating a list of all possible pairs of patients of whom one or both have died during follow-up. He considers a pair concordant if the model predicts a longer survival time for the patient who in fact lived longer or is known to have at least been still alive at a time the other patient had already died. Conversely, he considers a pair discordant if the model predicts a longer survival time for the patient who died first. Pairs are discarded if they cannot be ordered. This is the case, for example, if both patients died at the very same time, or if only one patient is still alive, but their follow-up is yet too short to tell if they will outlive the one who died. Ultimately, if the predicted survival times of two patient pairs are exactly the same, Harrell counts them as half a concordant pair. In this case, the concordant pair count is increased by ½, while the total number of pairs is still increased by 1. This is a compromise not to throw away the information gained from pairs where the predicted rank order is obviously inaccurate, yet not technically wrong.

The concordance is also a rescale of *Somers' D_{xy}* to the range of 0 to 1 (Harrell 2016):

$$c = \frac{D_{xy} + 1}{2} \Leftrightarrow D_{xy} = 2c - 1$$

Equation 18: Harrell's Concordance.

It follows that a concordance of 0.5 indicates no predictive discrimination since precisely half of the estimates are wrong. A value of 1 indicates that the model's prediction is always correct. Consistently, a concordance of 0 implies that the model always predicts the exact opposite order of death than the one observed. Usually, a concordance of about 0.8 or higher is considered indicative of a decent predictive power of a regression model (Harrell 2016).

In applied statistics, a combination of the concordance with a *pseudo-R^2* measure can be quite interesting. Since survival analysis shares many principles with logistic regression, there exists a multitude of *pseudo-R^2* measures for it as well. However, R^2-like measures for the goodness of fit assess if a prediction will be correct on average, not how precise a specific prediction for an individual patient will be (Korn and Simon 1990). For this reason, R^2-like measures may obviously drop to very low values in clinical survival analysis if an exact death time prediction is unrealistic. However, when interpreting survival regression models like the Cox proportional hazards model, effect size estimates like the hazard ratios are often much more interesting than the prediction of high-precision death times. For example, if a model can reliably predict whether a patient can expect to live significantly longer with or without treatment, this information is clinically invaluable even if the same model is somewhat imprecise at predicting the exact amount of weeks or months of expected survival. Therefore, a lack of precision in survival time estimation is often not that important.

To conclude, a low pseudo-R^2 indicates that concrete death time estimates should be handled with care and expected to be imprecise for the individual patient. However, if the concordance is quite high, a low R^2 does not diminish the value of other more interesting insight gained from the model.

2.2.7 Handling of Missing Data

2.2.7.1 MCAR, MAR, and MNAR

Missing information in clinical surveys and trials is ubiquitous and may have a grave negative impact on the predictive power of many statistical procedures.

Data is commonly classified into three categories that reach back to Rubin (1976): missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Rubin studied the nature of processes that lead to missing information in data to tell under which circumstances it is appropriate to *ignore* these processes and still attain valid inferences from said data. He argued that every observation in a data set possesses a specific likelihood of being missing. As summarized by van Buuren (2018), data is MCAR if this chance of being missing is constant for all the data. He states that it is MAR if the chance of being missing is constant within, but not between, specific subgroups defined by the *observed* data. Ultimately, if this constraint is breached, too, he classifies the data as MNAR.

When data are MCAR, incomplete observations may simply be removed. This is known as complete case (CC) analysis, or listwise deletion. In the unique situation of MCAR, this is expected to not introduce new biases into the data (Mukaka et al. 2016). It does however come at the expense of a loss of information, obviously. Depending on the fraction of missing information, this data reduction may

severely reduce the explanatory power of subsequent statistical analyses. That said, it is rarely justified to assume data to be MCAR in clinical studies anyway. If MAR is still plausible, though, statistical analyses may be performed using a multitude of different methodical approaches. Complete case analysis and "last observation carried forward" are commonly practiced techniques that unfortunately tend to introduce new biases and may thus yet again lead to invalid effect size estimates (Altman 2009). Multiple imputation (MI) is a younger approach that aims to produce better results in cases where data is incomplete, yet can be assumed to be at least MAR. It will be discussed in chapter 2.2.7.4.

When data is MNAR, the observation method should be revised to yield better data.

2.2.7.2 Little's Test

Many statistical procedures expect data to be MCAR to yield valid results and high power of effect size estimates. It is therefore important to analyze missing data patterns before applying statistical methods that may be inept for use with MNAR data.

Most methods used to assess the MCAR assumption rely on testing for either homogeneity of means or homogeneity of covariances, the latter also known as homoscedasticity (Jamshidian and Jalal 2010).

Little (1988) introduced a test for the assessment of homogeneity of means. His test statistic divides the analyzed data into subgroups based on their patterns of missing values. It then analyzes how much the means of the non-missing values inside these groups vary between the groups. Little argued that if the differences between the means are negligible, i.e. the means are homogenous, evidence against MCAR is weak. Therefore, if the test fails to reach significance, the null hypothesis of MCAR cannot be rejected. Likewise, if the test turns out to be significant, there is strong evidence against MCAR. The data may still be MAR, though.

As described earlier, *Little's test* assumes normality (Jamshidian and Jalal 2010, Little 1988). If the data consist of small group sizes or do not meet the assumption of normality, many tests, including Little's, may lead to falsely rejecting MCAR (Jamshidian et al. 2014). The reason for this is quite intuitive: If the missing data are in fact missing at random, but the data itself are not normally distributed, then the missing data will equally present patterns of deviance from normality. If this is the case, other tests are required that do not rely on the assumption of normality.

2.2.7.3 Hawkins' Test and Jamshidian's Test

Jamshidian and Jalal (2010) introduced a test algorithm for the assessment of homoscedasticity, multivariate normality, and MCAR. A major benefit of their method is that it can be used if normality is violated and *Little's test* is therefore ineligible. To explain the functionality of these tests in detail would require an elaborate introduction and discussion of higher mathematics that would exceed the intended scope of this thesis. In simplified terms, they impute missing data to be capable of applying a variety of established procedures for the assessment of homoscedasticity that are usually unfit for incomplete observations. Imputation will be explained in more detail in the next section.

The algorithm described by Jamshidian and Jalal (2010) centers around a test statistic previously introduced by Hawkins (1981) that assesses the homogeneity of covariances, i.e. homoscedasticity, in complete data. They first impute missing values based on the data's mean and covariance under the assumption of normality to generate a "complete" data set, which is then split into groups. They then apply *Hawkins' test statistic* to each group and combine the results into an overall *p* value by utilizing a rank statistic known as the *Anderson-Darling test,* which was introduced by Scholz and Stephens (1987) and assesses distributional uniformity between each group's Hawkins' statistic. They argue that if this test's result is rejected, so is the assumption of normality. To account for cases in which the data is not normally distributed, they repeat the entire routine as just described, only this time using an imputation method inspired by Srivastava and Dolatabadi (2009) that makes no distributional assumptions and therefore does not require normality. Ultimately, they conclude that if this second iteration of their test is rejected as well, evidence is strong for a violation of homoscedasticity and thus MCAR.

For the sake of clarity, the first iteration of the test will be referred to as *Hawkins' test* in this thesis, while the second iteration will be named *Jamshidian's test*. Both may be applied in *R* using the *MissMech* package provided by Jamshidian et al. (2014).

Figure 18 illustrates the complete MCAR test algorithm developed for this thesis. The reason it starts with *Little's test* to assess the homogeneity of means instead of always applying *Hawkins' test* for the homogeneity of covariances right away is that Little's test is a long known and well-established standard. For many applications, it is expected to suffice. If the tested data is a priori known or strongly suspected to lack normality, *Little's test* might as well be skipped, though.

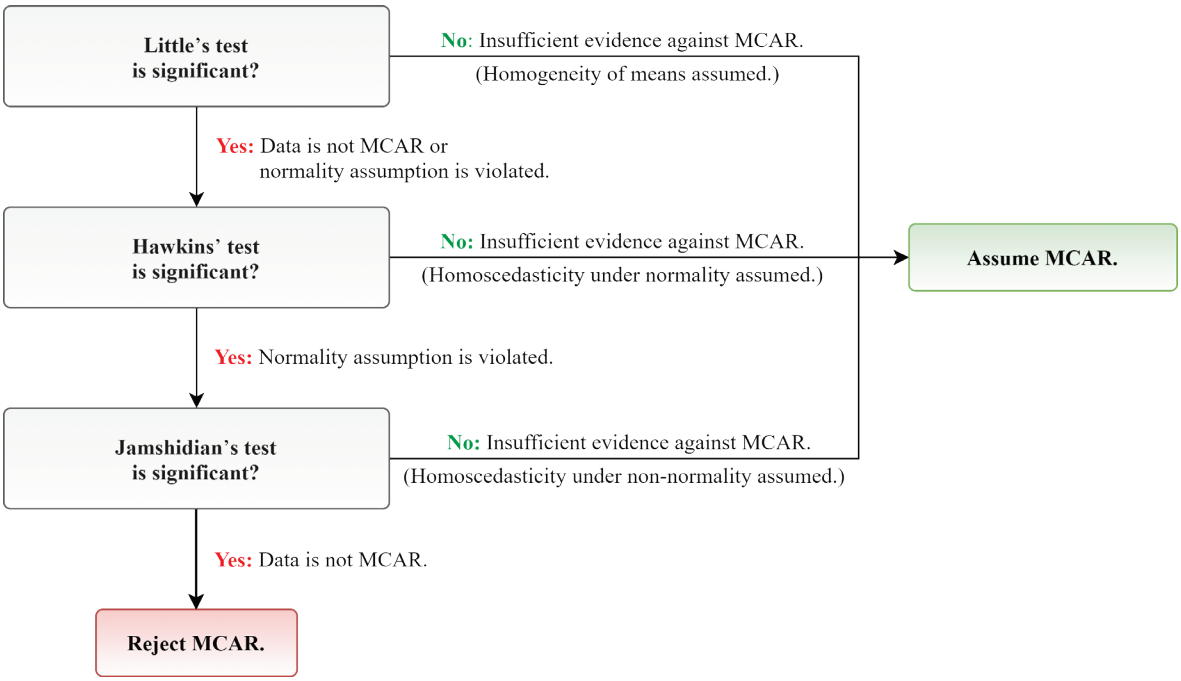


Figure 18: MCAR Test Algorithm. If either Little's test, Hawkins' test, or Jamshidian's test for homoscedasticity consecutively fails to reach significance, evidence against MCAR is weak.

2.2.7.4 Multiple Imputation

Multiple imputation is an advanced method for the handling of missing data introduced by Rubin (1987). Since its introduction, the increasing processing power of modern computer systems has allowed it to become the focus of attention of many researchers. It is a modern alternative to older strategies that required the removal of incomplete observations and thus forced a reduction of the available information.

Unlike methods such as complete case analysis, multiple imputation does not remove incompletely observed data records. Instead, a multitude of copies of the entire data set is created. In each of these, the missing observations are interpolated with plausible values based on the observed data. This factors in the uncertainty associated with the unknown values while at the same time avoiding an over-representation of the unknown values compared to the known ones. Multiple imputation keeps valuable partial information from incomplete records in the data and thus makes it available for the use in subsequent statistical models and procedures. As such, this approach is especially useful in clinical settings, where a drop-out of patients over time is quite common. In addition to that, it minimizes the introduction of new biases, which is usually expected from many other methods such as, again, complete case analysis.

There is a plethora of algorithms that can be used to impute missing data. One of the most common choices is *predictive mean matching*. As described by Harrell (2016), it replaces each missing observation with an actually observed so-called *donor observation*, which it samples from a distribution derived from the unimputed original data. As he points out, a major advantage of this method is its independence of distributional assumptions, which it maintains by not generating new values. This means that predictive mean matching yields good results even if the imputed data lacks normality. It is a robust method.

After multiple imputation has been performed and a multitude of imputed data sets have been generated, the statistical procedure of choice (e.g. linear regression, logistic regression, or Cox regression) is applied separately to each imputed data set. The resulting statistical measures of interest (e.g. effect sizes, standard errors, or *p* values) are then each pooled into a single estimate by following *Rubin's rules*, a defined set of mathematical algorithms also introduced by Rubin (1987). In the case of regression analysis, this results in a single pooled multiply imputed regression model.

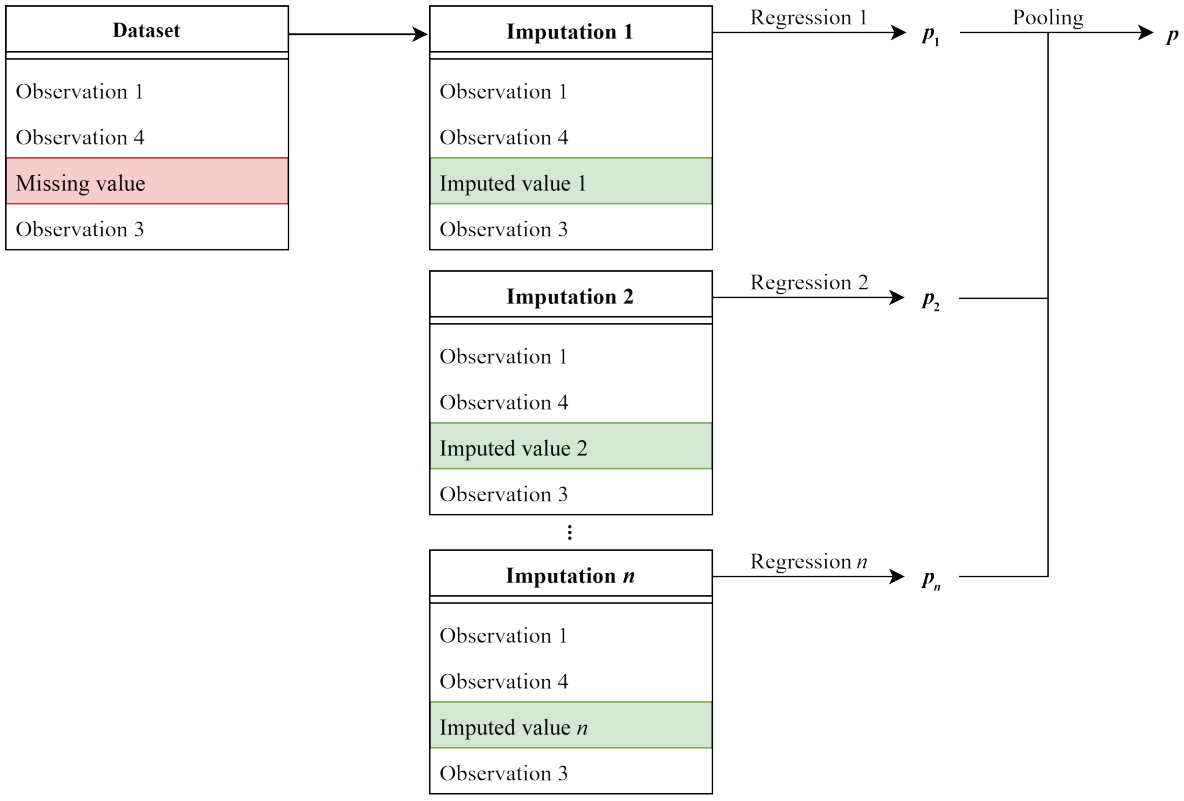


Figure 19: Basic Principles of Multiple Imputation. First, several copies of the original incomplete data set are generated. In each of these, missing observations are replaced by actually observed donor values taken from the observed data. For each imputed data set, the statistical procedure of choice is then applied to calculate statistical measures of interest, such as *p*. Each of these measures is eventually pooled into a single estimate using Rubin's rules.

Unfortunately, besides the basic effect size estimates of regression analysis, only a few combination rules for some very specific statistical measures have been developed to date (van Ginkel et al. 2020). Therefore, for many advanced modern statistical measures, such as the *AIC*, *concordance*, τ_a, γ, and D, as well as most routines, such as the *Hosmer-Lemeshow* test, there is little to no evidence available on how to adequately pool them. How to best deal with this problem is debated among researchers and in dire need of clarification. For now, van Buuren (2018) suggests to calculate such advanced statistical measures for a supplementary model built from the complete case data instead, and use these as rough surrogates for the unavailable measures of the pooled multiply imputed model. Another approach brought up by van Ginkel et al. (2020) is to simply pool statistical measures by calculating their means across the imputed models. They argue that even if there is no theoretical justification for this, the calculated means can still be expected to provide rough yet reasonable insight. They do emphasize, though, that one should always evaluate whether this approach is adequate for the concrete statistical measure in question, and to always be transparent with the fact that this approach is solely taken due to a lack of better alternatives. This thesis implements either approach: for each multiply imputed regression model, effect size estimates and other statistical measures are provided both calculated from comparative complete case models, as well as averaged across all imputed models. They are then discussed.

As explained in chapter 2.2.5, regression analysis tries to predict a dependent outcome variable as a function of independent predictor variables. There is little discord on the capability of multiple imputation to handle missing predictor observations. However, data may of course also be missing observations of the outcome variable itself. Imputing the outcome variable is controversial. It seems ill-advised to impute outcomes based on specific predictor values and then, during regression analysis, try to explain these same imputed outcomes based on the very same predictors. A common criticism is that these imputed values might tend to confirm the model that was built upon them. Harrell (2016) recommends removing records with missing outcomes from the data prior to imputation. In contrast, von Hippel (2007) explicitly argues in favor of imputing the outcome variable since it may provide additional information for the imputation of the predictors, or else the predictors would simply be imputed as if they were unrelated to the outcome. However, he suggests to *not* include these records with imputed outcomes in subsequent statistical analyses. Ultimately, van Ginkel et al. (2020) defend the idea of imputing the outcome variable *and* including the imputed records in subsequent statistical procedures. They claim that most common reservations against imputing outcomes are likely unfounded if there are no severe flaws in the study's design, such as ignoring evidence against MCAR. Overall, there is no conclusive evidence in favor of, or against, any of the available methodologies over the others. In this thesis, Harrell's approach of *not* imputing outcome variables was chosen for being the most conservative approach.

Ultimately, the final question to address is how many imputations to choose. Let m be the number of imputed data set and let γ denote the fraction of missing information in the data. According to Rubin (1987), the efficiency of an imputation estimate then approximates to $(1 + \gamma \times m^{-1})^{-1}$. This means that the estimation efficiency increases with a rising number of imputations, m, whereas it shrinks proportionally with an increasing fraction of missing information, γ. For best imputation results, it is therefore recommended to choose an imputation count of approximately equal to γ (Harrell 2016). As an example, if 20 % of the records contain missing data that need to be imputed, about 20 imputations ought to be used. Some publications suggest using an even higher number of imputations, especially when dealing with small effect sizes (Graham et al. 2007).

2.3 Study Design

2.3.1 Inclusion and Exclusion Criteria

In total, 445 patients were treated with POEM at our facility between June 30, 2010, and December 31, 2017. Out of these, 374 patients (84.04 %) were included in this thesis based on the following criteria:

1. Treated with POEM at the Department for Interdisciplinary Endoscopy, University Medical Center Hamburg-Eppendorf, Germany, between June 30, 2010, and December 31, 2017.
2. Diagnosed with achalasia type I, II, or III.
3. No achalasia-specific treatments before POEM except for drugs, botulinum toxin injections, and pneumatic balloon dilatations.
4. Eckardt score above 3 before POEM.
5. No non-achalasia related operations in the gastric or esophageal anatomical area before POEM.
6. No non-achalasia related pathological conditions in the gastric or esophageal anatomical area before POEM.

Criteria 1 to 3 are the main *inclusion* criteria. Criteria 4 to 6 are *exclusion* criteria. They aim to minimize biases in the statistical results and to maximize the explanatory power of the statistical models by eliminating patients with atypical pre-conditions suspected to have a great impact on the treatment outcome.

Since the threshold for treatment failure was defined as an Eckardt score above 3, treatment success cannot be measured for patients who already reported an Eckardt score below 4 before POEM. Such patients already fulfilled the criteria for a successful treatment even before the actual intervention. Therefore, any positive treatment response observed in them after POEM would not be attributable *to* POEM.

They were therefore excluded to avoid positive selection bias. Likewise, clinically relevant pre-conditions and pre-operations in the anatomical areas neighboring the esophagogastric structures are expected to have a significant influence on the clinical symptoms reported by patients, and consequently also on the treatment outcome. Such patients were removed to avoid negative selection bias.

Ultimately, two more patients with unclear achalasia types were also removed, as well as one teaching patient of whom no data was available in the records at all, except for their sex and age.

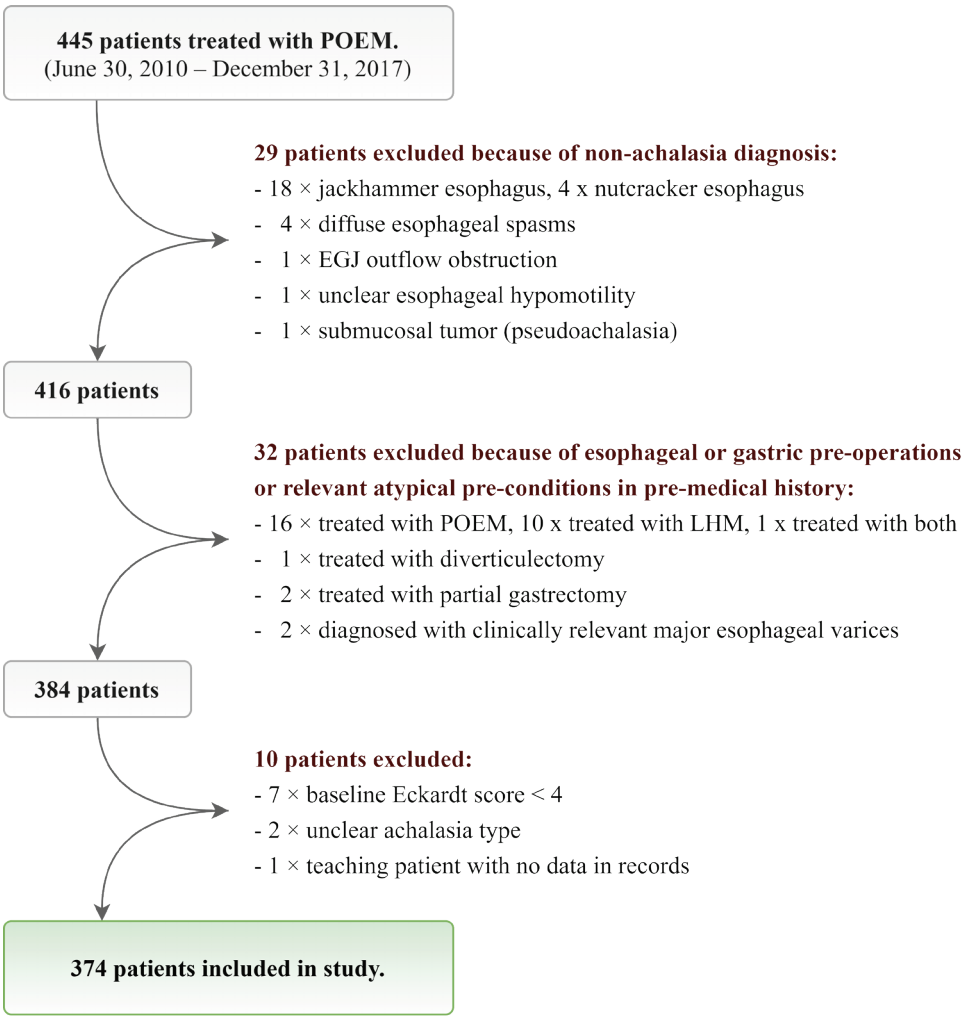


Figure 20: Patient Selection.

2.3.2 Clinical Study and Systematic Follow-Up

The approval of an ethics committee was not compulsory for the conduction of this study. Upon inquiry, this was confirmed by the ethics committee of the Medical Association of Hamburg, Germany, under reference number *WF-031/19*. A standardized questionnaire was used to ask the patients about their

symptoms according to the Eckardt score, their size, weight, reflux, and current anti-reflux medication. This questionnaire was used before POEM and during the systematic follow-up (FU) 3, 6, 12, 24, 36, and 60 months after POEM. To allow for some leeway in the patients' response behavior, a questionnaire was considered valid for a given follow-up time t if it reached us within $t \pm 3$ months.

Prior to POEM, patients underwent high-resolution manometry to secure their achalasia type diagnosis and to acquire a recent IRP. They also received upper endoscopy to objectify their reflux symptoms and to get a histopathological evaluation of reflux lesions, if any were found. Manometry and upper endoscopy were routinely repeated after 3, 24, and 60 months. Many patients underwent these follow-up diagnostics at a local hospital of their choice and willingly informed us about their results.

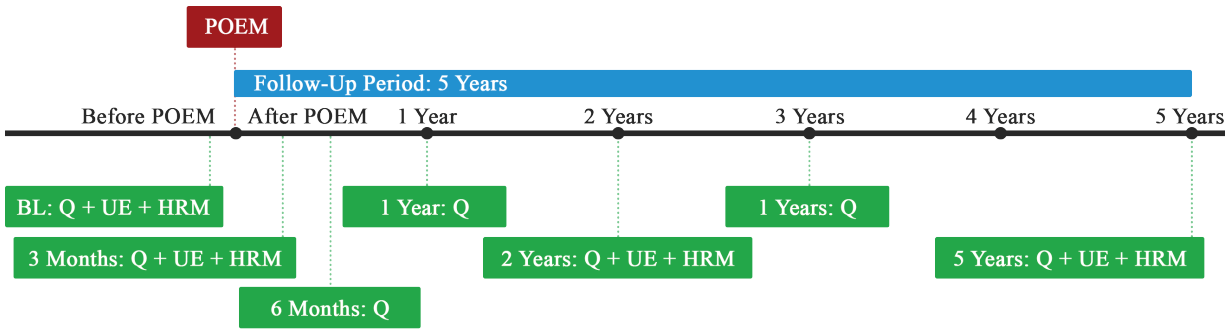


Figure 21: Clinical Study Design and Follow-Up Structure.
BL: baseline. Q: questionnaire. UE: upper endoscopy. HRM: high-resolution manometry.

2.3.3 Statistical Methodology

The data of all patients matching the inclusion criteria described in chapter 2.3.1 were statistically analyzed, as will be described in this section. A significance level of $\alpha = 0.05$ was chosen for all statistical models. Consequently, all confidence intervals (CI) are 95 % confidence intervals.

1) Clinical baseline data retrieved from the patients before POEM, together with some descriptive population data, were aggregated into table 6 in chapter 3.1. Follow-up data for 3, 6, 12, 24, 36, and 60 months after POEM were aggregated into the tables reported and discussed in chapter 3.2. Continuous variables were analyzed with *two-sample t tests*, factorial variables by *exact Fisher tests*.

   The distributions of important parameters later used as predictors in the statistical models were analyzed using the *Shapiro-Wilk test* provided by *R*. Skew and kurtosis were calculated using the *R* package *moments*. Missing data patterns were visualized and discussed using the *R* package *finalfit*. Chapter 3.3 provides the distribution analysis. Chapter 3.4 reports on missing data.

2) Multiply imputed multivariate logistic regression models were fit based on the two-, three- and five-year follow-up data available. They are reported in chapter 3.5. For each of these data sets, *Little's test*, provided by the *R* package *BaylerEdPsych*, was utilized to ensure MCAR. For this cause, factorial variables were remodeled as discrete integers. If necessary, *Little's test* was supplemented by *Hawkins' test* and *Jamshidian's test*, both provided by the *R* package *MissMech*.

   Assuming MCAR, patients with missing treatment outcome observations were removed and the data of the two-, three- and five-year follow-ups were each multiply imputed using predictive mean matching provided by the *R* package *mice*. Logistic regression provided by *R* was performed on the imputed datasets. The resulting statistical measures were pooled into final effect size estimates using *Rubin's rules*, provided once again by the R package *mice*. To assess the validity, quality, and predictive power gained by multiple imputation, the imputations were visualized, and supplementary complete case analyses were performed for each of the three data sets. Regression model validation was performed wherever possible by using common statistical measures provided by the *R* packages *rsm* and *generalhoslem*.

3) A multiply imputed multivariate Cox proportional hazards regression model was fit. It is reported in chapter 3.6. MCAR was assessed the same way as previously described for logistic regression. The validity and predictive power of the calculated regression model was assessed. The proportional hazards assumption was confirmed with the *Schoenfeld test* provided by the *R* package *survival*.

For all regression models, the primary outcome was *treatment failure*. It was defined as a patient reporting an Eckardt score above 3 or having undergone another treatment after POEM, except for drugs.

2.3.4 Reporting Conventions

In tables and graphs, effect sizes are rounded to three significant digits to avoid the implication of an unjustifiable precision. Otherwise, they are rounded to the number of significant digits that is deemed most adequate for their individual measure. The definition operator is denoted by "≔". Significance levels are reported accompanied by threshold markers as depicted in table 5.

Table 5: Reporting Conventions for *p* Values.

Marker	Condition	Significant ($p \leq \alpha$)
	$p > 0.1$	No
+	$p \leq 0.1$	No
*	$\mathbf{p \leq 0.05}$	**Yes**
**	$\mathbf{p \leq 0.01}$	**Yes**
***	$\mathbf{p \leq 0.001}$	**Yes**

3 Results

3.1 Population

3.1.1 Base Data

Table 6: Structural Base Data, Perioperative Data, and Clinical Baseline Parameters.

Variable	Treatment-Naïve	Pre-Treated	*p* [a]
Number of patients	191	183	
Follow-up (months), mean ± SD (range)	36.7 ± 17.7 (0 – 60)	36.9 ± 20.4 (0 – 60)	0.930
Age (years), mean ± SD (range)	44.4 ± 16.1 (12 – 83)	47.8 ± 15.9 (16 – 87)	**0.040** *
< 40, % (*n*)	38.2 % (73)	32.8 % (60)	
40 – 64, % (*n*)	48.7 % (93)	50.8 % (93)	
≥ 65, % (*n*)	13.1 % (25)	16.4 % (30)	
Male : female (% male)	96 : 95 (50.3 %)	107 : 76 (58.5 %)	0.120
BMI (kg/m²), mean ± SD (range), *n*	24.6 ± 5.2 (15.6 – 43.8), 168	26.0 ± 5.1 (15.6 – 57.9), 162	**0.013** *
Pre-treatment, % (*n*)			NA
Only botulinum toxin injection(s)	-	16.4 % (30)	
Only balloon dilatation(s)	-	73.2 % (134)	
Both	-	10.4 % (19)	
Diagnosis, % (*n*)			**0.028** *
Achalasia type I	21.5 % (41)	30.1 % (55)	
Achalasia type II	67.0 % (128)	53.6 % (98)	
Achalasia type III	11.5 % (22)	16.4 % (30)	
Days to discharge, mean ± SD (range) [b]	3.5 ± 2.2 (1 – 21)	3.5 ± 1.4 (2 – 13)	0.884
Immediate re-hospitalization, % (*n*) [c]	1.6 % (3)	2.2 % (4)	0.719
Eckardt Score, Mean ± SD (Range), *n*			
Score	6.9 ± 2.0 (4 – 12), 177	6.4 ± 1.9 (4 – 12), 169	**0.034** *
4 – 6, % (*n*)	46.9 % (83)	59.8 % (101)	
7 – 9, % (*n*)	40.7 % (72)	32.0 % (54)	
10 – 12, % (*n*)	12.4 % (22)	8.3 % (14)	
Dysphagia	2.7 ± 0.6 (0 – 3), 163	2.6 ± 0.7 (0 – 3), 151	0.161
Regurgitations	1.6 ± 0.8 (0 – 3), 163	1.5 ± 0.9 (0 – 3), 151	0.128
Retrosternal pain	1.2 ± 0.9 (0 – 3), 163	1.2 ± 0.9 (0 – 3), 151	0.911
Weight loss	1.3 ± 1.1 (0 – 3), 163	0.9 ± 1.1 (0 – 3), 151	**0.006** **
IRP (mmHg)			
Mean ± SD (range), *n*	30.0 ± 12.1 (0.2 – 68.8), 160	21.8 ± 11.8 (0.8 – 62.5), 135	**< 0.001** ***
0 – 5, % (*n*)	1.3 % (2)	4.4 % (6)	
6 – 10, % (*n*)	1.3 % (2)	7.4 % (10)	
11 – 15, % (*n*)	5.0 % (8)	18.5 % (25)	
16 – 20, % (*n*)	14.4 % (23)	19.3 % (26)	
21 – 25, % (*n*)	15.6 % (25)	17.0 % (23)	
26 – 30, % (*n*)	16.9 % (27)	13.3 % (18)	
> 30, % (*n*)	45.6 % (73)	20.0 % (27)	
Reflux esophagitis, % (*n* / total) [d]	2.6 % (5 / 189)	5.1 % (9 / 177)	0.280
Los Angeles grade A, % (*n*)	100.0 % (5)	88.9 % (8)	
Los Angeles grade B, % (*n*)	-	11.1 % (1)	
Los Angeles grades C and D, % (*n*)	-	-	

*[a] Fisher's exact test for factorial variables. Two-sample t test for continuous variables. [b] Days after POEM until discharge from hospital. [c] Re-hospitalization necessary shortly after discharge as a consequence of POEM. [d] Endoscopically diagnosed. + p ≤ 0.1, * p ≤ 0.05, ** p ≤ 0.01, *** p ≤ 0.001. NA: not applicable, SD: standard deviation, - none.*

Table 6 details the basic data of the study population divided into *treatment-naïve* and *pre-treated* patients. Continuous variables are compared by *two-sample t tests*, factorial variables by *exact Fisher tests*.

3.1.2 Between-Group Differences

As shown in table 6, treatment-naïve and pre-treated patients are structurally similar in regard to number, sex, mean follow-up duration, days to discharge after POEM, re-hospitalization rate after their initial discharge, and endoscopically diagnosed baseline reflux disease prevalence.

The groups differ significantly in some categories. Previously treated patients are on average about 3 years older than treatment-naïve patients (47.8 vs. 44.4 years, $p = 0.040$). They rate on average about 0.5 points lower in the Eckardt score (6.4 vs. 6.9, $p = 0.034$), which is primarily attributable to a 0.4 score points lower mean weight loss (0.9 vs. 1.3, $p = 0.006$). Additionally, they show a 1.4 kg/m² higher mean BMI (26.0 vs. 24.6, $p = 0.013$) and a significantly lower baseline mean IRP of about 8.2 mmHg below that of treatment-naïve patients (21.8 vs. 30.0 mmHg, $p < 0.001$). Upon comparing the IRP distributions of both groups, the IRP seems to be distributed quite evenly among the range of 0 – 30 mmHg in the pre-treated group. In contrast, the observations accumulate beyond 30 mmHg in the treatment-naïve group.

There is a significant difference in the achalasia type distribution between both groups ($p = 0.028$). Pre-treated patients show a higher percentage of type I achalasia (30.1 % vs. 21.5 %), a lower percentage of type II (53.6 % vs. 67.0 %), and a slightly higher percentage of type III (16.4 % vs. 11.5 %). However, the rank order of the types is homogeneous between both groups: type II achalasia is by far the most common one, followed by type I. Type III is the rarest of the three.

3.1.3 Previous Treatments

Table 7 breaks down the prior treatments of the patients in the pre-treatment group. Among all patients, 51.1 % did not receive any treatments before POEM except for medication, 35.8 % underwent balloon dilatations, 8.0 % underwent botulinum toxin injections, and 5.1 % were previously treated with both. Thus, most pre-treated patients received only balloon dilatations.

Table 7: Previous Treatments Prior to POEM.

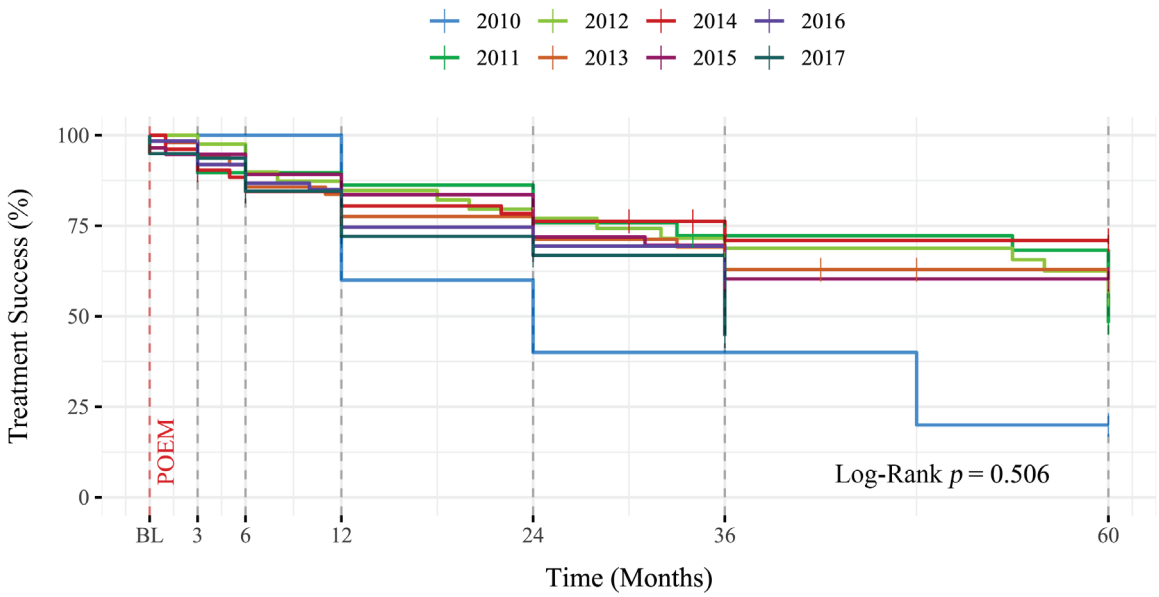
Group	Patients, % total (*n*)	Patients, % among group (*n*)
All patients	100 % (374)	-
Treatment-naïve	51.1 % (191)	-
Only balloon dilatation(s)	35.8 % (134)	-
⌀ < 30 mm [a]	-	9.0 % (12)
⌀ ≥ 30 mm [b]	-	63.4 % (85)
⌀ unknown [c]	-	27.6 % (37)
1 ×	-	32.1 % (43)
2 ×	-	27.6 % (37)
3 ×	-	20.1 % (27)
4 – 6 ×	-	14.2 % (19)
> 6 ×	-	6.0 % (8)
Only botulinum toxin injection(s)	8.0 % (30)	-
1 ×	-	40 % (12)
2 – 3 ×	-	40 % (12)
4 – 6 ×	-	13.3 % (4)
> 6 ×	-	6.7 % (2)
Both	5.1 % (19)	-
⌀ < 30 mm [a]	-	10.5 % (2)
⌀ ≥ 30 mm [b]	-	57.9 % (11)
⌀ unknown [c]	-	31.6 % (6)
1 balloon dilatation	-	26.3 % (5)
2 – 3 balloon dilatations	-	42.1 % (8)
4 – 6 balloon dilatations	-	26.3 % (5)
> 6 balloon dilatations	-	5.3 % (1)
1 botulinum toxin injection	-	68.4 % (13)
2 – 3 botulinum toxin injections	-	26.3 % (5)
4 – 6 botulinum toxin injections	-	5.3 % (1)

[a] All diameters known and all below 30 mm. [b] At least one known diameter of 30 mm or above. [c] At least one unknown diameter and no known diameter of 30 mm or above. ⌀: balloon diameter. - not applicable because of different scope.

As explained in chapter 1.8.1, balloon dilatation is recommended to be performed with a balloon of at least 30 mm diameter to warrant a sufficient treatment. This was done in 63.4 % of patients who were previously only treated with balloon dilatations, and in 57.9 % of patients who had undergone both balloon dilatations and botulinum toxin injections of in the past. Therefore, most patients with a past medical history of at least one balloon dilatation had been sufficiently pre-treated before POEM. On a side note, these patients had mostly undergone one to three dilatations. Among the small group of patients that had only received botulinum toxin injections before POEM, most had participated in one to three pre-treatment sessions.

3.1.4 Endoscopy Proficiency

All POEMs were performed or supervised by a highly skilled endoscopist with many years of experience in a variety of different endoscopic procedures. Figure 22 illustrates the patients' failure-free survival by year. The first cohort appears to have responded worse to treatment compared to the cohorts of all subsequent years. Also, the distinct three-year outcomes of the 2017 cohort seem to be exceptionally bad, being on par with the outcomes usually seen in the other cohorts after five years. However, when comparing the curves using the *log-rank test*, their overall differences are insignificant ($p = 0.506$).



Number at Risk (Number Censored):

5 (0)	5 (0)	5 (0)	3 (0)	3 (0)	2 (0)	2 (0)	2 (0)	2 (0)	1 (0)	1 (1)
29 (0)	26 (0)	26 (0)	25 (0)	25 (1)	21 (1)	19 (3)	18 (3)	18 (3)	18 (3)	17 (15)
41 (0)	38 (2)	34 (2)	33 (2)	31 (4)	27 (4)	26 (7)	22 (7)	22 (7)	22 (7)	20 (25)
49 (0)	45 (0)	41 (1)	37 (1)	37 (1)	34 (1)	33 (5)	26 (7)	24 (8)	23 (8)	23 (30)
52 (0)	45 (2)	42 (4)	38 (4)	37 (7)	33 (10)	29 (17)	21 (17)	21 (17)	21 (17)	21 (38)
57 (0)	51 (3)	48 (5)	43 (5)	43 (11)	31 (11)	30 (32)	5 (32)	5 (32)	5 (32)	5 (37)
62 (0)	53 (4)	49 (4)	43 (4)	43 (12)	32 (12)	32 (41)	0 (41)	0 (41)	0 (41)	0 (41)
79 (0)	62 (14)	54 (19)	41 (19)	41 (54)	3 (54)	3 (56)	0 (56)	0 (56)	0 (56)	0 (56)

Figure 22: Failure-Free Survival After POEM by Treatment Year. BL: baseline.

3.2 Outcome

In this chapter, the treatment effects observed throughout the systematic follow-up is reported. Some patients were treated with POEM less than three or five years ago. To account for this, the three- and five-year analyses are limited to the subsets of patients treated with POEM until December 31, 2016 (n = 295), and December 31, 2014 (n = 176), respectively.

3.2.1 Treatment Success

Table 8 and figure 23 depict the treatment success up to five years after POEM. As always, treatment success for each follow-up time was defined as reporting an Eckardt score below 4 at said time and not having undergone any re-treatment since POEM, except for drugs.

Table 8: Treatment Success After POEM.

Time	Treatment-Naïve % (*n* success / *n* known), *n* total (n unknown)		Pre-Treated % (*n* success / *n* known), *n* total (*n* unknown)		*p* [a]
3 months	97.0 % (159 / 164),	191 (27)	92.4 % (145 / 157),	183 (26)	0.082 +
6 months	91.2 % (125 / 137),	191 (54)	83.3 % (120 / 144),	183 (39)	0.051 +
1 year	84.8 % (128 / 151),	191 (40)	75.0 % (111 / 148),	183 (35)	**0.043 ***
2 years	81.6 % (129 / 158),	191 (33) [b]	74.3 % (113 / 152),	183 (31)	0.132
3 years [c]	76.0 % (92 / 121),	150 (29) [b]	64.0 % (73 / 114),	145 (31) [b]	**0.047 ***
5 years [d]	68.3 % (41 / 60),	83 (23)	51.4 % (37 / 72),	93 (21) [b]	0.053 +

[a] Fisher's exact test. [b] One patient censored by not achalasia-related death. [c] Based on the patients treated with POEM until December 31, 2016 (n = 295). [d] Based on the patients treated with POEM until December 31, 2014 (n = 176).
*+ p ≤ 0.1, * p ≤ 0.05, ** p ≤ 0.01, *** p ≤ 0.001.*

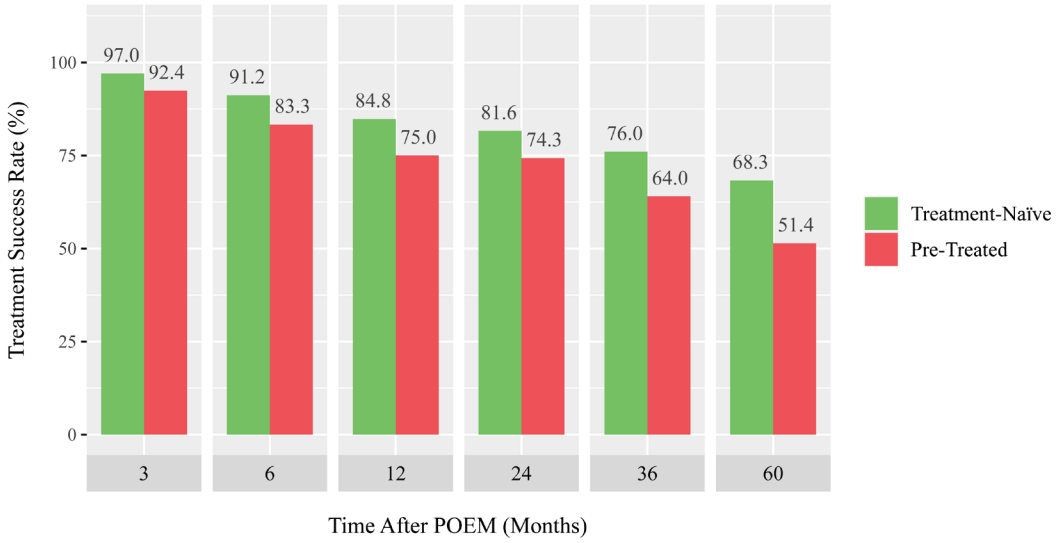


Figure 23: Treatment Success After POEM by Follow-Up. Treatment success was defined as an Eckardt score below 4 and no re-treatment since POEM, except for drugs.

When comparing treatment-naïve to previously treated patients by *exact Fisher tests*, treatment-naïve patients show significantly better treatment success rates after one year (84.8 % vs. 75.0 %, $p = 0.043$) and after three years (76.0 % vs. 64.0 %, $p = 0.047$). They also show better, yet by a narrow margin insignificant, treatment success rates after three months (97.0 % vs. 92.4 %, $p = 0.082$), six months (91.2 % vs. 83.3 %, $p = 0.051$), and five years (68.3 % vs. 51.4 %, $p = 0.053$). After two years, the overall treatment success rate is still high in both groups: more than 70 % of patients report favorable Eckardt scores with no need for another treatment. However, these promising numbers drop considerably after five years, where 31.7 % of treatment-naïve and 48.6 % of pre-treated patients experience failure.

Figures 24 to 29 illustrate the failure-free survival of the patients overall, as well as divided into different subgroups based on previous treatments, sex, achalasia type, and baseline IRP. The groups are compared by *log-rank tests*. These graphs are shown to grant an overview of the patients' follow-up development. In-depth analyses are not provided in favor of the more sophisticated multivariate regression models that will be reported in chapters 3.5 and 3.6.

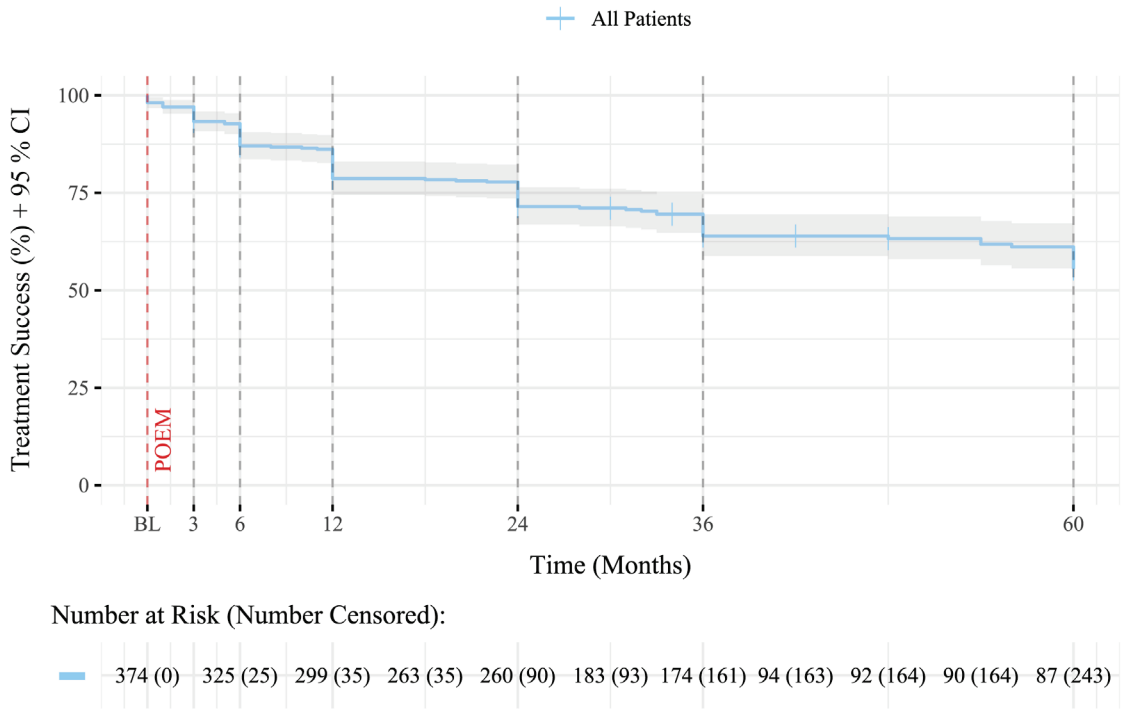


Figure 24: Failure-Free Survival After POEM. BL: baseline, CI: confidence interval.

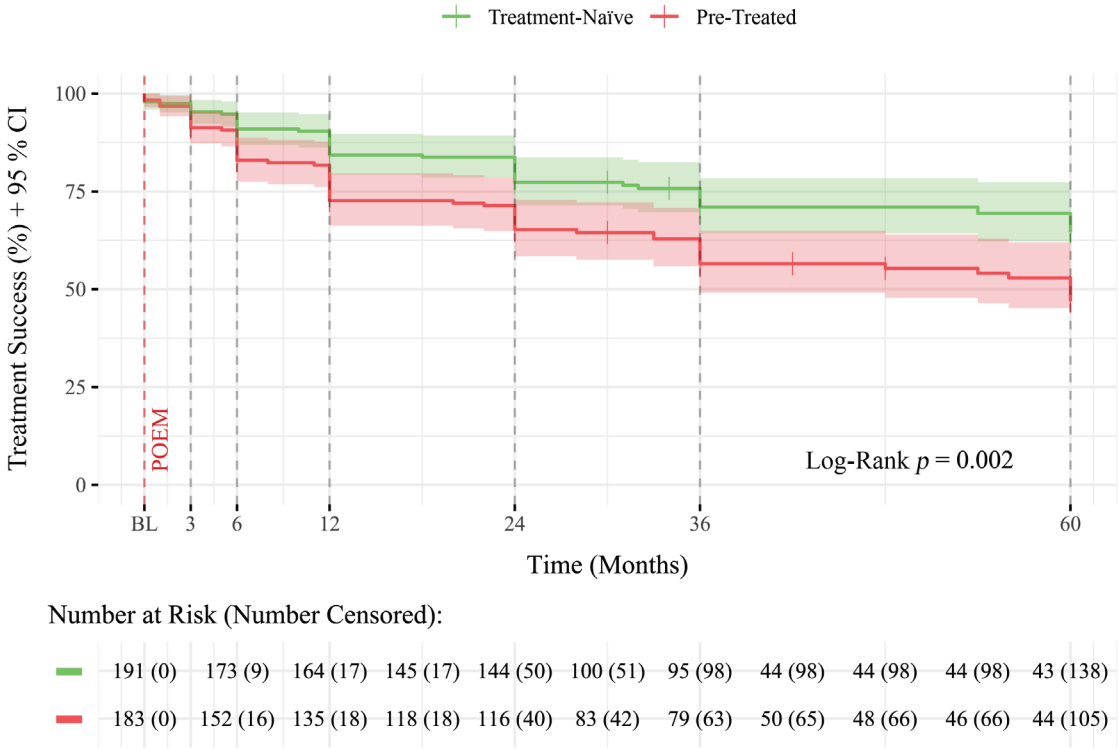


Figure 25: Failure-Free Survival After POEM by Pre-Treatment Group. BL: baseline, CI: confidence interval.

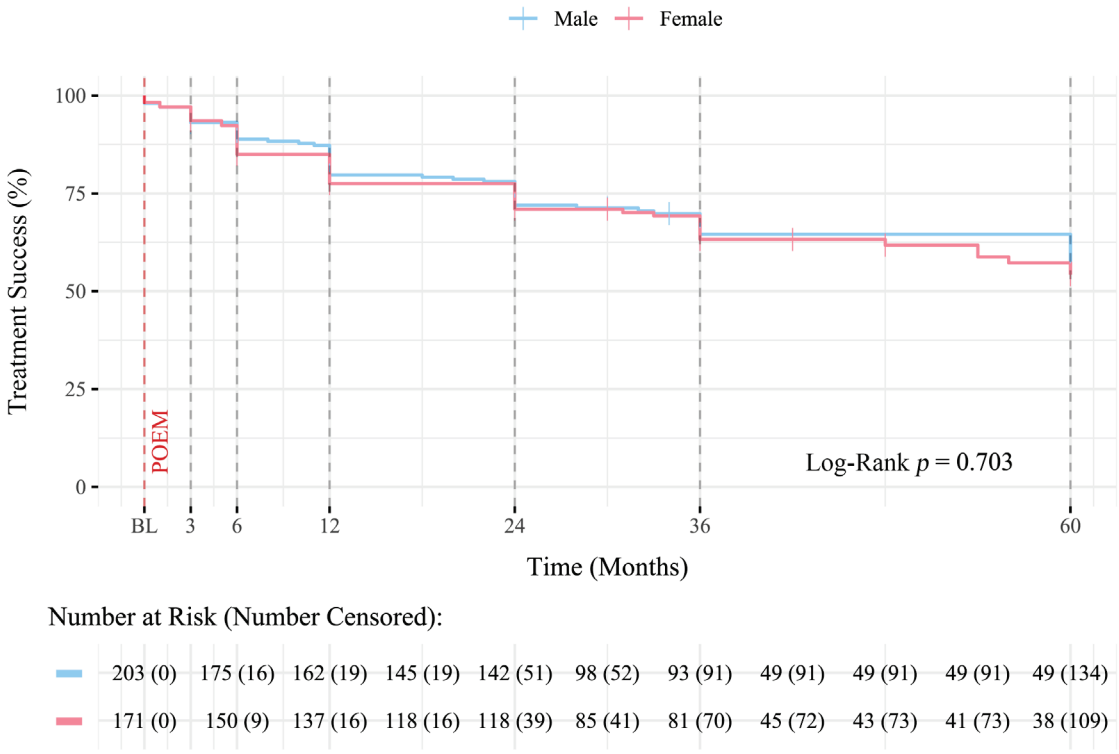


Figure 26: Failure-Free Survival After POEM by Sex. BL: baseline.

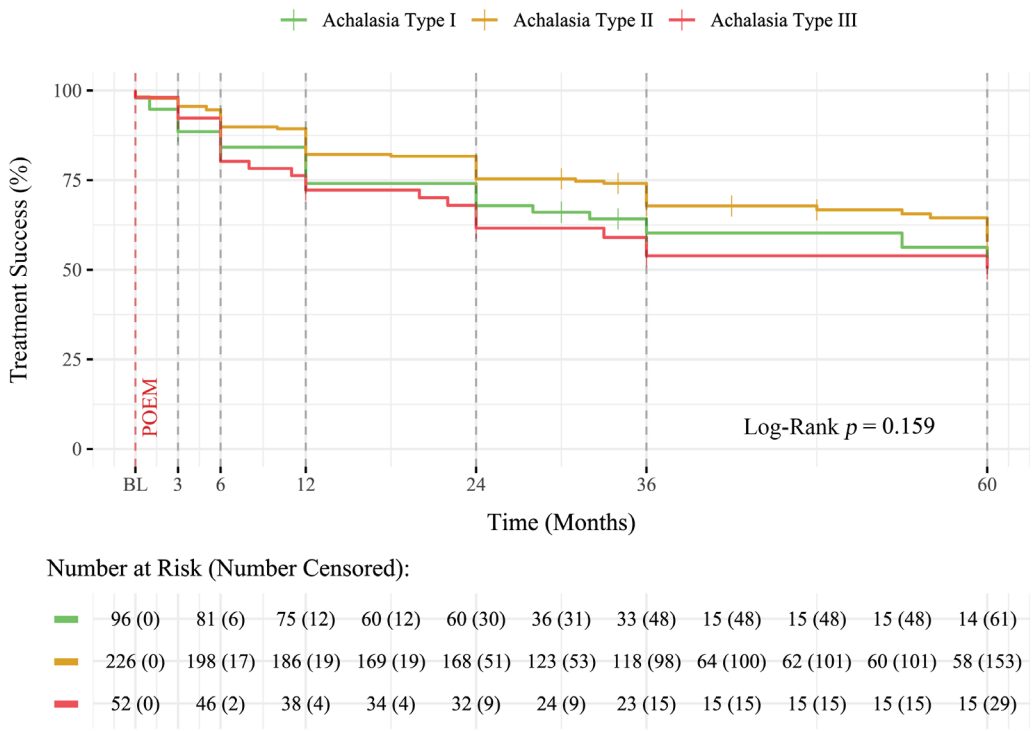


Figure 27: Failure-Free Survival After POEM by Achalasia Type. BL: baseline.

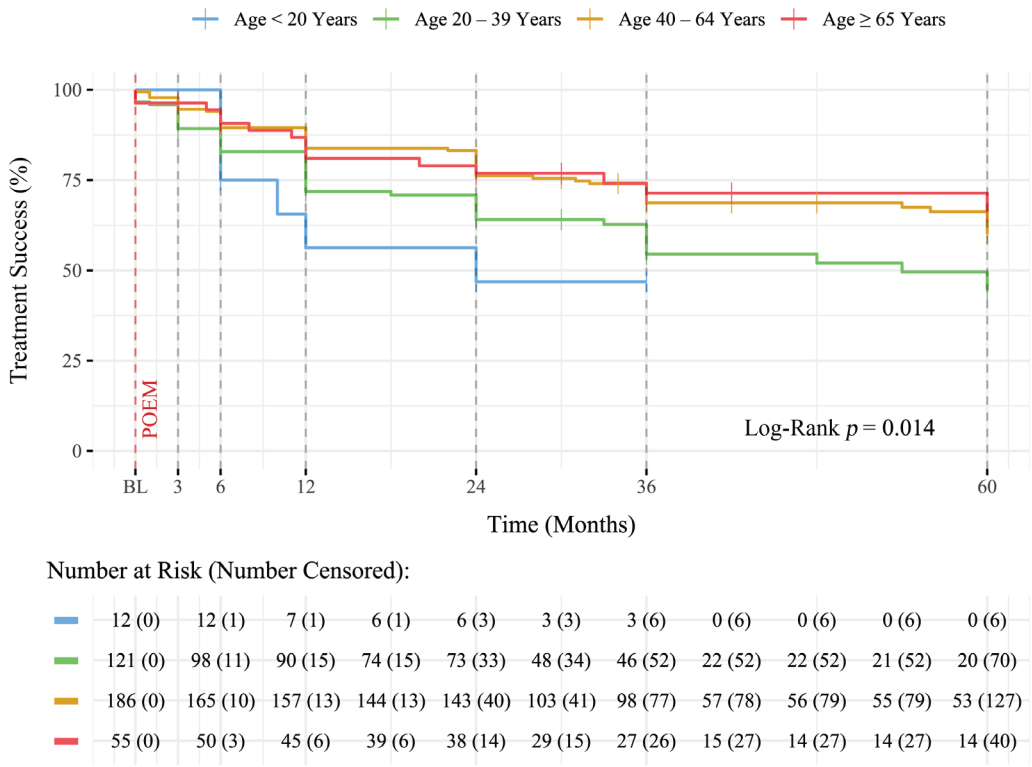


Figure 28: Failure-Free Survival After POEM by Age. BL: baseline.

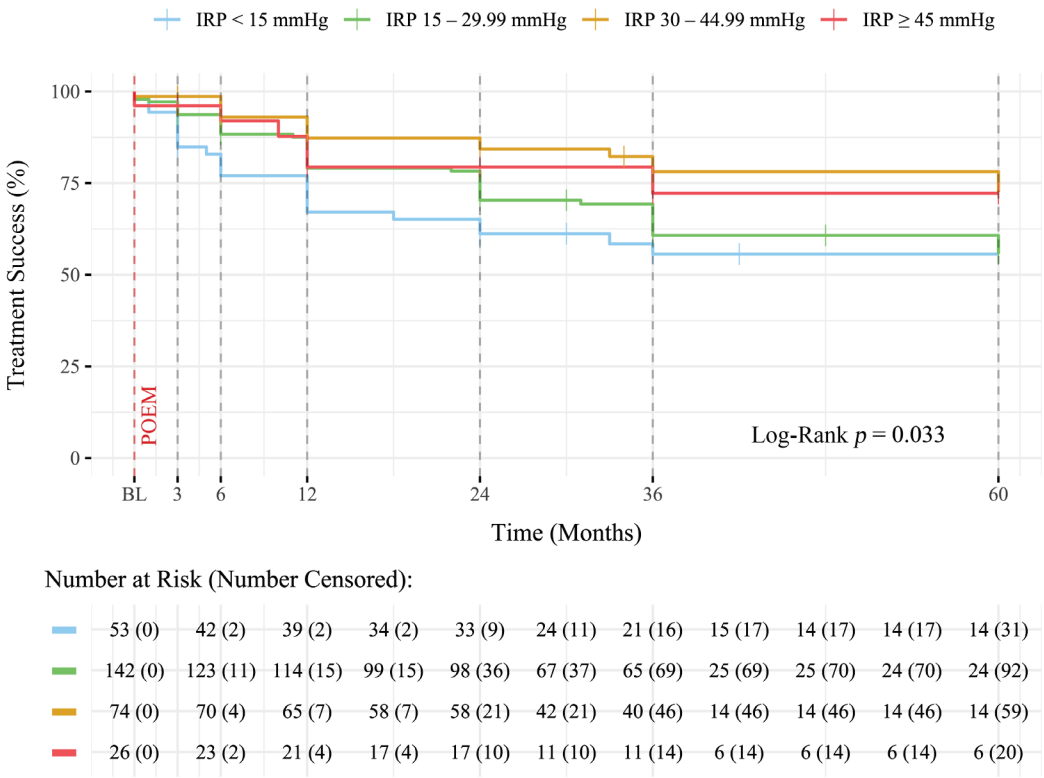


Figure 29: Failure-Free Survival After POEM by IRP. BL: baseline.

Slight differences between the survival curves and the numbers provided in table 8 can be explained by the fact that, as explained at the beginning of this chapter, some patients were excluded from the three- and five-year follow-ups in the table due to their recent POEM dates. In contrast, the survival curves have been drawn using *all* available data. Patients that were treated just after the cut-off dates of December 31, 2014, and December 31, 2016, may have already provided valid follow-up data that was excluded from the table, yet incorporated in the drawing of the survival curves.

Overall, almost 50 % of the patients experience treatment-failure after five years. Univariate comparisons using the *log-rank test* suggest that previously treated patients, compared to treatment-naïve patients, respond significantly worse throughout the entire follow-up ($p = 0.002$). Type II achalasia shows a better response than type I, and both fare better than type III. However, these differences fail to reach significance ($p = 0.159$). When dividing the patients into groups based on their baseline IRPs, as shown in figure 29, lower IRPs are associated with a significantly worse clinical response ($p = 0.033$). On a similar note, the younger a patient is, the worse their treatment response appears to be (see figure 28). Patients below 20 years of age experience particularly early treatment failures ($p = 0.014$). No sex-specific differences were found. For these models are univariate, their predictive power is limited. However, they hint at previous treatments and the IRP being candidates for significant predictors.

3.2.2 Eckardt Score Development

This chapter describes the development of the Eckardt score as it was reported by the patients during follow-up, as well as the contribution of its four sub-components: dysphagia, regurgitations, retrosternal pain, and weight loss. Table 9 and figures 30 to 35 summarize the acquired data. The score change Δ is the difference between the score at the respective follow-up time *t* and the pre-POEM baseline score: *Δ Eckardt Score: f(t) = Eckardt Score(t) – Eckardt Score(0)*.

Table 9: Eckardt Score Development After POEM.

Time	Eckardt Score	Treatment-Naïve: mean ± SD (range), *n*	Pre-Treated: mean ± SD (range), *n*	*p* [a]
3 months	Score	1.1 ± 1.3 (0 – 9), 164	1.3 ± 1.3 (0 – 6), 154	0.211
	4 – 6, % (*n*)	1.8 % (3)	5.8 % (9)	
	7 – 9, % (*n*)	1.2 % (2)	-	
	10 – 12, % (*n*)	-	-	
	Δ Eckardt Score	-5.8 ± 2.2 (-1 – -11), 154	-5.0 ± 2.2 (-11 – 0), 148	**0.005** **
	Dysphagia	0.6 ± 0.6 (0 – 3)	0.6 ± 0.8 (0 – 3)	0.246
	Regurgitations	0.2 ± 0.4 (0 – 3)	0.2 ± 0.4 (0 – 2)	0.739
	Retrosternal pain	0.4 ± 0.6 (0 – 3)	0.4 ± 0.5 (0 – 3)	0.371
	Weight loss	0.0 ± 0.3 (0 – 3)	0.0 ± 0.2 (0 – 1)	0.705
6 months	Score	1.4 ± 1.5 (0 – 8), 136	1.9 ± 2.0 (0 – 9), 142	**0.007** **
	4 – 6, % (*n*)	7.4 % (10)	9.2 % (13)	
	7 – 9, % (*n*)	0.7 % (1)	5.6 % (8)	
	10 – 12, % (*n*)	-	-	
	Δ Eckardt Score	-5.5 ± 2.3 (-11 – 0), 127	-4.4 ± 2.6 (-10 – -4), 133	**< 0.001** ***
	Dysphagia	0.6 ± 0.8 (0 – 3)	0.8 ± 0.9 (0 – 3)	0.153
	Regurgitations	0.2 ± 0.5 (0 – 3)	0.4 ± 0.7 (0 – 3)	**0.002** **
	Retrosternal pain	0.4 ± 0.6 (0 – 3)	0.6 ± 0.6 (0 – 3)	0.129
	Weight loss	0.1 ± 0.3 (0 – 1)	0.2 ± 0.4 (0 – 2)	0.056 +
1 year	Score	1.8 ± 1.5 (0 – 8), 150	2.1 ± 1.8 (0 – 7), 145	0.111
	4 – 6, % (*n*)	12.0 % (18)	18.6 % (27)	
	7 – 9, % (*n*)	0.7 % (1)	2.8 % (4)	
	10 – 12, % (*n*)	-	-	
	Δ Eckardt Score	-5.0 ± 2.4 (-11 – 0), 140	-4.2 ± 2.6 (-10 – 2), 136	**0.009** **
	Dysphagia	0.8 ± 0.8 (0 – 3)	0.9 ± 0.9 (0 – 3)	0.095
	Regurgitations	0.3 ± 0.5 (0 – 2)	0.4 ± 0.6 (0 – 2)	0.054 +
	Retrosternal pain	0.6 ± 0.6 (0 – 2)	0.6 ± 0.7 (0 – 3)	0.506
	Weight loss	0.1 ± 0.5 (0 – 3)	0.2 ± 0.5 (0 – 3)	0.726
2 years	Score	2.1 ± 1.5 (0 – 7), 155	2.4 ± 1.9 (0 – 9), 147	0.122
	4 – 6, % (*n*)	12.9 % (20)	16.3 % (24)	
	7 – 9, % (*n*)	0.6 % (1)	4.1 % (6)	
	10 – 12, % (*n*)	-	-	
	Δ Eckardt Score	-4.6 ± 2.3 (-11 – 1), 145	-3.9 ± 2.5 (-10 – 3), 138	**0.010** **
	Dysphagia	1.0 ± 0.8 (0 – 3)	1.0 ± 0.9 (0 – 3)	0.699
	Regurgitations	0.4 ± 0.5 (0 – 2)	0.5 ± 0.7 (0 – 3)	0.054 +
	Retrosternal pain	0.7 ± 0.6 (0 – 2)	0.7 ± 0.6 (0 – 3)	0.584
	Weight loss	0.1 ± 0.3 (0 – 2)	0.2 ± 0.5 (0 – 3)	**0.020** *

Time	Eckardt Score	Treatment-Naïve: mean ± SD (range), *n*	Pre-Treated: mean ± SD (range), *n*	*p* [a]
3 years [b]	Score	2.1 ± 1.6 (0 – 6), 114	2.6 ± 2.1 (0 – 10), 102	0.052 +
	4 – 6, % (*n*)	16.7 % (19)	16.7 % (17)	
	7 – 9, % (*n*)	-	4.9 % (5)	
	10 – 12, % (*n*)	-	1.0 % (1)	
	Δ Eckardt Score	-4.9 ± 2.5 (-11 – 1), 105	-3.9 ± 2.7 (-11 – 5), 94	**0.012 ***
	Dysphagia	0.9 ± 0.8 (0 – 3)	1.1 ± 0.9 (0 – 3)	0.106
	Regurgitations	0.4 ± 0.6 (0 – 3)	0.6 ± 0.8 (0 – 3)	0.075 +
	Retrosternal pain	0.6 ± 0.6 (0 – 2)	0.7 ± 0.6 (0 – 3)	0.589
	Weight loss	0.1 ± 0.4 (0 – 3)	0.2 ± 0.5 (0 – 3)	0.376
5 years [c]	Score	2.0 ± 1.5 (0 – 6), 56	3.0 ± 2.1 (0 – 9), 64	**0.006 ** **
	4 – 6, % (*n*)	17.9 % (10)	23.4 % (15)	
	7 – 9, % (*n*)	-	6.3 % (4)	
	10 – 12, % (*n*)	-	-	
	Δ Eckardt Score	-4.9 ± 2.2 (-10 – 1), 50	-4.0 ± 2.6 (-10 – 1), 57	0.074 +
	Dysphagia	0.9 ± 0.7 (0 – 3)	1.2 ± 0.9 (0 – 3)	0.062 +
	Regurgitations	0.4 ± 0.6 (0 – 2)	0.7 ± 0.8 (0 – 2)	**0.018 ***
	Retrosternal pain	0.5 ± 0.5 (0 – 2)	0.9 ± 0.8 (0 – 3)	**0.002 ** **
	Weight loss	0.2 ± 0.6 (0 – 3)	0.2 ± 0.4 (0 – 1)	0.645

*[a] Two-sample t test. [b] Based on the patients treated with POEM until December 31, 2016 (n = 295). [c] Based on the patients treated with POEM until December 31, 2014 (n = 176). + p ≤ 0.1, * p ≤ 0.05, ** p ≤ 0.01, *** p ≤ 0.001. SD: standard deviation, - none.*

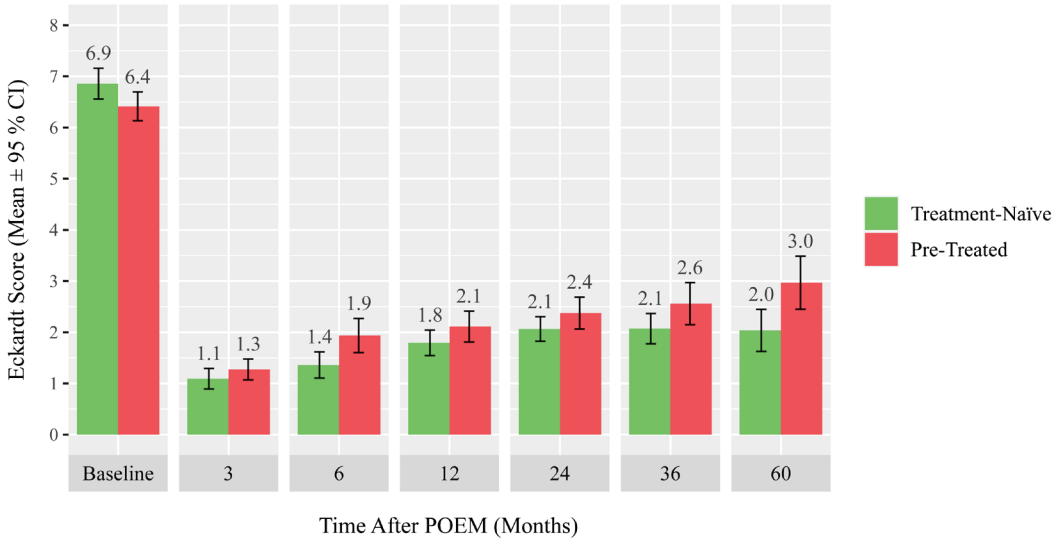


Figure 30: Eckardt Score After POEM. Shown is the Eckardt score as the sum of its four components: dysphagia, regurgitations, retrosternal pain, and weight loss. Each component can take a discrete value of either 0, 1, 2, or 3. Thus, the Eckardt score may range from 0 to 12. CI: confidence interval.

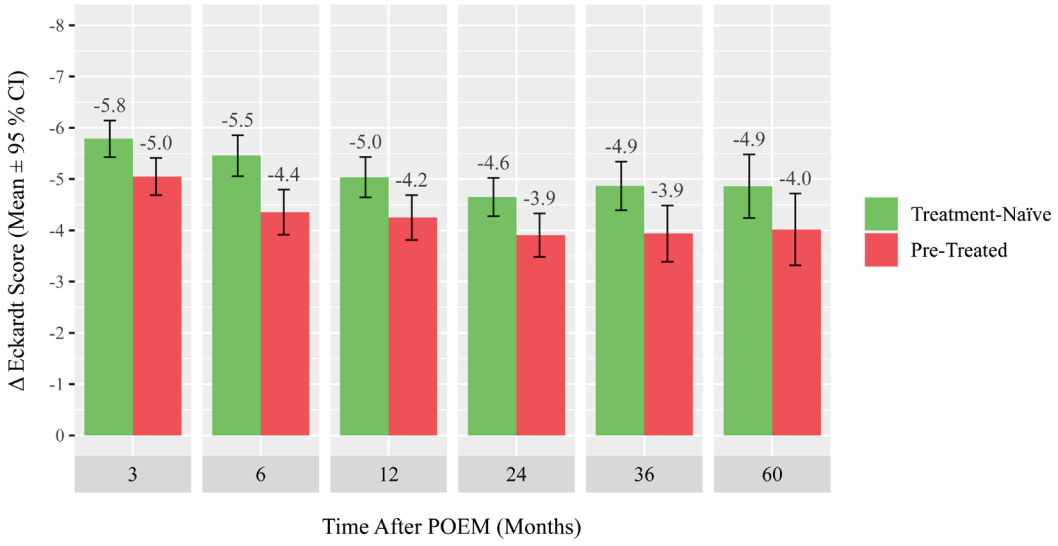


Figure 31: Δ Eckardt Score After POEM. Shown is the change in the Eckardt score. It is defined as the difference between the score at the time *t* after POEM and the baseline score reported before POEM. CI: confidence interval.

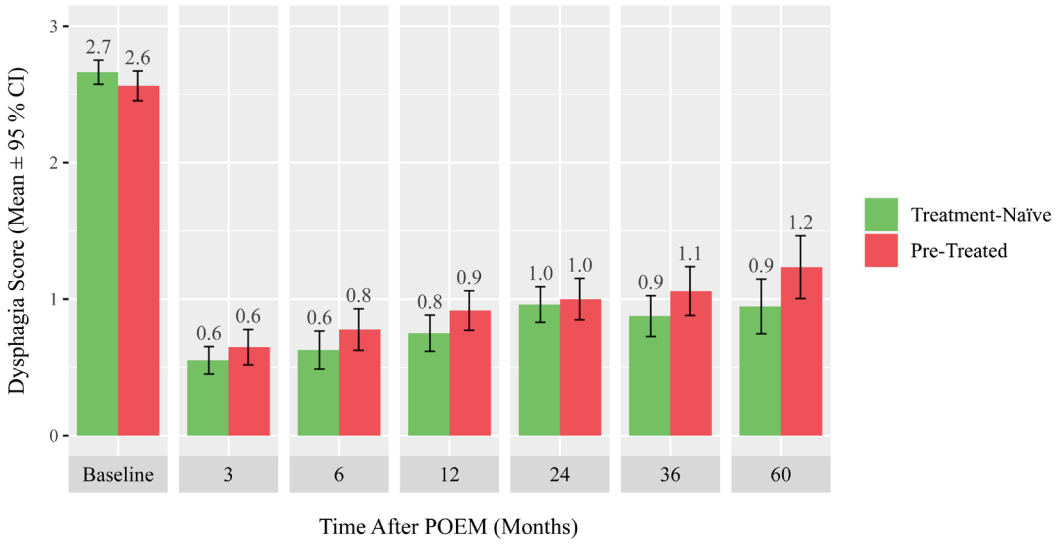


Figure 32: Dysphagia After POEM. Shown is the dysphagia component of the Eckardt score: no dysphagia ≙ 0, occasional dysphagia ≙ 1, daily dysphagia ≙ 2, dysphagia with each meal ≙ 3. CI: confidence interval.

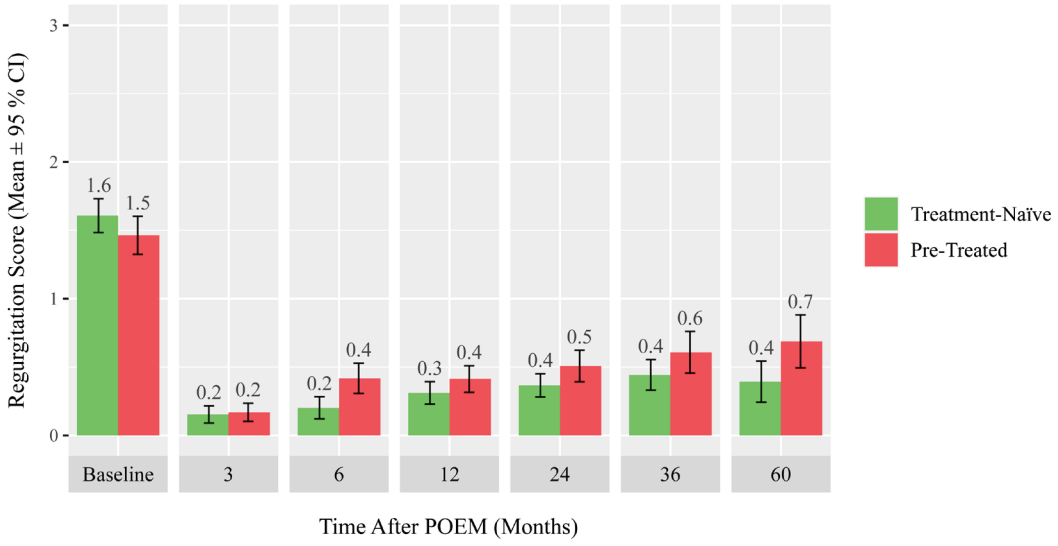


Figure 33: Regurgitations After POEM. Shown is the regurgitation component of the Eckardt score: no regurgitations ≙ 0, occasional regurgitations ≙ 1, daily regurgitations ≙ 2, regurgitations with each meal ≙ 3. CI: confidence interval.

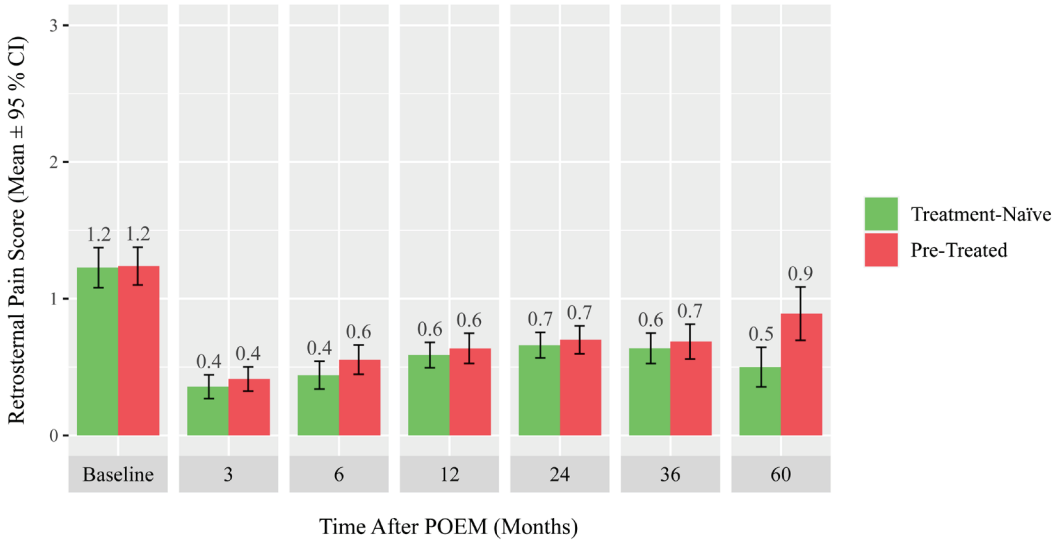


Figure 34: Retrosternal Pain After POEM. Shown is the retrosternal pain component of the Eckardt score: no retrosternal pain ≙ 0, occasional retrosternal pain ≙ 1, daily retrosternal pain ≙ 2, retrosternal pain with each meal ≙ 3. CI: confidence interval.

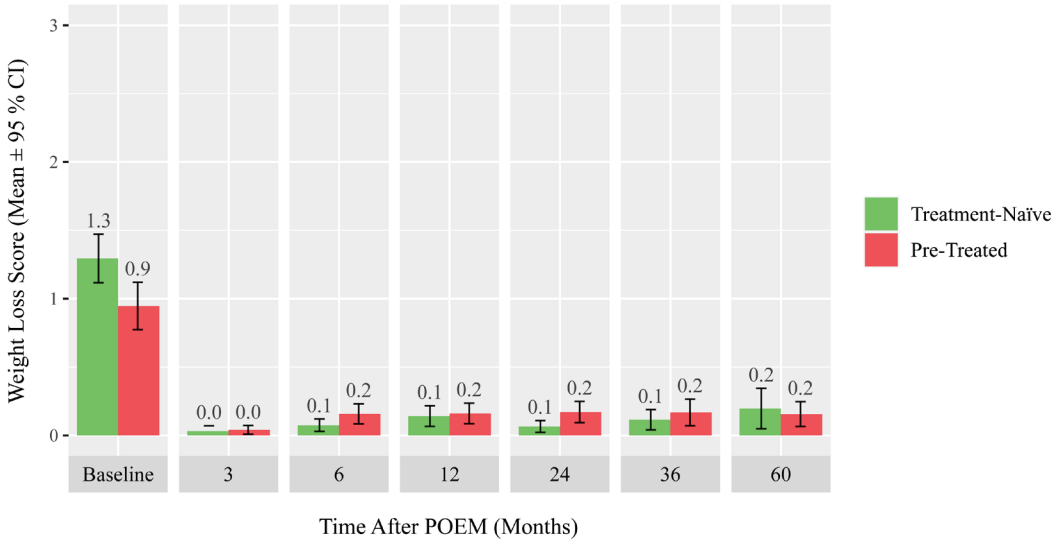


Figure 35: Weight Loss After POEM. Shown is the weight loss component of the Eckardt score: no weight loss ≙ 0, weight loss between 0 and 5 kg ≙ 1, weight loss between 5 and 10 kg ≙ 2, weight loss above 10 kg ≙ 3. CI: confidence interval.

Treatment-naïve patients, when compared to previously treated patients, report lower mean Eckardt scores after six months (1.4 vs. 1.9, $p = 0.007$) and five years (2.0 vs. 3.0, $p = 0.006$), rarer regurgitations after six months (0.2 vs. 0.4, $p = 0.002$), and less weight loss after two years (0.1 vs. 0.2, $p = 0.020$). After five years, they report rarer regurgitations (0.4 vs. 0.8, $p = 0.018$) and less retrosternal pain (0.5 vs. 0.9, $p = 0.002$) than previously treated patients.

Regarding differences that failed to reach significance by a narrow margin, treatment-naïve patients, compared to previously treated patients, show less regurgitations after one year (0.3 vs. 0.4, $p = 0.054$), two years (0.4 vs. 0.5, $p = 0.054$), and three years (0.4 vs. 0.6, $p = 0.075$). Their mean Eckardt score is lower after three years (2.1 vs. 2.4, $p = 0.052$), their mean weight loss score is lower after six months (0.1 vs. 0.2, $p = 0.056$), and they experience less severe dysphagia after five years (0.9 vs. 1.2, $p = 0.062$).

Figure 36 illustrates the differences in the means of the Eckardt score and its components between both groups, as well as the respective distributional differences assessed via *two-sample t tests*.

Figure 36: Between-Group Differences of the Eckardt Score Throughout the Follow-Up. A green circle indicates that the mean score is higher in treatment-naïve than in previously treated patients. A red circle indicates the opposite. A similar mean in both groups is illustrated by a gray circle. An empty circle shows that the score distributions between both groups are similar when compared by a two-sample t test ($p > 0.1$). If the test fails to reach significance only by a narrow margin, the circle is filled with bright green or red ($0.05 < p \leq 0.1$). Ultimately, a circle filled with dark green or red is indicative of significant differences between the distributions of both groups ($p \leq 0.05$).

In general, there are little significant differences between treatment-naïve and previously treated patients regarding the Eckardt score and its components. Interestingly, Δ Eckardt is significantly larger for previously treated than for treatment-naïve patients at every follow-up except after five years, where it slightly misses the threshold for significance ($p = 0.074$). As a side note, because of its negative scale, the larger Δ Eckardt out of two values is the one that is closer to 0. Thus, a larger Δ Eckardt indicates a *smaller* change between the scores of the baseline and the follow-up. Pre-treated patients redevelop more severe regurgitations as early as six months after POEM. This tendency persists for the entirety of all subsequent follow-ups. It ultimately reaches significance after five years. The Eckardt score difference between the groups starts to strive toward significance after three years. It then grows significantly larger in pre-treated patients after five years. Until three years after POEM, dysphagia and retrosternal pain tend to be either insignificantly higher in previously treated patients, or on the same level in both groups. After five years, dysphagia shows a clear trend toward significantly higher values in pre-treated patients, and their retrosternal pain *is* in fact significantly more severe at that time.

In summary, the Eckardt score and most of its separate components tend to be higher in previously treated patients throughout the entire follow-up. This trend reaches significance after three to five years.

3.2.3 Reflux Development

Table 10 and figure 37 show the reflux development after POEM.

Table 10: Reflux After POEM.

Time	Reflux Esophagitis [a]	Treatment-Naïve		Pre-Treated		*p* [b]
3 months	**Reflux, % (*n* / *n* known)**	62.4 %	(106 / 170)	59.9 %	(94 / 157)	0.352
	Los Angeles grade A, % (*n*)	57.6 %	(61)	51.1 %	(48)	
	Los Angeles grade B, % (*n*)	30.2 %	(32)	39.4 %	(37)	
	Los Angeles grade C, % (*n*)	7.6 %	(8)	5.3 %	(5)	
	Los Angeles grade D, % (*n*)	-		1.1 %	(1)	
	Unclassified, % (*n*)	4.7 %	(5)	3.2 %	(3)	
2 years	**Reflux, % (*n* / *n* known)**	52.0 %	(39 / 75)	58.4 %	(45 / 77)	0.564
	Los Angeles grade A, % (*n*)	66.7 %	(26)	66.7 %	(30)	
	Los Angeles grade B, % (*n*)	17.9 %	(7)	28.9 %	(13)	
	Los Angeles grade C, % (*n*)	5.1 %	(2)	4.4 %	(2)	
	Los Angeles grade D, % (*n*)	2.6 %	(1)	-		
	Unclassified, % (*n*)	7.7 %	(3)	-		
5 years [c]	**Reflux, % (*n* / *n* known)**	53.6 %	(15 / 28)	69.0 %	(20 / 29)	0.847
	Los Angeles grade A, % (*n*)	53.3 %	(8)	35.0 %	(7)	
	Los Angeles grade B, % (*n*)	33.3 %	(5)	35.0 %	(7)	
	Los Angeles grade C, % (*n*)	6.7 %	(1)	5.0 %	(1)	
	Los Angeles grade D, % (*n*)	-		-		
	Unclassified, % (*n*)	6.7 %	(1)	25.0 %	(5)	

[a] Endoscopically diagnosed. [b] Fisher's exact test on the distribution of the five subgroups. [c] Based on the patients treated with POEM until December 31, 2014 (n = 176). - none.

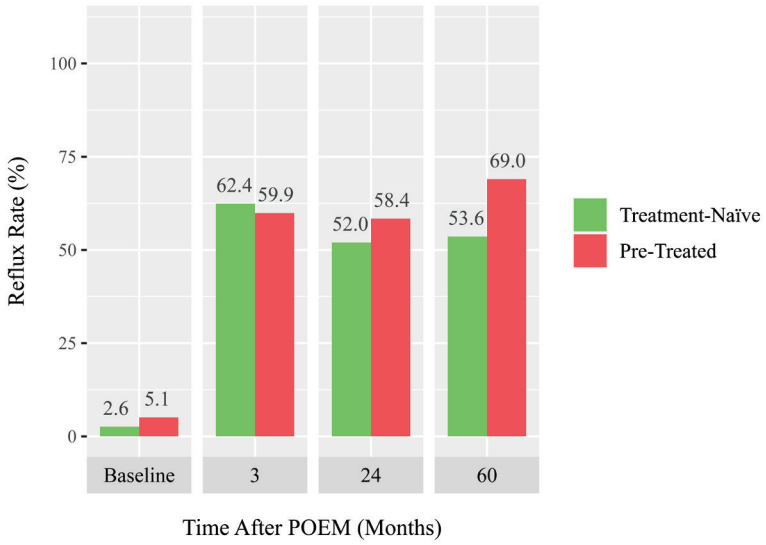


Figure 37: Endoscopically Diagnosed Gastroesophageal Reflux Disease After POEM. Reports of reflux symptoms without endoscopic confirmation were not considered.

Overall, post-interventional reflux is highly prevalent among both groups. With about 70 %, the reflux prevalence is exceptionally high in previously treated patients after five years. Besides that, it fluctuates between about 50 and 60 % in either group at any given follow-up time. When compared using *exact Fisher tests*, no significant differences between treatment-naïve and pre-treated patients were found.

Most patients show LA grade A or B reflux. For both groups and at any follow-up time, LA grade A is most common. It is found in 50 to 70 % of patients. LA grade B is the second most common grade, which ranges from 20 to 50 %. Far behind, LA grades C and D are very rare in both groups. When comparing the three-month follow-up to the two- and the five-year follow-ups, there seems to be no apparent tendency for a progression from lower to higher LA grades. However, the numbers of patients who underwent upper endoscopy at the later follow-ups are way too small to pass a verdict.

3.2.4 IRP Development

Table 11 and figure 38 show the IRP development after POEM.

Table 11: IRP After POEM.

Time	IRP (mmHg)	Treatment-Naïve	Pre-Treated	*p* [a]
3 months	**Mean ± SD (range), *n***	11.0 ± 6.0 (2.0 – 52.0), 113	9.6 ± 4.9 (1.0 – 25.9), 103	0.072 +
	0 – 5, % (*n*)	11.5 % (13)	13.6 % (14)	
	6 – 10, % (*n*)	37.2 % (42)	48.5 % (50)	
	11 – 15, % (*n*)	39.8 % (45)	24.3 % (25)	
	16 – 20, % (*n*)	8.0 % (9)	8.7 % (9)	
	21 – 25, % (*n*)	1.8 % (2)	3.9 % (4)	
	26 – 30, % (*n*)	0.9 % (1)	1.0 % (1)	
	> 30, % (*n*)	0.9 % (1)	-	
2 years	**Mean ± SD (range), *n***	11.5 ± 6.5 (2.7 – 26.7), 16	11.9 ± 6.9 (0.0 – 25.0), 19	0.861
	0 – 5, % (*n*)	18.8 % (3)	15.8 % (3)	
	6 – 10, % (*n*)	31.3 % (5)	21.1 % (4)	
	11 – 15, % (*n*)	25.0 % (4)	26.3 % (5)	
	16 – 20, % (*n*)	12.5 % (2)	26.3 % (5)	
	21 – 25, % (*n*)	6.3 % (1)	10.5 % (2)	
	26 – 30, % (*n*)	6.3 % (1)	-	
	> 30, % (*n*)	-	-	
5 years [b]	**Mean ± SD (range), *n***	5.9 ± 3.3 (2.9 – 11.0), 5	10.6 ± 5.1 (6.0 – 23.0), 10	0.055 +
	0 – 5, % (*n*)	40.0 % (2)	-	
	6 – 10, % (*n*)	40.0 % (2)	60.0 % (6)	
	11 – 15, % (*n*)	20.0 % (1)	30.0 % (3)	
	16 – 20, % (*n*)	-	-	
	21 – 25, % (*n*)	-	10.0 % (1)	
	26 – 30, % (*n*)	-	-	
	> 30, % (*n*)	-	-	

[a] Two-sample t test. [b] Based on the patients treated with POEM until December 31, 2014 (n = 176).
*+ p ≤ 0.1, * p ≤ 0.05, ** p ≤ 0.01, *** p ≤ 0.001. SD: standard deviation, - none.*

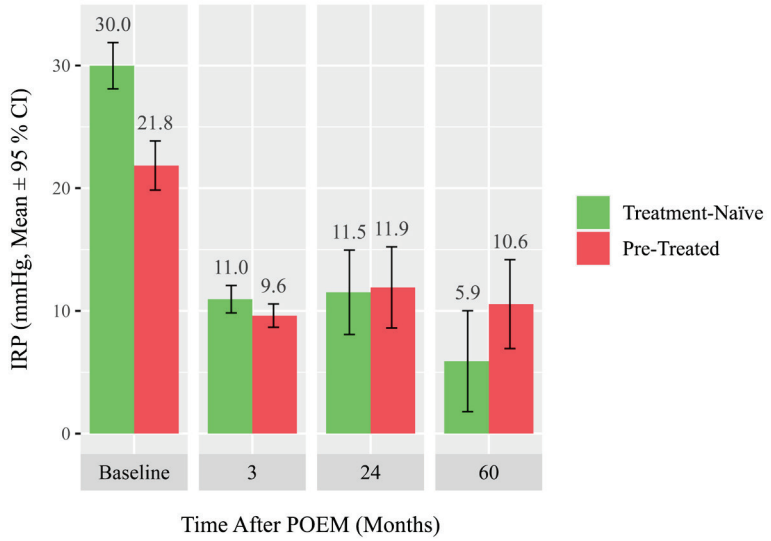


Figure 38: IRP Development After POEM. An IRP below 15 mmHg is considered normal. CI: confidence interval.

Treatment-naïve patients show significantly higher baseline IRPs compared to previously treated patients. Subsequent high-resolution manometries performed after three months, two years, and five years did not reveal any significant differences between the IRPs measured in both groups. Treatment-naïve patients, in comparison to previously treated patients, show a trend toward slightly higher IRPs after three months (11.0 vs. 9.6 mmHg, $p = 0.072$) and toward much lower IRPs after five years (5.9 vs. 10.6 mmHg, $p = 0.055$). Among both groups, the observed IRPs seem to be distributed quite evenly between 0 and about 20 mmHg in the three-month and the two-year follow-up. The five-year results cannot be interpreted because only five treatment-naïve and ten pre-treated patients underwent follow-up manometry at that time.

3.2.5 Re-Treatments

Known re-treatments after POEM are depicted in table 12.

Table 12: Re-Treatments After POEM.

Re-Treatment [b]	**Treatment-Naïve, % (*n*)**	**Pre-Treated, % (*n*)**	***p*** [a]
All re-treatments	9.4 % (18)	15.3 % (28)	0.115
Botulinum toxin injection	5.6 % (1)	3.6 % (1)	
Balloon dilatation	16.7 % (3)	21.4 % (6)	
Laparoscopic Heller myotomy	27.8 % (5)	21.4 % (6)	
Per-oral endoscopic myotomy	50.0 % (9)	50.0 % (14)	
Esophagectomy	-	3.6 % (1)	

The percentages of the individual re-treatments are relative to the number of patients in the respective pre-treatment group.
[a] Fisher's exact test on the overall re-treatment rates. [b] If multiple re-treatments are known, only the first one is considered.

Re-treatment rates after POEM seem to be similar in treatment-naïve and previously treated patients. The most frequently reported re-treatment is another POEM, which makes up for 50 % of re-treatments in both groups. Heller myotomy was performed in 27.8 % of the re-treated patients in the treatment-naïve group and in 21.4 % of cases in the previously treated group. It is followed by balloon dilatation with 16.7 % in treatment-naïve and 21.4 % in previously treated patients. Botulinum toxin injections are infrequent. One pre-treated patient underwent esophagectomy in an external facility after being diagnosed with esophageal perforation following their initial discharge.

3.2.6 Perioperative Biochemical Laboratory Markers

Table 13 shows the *C-reactive protein* (CRP) and the *Leucocytes*, as well as perioperative blood loss as indicated by the *hemoglobin* concentration, in the patients' blood before and after POEM.

Table 13: Perioperative Markers.

Variable	Treatment-Naïve	Pre-Treated	*p* [a]
Preoperative markers, mean ± SD (range) [b]			
Leucocytes (10^9/l)	7.4 ± 2.4 (3.3 – 16.6)	7.0 ± 2.1 (3.7 – 14.5)	0.171
C-reactive protein (mg/l)	9.5 ± 15.4 (5 – 120)	6.1 ± 3.8 (5 – 32)	**0.017** *
Hemoglobin (g/dl)	14.1 ± 1.5 (8.0 – 18.3)	14.0 ± 1.5 (10.2 – 17.1)	0.603
Postoperative markers, mean ± SD (range) [c]			
Leucocytes (10^9/l)	10.8 ± 3.3 (3.8 – 26.8)	10.6 ± 3.3 (4.2 – 28.0)	0.638
C-reactive protein (mg/l)	76.6 ± 38.1 (5 – 182)	78.2 ± 44.0 (5 – 246)	0.709
Hemoglobin (g/dl)	12.7 ± 1.3 (8.6 – 17.3)	12.7 ± 1.3 (9.8 – 16.0)	0.709

*[a] Two-sample t test. [b] Trough level observed up to three days before POEM. [c] Peak level observed up to five days after POEM. + p ≤ 0.1, * p ≤ 0.05, ** p ≤ 0.01, *** p ≤ 0.001. SD: standard deviation.*

Pre-treated patients show a nearly 3.5 mg/l lower pre-POEM mean CRP concentration when compared to treatment-naïve patients ($p = 0.017$). No other significant differences were found.

3.3 Distribution Analyses

The multivariate regression models that will be reported in chapters 3.5 and 3.6 include five predictors: *pre-treatment*, *sex*, *age*, *IRP*, and *achalasia type*. Among these, only *age* and *IRP* are continuous variables. To attain reliable effect size estimates, the plausibility of their observed values is particularly relevant. Their distributions were analyzed and screened for outlying observations. The variables *pre-treatment*, *sex*, and *achalasia type* are not discussed here because they are factorial. As such, they cannot have outliers because each factor level is plausible by definition.

3.3.1 IRP Distribution

Figures 39 to 42 and table 14 detail the patients' baseline IRP distribution.

Table 14: IRP Distribution.

Subgroup	Mean ± SD (mmHg)	Median (mmHg)	Skew / Kurtosis	*p* [a]
All	26.3 ± 12.6	25.0	0.619 / 3.19	**< 0.001** ***
Male	24.5 ± 11.8	22.0	0.775 / 3.88	**< 0.001** ***
Female	28.2 ± 13.2	26.9	0.437 / 2.71	**0.031** *
Treatment-naïve	30.0 ± 12.1	28.9	0.415 / 3.00	0.086 +
Pre-treated	21.8 ± 11.8	20.1	1.09 / 4.56	**< 0.001** ***
Achalasia type I	26.3 ± 12.6	25.0	0.619 / 3.19	**< 0.001** ***
Achalasia type II	28.1 ± 12.6	26.7	0.506 / 3.10	**0.018** *
Achalasia type III	25.8 ± 11.6	23.0	1.20 / 4.14	**< 0.001** ***

[a] *Shapiro-Wilk test. + p ≤ 0.1, * p ≤ 0.05, ** p ≤ 0.01, *** p ≤ 0.001. SD: standard deviation.*

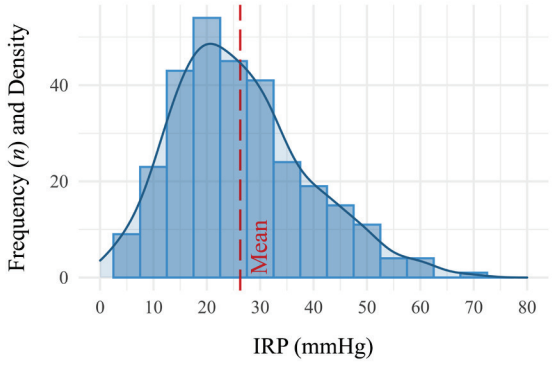


Figure 39: Overall IRP Distribution.

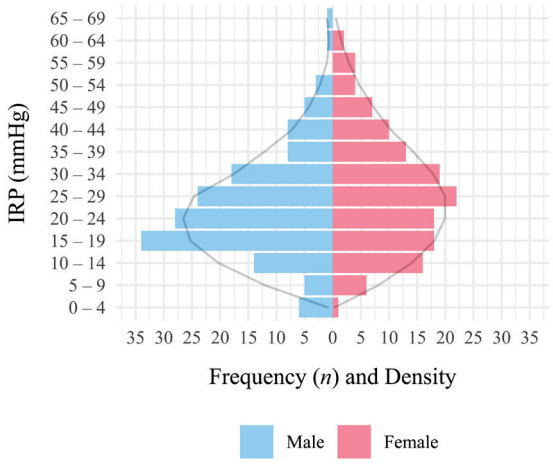


Figure 40: IRP Distribution by Sex.

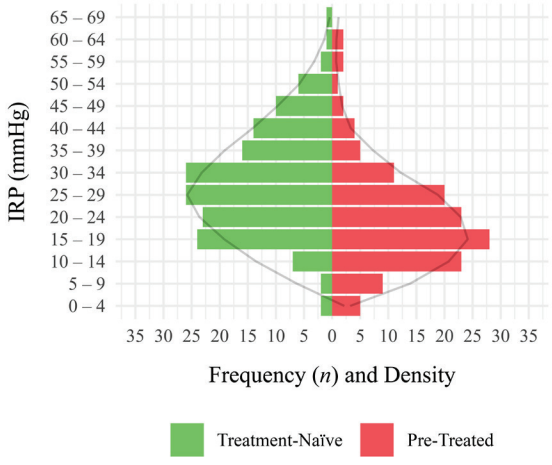


Figure 41: IRP Distribution by Pre-Treatment Group.

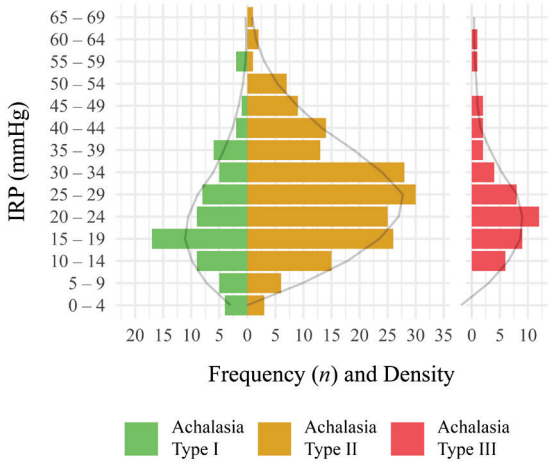


Figure 42: IRP Distribution by Achalasia Type.

The overall distribution and the distributions of each subgroup show a positive skew. This indicates an accumulation of values at the lower-value side of the mean. Male and pre-treated patients, as well as patients diagnosed with type III achalasia, show elevated kurtoses of 3.9, 4.6, and 4.1, respectively. Therefore, these subgroups' distributions have higher number of values in their tails than a normal distribution would.

Male patients show a slightly lower mean IRP compared to female patients (24.5 vs. 28.2 mmHg). Also, as already discussed in chapter 3.1, pre-treated patients have a much lower mean IRP compared to treatment-naïve patients (21.8 vs. 30.0 mmHg).

The *Shapiro-Wilk test* turned out significant for the overall distribution and every single subgroup's distribution, except for the treatment-naïve patients, for which the test failed to reach significance by a narrow margin ($p = 0.087$). To summarize, the IRP does not follow a normal distribution.

3.3.2 Age Distribution

Figures 43 to 46 and table 15 detail the patients' age distribution.

Table 15: Age Distribution.

Subgroup	Mean ± SD (Years)	Median (Years)	Skew / Kurtosis	p [a]
All	46.0 ± 16.1	45.0	0.218 / 2.38	0.559
Male	46.2 ± 16.5	44.0	0.346 / 2.29	**< 0.001** ***
Female	45.8 ± 15.5	47.0	0.027 / 2.48	0.188
Treatment-naïve	44.4 ± 16.1	44.0	0.244 / 2.51	**0.038** *
Pre-treated	47.8 ± 15.9	47.0	0.209 / 2.25	**0.008** **
Achalasia type I	44.9 ± 14.6	44.0	0.230 / 2.38	0.184
Achalasia type II	44.0 ± 15.2	44.0	0.186 / 2.43	**0.018** *
Achalasia type III	56.8 ± 18.1	60.5	-0.331 / 2.17	0.094 +

[a] *Shapiro-Wilk test. + $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$. SD: standard deviation.*

Again, the overall distribution as well as each subgroup's distribution show a positive skew, and as such an accumulation of values at the lower-value side of the mean. Therefore, younger patients are more frequent than older patients. The only exception to this rule is found in patients diagnosed with type III achalasia: for them, with a negative skew of -0.3, the opposite holds true. The kurtoses of the different groups range between about 2.1 and 2.5. This indicates generally slightly reduced numbers of values in the distributions' tails, compared to a normal distribution.

Ranging from 44.0 to 47.8 years, the mean of almost every subgroup's distribution is largely similar to the overall distribution's mean of 46.0 years. Only patients diagnosed with type III achalasia are generally much older, as it is indicated by their exceptionally high mean age of 56.8 years

With the *Shapiro-Wilk test* reaching significance, the age distributions of males, both treatment-naïve and pre-treated patients, as well as patients diagnosed with type II achalasia have been shown to not be normally distributed. The age distributions of the complete data set, females, and patients diagnosed with the achalasia types I and III, in contrast, are.

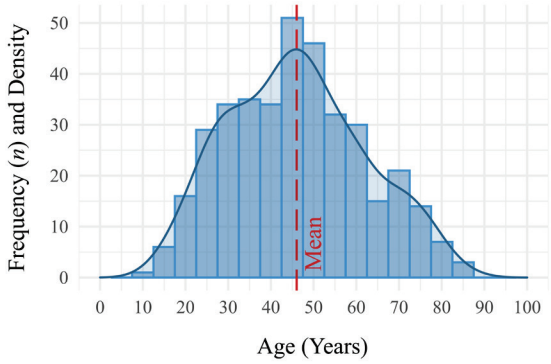


Figure 43: Overall Age Distribution.

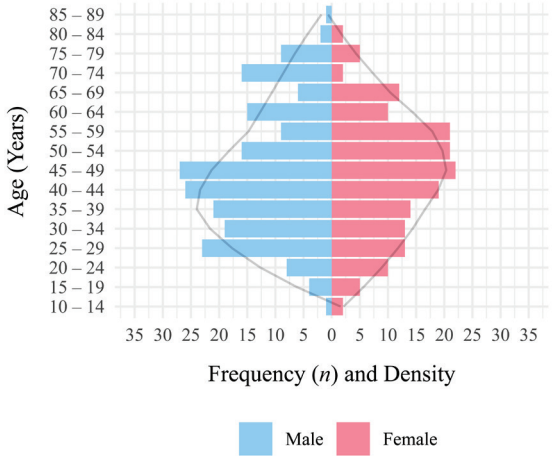


Figure 44: Age Distribution by Sex.

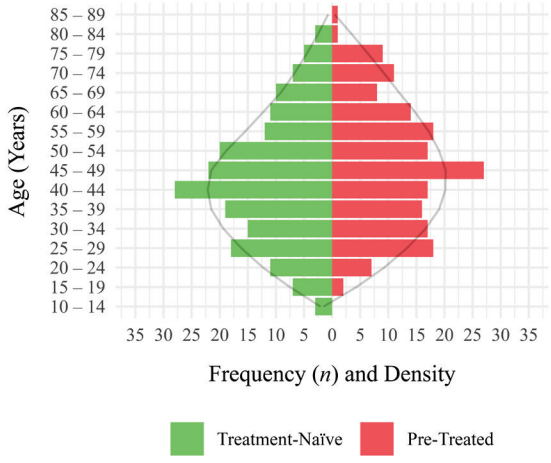


Figure 45: Age Distribution by Pre-Treatment Group.

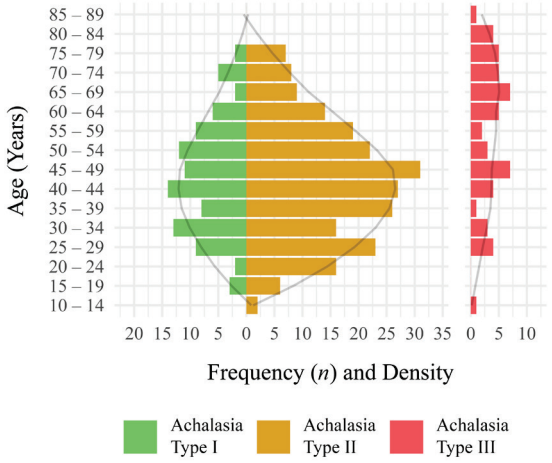


Figure 46: Age Distribution by Achalasia Type.

3.3.3 Outlier Analysis

The results of the outlier analysis of the only continuous predictors of interest, *age* and *IRP*, are shown in figure 47. Outliers were detected via *interquartile range analysis*, as explained in chapter 2.2.4.

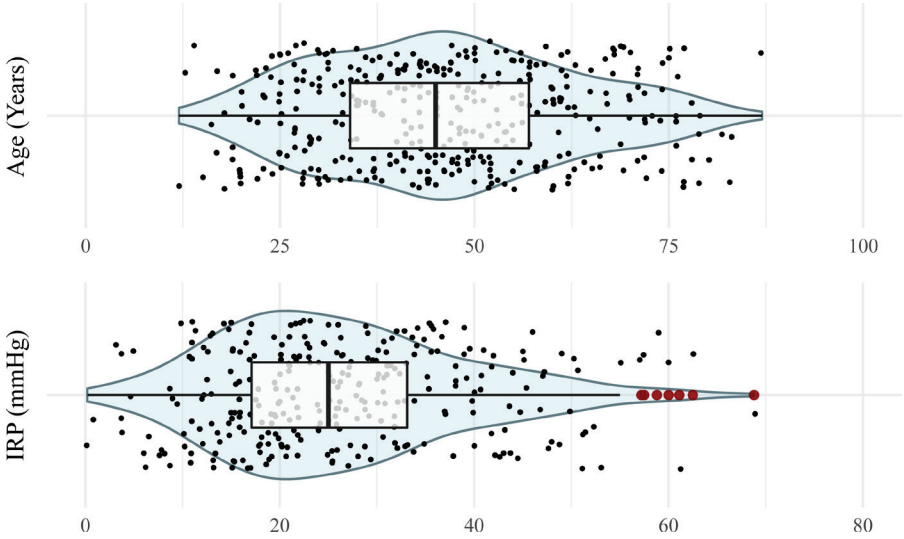


Figure 47: Outlier Analysis. Each observation is depicted as a black dot, transposed by a random vertical offset. The underlying blue curve is the density. As usual, the Tukey box plot shows the first and third quartile (Q_1 and Q_3, left and right box borders), the interquartile range (box width), and the median (vertical line inside the box). The horizontal line extends along both sides to the lower and upper extremes of the distribution, that is up to $Q_1 - 1.5\ IQR$ and $Q_3 + 1.5\ IQR$, respectively. Observations beyond these thresholds are considered potential outliers. They are depicted as thick red dots.

For *age*, Q_1 = 34.0 years, Q_3 = 57.0 years, and IQR = 23.0 years. The cut-off values for the outlier detection therefore are -0.5 years and 91.5 years. No outliers beyond these thresholds were found.

For the *IRP*, Q_1 = 17.1 mmHg, Q_3 = 33.1 mmHg, and IQR = 16.0 mmHg. The cut-off values for the outlier detection therefore are -6.9 mmHg and 57.1 mmHg. Seven outlying IRPs were found in the data: 57.2, 57.5, 58.8, 60.0, 61.1, 62.5, and 68.8 mmHg. These observations are high, yet plausible. They are, together with their associated observed values of the other relevant predictors, shown in table 16.

Table 16: Outlying IRP Observations.

#	IRP (mmHg)	Sex	Age (Years)	Achalasia Type	Pre-Treatment Group
1	**57.2**	Female	55	II	Pre-treated
2	**57.5**	Female	58	I	Pre-treated
3	**58.8**	Female	47	III	Treatment-naïve
4	**60.0**	Male	83	III	Pre-treated
5	**61.1**	Female	63	II	Treatment-naïve
6	**62.5**	Female	48	II	Pre-treated
7	**68.8**	Male	51	II	Treatment-naïve

Most outlying patients are female, as indicated by a sex ratio of 5 to 2. With 83 years, one male patient is much older than the others, who are aged between 47 and 63 years. Overall, the observed predictor values seem random among all records that contain outlying IRPs. No apparent trends were found. This indicates that the outliers most likely appeared by chance. There is no evidence for an underlying systematic error. Since all observed values are plausible, no further action is required to avoid bias.

3.4 Analysis of Missing Information

3.4.1 Missing Baseline Data

The completeness of the baseline data used in the statistical models that will be described in chapters 3.5 and 3.6 are summarized in table 17. The temporal distribution of missing IRP observations is illustrated in figure 48.

Table 17: Baseline Data Completeness.

Variable	Measure	Completeness, % (*n* / total)	Missingness, %
Pre-treatment	Treatment-naïve or pre-treated	Complete	0
Sex	Male or female	Complete	0
Age	Years	Complete	0
Achalasia type	I, II, or III	Complete	0
IRP	mmHg	78.9 % (295 / 374)	21.1 %

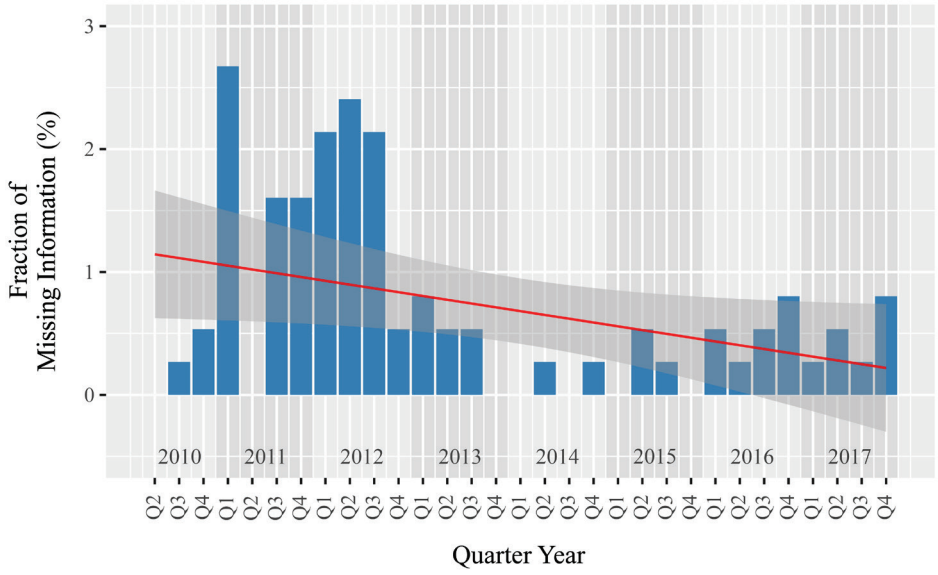


Figure 48: Temporal Distribution of Missing Baseline IRP Observations. Shown are the fractions of missing information per quarter year, relative to the total amount of missing data. The trend over time is visualized by a linear regression line with a 95 % confidence interval.

Overall, 21.1 % of IRP observations are missing. All other variables were completely observed. A confined cluster of missing observations stretches from 2010 to the third quarter of 2012. Until the end of the latter year, HRM was much less available than it is nowadays. It was not routinely performed, especially not if patients had already undergone conventional manometry in external facilities. Because high-resolution manometry is required to determine the IRP, most patients with missing IRP observations were treated during these early infancy years of POEM. Considering this, the patients who underwent POEM in first year, that is 2010, appear to have an unusually low fraction of missing IRP information. This is merely an artifact: only five patients were treated that year.

3.4.2 Patient Compliance and Missing Follow-Up Data

Table 18 summarizes the patients' compliance in attending the systematic follow-ups that have been pictured in chapter 2.3.1. The knowledge of a patient's treatment response is especially relevant for the quality of the regression models that will be reported in chapters 3.5 and 3.6. As explained in chapter 2.2.7.4, treatment failure must not be imputed because it is the primary outcome of these models. Because of this, patients whose treatment response is unknown for a specific follow-up cannot be included in that same follow-up's regression model.

Because undergoing another treatment after POEM is considered failure, a negative treatment response of some patients can be inferred even if they are lost to follow-up. This is the *inferred treatment state* that is reported in table 18. It does not represent the actual patient participation in the respective follow-up. Instead, it describes the fraction of known response states, either observed *or* inferred from known re-treatments. These states represent the data that can ultimately be included in the statistical model. As such, they represent the actual completeness of the data these models are built upon.

The overall follow-up participation regarding the questionnaire is 80.7 % after two years, 73.2 % after three years, and 68.2 % after five years. The questionnaire of course enquires about the Eckardt score of a patient, and thus delivers knowledge about their treatment response. By inferring treatment failures from known re-treatments after POEM, these percentages rise to 82.9 % after two years, 79.7 % after three years, and 75.0 % after five years. This indicates a good follow-up adherence of the patients and a solid foundation for the application of the regression analyses provided by the upcoming chapters.

Participation in upper endoscopy, and even more so in manometry is significantly worse, though. While about 40 % of the patients provided us with endoscopy reports after two years, as did about 30 % after five years, not even 10 % underwent follow-up manometry at either time.

Table 18: Follow-Up Compliance and Data Completeness.

Procedure	Follow-Up Time	Completeness, % (*n* / total)		Missingness, %
Questionnaire: Eckardt score	3 months	85.0 %	(318 / 374)	15.0 %
	6 months	74.3 %	(278 / 374)	25.7 %
	1 year	78.9 %	(295 / 374)	21.1 %
	2 years	80.7 %	(302 / 374)	19.3 %
	3 years [a]	73.2 %	(216 / 295)	26.8 %
	5 years [b]	68.2 %	(120 / 176)	31.8 %
Inferred treatment state [c]	3 months	85.8 %	(321 / 374)	14.2 %
	6 months	75.1 %	(281 / 374)	24.9 %
	1 year	79.9 %	(299 / 374)	20.1 %
	2 years	82.9 %	(310 / 374)	17.1 %
	3 years [a]	79.7 %	(235 / 295)	20.3 %
	5 years [b]	75.0 %	(132 / 176)	25.0 %
Manometry: IRP	3 months	57.8 %	(216 / 374)	42.3 %
	2 years	9.4 %	(35 / 374)	90.6 %
	5 years [b]	8.5 %	(15 / 176)	81.5 %
Upper endoscopy: gastroesophageal reflux disease	3 months	87.2 %	(326 / 374)	12.8 %
	2 years	40.6 %	(152 / 374)	59.4 %
	5 years [b]	32.4 %	(57 / 176)	67.6 %

[a] Based on the patients treated with POEM until December 31, 2016. [b] Based on the patients treated with POEM until December 31, 2014. [c] Data provided by the patients, supplemented with treatment responses inferred from known re-treatments.

3.4.3 Patterns and Correlations

Figure 49 shows an intersectional matrix. It illustrates potential correlations between the missing information of all predictors that will be included in the regression models described in the next two chapters.

It needs to be pointed out once again that the data contains patients who were not able to contribute to the three- or five-year follow-ups simply because they underwent POEM less than three or five years ago, respectively. If they would have been included in the graph, it would hardly be interpretable. This is because it would be unclear what fraction of missing data would be *actually* missing, rather than being censored by some of the patients' too recent POEM dates. Therefore, correlations regarding treatment failure after three and five years were assessed based on the subsets of patients that actually *could* contribute to the respective follow-ups. In the graph, this is indicated by their different colors.

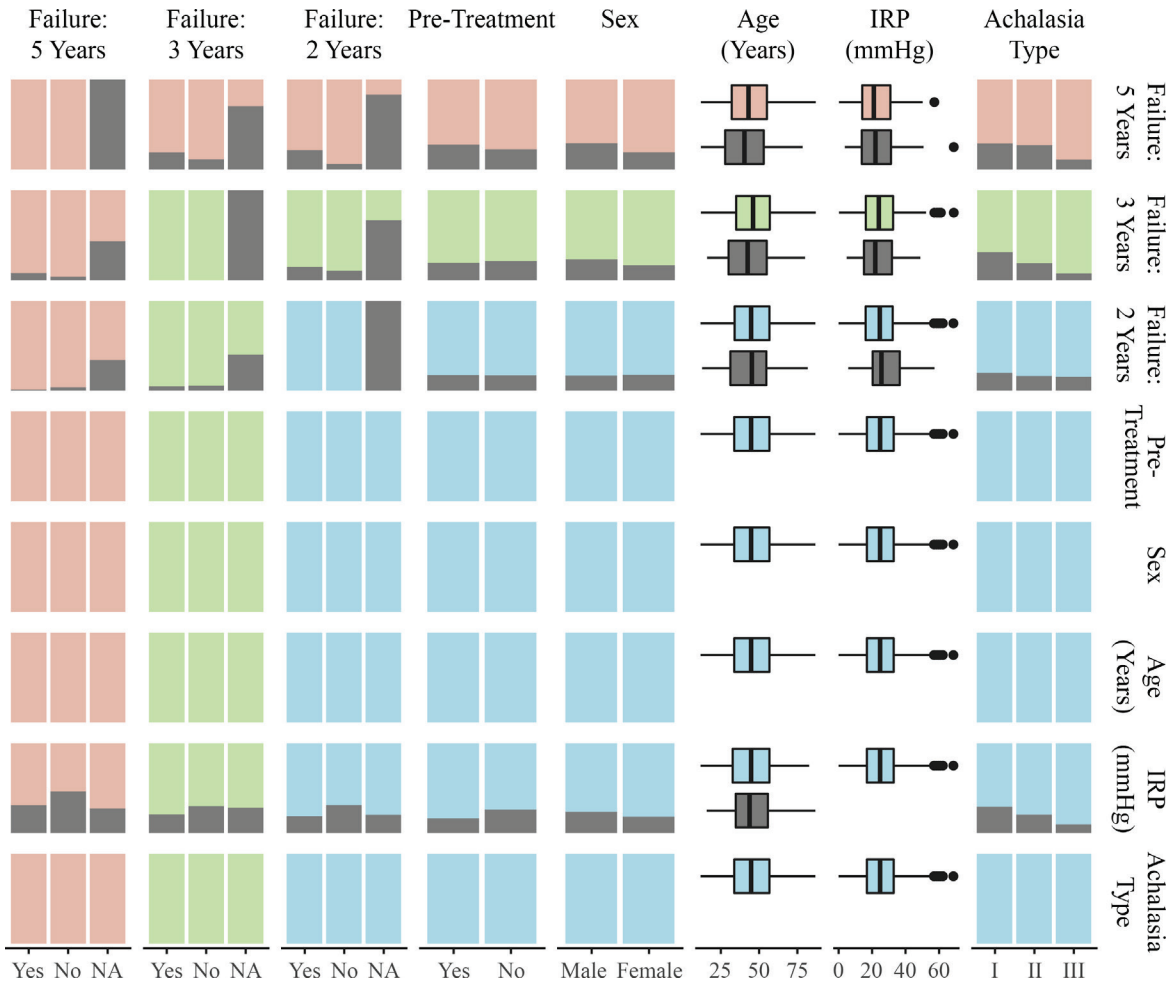


Figure 49: Matrix of Missing Information. Factorial variables are given as fractions. Continuous variables are illustrated as Tukey box plots, as explained in chapters 2.2.4 and 3.3.3. Blue: based on all data. Green: based on the patients treated with POEM until December 31, 2016 (n = 295). Red: based on the patients treated with POEM until December 31, 2014 (n = 176). Colored bars: proportions of known data. Gray bars: proportions of missing data. NA: not available (missing).

The matrix hints at correlations at best. However, a few observations should be mentioned. Most notably, type I achalasia as well as treatment failure after two and five years are correlated to a disproportionally large portion of missing baseline IRP observations. Treatment-naïve patients, compared to previously treated patients, are missing baseline IRP observations slightly more often as well. More so than type III achalasia, the types I and II are correlated with higher chances of missing out on the three- and five-year follow-ups. The same seems to apply to males compared to females. It is no surprise that not participating in any one follow-up is also tightly correlated to missing out on the other follow-ups. Besides these tendencies, no evidence for major correlations between the missing data was found.

3.5 Logistic Regression

3.5.1 Regression Model

The following logistic regression model was fit for each long-term follow-up, that is for the follow-ups conducted after two, three, and five years. The equation estimates the probability of treatment failure:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1\, pt + \beta_2\, sex + \beta_3\, age + \beta_4\, irp + \beta_5\, at_2 + \beta_6\, at_3)}}$$

Equation 19: Logistic Regression Model for the Probability of Treatment Failure.

For ease of view, the model is rewritten as a function of the *log-odds* (see chapter 2.2.5.2):

$$logit(p) = \beta_0 + \beta_1\, pt + \beta_2\, sex + \beta_3\, age + \beta_4\, irp + \beta_5\, at_2 + \beta_6\, at_3$$

Equation 20: Logistic Regression Model for the Log-Odds of Treatment Failure.

The coefficients and the independent predictor variables used in the model are depicted in table 19. The same parameters, except for the residual intercept β_i, will also be used for the Cox model, which will be presented in chapter 3.6.

Table 19: Logistic Regression Model Parameters.

Coefficients $\boldsymbol{\beta}$	**Meaning**	**Type and Definition**	
$\boldsymbol{\beta_0}$	Residual intercept estimate	metric	(dimensionless)
$\boldsymbol{\beta_i}$	Effect size estimate for x_i	metric	(dimensionless)
Predictors $\boldsymbol{x}$			
$\boldsymbol{x_1 = pt}$	Pre-treatment	factor	(pre-treated ≔ 1, treatment-naïve ≔ 0)
$\boldsymbol{x_2 = sex}$	Sex	factor	(female ≔ 1, male ≔ 0)
$\boldsymbol{x_3 = age}$	Age	metric	(years)
$\boldsymbol{x_4 = irp}$	IRP	metric	(mmHg)
$\boldsymbol{x_5 = at_2}$	Achalasia type II	factor	(achalasia type II ≔ 1, else ≔ 0)
$\boldsymbol{x_6 = at_3}$	Achalasia type III	factor	(achalasia type III ≔ 1, else ≔ 0)

As stated in chapter 2.2.5.2, the odds ratio of a predictor in a logistic regression model is the change in odds per 1-unit increase of said predictor. Obviously, a 1-year increase in *age* is not expected to have a large effect on the odds, nor is a 1-mmHg increase in the *IRP*. For this reason, the odds ratios of the aforementioned predictors may turn out unhandily small. To report more practical values, the predictors *age* and *IRP* are rescaled to estimate the odds ratio per 10 years and per 5 mmHg, respectively.

3.5.2 Treatment Failure After Two Years

3.5.2.1 Data Analysis and Imputation

For the two-year follow-up analysis, 64 data sets were removed because of unknown treatment response states. Subsequently, 310 data sets were included in the analysis. Baseline IRPs were observed for 244 out of these 310 data sets (21.3 % missing, mean: 25.8 mmHg). No other variables had missing data.

Little's test reached significance, indicating a potential MCAR violation ($\chi^2 = 12.8$, df = 5, $p = 0.026$). However, *Hawkins' test* reached significance as well ($p < 0.001$), whereas *Jamshidian's test* turned out insignificant ($p = 0.151$). These results imply homoscedasticity under non-normality. Therefore, evidence against MCAR is weak, and preliminary requirements for imputation were satisfied.

A total of 25 data sets with imputed IRP values were calculated by predictive mean matching. The imputations are visualized in figure 50.

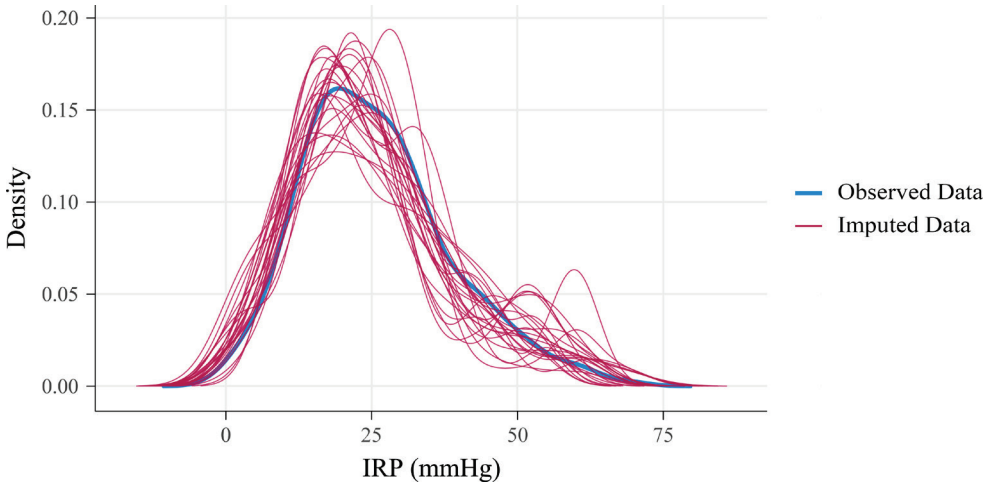


Figure 50: Imputed Baseline IRP Data of the Two-Year Logistic Regression Model.
Each red line represents the density of one imputed data set.

The densities of the imputed data match the density of the observed data well. This indicates a plausible distribution of the imputed values, and therefore speaks in favor of an overall good imputation model.

3.5.2.2 Regression Estimates

The regression estimates of the multiply imputed logistic regression model for the prediction of treatment failure two years after POEM are reported in table 20 and visualized in figure 51.

Table 20: Pooled Regression Estimates for Treatment Failure After Two Years.

Parameter	β	SE β	Odds Ratio e^{β} (95 % CI)	p
Intercept	1.10	0.673	NA	NA
Pre-treatment	0.239	0.322	NA	0.459
Sex	0.262	0.299	NA	0.380
Age / 10	-0.409	0.102	0.664 (0.543 – 0.813)	**< 0.001** ***
IRP / 5	-0.157	0.080	0.855 (0.730 – 1.002)	0.052 +
Achalasia type II	-0.358	0.350	NA	0.308
Achalasia type III	0.913	0.460	2.49 (1.01 – 6.16)	**0.048** *

Mean deviance: 296 (303 df), null deviance: 326 (309 df), mean AIC: 310, null AIC: 328, mean R^2_N: 0.144.

*+ $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$. CI: confidence interval, NA: not assessed, SE: standard error.*

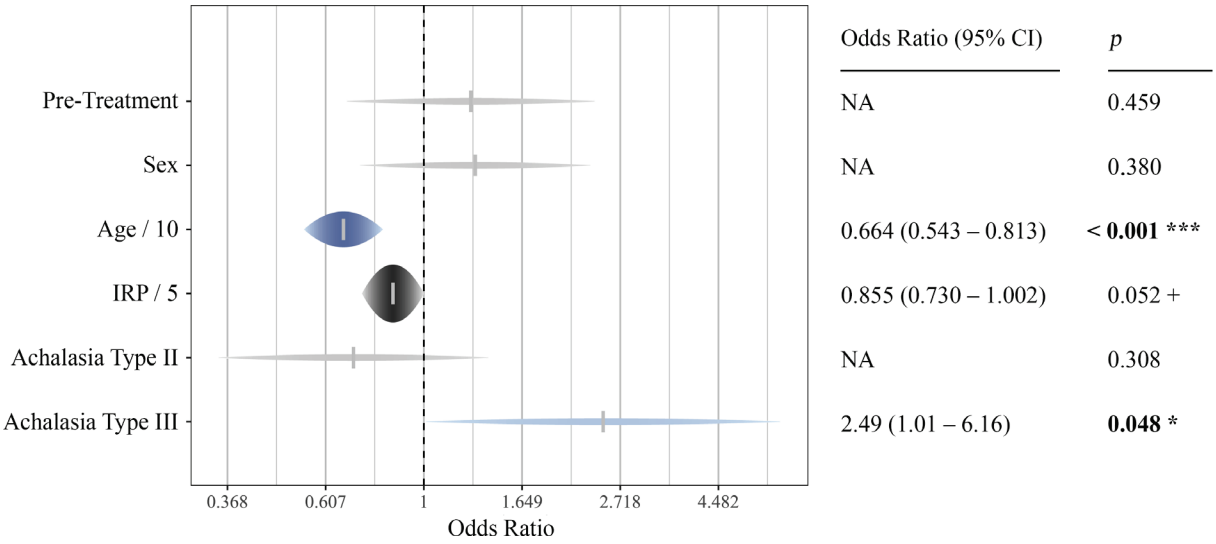


Figure 51: Logistic Regression Estimates for Treatment Failure After Two Years. Vertical bars: odds ratios, raindrop width: 95 % CI, raindrop height and color intensity: likelihood of the respective odds ratio within the CI. No odds ratios are given for insignificant predictors, and their raindrops are grayed out. + $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$. CI: confidence interval, NA: not assessed.

The estimated regression equation with all predictors as defined in table 19 has the following form:

$$logit(p) = 1.10 + 0.239\,pt - 0.262\,fem - 0.041\,age - 0.031\,irp - 0.358\,at_2 + 0.913\,at_3$$

Equation 21: Fit Logistic Regression Model for Treatment Failure After Two Years.

Age and type III achalasia are significant predictors of treatment failure two years after POEM. Higher age has a protective effect: the older a patient is at the time of POEM, the less likely they are to experience treatment failure. For every ten years of higher age, the odds of treatment failure decrease by factor 0.66 ($p < 0.001$). Type III achalasia on the other hand is a risk factor that, compared to type I, increases the odds by factor 2.5 ($p = 0.048$). No other predictors reached significance. However, with its p value trending toward α, the IRP can be considered another *potential* protective factor (OR: 0.86, $p = 0.052$).

3.5.2.3 Model Validation

To validate the imputed model, another regression model with the same set of predictors was fit without imputation, based on the 244 completely observed cases. The regression estimates and additional measures for the goodness-of-fit are detailed in table 21.

Table 21: Complete Case Regression Estimates for Treatment Failure After Two Years.

Parameter	β	**SE** β	**Odds Ratio** e^{β} **(95 % CI)**	p
Intercept	1.60	0.797	NA	NA
Pre-treatment	0.296	0.382	NA	0.440
Sex	0.393	0.360	NA	0.274
Age / 10	-0.626	0.134	0.535 (0.406 – 0.687)	**< 0.001** ***
IRP / 5	-0.178	0.085	0.837 (0.704 – 0.983)	**0.036** *
Achalasia type II	-0.081	0.436	NA	0.854
Achalasia type III	1.11	0.562	3.04 (1.02 – 9.37)	**0.048** *

Deviance: 205 (237 df), null deviance: 239 (243 df), AIC: 219, null AIC: 241.

R_N^2: 0.212, concordance: 0.756, D_{xy}: 0.511, γ: 0.512, τ_a: 0.160. Hosmer-Lemeshow test: $\chi^2 = 5.85$ (8 df, $p = 0.664$).

*+ $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$. CI: confidence interval, NA: not assessed, SE: standard error.*

Results and implications will be discussed in chapter 4.

3.5.3 Treatment Failure After Three Years

3.5.3.1 Data Analysis and Imputation

For the three-year follow-up analysis, 79 patients were excluded because they had undergone POEM less than three years ago. Out of the remaining 295 data sets, 60 were removed because of unknown treatment response states. This yielded a total of 235 data sets that were included in the analysis. *Little's test* did not reach significance ($\chi^2 = 10.1$, df = 5, $p = 0.072$). Therefore, evidence against MCAR is weak, and preliminary requirements for imputation were satisfied.

Similar to the two-year follow-up, the only variable in the data set that contained missing values is the baseline IRP. It was observed in 180 out of 235 data sets (23.4 % missing, mean: 25.8 mmHg).

A total of 25 data sets with imputed IRP values were calculated by predictive mean matching. The imputations are visualized in figure 52.

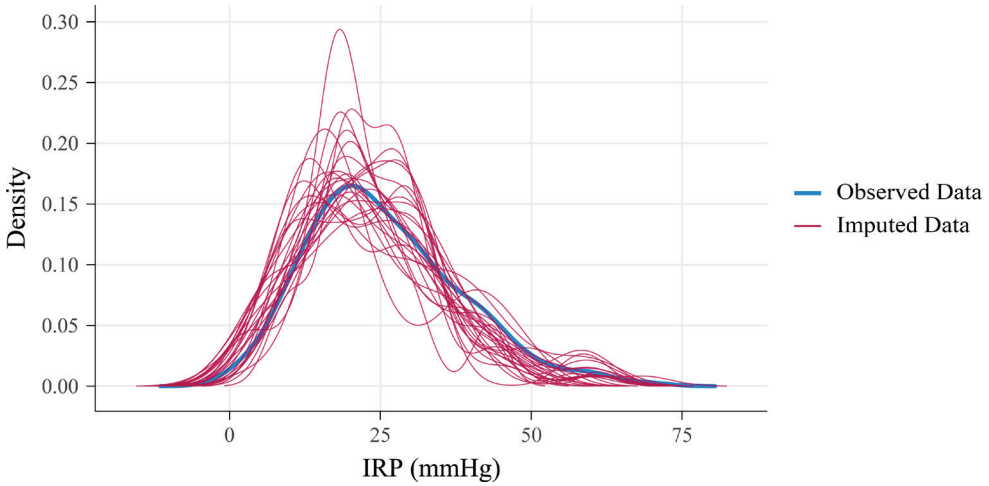


Figure 52: Imputed Baseline IRP Data of the Three-Year Logistic Regression Model.
Each red line represents the density of one imputed data set.

The imputations overall seem to have a slightly higher variance compared to the two-year model. Yet, the densities of the imputed data match the density of the observed data still well. Again, this is evidence of a plausible distribution of the imputed values and indicates an overall good imputation model.

3.5.3.2 Regression Estimates

The regression estimates of the multiply imputed logistic regression model for the prediction of treatment failure three years after POEM are reported in table 22 and visualized in figure 53.

Table 22: Pooled Regression Estimates for Treatment Failure After Three Years.

Parameter	β	SE β	Odds Ratio e^{β} (95 % CI)	p
Intercept	1.28	0.713	NA	NA
Pre-treatment	0.364	0.330	NA	0.271
Sex	-0.026	0.313	NA	0.934
Age / 10	-0.328	0.105	0.720 (0.585 – 0.886)	**0.002** **
IRP / 5	-0.118	0.079	NA	0.137
Achalasia type II	-0.652	0.374	0.521 (0.249 – 1.089)	0.083 +
Achalasia type III	0.718	0.479	NA	0.135

Mean deviance: 261 (228 df), null deviance: 286 (234 df), mean AIC: 275, null AIC: 288, mean R^2_N: 0.143.

*+ $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$. CI: confidence interval, NA: not assessed, SE: standard error.*

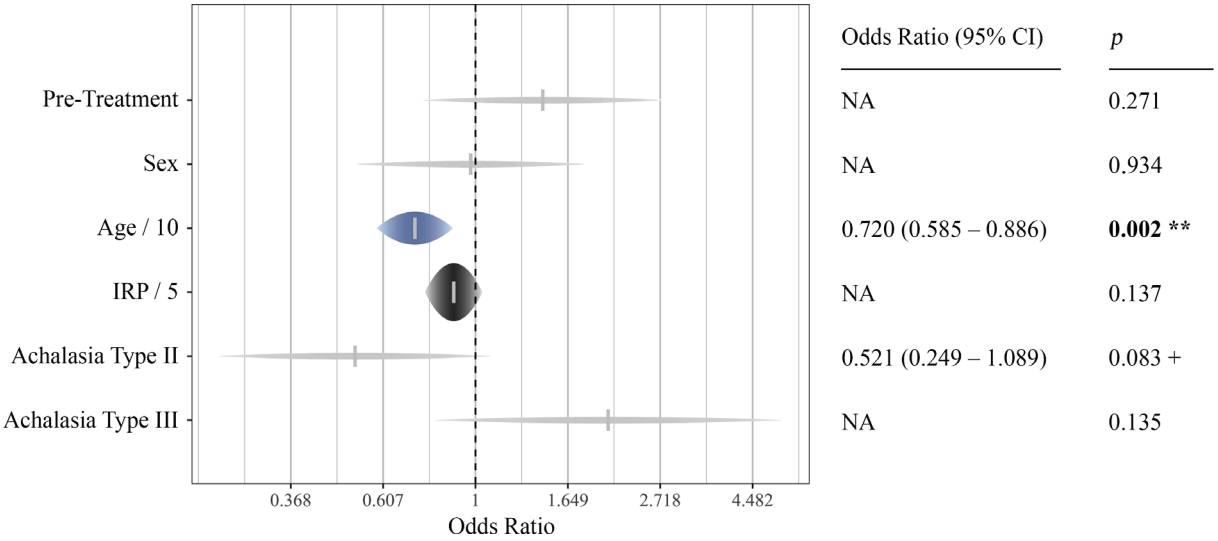


Figure 53: Logistic Regression Estimates for Treatment Failure After Three Years. Vertical bars: odds ratios, raindrop width: 95 % CI, raindrop height and color intensity: likelihood of the respective odds ratio within the CI. No odds ratios are given for insignificant predictors, and their raindrops are grayed out. + $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$. CI: confidence interval, NA: not assessed.

The estimated regression equation with all predictors as defined in table 19 has the following form:

$$logit(p) = 1.28 + 0.364\, pt - 0.026\, fem - 0.033\, age - 0.024\, irp - 0.652\, at_2 + 0.718\, at_3$$

Equation 22: Fit Logistic Regression Model for Treatment Failure After Three Years.

As in the two-year model, higher age is yet again a significant predictor of treatment failure. For every ten years of higher age, the odds of treatment failure after three years decrease by factor 0.72 (p = 0.002). No other predictors reached significance. Type II achalasia appears to be another *potential* protective factor (OR: 0.52, p = 0.083), though.

3.5.3.3 Model Validation

A comparative regression model was fit with the same set of predictors, but without imputation, based on the 180 complete cases. The regression estimates and additional measures for the goodness-of-fit are detailed in table 23.

Table 23: Complete Case Regression Estimates for Treatment Failure After Three Years.

Parameter	β	SE β	Odds Ratio e^{β} (95 % CI)	p
Intercept	0.908	0.827	NA	NA
Pre-treatment	0.568	0.382	NA	0.137
Sex	0.001	0.362	NA	0.997
Age / 10	-0.368	0.124	0.692 (0.538 – 0.876)	**0.003** **
IRP / 5	-0.131	0.082	NA	0.113
Achalasia type II	-0.134	0.470	NA	0.774
Achalasia type III	0.975	0.571	2.65 (0.88 – 8.38)	0.088 +

Deviance: 261 (228 df), null deviance: 286 (234 df), AIC: 275, null AIC: 288.

R_N^2: 0.147, concordance: 0.717, D_{xy}: 0.435, γ: 0.435, τ_a: 0.173. Hosmer-Lemeshow test: χ^2 = 11.3 (8 df, p = 0.187).

*+ $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$. CI: confidence interval, NA: not assessed, SE: standard error.*

Results and implications will be discussed in chapter 4.

3.5.4 Treatment Failure After Five Years

3.5.4.1 Data Analysis and Imputation

For the five-year follow-up analysis, 198 patients were excluded because five years had not yet passed since they had undergone POEM. Out of the remaining 176 data sets, 44 were removed because of unknown treatment response states. Subsequently, 132 data sets were included in the analysis. *Little's test* did not reach significance ($\chi^2 = 9.41$, df = 5, $p = 0.094$). Therefore, evidence against MCAR is yet again weak, and preliminary requirements for imputation were satisfied.

Similar to the two- and three-year follow-ups, the only variable with missing data is the baseline IRP. It was observed in 83 out of 132 data sets (37.1 % missing, mean: 23.5 mmHg).

A total of 40 data sets with imputed IRP values were calculated by predictive mean matching. The imputations are visualized in the following figure

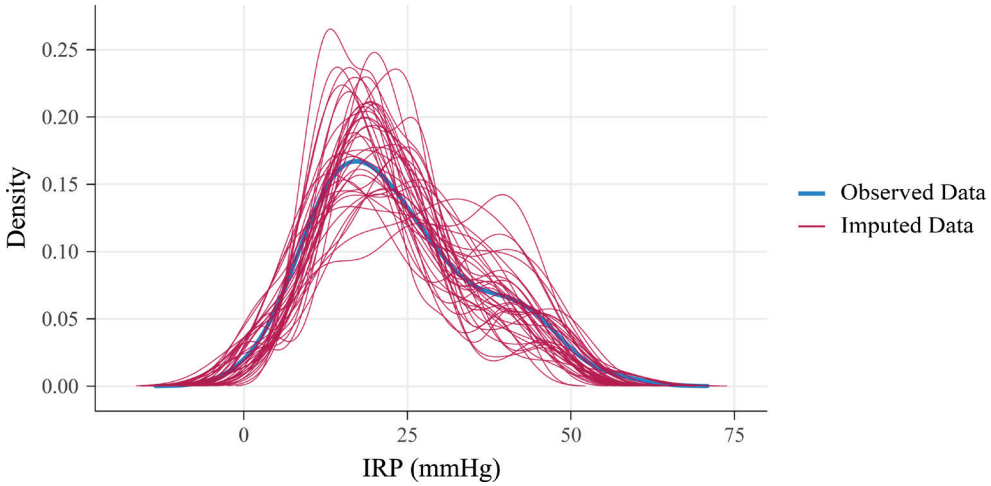


Figure 54: Imputed Baseline IRP Data of the Five-Year Logistic Regression Model.
Each red line represents the density of one imputed data set.

The imputations seem to have a higher variance than the imputations of the two- and three-year models. However, the densities of the imputed data match the density of the observed data still decently. Again, this is evidence of a plausible distribution of the imputed values and indicative of an overall good imputation model. The slightly more pronounced deviations of the imputed from the observed data compared to the two- and three-year models were expected, considering that the imputations of the five-year model are based on much less data.

3.5.4.2 Regression Estimates

The regression estimates of the multiply imputed logistic regression model for the prediction of treatment failure five years after POEM are reported in table 24 and visualized in figure 55.

Table 24: Pooled Regression Estimates for Treatment Failure After Five Years.

Parameter	β	SE β	Odds Ratio e^{β} (95 % CI)	p
Intercept	1.13	0.879	NA	NA
Pre-treatment	0.754	0.396	2.13 (0.97 – 4.66)	0.059 +
Sex	0.052	0.398	NA	0.896
Age / 10	-0.304	0.131	0.738 (0.569 – 0.956)	**0.022** *
IRP / 5	-0.016	0.110	NA	0.886
Achalasia type II	-0.790	0.488	NA	0.108
Achalasia type III	0.099	0.598	NA	0.869

Mean deviance: 165 (125 df), null deviance: 179 (131 df), mean AIC: 179, null AIC: 181, mean R^2_N: 0.132.

*+ $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$. CI: confidence interval, NA: not assessed, SE: standard error.*

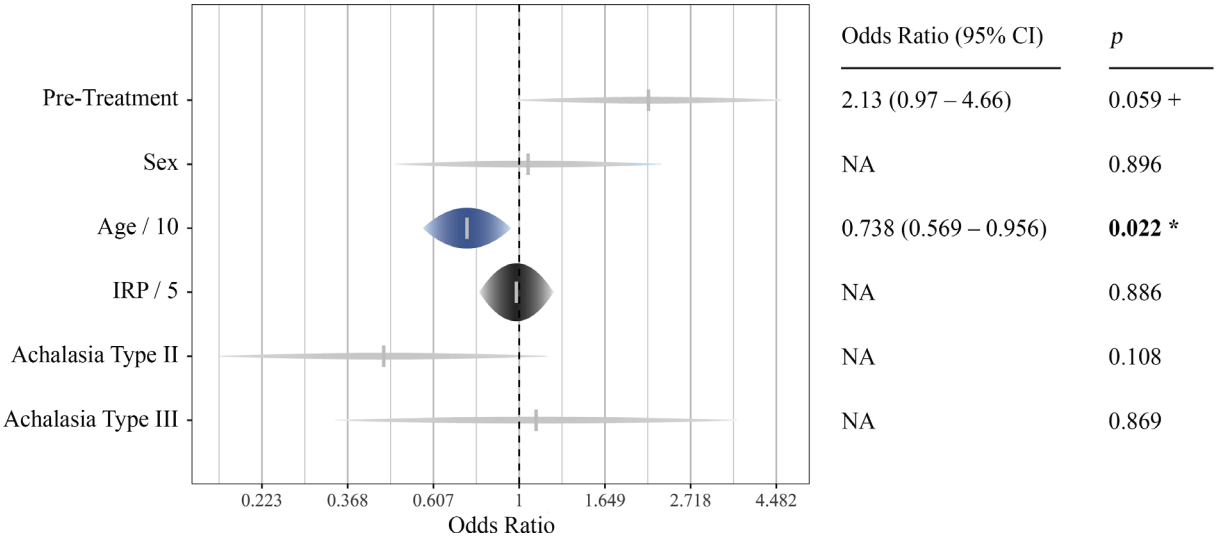


Figure 55: Logistic Regression Estimates for Treatment Failure After Five Years. Vertical bars: odds ratios, raindrop width: 95 % CI, raindrop height and color intensity: likelihood of the respective odds ratio within the CI. No odds ratios are given for insignificant predictors, and their raindrops are grayed out. + $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$. CI: confidence interval, NA: not assessed.

The estimated regression equation with all predictors as defined in table 19 has the following form:

$$logit(p) = 1.13 + 0.754\, pt - 0.052\, fem - 0.030\, age - 0.003\, irp - 0.790\, at_2 + 0.099\, at_3$$

Equation 23: Fit Logistic Regression Model for Treatment Failure After Five Years.

Similar to the two- and three-year models, higher age is yet again a significant protective predictor of treatment failure. For every ten years of higher age, the odds of treatment failure after five years decrease by factor 0.74 (p = 0.022). No other predictors reached significance. However, in this model, previous treatments are a *potential* risk factor (OR: 2.1, p = 0.059).

3.5.4.3 Model Validation

A comparative regression model was fit based on the same set of predictors but without imputation, based on the 83 complete cases. The regression estimates and additional measures for the goodness-of-fit are detailed in table 25.

Table 25: Complete Case Regression Estimates for Treatment Failure After Five Years.

Parameter	β	**SE** β	**Odds Ratio** e^{β} **(95 % CI)**	p
Intercept	0.60	1.13	NA	NA
Pre-treatment	0.750	0.516	NA	0.146
Sex	-0.149	0.510	NA	0.771
Age / 10	-0.291	0.164	0.748 (0.535 – 1.024)	0.077 +
IRP / 5	-0.022	0.108	NA	0.837
Achalasia type II	-0.360	0.681	NA	0.597
Achalasia type III	0.438	0.753	NA	0.560

Deviance: 100 (76 df), null deviance: 107 (82 df), AIC: 114, null AIC: 109.

R_N^2: 0.115, concordance: 0.684, D_{xy}: 0.368, γ: 0.435, τ_a: 0.169. Hosmer-Lemeshow test: χ^2 = 10.2 (8 df, p = 0.253).

*+ $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$. CI: confidence interval, NA: not assessed, SE: standard error.*

Results and implications will be discussed in chapter 4.

3.6 Cox Proportional Hazards Regression

3.6.1 Regression Model

A multiply imputed Cox regression model was fit using the same parameters as described for logistic regression in table 19, only without the residual intercept β_0. As explained in chapter 2.2.5.3, H estimates the *hazard* after t months under the assumption of *proportional hazards*:

$$H(t) = \lambda(t)\, e^{\beta_1\, pt + \beta_2\, sex + \beta_3\, age + \beta_4\, irp + \beta_5\, at_2 + \beta_6\, at_3}$$

Equation 24: Cox Proportional Hazards Regression Model.

3.6.2 Data Analysis and Imputation

The entire data, which comprised 374 records, were included in the model. As it was shown in chapter 3.3.1, the IRP is not normally distributed. *Little's test* was therefore omitted. *Hawkins' test* did not reach significance (p = 0.875). Thus, evidence against MCAR is weak, and preliminary requirements for imputation were therefore satisfied.

A total of 79 out of 374 records contained missing IRP observations (21.1 %, mean: 26.3 mmHg). All other records were complete. Conclusively, 25 data sets with imputed IRP values were calculated by predictive mean matching. The densities of the imputations are shown in figure 56.

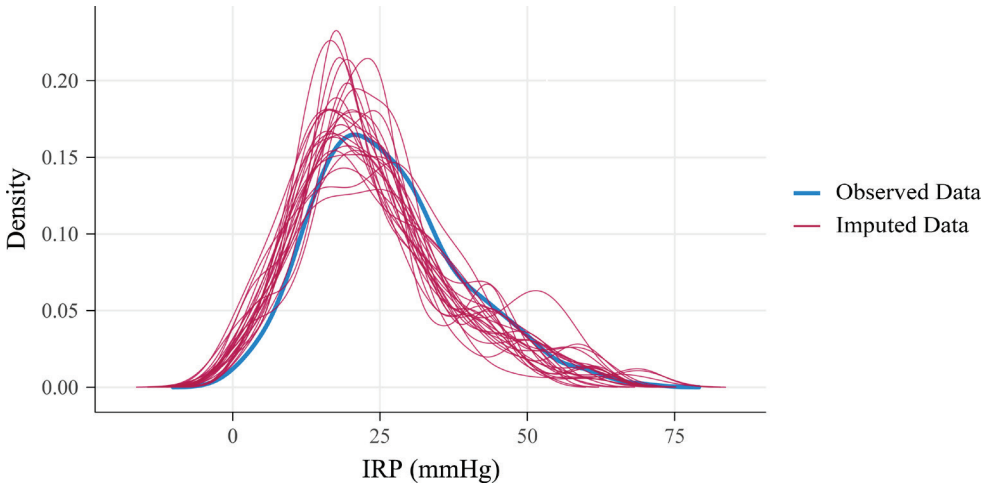


Figure 56: Imputed Baseline IRP Data of the Cox Regression Model.
Each red line represents the density of one imputed data set.

The densities of the imputed data match the density curve of the observed data well. This indicates a plausible distribution of imputed values and thus an overall good imputation model. The imputations show a slight trend towards lower IRPs, though. This will be addressed in chapter 4.

3.6.3 Regression Estimates

The regression estimates of the multiply imputed Cox regression model for the prediction of treatment failure are reported in table 26 and visualized in figure 57.

Table 26: Pooled Cox Regression Estimates for Treatment Failure.

Parameter	β	SE β	Hazard Ratio e^{β} (95 % CI)	p
Pre-treatment	0.414	0.192	1.51 (1.04 – 2.21)	**0.031** *
Sex	0.156	0.178	NA	0.382
Age / 10	-0.217	0.060	0.805 (0.716 – 0.906)	**< 0.001** ***
IRP / 5	-0.102	0.049	0.903 (0.820 – 0.994)	**0.038** *
Achalasia type II	-0.152	0.215	NA	0.480
Achalasia type III	0.372	0.279	NA	0.182

Mean concordance: 0.653 ± 0.03 (mean SE).

*+ $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$. CI: confidence interval, NA: not assessed, SE: standard error.*

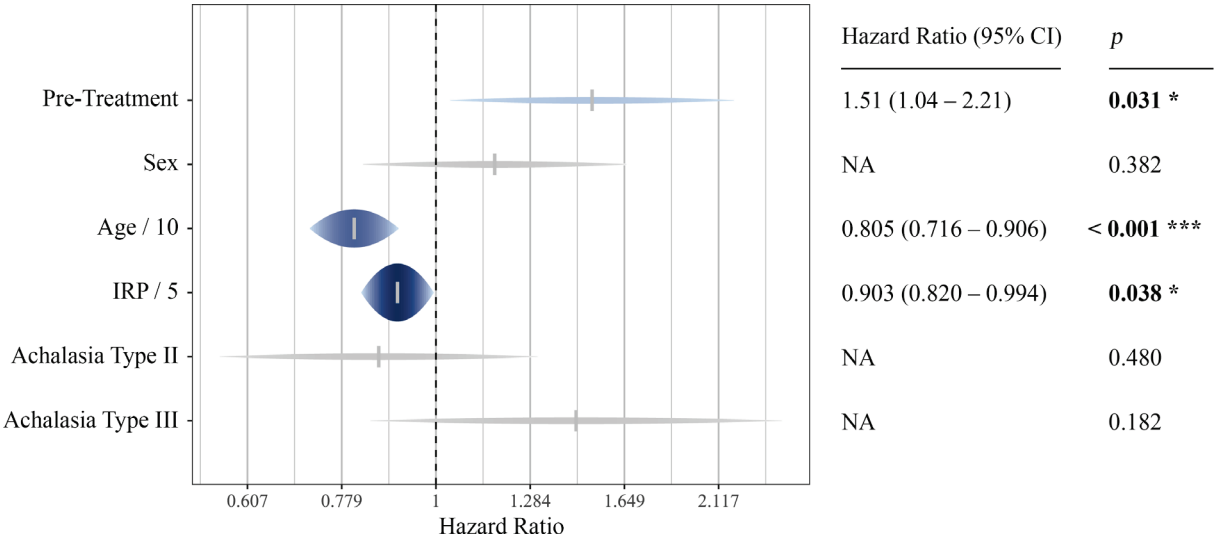


Figure 57: Cox Regression Estimates for Treatment Failure. Vertical bars: hazard ratios, raindrop width: 95 % CI, raindrop height and color intensity: likelihood of the respective hazard ratio within the CI. No hazard ratios are given for insignificant predictors, and their raindrops are grayed out. + $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$. CI: confidence interval, NA: not assessed.

The estimated regression equation with all predictors as defined in table 19, except that there is no residual intercept, has the following form:

$$H(t) = \lambda(t)\, e^{0.414\, pt + 0.156\, sex - 0.022\, age - 0.020\, irp - 0.152\, at_2 + 0.372\, at_3}$$

Equation 25: Fit Cox Proportional Hazards Regression Model.

For every ten years of higher age, a patient's hazard decreases by factor 0.81 ($p < 0.001$). This conforms to the logistic regression models that were discussed in chapter 3.5. For every 5 mmHg of higher baseline IRP, the hazard decreases by factor 0.90 ($p = 0.038$). Ultimately, for previously treated patients, the hazard increases by factor 1.5 ($p = 0.031$). No other predictors reached significance in the Cox model.

3.6.4 Assessment of the Proportional Hazards Assumption

For each of the Cox regression models built upon the 25 imputed data sets, its *Schoenfeld test's* result is shown in table 27. The 25 models were then pooled into the final model reported in the previous section.

Table 27: Global Schoenfeld Test Results for the Unpooled Imputed Cox Regression Models.

#	*p*	#	*p*	#	*p*	#	*p*	#	*p*
1	0.737	**6**	0.746	**11**	0.558	**16**	0.540	**21**	0.815
2	0.792	**7**	0.696	**12**	0.705	**17**	0.838	**22**	0.860
3	0.807	**8**	0.683	**13**	0.678	**18**	0.656	**23**	0.437
4	0.505	**9**	0.865	**14**	0.312	**19**	0.378	**24**	0.718
5	0.828	**10**	0.760	**15**	0.685	**20**	0.575	**25**	0.688

No *Schoenfeld test* of any imputed Cox model reached significance. This suggests that the proportional hazards assumption holds true for all of them.

3.6.5 Model Validation

A comparative Cox regression model was fit with the same set of predictors, but without imputation, based on the 295 completely observed cases. The complete case model's regression estimates and its *log-rank test* result are given in table 28.

Table 28: Complete Case Cox Regression Estimates for Treatment Failure.

Parameter	β	SE β	Hazard Ratio e^{β} (95 % CI)	*p*
Pre-treatment	0.446	0.223	1.56 (1.01 – 2.42)	**0.046 ***
Sex	0.175	0.210	NA	0.405
Age / 10	-0.249	0.071	0.780 (0.679 – 0.895)	**< 0.001 *****
IRP / 5	-0.112	0.049	0.894 (0.813 – 0.983)	**0.021 ***
Achalasia type II	0.081	0.269	NA	0.762
Achalasia type III	0.571	0.330	1.77 (0.93 – 3.38)	0.084 +

Concordance: 0.658 ± 0.03 (SE), log-rank test: $\chi^2 = 31.7$ (on 6 df, $p < 0.001$).

*+ $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$. CI: confidence interval, NA: not assessed, SE: standard error.*

Results and implications will be discussed in chapter 4.

4 Discussion

4.1 Primary Results

As already briefly mentioned in chapter 1, being introduced in 2010, POEM is still quite a novel approach to the treatment of achalasia. Even though more and more studies on the outcome after POEM are recently being published, the distinct influence of prior treatments remains a virtually uncharted field of research. By contrast, the older methods, such as balloon dilatation and Heller myotomy, look back on decades-long histories. For this reason, many patients who may consider POEM as a treatment tend to have already undergone one or multiple of these older procedures in the past. It is therefore of vital interest to study the influence of these traditional treatments on the expected outcome after POEM to be able to deliver expert advice when counseling these patients. For this sake, it was decided to choose treatment effects as the primary topic of this thesis, and prior treatments as its main predictor of interest.

4.1.1 Treatment Effects

The primary statistical results of this thesis are reported in detail in chapters 3.2, 3.5 and 3.6, and in particular in the tables 8, 20, 22, 24, and 26. Their interpretations and limitations will now be discussed.

Monovariate comparisons using *exact Fisher tests* revealed that treatment-naïve patients, compared to pre-treated patients, experience a failure-free survival rate of 97.0 % vs. 92.4 % after three months ($p = 0.082$), 91.2 % vs. 83.3 % after six months ($p = 0.051$), 84.8 % vs. 75.0 % after one year ($p = 0.043$), 81.6 % vs. 74.3 % after two years ($p = 0.132$), 76.0 % vs. 64.0 % after three years ($p = 0.047$), and 68.3 % vs. 51.4 % after five years ($p = 0.053$). Thus, pre-treated patients present unanimously worse outcomes among all follow-ups. Except after two years, this difference is always either significant, or it at least trends toward significance in the sense that its p value just slightly fails to undercut α.

Multiply imputed logistic regression models were fit for the prediction of post-interventional treatment failure after two, three, and five years. They include the independent predictors *pre-treatment*, *sex*, *age*, *IRP*, and *achalasia type,* as defined in table 19 in chapter 3.5. The two-year model identifies a higher age as a protective factor (OR per 10 years: 0.66, $p < 0.001$) and type III achalasia as a risk factor (OR: 2.5, $p = 0.048$). The three-year model identifies a higher age as a protective factor as well (OR per 10 years: 0.72, $p = 0.002$), as does the five-year model (OR per 10 years: 0.74, $p = 0.022$). The two- and three-year models do not confirm prior treatments as a risk factor, though, whereas the five-year model again hints at them being a risk factor by only missing the threshold for significance by a narrow margin (OR: 2.1, $p = 0.059$).

To supplement these findings and to further research the covariant relationships among the predictors suggested by the previously described analyses, a multiply imputed Cox proportional hazards regression model with the same set of predictors was fit for treatment failure after POEM. It identifies previous treatments as a risk factor (HR: 1.5, $p = 0.046$), higher age yet again as a protective factor (HR per 10 years: 0.78, $p < 0.001$), and a higher IRP as another protective factor (HR per 5 mmHg: 0.89, $p = 0.021$).

4.1.2 Model Differences and Conflicting Results

The results of the monovariate comparisons, the logistic regression models, and the multivariate Cox model diverge considerably. The monovariate between-group comparisons using *exact Fisher tests* implicate a higher rate of treatment failure among pre-treated patients compared to treatment-naïve patients, which falls in line with the results of the Cox regression model. However, supporting evidence from the logistic regression models is rather weak. These divisive findings are not unexpected, as will be clarified later in this section. They might be caused by a variety of different factors.

Monovariate analyses provide mere descriptions of between-group differences among the observed treatment failure rates. They are based on very little information and do not consider the influence of potential confounding factors. As such, they cannot describe nor attribute causality. In the chapters 3.1.1 and 3.3.2, pre-treated patients were found to be significantly older than treatment-naïve patients. Considering that age was then identified as a protective factor in every single multivariate model, this age-discrepancy alone might have biased the results of these basic statistical tests. Their results can therefore only serve as rough indicators of significantly disproportionate failure rates between the two groups.

The logistic regression estimates of course suffer from a limited number of patients and a varying set of observations included in each model. As emphasized many times before, the models for the prediction of treatment failure after three and five years obviously needed to exclude patients who underwent POEM less than three and five years ago, respectively. Contrary to that, the Cox proportional hazards regression model utilizes the combined data of all patients acquired during all follow-ups. It can therefore be considered the most robust model built upon the least incomplete data set. As such, it is no surprise that it identifies the highest number of significant predictors at once. However, the logistic regression models as point-in-time analyses might yet be able to unveil correlations that remain undetected in the Cox model. Since Cox models time-to-event data, it is blind to subsequent developments of patients *after* an event occurred. Furthermore, in many study designs, patients who experience an event are not registered until the next pre-determined follow-up time. Thus, the Cox model may overestimate survival times unless exhaustive effort is put into the gathering of the actual event times. This is often unrealistic in many clinical settings, and especially in retrospective studies such as this thesis. Last but not least,

the stratification by time, which is inherent to the Cox model, might obscure vague effects in the data that become observable only at certain moments throughout the follow-up. The clinical relevance of such effects may well be open to debate. However, they may still provide valuable information for the conceptualization of future studies, and especially for potential optimizations to the model selection process. Thus, logistic regression models and Cox regression models can in fact complement each other and may yield benefits when looked at in conjunction. In the case of this study, some of the logistic regression models hint at type II achalasia, compared to type I, being a potential protective factor for treatment failure, as well as type III being a potential risk factor. Even if the overall evidence provided for these assumptions remains weak based on the data currently available, these are both effects utterly invisible in the Cox model. And yet, scientists might want to consider them in their upcoming research. In a nutshell, the inclusion of the logistic regression models as complements to the Cox model yielded additional insight and was therefore well worth the additional effort.

Overall, there are no actually conflicting findings among the models. The statistically most extensive and best-substantiated Cox model identifies significant predictors that are compliant with the results of the three point-in-time logistic regression models. The latter were not capable of identifying all the same predictors in every model, and they even identify some additional predictors at specific points in time. However, no straight out contradicting effect size estimates exist between any two models.

From a methodical perspective, it should be noted that logistic regression aims to predict the odds, whereas Cox regression of course intends to predict the hazard. Changes in odds ratios and changes in hazard ratios often being interpreted in a quite similar fashion may distract from the fact that the odds and the hazard are fundamentally different concepts, as explained in chapters 2.2.5.2 and 2.2.5.3. As such, logistic regression and Cox regression *should* be expected to show some variance in their results, even more so when based upon a limited or even varying pool of observations. The fact that all models still essentially agree without major contradictions may be considered evidence for a robust statistical foundation in all models and substantial predictive powers. Future studies will likely unveil a higher grade of convergence between these different model types as the numbers of included patients grows.

4.1.3 Comparative Literature Review

4.1.3.1 Previous Treatments as a Risk Factor

Since the inception of POEM in 2010, only a few studies have been published on the distinct question of whether and how prior treatments influence the treatment outcome. They are outlined in table 29. A few additional papers, which compare the outcomes of treatment-naïve patients with patients pre-treated

with Heller myotomy, are not referred to here. They were not considered because Heller myotomy was distinctly excluded from this study. The referenced publications were all that could be found by searching for *"POEM"* in the PubMed database as of May 2020. The search yielded 1,436 results. They were manually assessed by their titles and, if deemed relevant to the topic of this thesis, by their abstracts. Where possible, table 29 gives the relevant data separately for treatment-naïve and pre-treated patients. If it does not provide distinct values for both groups, the respective study does not disclose such data.

Table 29: Publications on the Effects of Previous Treatments on the Outcome after POEM.

Study	Follow-Up (mean years)	Patients (*n*), % treatment-naïve	Males (%), Age (Mean Years): treatment-naïve / pre-treated	Treatment Success (%): treatment-naïve / pre-treated
Yeniova et al. (2020)	0.5	209, 54.0	47.8, 43.2 / 45.8, 44.2	94.6 / 94.7, $p = 0.978$
Zou et al. (2020)	0.5 – 1	43, 55.8	58.3, 28 $\tilde{x}$ / 47.4, 36 $\tilde{x}$	100 / 94.7, $p = 0.442$
Liu et al. (2019)	1.9 $\tilde{x}$	849, 71.1	48.2, 38 $\tilde{x}$ / 53.9, 38 $\tilde{x}$	1 y: 95.0 / 88.6, $p = 0.001$ 2 y: 93.5 / 86.5, $p = 0.001$ 5 y: 91.7 / 82.0, $p < 0.001$
Li et al. (2018)	4.1 $\tilde{x}$	564, 65.8	48.6, 38 $\tilde{x}$	1 y: 94.2 2 y: 92.2 3 y: 91.1 4 y: 88.6 5 y: 87.1
Nabi et al. (2018)	1.9 $\tilde{x}$	502, 51.8	54.6, 38.0 / 56.6, 42.4	6 m: 92.4 / 92.5, $p = 0.95$ 1 y: 90.7 / 91.2, $p =$ NA 2 y: 87.5 / 84.2, $p =$ NA 3 y: 87.1 / 76.3, $p =$ NA
Louie et al. (2017)	0.5 $\tilde{x}$	38, 50.0	57.9, 58 $\tilde{x}$ / 47.4, NA	NA
Jones et al. (2016)	0.8 $\tilde{x}$	45, 66.7	66.7, 46.2 / 60.0, 64.4	NA
Orenstein et al. (2015)	0.8 $\tilde{x}$	40, 60.0	NA	NA
Ling et al. (2014)	1	51, 58.8	33.3, 42.5 / 38.1, 43.2	87.5
Sharata et al. (2013)	0.5	40, 70.0	42.8, 48 $\tilde{x}$ / 41.7, 55 $\tilde{x}$	100

If no distinct values are given for pre-treated and treatment-naïve patients, the cited study only provides an overall value for both groups. NA: not assessed or not reported in the study, $\tilde{x}$: median instead of the mean.

Even though the insight gained from these studies is very valuable, many of the publications have a restricted scope and offer reasons for methodical critique.

- Except for Li et al. (2018), Liu et al. (2019), and Nabi et al. (2018), the follow-up durations of each study is limited to mostly less than a full year after POEM.
- Excluding Li et al. (2018), Liu et al. (2019), Nabi et al. (2018), and Yeniova et al. (2020), each study comprises only about 50 patients or fewer.

- Jones et al. (2016) use a Likert scale for dysphagia severity instead of the well-established Eckardt score. Also, they only assess reflux based on a questionnaire, as does Orenstein et al. (2015).
- Louie et al. (2017) divide the patients into three groups instead of two: treatment-naïve patients, patients with simple pre-treatments, and patients with complex pre-treatments. Among others, they consider dilatations with a balloon diameter below 30 mm and botulinum toxin injections as simple. Examples for complex pre-treatments in their proposed classification are surgery and dilatations with a diameter of 30 mm or above.
- As usual, treatment success is commonly defined as a post-POEM Eckardt score below 4 among all major studies. There are a few exceptions, though. Zou et al. (2020) consider an Eckardt score above 3 still as treatment success if it is 3 points or more below the pre-POEM baseline score of the same patient. Jones et al. (2016), Louie et al. (2017), and Orenstein et al. (2015) do not provide treatment success data that is based on the Eckardt score at all.

Especially the complex pre-treatment classification system introduced by Louie et al. (2017) is indeed an interesting and novel approach that appears worthy of further exploration in the future. However, the overall heterogeneity between their study and most others renders their results difficult to compare.

The publications by Zou et al. (2020), Louie et al. (2017), Jones et al. (2016), Orenstein et al. (2015), Ling et al. (2014), and Sharata et al. (2013) provide very limited informative value due to their short follow-up durations, small patient numbers, and at least partially disparate study designs. The following discussion will therefore focus on an in-depth exploration of the findings reported by Yeniova et al. (2020), Liu et al. (2019), Li et al. (2018), and Nabi et al. (2018).

Yeniova et al. (2020) report treatment success in 94.6 % of pre-treated patients and 94.7 % of treatment-naïve patients six months after POEM. Consequentially, they argue that there is no observable short-term difference between both groups, thus dissenting from the nigh-significant findings of this thesis for the same follow-up period (91.2 % vs. 83.3 %, $p = 0.051$; see chapter 3.2). With 209 included patients, their study is one of the very few on the topic with a sufficiently large patient pool. Unfortunately, the extremely short follow-up duration of only six months lends it a very restricted scope. Their methodology is limited to basic monovariate statistics in the like of χ^2 *tests* and *exact Fisher tests*. Also, non-achalasia motility disorders make up about 10 % of their patients in each group, which might have biased their results. Future studies with higher numbers of included patients and, especially, a longer follow-up duration might provide results that are better comparable to this thesis.

Liu et al. (2019) published the most extensive study on the topic to date. It is based on 849 patients that were followed-up on over a proper median duration of 23 months. The authors report treatment success in 88.9 % of all patients, and a reflux prevalence of 23.9 %. The Cox regression model they provide identifies prior treatments as a significant risk factor for treatment failure (HR: 1.9, $p = 0.002$). This coincides with this thesis' Cox regression model, although the latter estimates the hazard ratio to be slightly lower (1.5, $p = 0.031$, see chapter 3.6). Interestingly, logistic regression models for the prediction of major adverse events and clinical reflux that were also fit by Liu et al. (2019) do not show any significant effect of prior theories. On a general note, their follow-up reaches up to five years after POEM, but data after two years appears to be quite limited. A major weakness of their study is the preliminary exclusion of 535 of their initially treated 1384 patients. Thus, they removed 38.7 % of all patients without providing an analysis of missing information. Their statistical procedures might therefore be selection biased. Also, no assessment of the goodness of fit is reported for any of their models. Despite these eye-catching limitations, Liu et al. (2019) still provide one of the very few available studies to incorporate regression analysis *at all*. Ultimately, their findings confirm the findings of this thesis.

Li et al. (2018) report a multiply imputed Cox regression analysis based on 564 patients with a median follow-up duration of 49 months. Their study does not focus specifically on the influence of prior treatments on the clinical outcome. However, it still reveals very interesting results on this topic. Their model includes many predictors, amongst others *age* (above or below 60 years), *sex*, *disease duration* (above or below 10 years), *sigmoid esophagus* (yes or no), *pre-POEM Eckardt score* (above 7 or below), and *prior treatments* (yes or no). It identifies a disease duration of 10 years or longer as a significant risk factor for treatment failure (HR: 2.5, $p < 0.01$), as well as previous treatments (HR: 1.1, $p = 0.02$). The latter once again coincides with the findings of this thesis, which estimates the hazard ratio to be a bit higher (HR: 1.5, $p = 0.031$). With its well-founded study design and its ambitious statistical approach, their publication stands out from many others. It is yet susceptible to criticism, for it provides neither an assessment of missingness patterns prior to imputation, nor does it validate the imputation quality or the final regression model. Besides that, Li et al. (2018) provide one of the most substantial studies to date in terms of a large number of included patients, a decent follow-up duration, and with the most sophisticated statistics among all publications discussed in this chapter. As such, its findings regarding the influence of prior treatments on the outcome after POEM provide strong evidence in support of this thesis' primary result. Interestingly, they find the disease duration also to be a significant risk factor – a predictor not included in this thesis' regression models. Its evaluation as an additional parameter might be considered for future studies. Also, they do not find age to be a significant predictor at all. This is quite surprising considering that age turned out to be a unanimously significant parameter among all models reported by this thesis. The implications will be further discussed in the next chapter.

Ultimately, Nabi et al. (2018) report a multivariate regression analysis built upon data acquired from 502 patients during a median follow-up time of 1.9 years. Even though their model predicts the operation time, not treatment failure, data on treatment success is still provided. It is found to be similar between treatment-naïve and pre-treated patients. Overall treatment success is observed in 90.9 %, 86.0 %, and 81.2 % of patients after one, two, and three years, respectively. Treatment-naïve patients are reported to show a six-month treatment success rate of 92.4 %, compared to 92.5 % in pre-treated patients ($p = 0.95$). Additional treatment success rates are depicted as 90.7 % vs. 91.2 % after one year, 87.5 % vs. 84.2 % after two years, and 87.1 % vs. 76.3 % after three years. Unfortunately, no statistical tests are provided for these later follow-ups. However, even without objective evidence, the reported fractions quite strongly imply that pre-treated patients indeed experience long-term treatment failure more frequently than treatment-naïve patients. This would in fact confirm the results of this thesis. Another limitation of their study is that only 69 patients provided follow-up data after more than two years. Also, it provides a list of exclusion criteria, yet it gives no information about how many patients were subsequently removed. Therefore, the risk for selection bias cannot be judged. Compared to Nabi et al. (2018), the patients observed in this thesis show lower treatment success rates at all respective follow-up times. Also, in this study, pre-treated patients usually fare much worse than treatment-naïve patients (1 year: 84.8 % vs. 75.0 %, $p = 0.043$; 2 years: 81.6 % vs. 74.3 %, $p = 0.132$; 3 years: 76.0 % vs. 64.0 %, $p = 0.047$; see chapter 3.2). The cause of these early follow-up differences remains to be determined.

Overall, this thesis' results align well with the only two other studies with comparable scopes and statistical foundations published to date: it confirms pre-treatment as a risk factor with a hazard ratio of 1.5 ($p = 0.031$), which blends in perfectly right in the middle between the hazard ratio of 1.1 ($p = 0.02$) reported by Li et al. (2018) and the hazard ratio of 1.9 ($p = 0.002$) reported by Liu et al. (2019). Yeniova et al. (2020) find no influence of previous treatments on the outcome after POEM, but their follow-up duration of six months seems way too short to tell. Nabi et al. (2018), unfortunately, do not report statistical test results for the comparison between treatment-naïve and pre-treated patients after more than six months. However, the numbers they provide for treatment success after three years are highly suggestive of also showing a better outcome for treatment-naïve patients compared to pre-treated patients (87.1 % vs. 76.3 %). All other available publications on the topic are conceptually incapable of assessing middle- or long-term predictors of treatment failure due to either too few included patients or too short follow-up durations. Ultimately, the only two large-scale studies available to date unanimously identify previous treatments as a significant risk factor for treatment failure – just like this thesis. Thus, evidence is strong in support of this thesis' primary results.

4.1.3.2 Age as a Protective Factor

As reported in chapters 3.5 and 3.6, age was identified as a protective factor in every statistical model provided by this thesis. The multiply imputed two-, three- and five-year logistic regression models estimate the odds ratios for treatment failure per 10 years of higher age to be 0.64 ($p < 0.001$), 0.72 ($p = 0.002$), and 0.74 ($p = 0.022$), respectively. The multiply imputed Cox regression model yields a hazard ratio per 10 years of higher age of 0.81 ($p < 0.001$). Overall, these results provide strong evidence for younger patients having a significantly higher risk of treatment failure after POEM. This is an interesting and, to the knowledge of the author of this thesis, unprecedented observation.

Liu et al. (2020) report no evidence for a significant influence of age on the treatment outcome in 849 patients during a median follow-up of 1.9 years. In a preliminary univariate Cox regression analysis, which precedes their multivariate model discussed in the previous section, they report a hazard ratio of 1.1 for patients of 60 years or older ($p = 0.89$). In their multivariate model, age is then no more included as a predictor. Li et al. (2018) find no evidence for an influence of the patients' age as well in their multivariate Cox regression model based on 567 patients and a median follow-up of 4.1 years. They report a hazard ratio of 1.2 ($p = 0.67$) for patients that are 60 years or older. The findings by both studies, however, do not necessarily contradict the results of this thesis. The reason for that is that they both remodeled the continuous variable *age* as a binary category. Such a procedure always involves a loss of discrimination. It might have reduced the information in their data to such an extent that it obfuscated a potential effect of the patients' younger age. The findings of this thesis suggest that a higher risk of treatment failure manifests particularly in patients younger than 40 years (see chapter 3.2.1, in particular figure 28). If similar correlations existed in the data of the studies provided by Liu et al. (2020) and Li et al. (2018), they might well have been obscured by the reduction of *age* to a binary category based around a threshold of 60 years.

No other distinct studies on the influence of *age* on the outcome after POEM were found. It remains unclear why younger patients respond worse to POEM in this thesis. One may theorize that the circular muscle layer in older patients might be less prone to react with excessive scarring after being cut, maybe due to reduced regenerative capabilities or a higher degree of atrophy. It is also possible that young patients might perceive their illness as more menacing than older patients. The Eckardt score is commonly utilized to determine treatment failure, yet it is a self-assessment scale based on subjective symptoms reported by the patients themselves. As such, the score might either facilitate an exaggeration of symptoms by young patients, or an understatement by the elderly. Supplementary studies will be required to garner further evidence in favor of these hypotheses, or against them.

4.1.3.3 Influences of the Achalasia Types

Type III achalasia was identified as a significant risk factor in the logistic regression model for treatment failure after two years (OR: 2.5, $p = 0.048$). In contrast, type II shows a vague trend toward being a protective factor in the three-year model (OR: 0.52, $p = 0.083$). It becomes insignificant after five years (OR: 0.49, $p = 0.108$). No significant effect of any achalasia type was found in the Cox regression model.

There is still little consensus on if and to what degree the distinct achalasia types take a toll on the treatment success after POEM. Greene et al. (2015) argue against such correlation existing, but they provide limited evidence with only 49 included patients and a median follow-up of just 16 months. Zheng et al. (2019) compare the outcomes of type I achalasia with those of type II. They find no difference in the clinical response as well. Yet again, their study has similar restrictions since it only includes 40 patients and follow-ups until one year after POEM. Louie et al. (2017) divide their pre-treated patients into two groups: “simple” and “complex” achalasia. They utilize quite complicated criteria for the group assignment, such as prior treatments, balloon diameters, and the shape of the esophagus (i.e. straight or sigmoid). While they report “complex” achalasia to present with more severe pre-operative dysphagia and to require a significantly longer operation time, dysphagia after POEM does not seem to differ between the compared groups. This is a unique approach that imposes difficulties when trying to compare their results to other studies and this thesis. Again, only 38 patients were included in their study, and the median follow-up was limited to six months. Li et al. (2018) include the esophagus’ shape as a predictor in their statistical models as well, which are based on 464 patients and a median follow-up duration of 4.1 years. While they identify a sigmoid esophagus as a significant risk factor in a univariate Cox regression model (HR: 2.4, $p = 0.03$), it is insignificant in a subsequently fit multivariate model. However, even if their results were significant, the shape of the esophagus as an indicator of a late disease stage cannot be expected to translate well to the Chicago classification’s achalasia types. Recently, Podboy et al. (2020) identified type III achalasia as a risk factor for treatment failure in a monovariate Cox regression model based on 98 patients and a mean follow-up of 3.9 years (HR: 2.3, $p = 0.029$). Once again, the predictor becomes insignificant in a multivariate follow-up model. Interestingly, though, type II achalasia shows a very vague trend toward being a protective factor in their study (HR: 0.50, $p = 0.116$).

In a systematic meta-analysis by Pandolfino and Gawron (2015), which includes 93 articles and 734 patients, they report type II achalasia to generally show the best treatment response, followed by type I, and both far ahead of type III. This matches the results of this thesis. However, their meta-analysis includes patients treated with either POEM, surgical myotomy, or balloon dilatations. Therefore, the results are once again difficult to interpret.

Overall, the influence of the achalasia type on treatment failure after POEM remains to be determined. There are some vague hints at type II faring better than type I. Some more conclusive evidence exists for type III to respond worst to POEM among the three types. Such trends are consistent with the findings of this thesis. However, the overall evidence in the body of literature remains rather weak.

4.2 Secondary Results

4.2.1 Eckardt Scores

The Eckardt score before and after POEM is depicted in the chapters 3.1.1 and 3.2.2. Before POEM, treatment-naïve patients show significantly higher Eckardt scores compared to pre-treated patients (mean: 6.9 vs. 6.4, $p = 0.034$). Among both groups, the most severe component by far is the dysphagia score (mean: 2.7 vs. 2.6, $p = 0.161$). The higher baseline Eckardt scores seen in treatment-naïve patients can mostly be attributed to a higher weight-loss component (mean: 1.3 vs. 0.9, $p = 0.006$). If weight loss is ignored, the mean Eckardt score is nigh on similar in both groups. Regarding regurgitations and retrosternal pain, there are no significant differences either.

After POEM, the observed Eckardt scores decrease severely among both groups. Throughout the entire follow-up, previously treated patients consistently show higher Eckardt scores compared to treatment-naïve patients (see chapter 3.2, table 9 and figure 30). These differences reach significance at six months and five years after POEM, and they just slightly miss the threshold at three years. At three and five years, treatment-naïve and pre-treated patients report mean Eckardt scores of 2.1 vs. 2.6 ($p = 0.052$), and 2.0 vs. 3.0 ($p = 0.006$), respectively. Thus, the differences between the means of the groups have returned to quite impressive values of 0.5 and 1 score points, respectively. The *absolute* score values are still far below the pre-POEM baseline scores. However, the score *differences* between both groups have largely surpassed the baseline difference by a large margin. This indicates a diverging long-term score rebound at the expense of previously treated patients. As time goes by, both groups show a resurgence of symptoms, but it is significantly more severe in pre-treated patients.

At each follow-up, the dysphagia component contributes by far the most to the overall Eckardt score. Reaching a mean score of 0.9 in treatment-naïve patients and 1.2 in pre-treated patients after five years ($p = 0.062$), dysphagia is also the only score component that shows a distinct post-interventional regrowth over time. Interestingly, even though weight loss is a major discriminatory component before POEM, it plays no relevant role in any follow-up of either group. Its mean never grows back to above 0.2 score points. Regurgitations and retrosternal pain generally contribute much less to the overall

Eckardt score. With their means fluctuating between 0.4 and 0.7 score points among both groups at each follow-up, they are usually quite evenly matched.

Conclusively, these observations hint at a possible translational problem with the Eckardt score as the tool of choice for the assessment of post-POEM treatment failure. Each of its four components – dysphagia, regurgitations, retrosternal pain, and weight loss – contributes up to 3 points to the overall score. As such, they are evenly weighted. Yet, the discriminatory contribution of dysphagia seems to be disproportionally strong, whereas weight loss appears to be negligible. These remarks are not intended to discredit the Eckardt score, especially since it is the only established scale for the classification of achalasia severity. However, weight loss seems to be a relevant factor prior to POEM, but not so much thereafter. Thus, the Eckardt score might not be as equally suitable for the discrimination between post-operative treatment response as it is for the initial assessment of pre-operative achalasia severity. It may therefore be questioned whether it is ideal to use the same threshold of 4 score points for both causes. To the knowledge of the author of this thesis, this issue has not been addressed in a large-scale study so far. Comparative Eckardt scores reported in the body of literature are summarized in table 30.

Table 30: Publications on the Eckardt Score Before and After POEM.

Study	Eckardt Score Before POEM (Mean): treatment-naïve / pre-treated	Eckardt Score After POEM (Mean): treatment-naïve / pre-treated	
Shiwaku et al. (2020)	6.1	1.1	after 1 year
Yeniova et al. (2020)	6.4 / 6.4	1.3 / 1.4	
Zou et al. (2020)	7 / 7 $\tilde{x}$	2 / 1 $\tilde{x}$	
Liu et al. (2019)	7 / 8 $\tilde{x}$	1.4 / 1.7	
Li et al. (2018)	8 $\tilde{x}$	2 $\tilde{x}$	
Nabi et al. (2018)	7.1 / 7.0	1.1 / 1.1	after 3 months
Inoue et al. (2015)	6 $\tilde{x}$	1 $\tilde{x}$	after 2 months
		1 $\tilde{x}$	after 1 – 2 years
		1 $\tilde{x}$	after 3 years
Ling et al. (2014)	7.3 / 6.8	0.5 / 0.7	
Sharata et al. (2013)	6 / 5 $\tilde{x}$	1 / 1 $\tilde{x}$	

If no distinct values are given for pre-treated and treatment-naïve patients, the cited study only provides an overall value for both groups. If no time is given for the post-POEM Eckardt score, no precise temporal data is provided by the cited study. $\tilde{x}$: median instead of the mean.

In general, there appears to be little difference between pre- and post-POEM Eckardt scores of treatment-naïve and pre-treated patients in these studies. Both groups respond well to treatment with a considerable decrease in score points. Unfortunately, the four individual score components are usually not reported. Also, only Inoue et al. (2015) provide repeated scores for subsequent follow-ups. In contrast to this thesis, their patients seem to experience no resurgence of symptoms up to the latest follow-up included in their study, which is three years after POEM. It needs to be noted, though, that the study reports

median scores, whereas this thesis uses the mean. Thus, the values are hard to compare. Overall, the pre- and post-POEM Eckardt scores reported in the literature comply with the observations of this thesis.

4.2.2 Reflux

The results of the reflux analysis three months after POEM are presented in chapter 3.2.3. With 62.4 % of treatment-naïve patients and 59.9 % of pre-treated patients showing signs of mucosal erosion, endoscopically evidenced gastroesophageal reflux disease is a severe problem after POEM. Fortunately, most patients seem only to develop mild forms. Among all patients positive for reflux disease after 3 months, treatment-naïve patients and pre-treated patients were classified as 57.6 % vs. 51.1 % LA grade A, 30.2 % vs. 39.4 % LA grade B, 0 % vs. 1.1 % LA grade C, and 4.7 % vs. 3.2 % LA grade D, respectively. When compared with *Fisher's exact test*, the overarching distributional differences between both groups are insignificant (p = 0.352). Endoscopically evidenced pre-POEM baseline reflux was negligible. It affected 2.6 % of treatment-naïve and 5.1 % of pre-treated patients (p = 0.280, see chapter 3.1, table 6). Thus, there is little evidence to suggest that these short-term reflux observations after POEM might have been biased by reflux that was already present before. In fact, the data suggest quite the opposite: most cases of reflux disease appear to have developed de-novo after POEM. It may be possible, though, that patients used to respond better to anti-reflux medication prior to the procedure. Thus, a final verdict cannot be passed.

Reflux incidences after two and five years are very high, ranging from about 50 % to 70 %. However, only 40.6 % and 32.4 % of patients underwent upper endoscopy after two and five years, respectively (see chapter 3.4.2). No reliable conclusions can be derived from such incomplete data. Therefore, no further analyses were conducted. Future studies may think about incentives to increase the patients' long-time follow-up compliance regarding upper endoscopy.

Liu et al. (2019) report endoscopically evidenced reflux disease in 17.3 % of treatment-naïve patients and in 22.8 % of pre-treated patients (p = 0.10). They performed endoscopic reflux surveillance on 664 patients over a median follow-up time of 23 months. Nabi et al. (2018) endoscopically assessed reflux in 247 patients one year after POEM. They diagnosed reflux disease in 22.1 % of treatment-naïve patients and in 20.7 % of previously treated patients (p = 0.88). Li et al. (2018) describe a reflux prevalence of 37.3 % over a median follow-up time of 49 months. Sanaka et al. (2020) compare post-POEM reflux of non-obese and obese patients. For these groups, they report a similar prevalence of reflux symptoms in 17.8 % vs. 20.0 % of patients two months after POEM, respectively. Finally, Shiwaku et al. (2020) assessed reflux in 1176 patients as a part of their large multi-center study. They report endoscopic

evidence of reflux disease in 63.1 % of all patients during the first six months after POEM. Figure 58 illustrates LA grade distributions of reflux disease provided in recent studies.

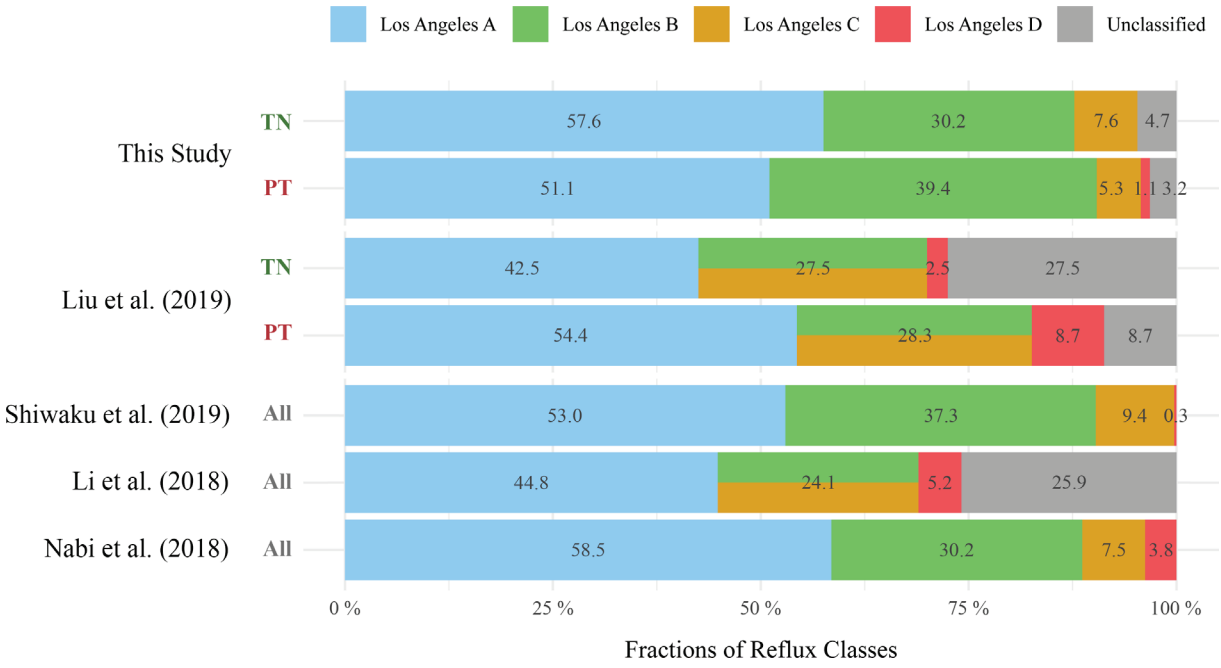


Figure 58: Los Angeles Grade Distributions of Post-POEM Reflux Disease in the Literature. The sum of fractions might not always add up to 100 % as a consequence of rounding. In three studies, the LA grades B and C were not differentiated. This is indicated by the two-colored bars. Some publications provide the LA grade fractions relative to the number of LA classified cases instead of relative to all patients. To allow for a better comparison, these percentages were recalculated to also respect the unclassified cases. TN: treatment-naïve, PT: pre-treated.

Comparing the post-POEM reflux prevalences provided by the body of literature is complicated by varying follow-up durations and by the different ways in which reflux is assessed and reported. Some authors utilize standardized questionnaires. Others rely on non-standardized patient-reported symptom assessment. Many studies incorporate combinations of reflux symptoms, upper endoscopy, and 24-hour pH measurement. Even more confusing, some publications report their reflux rates as “observed during follow-up”. This is too ambiguous and provides no clear information about how long after POEM reflux disease was actually diagnosed. Ultimately, the original LA classification system and the modified version, which is somewhat prevalent in Japan, use slightly different criteria (see chapter 1.7). This renders the reported reflux data of various studies even more difficult to compare. A standardized follow-up scheme would undoubtedly provide benefits for future studies.

This thesis focuses on endoscopically evidenced reflux disease. As far as there are tendencies detectable in the literature, reflux appears to be highly prevalent after POEM. Fortunately, as it was found in this thesis, most cases of reflux disease reported by other authors are rather mild in the sense that they correspond to the LA grades A or B. Grades C and D are rare. However, grade C and B are sometimes reported as one. Thus, their relevance remains slightly more nebulous.

Keeping in mind the severely limited comparability between the different studies, the overall tendencies regarding the reflux disease prevalence after POEM and the distinct LA grade distributions reported in the literature match well with the findings of this thesis. Additionally, no significant differences between treatment-naïve and previously treated patients regarding post-POEM reflux disease were found by this study, nor in recent literature.

4.2.3 IRPs

The results of the patients' IRP measurements before and after POEM are depicted in chapter 3.1, table 6, and in chapter 3.2.4, table 11. The baseline IRP distribution is depicted in chapter 3.3.1. Treatment-naïve patients show significantly higher IRPs before POEM compared to pre-treated patients (mean: 30.0 vs. 21.8 mmHg, $p < 0.001$). Furthermore, the baseline IRPs of treatment-naïve patients are normally distributed around their mean. In contrast, the distribution of the pre-treated patients is positively skewed, which means that observations accumulate below the mean. This is not an unexpected discovery: a higher IRP is an expression of pathologically increased LES contractility, of which a reduction is a consequence of prior treatments. Therefore, it is reasonable to expect lower IRPs in pre-treated patients.

Three months after POEM, the mean IRP has shrunk considerably, with no significant differences between treatment-naïve and pre-treated patients (11.0 vs. 9.6 mmHg, $p = 0.072$). Among both groups, most patients present a post-operative IRP of 15 mmHg or below. This indicates an overall favorable pressure reduction in most patients and is a tendency that appears to remain unchanged after two and five years. However, only 9.4 % and 8.5 % of eligible patients underwent high-resolution manometry again after two and five years, respectively (see chapter 3.4.2). Therefore, no reliable conclusions can be derived from the observations of these later follow-ups, and no further analyses were justified. The patients' tenacious reluctance to undergo follow-up manometry is certainly evoked by the discomfort the procedure tends to inflict upon them. Furthermore, the still fairly limited availability of high-resolution manometry might necessitate long and possibly cumbersome travels for many patients. Future studies may want to think about ways to improve the incentives for their patients to engage in these diagnostics. For now, no reliable conclusions on the long-term development of the IRP after POEM can be drawn.

There are little comparative reports available in the literature. Nearly all of the few studies published on this topic report the LES pressure rather than the newer IRP. Zou et al. (2020) assess 43 patients. Treatment-naïve and pre-treated patients are reported to show a mean IRP of 27 vs. 24 mmHg before POEM, respectively, which shrink to 4.5 vs. 5 (*sic*) mmHg after POEM. Yeniova et al. (2020) report the mean IRP of treatment-naïve vs. pre-treated patients as 31.2 vs. 20.3 mmHg before POEM, and 12.1 vs. 11.7 mmHg after six months. Their study comprises 209 patients. No statistical comparisons between the groups are provided by these publications. However, both studies depict a pronounced IRP decrease after POEM in both treatment-naïve and pre-treated patients to a similar level. This aligns well with the findings of this thesis and confirms that both groups respond equally well to POEM regarding a reduction of their IRPs. Interestingly, the baseline IRPs of both groups reported by Zou et al. (2020) are quite similar, even though treatment-naïve patients still seem to present slightly higher pressures. In contrast, Yeniova et al. (2020) report much higher baseline IRPs among treatment-naïve patients, which is the same finding as reported in this thesis. The overall evidence remains limited due to a lack of publications reporting on the IRP. Yet, it appears that the hints found in the literature fit the findings of this thesis rather well.

4.2.4 Re-Treatments

Re-treatments after POEM are rare among all patients. No significant differences regarding the fractions of patients who underwent another treatment could be found between treatment-naïve and pre-treated patients (9.4 % vs. 15.3 %, $p = 0.115$; see chapter 3.2.5). However, the latter still seem to trend toward a higher number of incidents. This coincides with pre-treated patients also experiencing treatment failure at a significantly higher rate (see chapter 4.1). Possibly, patients who already underwent a multitude of treatments in the past might have lower inhibitions to undergo yet another operation.

4.2.5 Peri-Operative Inflammation and Blood Loss

Peri-operative blood-loss is similar in treatment-naïve and pre-treated patients (see chapter 3.2.6). The markers of peri-operative inflammation (CRP and Leucocytes) provide no evidence for a significant difference between the groups as well. That is, with one exception: treatment-naïve patients have slightly higher pre-operative CRPs compared to pre-treated patients (mean: 9.5 vs. 6.1 mg/l, $p = 0.017$).

Further analysis revealed that the pre-POEM median CRP is 5 mg/l for both groups. The lower CRP detection limit was 5 mg/l. For technical reasons, patients with blood values below this threshold were always recorded as exactly 5 mg/l. Thus, the most obvious explanation is that more treatment-naïve than pre-treated patients were measured slightly above this threshold. However, one may also argue that a

marginally lower baseline CRP in previously treated patients could be an expression of reduced esophageal inflammation that possibly resulted from the previous treatment. This might seem a logical thought, considering that the measured CRPs reach up to 120 mg/l in treatment-naïve patients, whereas the highest observed CRP in pre-treated patients is 32 mg/l.

Overall, prior treatments seem to influence neither peri-operative blood loss nor system inflammation. Since peri-operative markers are not the focus of this thesis, though, no further analyses were conducted.

4.3 Comparative Demographics

The previous chapters provided explanations and interpretations of this thesis' research results. They were supplemented by in-depth literature comparisons that focused mainly on procedural and statistical aspects. This chapter will now follow up with an assessment of how well the patient population of this thesis matches the demographic of achalasia patients reported in the literature. Such analyses aid to detect, explain, or – ideally – rule out potential between-population variances that might bias the data.

Table 31 shows exemplary patient data assembled from the most extensive studies on POEM available. The table does not claim to be complete, but the publications were chosen to be as relevant as possible. Thus, only studies with a focus on POEM and a patient count above 200 where considered. In contrast to chapter 4.1, the studies where not required to assess outcome differences based on previous treatments. While there is a plethora of case series with usually less than 50 patients, large publications are still rare.

Table 31: The Most Extensive Studies on POEM.

Study	Patients (*n*)	Age (Mean)	Male Fraction (%)
Shiwaku et al. (2020)	1,346	47.2	45.8
Yeniova et al. (2020)	209	43.6	46.9
Liu et al. (2019)	849	38 $^{\tilde{x}}$	48.2
Feng et al. (2018)	568	43 $^{\tilde{x}}$	47.3
Li et al. (2018)	564	38 $^{\tilde{x}}$	48.6
Nabi et al. (2018)	502	40.1	55.6
Wu et al. (2017)	1,693	38 $^{\tilde{x}}$	49.5
Inoue et al. (2015)	500	43 $^{\tilde{x}}$	43.2

$\tilde{x}$: median instead of the mean.

4.3.1 Sex and Age

While treatment-naïve patients present a male fraction of 50.3 % and a mean age of 44.4 years in this thesis, pre-treated patients have a male fraction of 58.5 % and a mean age of 47.8 years (see chapter 3.1). Comparing this to the data presented in table 31 with no regards to prior treatments, the age and sex distributions of this thesis' patients blend in well with the data found in the body of literature.

The sexes are distributed evenly in the literature. The mean or median ages vary between about 38 and 47 years. This seems to correlate to a certain degree with the differences in the population distributions between the studies' respective countries of origin. As estimated by the United Nations (2019), the median age is higher in Japan, Korea, and Germany (48.4, 43.7, and 56.7 years), whereas it is lower in China and India (38.4 and 28.4 years). Local differences between the medical systems, the diagnostic evaluation processes, and the patients' desire or capability to afford treatment may play a role as well.

Interestingly, in this study, pre-treated patients were found to be on average about three years older than treatment-naïve patients (see chapter 3.1). This may be a consequence of the former having already undergone previous treatments and, thus, most likely more diagnostics, both of which take time.

4.3.2 Weight and Body Mass Index

This study found a mean BMI of 24.6 kg/m² in treatment-naïve patients, whereas it was 26.0 kg/m² in pre-treated patients. This difference is significant ($p = 0.013$, see chapter 3.1, table 6). Since BMIs are not mentioned frequently in the literature, smaller publications need to be assessed as references. Mean BMIs have been reported for treatment-naïve and pre-treated patients, respectively, by Yeniova et al. (2020) as 22.6 and 22.7 kg/m² ($n = 113$ and 96), by Jones et al. (2016) as 30.8 and 29 kg/m² ($n = 30$ and 15), and by Ling as 23.3 and 22.1 kg/m² ($n = 30$ and 21). Nabi et al. (2020) report a mean BMI of 22.2 kg/m² with no regard for prior treatments ($n = 209$).

Interestingly, most BMIs seem to trend slightly toward the upper cut-off value of normal weight. This is counterintuitive to weight loss being considered a cardinal symptom of achalasia, which is also reflected by its prominent role in the Eckardt score. Due to their inability to swallow, patients lose weight. Even if only temporarily, swallowing should get better after treatment and allow them to regain some of their lost mass. As observed in this thesis, one would therefore expect treatment-naïve patients to have lower BMIs than pre-treated patients. Yet, there appears to be no evidence for a significant weight difference between these groups in the literature. As discussed in chapter 4.2, weight loss is in fact a grave contributor to the pre-POEM Eckardt scores of both groups. This means that although many patients tend to lose weight due to their achalasia, underweight seems not to be a condition that is typically observed.

As stated before, the BMIs observed in this study show a clear trend toward overweight. Concordantly, the 45 patients described by Jones et al. (2016) are pre-obese to obese. This might be an expression of the populations in Germany and the United States generally trending toward higher weights. As it stands, the patients in this thesis seem to weigh more than the patients reported in most referenced studies. Sanaka et al. (2020) assessed whether obesity influences the risk of treatment failure and the

development of gastroesophageal reflux disease after POEM in 89 patients. They find no significant difference between patients with a BMI of either below or above 30 kg/m². However, a general tendency toward obesity among a study population might still affect the observed treatment success if patients decide to lose weight after POEM to follow a healthier lifestyle. In such cases, the Eckardt score might falsely label them as treatment failures. It remains to be determined if the BMIs of this thesis' patients should be considered an outlying condition that may negatively affect this study's comparability with other publications. The evidence in favor of this is weak for now, but future studies might want to dis-ambiguate this question.

4.3.3 Achalasia Type Distribution

As shown in chapter 3.1, table 6, treatment-naïve patients were diagnosed with achalasia type I, II, and III in 21.5 %, 67.0 %, and 11.5 % of cases, respectively. For pre-treated patients, the fractions were 30.1 %, 53 %, and 16.4 %. In both groups, type II achalasia is the most frequent type, followed by type I, and both far ahead of type III. Compared to treatment-naïve patients, pre-treated patients are more often diagnosed with type I achalasia and less frequently with type II. Figure 59 shows the achalasia type distribution reported in recent and representative literature. Feasible studies are once again rare since many publications do not disclose achalasia types according to the Chicago classification.

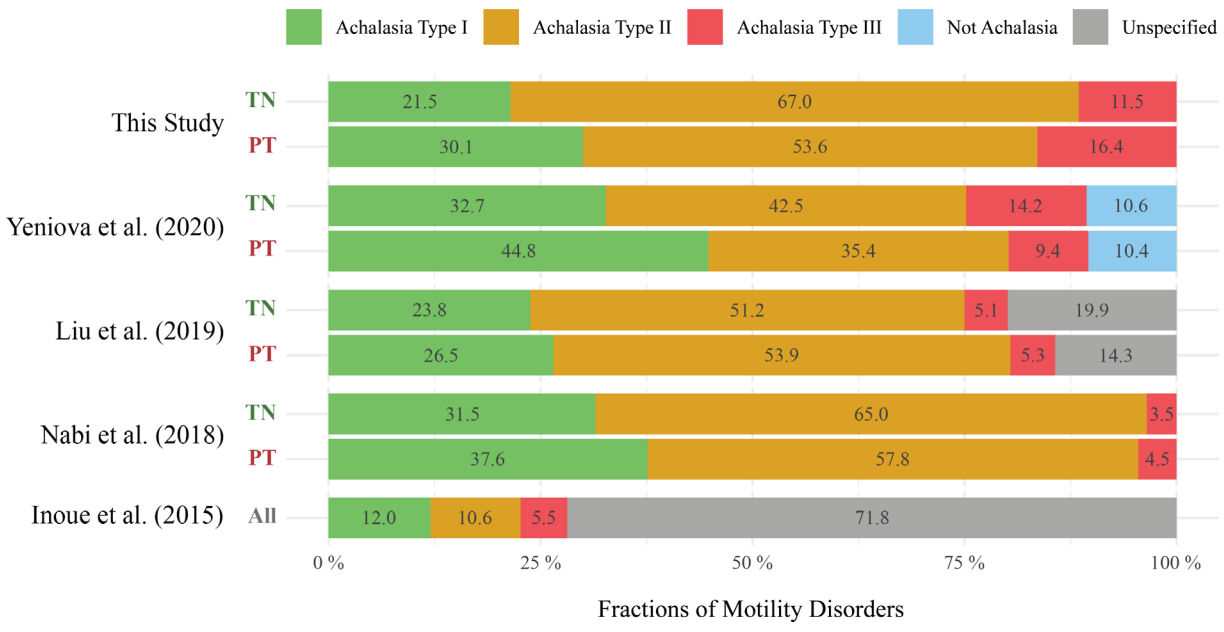


Figure 59: Achalasia Type Distributions in the Literature. The sum of fractions might not always add up to 100 % as a consequence of rounding. TN: treatment-naïve, PT: pre-treated.

The fractions differ quite a lot between studies, but some general tendencies seem to hold true and largely comply with the findings of this thesis. Type II achalasia is usually the most frequent type, and type III is very rare among all studies. Yeniova et al. (2020) and Nabi et al. (2018) both report lower percentages of type II and higher percentages of type I in pre-treated patients compared to treatment-naïve patients. This is again in agreement with this thesis and might be explained by a theory formulated by Sodikoff et al. (2016). They observed that the degree of ganglion cell loss in the myenteric plexuses in patients diagnosed with achalasia type I is higher than that of type II patients, but both types ultimately share a similar pattern. This leads them to suggest that type I achalasia might be a progression from type II. As discussed in chapter 4.3.1, previously treated patients are on average slightly older than treatment-naïve patients. As such, their higher prevalence of type I achalasia might be an expression of their illness having had more time to develop from a hypothetical earlier type II. Pre-treated patients might also look back on a longer disease history, which would fit in well with this theory. However, this remains speculation for now since the patients' disease durations were not assessed in this thesis. Liu et al. (2020) report an only slightly higher fraction of type I achalasia in pre-treated patients compared to treatment-naïve patients. This provides little to no additional evidence, but it does not contradict the theory in any way either.

Overall, the achalasia type distributions commonly reported in other studies match the distribution of this thesis' patients very well.

4.3.4 Summary

This chapter thoroughly assessed the age, sex, BMI, and achalasia type distributions of this thesis' patients. They show slightly higher weights compared to most other studies, likely an indicator of regional prosperity. Besides that, there is little evidence to suggest the existence of major demographic biases. The results of this thesis can therefore be expected to transfer well to other studies, and vice versa.

As a side-note, a few publications by *other* authors are based on unusually young populations. The regression models that were fit in this thesis strongly suggest a higher age to have a significant protective effect on post-interventional treatment failure. Therefore, these studies should be interpreted with caution, because their patients' younger demographics might facilitate an underestimation of their treatment success rates after POEM.

4.4 Methodical Validity

4.4.1 Follow-Up Compliance

The overall follow-up compliance is excellent up until two years after POEM with a participation of 80.7 % of eligible patients. It then remains decent throughout the subsequent three- and five-year follow-ups, with 73.2 % and 68.2 % of patients participating, respectively. Known re-treatments after POEM allow for the inference of treatment failure in some patients that are lost to follow-up. Thus, the completeness of the data sets used to nurture the regression models is even higher: it reaches 82.9 % after two years, 79.7 % after three years, and 75.0 % after five years. Considering the small number of studies available to date that incorporate a consistent follow-up of more than two years (see chapter 4.1.3), these results are most definitely well presentable. They indicate remarkable patient compliance despite the long follow-up duration.

In contrast, as already discussed in the chapters 4.2.2 and 4.2.3, the patients' abysmal participation in upper endoscopy and, especially, high-resolution manometry after three and five years is insufficient. For this reason, the prevalence of reflux disease and the IRP development after three and five years remain unfortunately pretty much uninterpretable.

Significant inter-group differences in the follow-up duration could introduce severe biases into the observed treatment outcomes. Fortunately, this is not the case in this study: the follow-ups time are similar between treatment-naïve and pre-treated patients (mean: 36.7 vs. 36.9 years, see chapter 3.1, table 6).

4.4.2 Missing Data

The analysis provided in chapter 3.4 unveiled overall high data completeness. The only relevant missing baseline data are 21.1 % of the IRP observations. These are mostly missing from the earliest years after the inception of POEM. High-resolution manometry was not as readily available then as it is nowadays. Many patients underwent the older conventional manometry instead, which is incapable of determining the IRP. Besides that, other conceivable reasons might be responsible for missing IRP observations. Patients may refuse the invasive and often displeasing manometry. They may also not tolerate the procedure, thereby necessitating an early abortion. The endoscopist may be unable to position the probe correctly. Finally, the measurement may fail due to mechanical breakdown.

No conspicuous patterns of missing information were found in the data (see chapter 3.4.3). However, correlation analysis revealed that patients diagnosed with type I achalasia are most frequently missing a baseline IRP observation among all three achalasia types. In contrast, patients diagnosed with type III are least likely to be missing this data. The reasons for this observation remain unclear. Patients

diagnosed with type III might tend to have an exceptionally long history of struggling with their disease. They may therefore have undergone multiple high-resolution manometries before POEM, rendering it more likely that a recent report was available to the author. Besides all that, further analysis revealed that 15 % of patients diagnosed with type III achalasia were treated between 2010 and 2012. As explained before, high-resolution manometry was hardly available at that time. In contrast, 20 % of patients diagnosed with type I achalasia and 22 % of type II patients were treated during these early years. This disproportion may have contributed to the uneven distribution of missing baseline IRP data as well.

No implausible outlying values were found for the IRP observations. Evidence against MCAR is weak for each individual model's data set, as evidenced by distributional analyses and the results of *Little's*, *Hawkins'*, and *Jamshidian's tests* (see chapters 3.5.2.1, 3.5.3.1, 3.5.4.1, and 3.6.2). Thus, preconditions were satisfied to allow for the imputation of missing IRP data for every model.

4.4.3 Endoscopy Proficiency

Chapter 3.1.4 reports the patients' treatment response rates per year. The 2010 cohort shows the worst outcomes by far. However, only five patients were treated that year. Thus, this deviation is negligible. It might well be an artifact caused by the small group size, and no harm to the quality of the study should be caused by 5 out of 374 patients. Additionally, the 2017 cohort seems to show an exceptionally bad three-year treatment response that is comparable to the outcomes usually seen in the other cohorts after five years. Since there is no obvious explanation for this observation, its cause remains to be determined. Keeping this exception in mind, patients treated in the earlier years do not show different outcomes compared to more recently treated patients. This indicates a homogenous procedural quality and speaks against a potential bias that could have been caused by varying levels of experience or different learning curves of the involved endoscopists.

4.4.4 Multiple Imputation

Multiple predictive mean matching imputation was heavily utilized in this thesis. As already mentioned in chapter 2.2.7.4, this method interpolates missing values with real observed donor values taken from the complete cases of the data set. This imputation method is practically free of distributional requirements because it does not generate new values. However, pre-existing biases in the data might still be aggravated under certain circumstances that will now be specified.

As explained in the previous chapter, achalasia type I patients were found to be more likely to be missing IRP observations. It remains unknown if this observation calls for a slight bias in the data. However,

distribution analysis of the known IRP values revealed that patients that were diagnosed with type I achalasia tend to present lower IRPs compared to patients diagnosed with type II or III (see chapter 3.3.1). The multiply imputed regression models might therefore slightly underestimate the predictive contribution of the IRP. The reason for that is rather technical, yet quite intuitive. If type I achalasia is associated with lower IRP values compared to the types II and III, and it is at the same time more likely to be missing IRP observations, predictive mean matching might disproportionally often impute missing and presumably lower observations in achalasia type I patients with presumably higher donor observations taken from type II or III patients. This is mere speculation. However, even if it were true, it would not diminish the quality of the regression models in this study. This is because it would be expected to *reduce* their sensitivity to the IRP as a predictor, not to falsely increase it. However, the Cox model identifies the IRP as a significant predictor, anyway. It might therefore even be a little bit *more* important than estimated by the model, which would not affect the conclusion that it *is* important. Also, as shown in chapter 3.6, figure 56, the density of the Cox model's imputed IRP values trends slightly towards lower values compared to the density of the completely observed data. This strongly hints at predictive mean matching having imputed correctly regarding the overrepresentation of missing values in achalasia type I patients.

Overall, the densities of all imputations match the densities of their respective observed data well. This indicates overall plausible imputations and good imputation qualities (see chapter 3.5, figures 50, 52, 54, and chapter 3.6, figure 56).

On a general note, multiple imputation increases the total variance in the data. It augments the conventional variability of completely observed data by additional variance that is caused by both the fact that there *are* missing observations, and the estimation process itself (van Buuren 2018). In contrast to that, the standard errors of each imputed regression model's effect size estimates (see chapter 3.5, tables 20, 22, 24, and chapter 3.6, table 26) are *lower* than the ones found in their respective complete case counterparts (see chapter 3.5, tables 21, 23, 25, and chapter 3.6, table 28). This indicates that in every imputed model, the effects of the information provided by the additionally included incomplete records outweighed the impact of the increased variance caused by multiple imputation. This yet again speaks in favor of a good imputation method and strengthens the argument to prefer the pooled multiply imputed models over their complete case counterparts. Overall, there is no evidence to suggest the presence of any detrimental statistical side effects caused by multiple imputation.

For the imputed model estimates, their respective deviances, AIC, and R_N^2 were each averaged as a surrogate for pooling. As explained in chapter 2.2.7.4, these measures cannot be pooled because Rubin's

rules do not apply to them, and there is no adequate alternative known to date. Hence it follows that the exact explanatory power of these resulting means remains controversial. They should therefore only serve as rough criteria to assess the models' quality, for a lack of alternatives. However, it seems very reasonable to assume that if the goodness-of-fit measures of most individual imputed models indicate good fits, then the pooled model should also be well fit. That said, it was decidedly refrained from averaging the more complex goodness-of-fit measures: D_{xy}, γ, τ_a, and the *Hosmer-Lemeshow test*. Trying to interpret their means would indeed have seemed rather far-fetched. They were calculated for the complete case models, though, which are not affected by this issue since they do not involve imputation. Since the null models do not include any predictor variables, their respective regression estimates are unaffected by predictor imputation. Accordingly, all null deviances and null AIC are identical for each imputation and thus do not require pooling.

4.4.5 Model Building

Model building is a complex and challenging part of regression analysis. It includes the essential question of which parameters to incorporate into the model.

A common approach is stepwise regression. A model may be chosen that includes either no parameters, or all parameters imaginable. Subsequently, predictors are added or removed from the model, which is then refit and compared to the previous one. If it is deemed worse, it is discarded. Else, the procedure is repeated until a model is found that is considered to be good enough. Another approach quite common in medical research is to perform univariate regression analyses on a variety of plausible predictors, of which the ones that reach significance are then combined into a multivariate follow-up model. This methodology was chosen by Liu et al. (2019) and Li et al. (2018), for example. A limitation of procedural approaches like these is that they may favor a model that does ultimately not measure up to the complexity of reality. Just because a predictor fails to reach significance, it does not necessarily need to be irrelevant to the model. Quite the contrary: the systematic removal of insignificant predictors might be frowned upon as a way of artificially inflating the *p* values of the remaining predictors.

There is no gold standard of parameters to include in a regression model. As explained, too few parameters bear the apparent risk of oversimplification. Too many parameters, on the other hand, may produce an overfitted model whose predictions might end up so close to its underlying data that it starts to reproduce deviations and errors contained in it. The models depicted in this thesis were carefully selected to compromise between these extremes. It was therefore decided to focus on a limited set of manually picked parameters that were expected to have a relevant clinical influence on the patients' treatment response after POEM. Some parameters were chosen based on clinical experiences, such as the *IRP* and

the *achalasia type.* Others, like *sex* and *age*, were included in the models because, in addition to being perfectly reasonable, it is considered to be good practice to do so. This predictor scheme was well thought out. Yet, of course, it retains a certain degree of arbitrariness; albeit it one that is shared by other studies.

Assessing the parameter selection of this thesis' models raises another interesting question. Previous treatments are tightly correlated to a lower baseline IRP (see chapter 3.1, table 6). Therefore, the latter may function as a surrogate parameter for prior treatments. This may also apply to other predictors, such as the pre-operative Eckardt score included in the regression model reported by Li et al. (2018). Preliminary covariance-sensitive models were fit during the early conceptualization phase of this thesis. They produced high *p* values and contained strong evidence for bad model fits. Thus, they were inconclusive and therefore ultimately discarded. To attain reliable results on this matter, it would likely have been necessary to include way more patients in the models than currently possible. Interestingly, previous treatments were identified as a significant risk factor by this thesis' Cox model. In contrast, a higher IRP was identified as a protective factor. If these predictors were correlated in a significant way, they should be expected to have a concordant effect on the predicted hazard. As it stands now, their effects are discordant. Therefore, there might be no relevant correlation, after all. Future studies may want to incorporate covariance analyses to shed light on how deep parameters like these are actually intertwined.

4.4.6 Logistic Regression Models

Each imputed logistic regression model has a reduced mean deviance compared to the deviance of its respective null model (2 years: 296 vs. 326, 3 years: 261 vs. 286, 5 years: 165 vs. 179; see chapter 3.5, tables 20, 22, and 24). This indicates that the parameters used to fit the models were chosen well. This assumption is further supported by a general decrease of the two- and three-year models' mean AICs compared to their respective null AICs (2 years: 310 vs. 328, 3 years: 275 vs. 288; see chapter 3.5, tables 20 and 22). However, the mean and the null AIC of the five-year model are similar (179 vs. 181, see chapter 3.5, table 24). This suggests that the two- and three-year models are well fit, while the five-year model is at least not worse fit than its null model. Since the five-year model is also the one build upon the least amount of observations, its estimates should be interpreted a little bit more carefully. The mean R_N^2 for each imputed model is generally quite low (2 years: 0.14, 3 years: 0.14, 5 years: 0.13; see chapter 3.5, tables 20, 22, and 24). Individual predictions may therefore lack precision.

To further assess the quality of the pooled imputed models, each of them can be compared to its respective complete case counterpart, which was fit with the same set of predictors, but based upon the complete-case data set without imputation. As explained in chapter 4.4.4, advanced measures for the assessment of the goodness of fit are available for complete case studies that cannot be safely applied after

multiple imputation because there is no known method to pool them adequately. One solution to this problem is to fit a supplementary complete case model, calculate the measures for this model, and discuss if the insight gained from that can be expected to transfer well to the imputed model (see chapter 2.2.7.4).

The goodness-of-fit measures discussed in this paragraph are provided in chapter 3.5, tables 21, 23, and 25. They depict the quality of the complete case models. Overall, their concordance is pretty high in the two-year model and borderline good in the three- and five-year models (2 years: 0.76, 3 years: 0.72, 5 years: 0.68). This indicates that the two-year model is quite decent at distinguishing between treatment failure and success. Predictions by the three- and five-year models should be interpreted a little bit more carefully, though. Goodman and Kruskal's γ (2 years: 0.51, 3 years: 0.44, 5 years: 0.44), and especially Somers' D_{xy} (2 years: 0.51, 3 years: 0.44, 5 years: 0.37), support these assumptions. As explained in chapter 2.2.6.5, D_{xy} is an especially well-equipped indicator for the goodness of fit of logistic regression models. Incidentally, D_{xy} is especially high in the two-year model. In contrast, Kendall's τ_a shrinks to low values due to the total number of possible pairs in its denominator (2 years: 0.16, 3 years: 0.17, 5 years: 0.17; see chapter 2.2.6.5, equation 14). Its application is therefore limited in this case. Both D_{xy} and γ indicate that each model can decently discriminate between treatment failure and success. However, the goodness of fit slightly declines after three and especially after five years. It stays in a quite decent range, though. Ultimately, the *Hosmer-Lemeshow test* does not reach significance for any model, providing even more persuasive evidence for the models being generally well fit (2 years: p = 0.664, 3 years: p = 0.187, 5 years: p = 0.253).

Assuming a certain degree of comparability between each multiply imputed model and its respective complete case counterpart, these findings can be expected to translate at least decently well to the pooled imputed models. However, as always, caution is advised.

4.4.7 Cox Proportional Hazards Regression Model

The multiply imputed cox regression model was fit under the premise that the proportional hazards assumption holds true. *Schoenfeld tests* of the imputed models prior to pooling revealed no evidence for a correlation between any two predictors, or any predictor and time, for any of the imputed models (see chapter 3.6.4). Thus, preliminary conditions for the fitting of the models using the proportional hazards formula were sufficiently satisfied. This is expected to translate decently well to the pooled final model, too, for similar reasons as discussed in the previous chapter for the logistic regression models.

The pooled multiply imputed Cox model has a borderline good mean concordance (0.65; see chapter 3.6, table 26). As already discussed, this indicates an overall good predictive distinction regarding the

outcomes. The model can therefore be expected to be, on average, pretty good at estimating which patient will experience treatment failure first. However, concrete death time estimates should not be calculated for an *individual* patient using this model, as those might lack sufficient precision.

The imputed model can be compared to its complete case counterpart. The concordance of the latter is pretty much the same as the mean concordance of the imputed models (0.66 vs. 0.65; see chapter 3.6, table 26 and 28). The *log-rank test* for the complete case model turned out highly significant ($p < 0.001$; see chapter 3.6, table 28). This indicates a good model fit. Just like the advanced goodness-of-fit measures for logistic regression models, this test cannot be applied after imputation because it is not known to date how to pool its results. However, assuming a decent enough comparability between the complete case model and the imputed model, the latter can be expected to be comparably well fit, too.

4.4.8 Summary

All of the pooled multiply imputed regression models are statistically well-founded, have been exhaustively validated and were found to be well fit. As mentioned several times before, one should however refrain from over-relying on exact risk estimations for individual patients. The general correlations identified by these models can be considered robust, though. Among the logistic models, the five-year model suffers from higher insecurities likely caused by the limited number of patients it was built upon.

4.5 Difficulties in the Comparison of Studies

Summarizing the discussions of chapters 4.1 to 4.24.3, studies on the clinical outcome after POEM are difficult to compare due to a multitude of factors that are not all apparent enough to not be overlooked.

Only a few papers take a closer look at the influence of the distinct achalasia types on clinical treatment response. Other do not differentiate between the types at all. Many publications throw together a diverse mixture of achalasia, spastic motility disorders like jackhammer esophagus, and other entities, such as EGJ outflow obstruction. Some authors stratify by achalasia types based on the Chicago classification, while others rather differentiate between straight and sigmoid type achalasia as indicators of the general disease progress. The assessment of reflux is hindered by the fact that some publications utilize the original Los Angeles classification system. In contrast, Japanese publications tend to use a modified version with slightly different grading criteria.

Furthermore, it is yet unclear how to best incorporate previous treatments in statistical models. Some studies differentiate between any pre-treatment or none, like this thesis. Others split their patients depending on the kind of previous treatments, such as botulinum toxin injections, balloon dilatations,

Heller myotomies, or even prior POEMs. Some even attempt to introduce complex classifications for prior treatments, like Louie et al. (2017). It is also unclear how to account best for varying amounts of previous treatments a patient has undergone, combinations of different treatments, and both in conjunction. On a side note, most papers do not report data on the endoscopists' experience, which might contribute substantially to the observed treatment success as well.

Ultimately, the study designs and follow-up protocols vary extensively between different publications. The decision about when to perform follow-ups seems arbitrary and may affect the reported results. There is barely more than a handful of reports available to date that include more than 200 patients or that are based on a consistent follow-up duration of more than two years. Few studies go beyond the scope of basic monovariate group comparisons by *t tests*, χ^2 *tests,* or *exact Fisher tests*. Advanced methods, such as multivariate regression analysis aided by multiple imputation, or even covariance analyses, are practically non-existent except for some very few pioneering publications. Overall, future studies would certainly benefit substantially from more standardized procedures, higher case numbers, and more elaborate statistical methodologies.

4.6 Implications and Limitations

This thesis focuses on achalasia patients only. For this reason, other motility disorders of the esophagus have been removed prior to analysis. This is a decision that is not shared by most other publications. In the same vein, this study only addresses patients pre-treated with either botulin toxin injections or balloon dilatations. All of this might hinder the transfer of results to a certain extent.

Besides these limitations, additional exclusion criteria were applied, which have been described in chapter 2.3.1. Overall, out of the 445 patients consecutively treated at our facility, 374 were ultimately included in this study. The removal of 71 patients (16.0 %) may have introduced selection bias into the data. However, 29 of these patients were removed due to having been diagnosed with non-achalasia motility disorders of the esophagus, as previously stated. Another 27 patients were excluded because they were pre-treated with Heller myotomy or POEM. This data reduction most likely increased the quality of the statistical models reported in this thesis. This assumption is justified as long as the conclusions drawn from this study are only applied to patients that satisfy the same set of inclusion and exclusion criteria. Therefore, the correlations identified by this thesis are most likely barely applicable to patients who are suffering from non-achalasia motility disorders of the esophagus. On a similar note, the study results that are specific to pre-treated patients should only be applied to other patients if they were pre-treated with botulinum toxin injections or balloon dilatations.

This thesis is susceptible to the same limitations as all retrospective studies. The fact that many patients are admitted to our hospital from far away might have increased heterogeneity in the data, especially if patients brought external reports with them. There appears to be a high inter-personal variance regarding the conduction of high-resolution manometry. This is indicated by occasionally dramatic differences between the IRPs measured at our facility and the ones previously determined at external facilities. Often, these diagnostics were performed less than a few weeks apart. Before POEM, most patients underwent control manometry at our facility, which was mostly conducted by one physician. Thus, even if inter-personal bias exists in the data, every endeavor has been made to minimize it.

The lack of an objective measure for the assessment of achalasia severity is a general problem. The only available scale capable of discriminating between treatment failure and success is the Eckardt score, which is highly subjective. Being assessed by the patients themselves, reported scores might fluctuate considerably on a weekly or even daily basis. This may cause bias in many studies, not just in this one.

Much effort was put into the structural analysis of this thesis' patient population and its internal dynamics. Fortunately, many potential types of bias could be ruled out in chapters 4.1 to 4.4. Nevertheless, a certain amount of heterogeneity among the patients remains. For example, the three achalasia types are not evenly distributed. Also, the pre-treated patients' past medical histories involve many different combinations of treatment methods and frequencies. A different study design, such as a prospective approach, would most likely be better capable of avoiding these limitations to a certain extent. Then again, the retrospective nature of this thesis allowed for the inclusion of a very high proportion of eligible patients. This, in conjunction with the utilization of multiple imputation, renders a detrimental influence of potential selection bias highly unlikely. That said, the limited total number of patients still poses a conceptual limitation to this thesis, naturally. To summarize, even in the case of a well-thought-out study, unavoidable factors may facilitate covert noise and bias in the data. Therefore, statistical results, however well-funded they may seem, should always be interpreted with caution.

Ultimately, the Cox model estimates the hazard, which is a function of time. Since the follow-up was limited to five years, the model should not be used to predict treatment failure beyond that time. Furthermore, many included patients were censored in the three- or five-year follow-ups because they had undergone POEM less than three or five years ago, respectively. Therefore, prognoses for treatment failure between two and five years after POEM should be expected to be less precise than predictions for earlier times. However, these considerations do not invalidate the significant correlations identified by this thesis in any way.

4.7 Conclusions and Outlook

Amid the very few publications available to date on the question of previous treatments' effects on the outcome after POEM, this study excels in a number of ways. It is characterized by an advanced statistical methodology, a long follow-up duration and a considerable number of patients with a notably high follow-up compliance. Therefore, it is reasonable to believe that this thesis contributes vital information to a largely uncharted field of clinical research that is all the more important for many achalasia patients.

Previous treatments, age, and the IRP have a significant effect on the clinical outcome after POEM. Whereas both a higher age and a higher IRP are associated with better outcomes, previous treatments pose a significant risk factor for treatment failure. This aligns well with the other few more extensive studies available on this topic. In contrast, the influence of the IRP has rarely been studied before since most recent studies still utilize the older LES instead. Higher age as a protective factor is an unprecedented correlation that, to the knowledge of the author, has not been proposed before. Last but not least, there is some evidence to suggest that the achalasia types II and III are associated with lower and higher risks of treatment failure, respectively. Since the respective models failed to reach significance by narrow margins, though, these correlations cannot be confirmed unerringly for now.

It can be recommended that future studies include more patients throughout longer systematic follow-ups. To achieve a higher grade of comparability between studies, scientists should adhere to the same classification systems, and standardize both scope and schedule of their follow-ups. Overall, a distinct lack of methodical proficiency can be observed among the current body of literature. As it stands, most studies are limited to elementary statistics in the like of barebone monovariate group comparisons. Researchers should be encouraged to embrace more sophisticated statistical procedures. As such, regression models should always be multivariate and include at least the dependent predictors *sex*, *age*, *achalasia type*, *IRP* or *LES*, and *pre-treatments*. Considering the patients' diverse past medical histories, the introduction of a standardized pre-treatment classification system would undoubtedly be beneficial to future research. Overweight, which is predominant among the populations of some western countries, may affect the treatment's outcome as well. Because of this, the *BMI* should be considered as another predictor. In addition to that, the patients' *disease duration* might be equally worth considering. However, scientists should avoid overfitting their models. Furthermore, parameters such as previous therapies, the IRP or LES, and the disease duration, might be correlated. To assess potential predictor interdependencies, the conduction of covariance analyses seems promising. Unfortunately, such studies will most likely require the inclusion of patient numbers that are, due to the young age of POEM, unrealistic for the near future. Ultimately, since age has a significant effect on post-interventional outcome, scientists should be aware of regional demographic differences when discussing treatment failure.

Obviously, the conduction of randomized rater-blinded controlled trials is the long-term objective. However, such studies will probably need many more years until enough patients are eligible for inclusion, and until enough time has passed to provide sufficient prospective long-term data. Until then, the available retrospective approaches should be refined to utilize the advantages of modern computational statistics to the highest degree possible. It is the belief of the author that this study manages to contribute to this goal, for it delivers well-founded suggestions to help with the conceptualization of future studies.

With the additional insight gained from this thesis, achalasia patients can be better counseled during pre-operation discussions, and based on a higher degree of evidence. Thus, they can hopefully give better-informed consent. Considering the currently available evidence, pre-treated patients must expect a significantly higher risk of post-interventional treatment failure than treatment-naïve patients.

To conclude, this thesis contributes to a better understanding among physicians of the dynamics between the patients, their illness, and the currently available treatment approaches. It sheds light on the relevance of a variety of known and even some newly discovered effectors on the outcome after POEM.

5 Summary

5.1 English

Per-oral endoscopic myotomy was introduced in 2010 as a novel treatment for achalasia. Middle- and long-term outcomes have been sparsely reported in the body of literature so far, and especially the impact of previous treatments is in dire need of clarification. A retrospective monocentric study was conducted to assess whether preceding botulinum toxin injections and pneumatic balloon dilatations affect the outcome after POEM. The study included 374 patients. Among these, 191 were treatment-naïve, and 183 were pre-treated. A systematic follow-up with a mean duration of 36.8 months was conducted. After two, three, and five years, 80.7 %, 73.2 %, and 68.2 % of the patients participated, respectively. Treatment failure was defined as an Eckardt score above 3 or having undergone another treatment.

Multiply imputed multivariate logistic regression models were fit for the prediction of treatment failure after two, three, and five years. They include previous treatments, sex, age, the IRP, and the achalasia type as predictors. The two-year model identified higher age as a protective factor (OR per 10 years: 0.66, $p < 0.001$) and type III achalasia as a risk factor (OR: 2.5, $p = 0.048$). Concordantly, higher age was also found to be a significant protective factor in the three-year model (OR per 10 years: 0.72, $p = 0.002$) and in the five-year model (OR per 10 years: 0.74, $p = 0.022$). A multiply imputed multivariate Cox proportional hazards regression model was fit with the same set of parameters as previously depicted. It identified prior treatments as a risk factor (HR: 1.5, $p = 0.031$), higher age yet again as a protective factor (HR per 10 years: 0.81, $p < 0.001$), and ultimately a higher IRP as another protective factor (HR per 5 mmHg: 0.90, $p = 0.038$). All models were thoroughly validated and deemed well fit.

This study provides strong evidence to suggest that the risk of treatment failure after POEM is significantly higher in previously treated patients than it is in treatment-naïve patients. A literature review is provided. Implications and limitations are discussed.

5.2 Deutsch

Die perorale endoskopische Myotomie wurde 2010 als neuartige Behandlung für Achalasie vorgestellt. Mittel- und langfristige Therapieergebnisse wurden bislang jedoch nur spärlich in der Literatur beschrieben. Insbesondere der Einfluss früherer Therapien ist nach wie vor unklar. Um zu klären, ob und inwieweit frühere Botoxinjektionen und pneumatische Ballondilatationen den Therapieerfolg nach POEM beeinflussen, wurde eine retrospektive monozentrische Studie durchgeführt. Sie umfasste 374 Patienten, darunter 191 therapienaive und 183 vortherapierte. Es wurde ein systematisches Follow-Up mit einer mittleren Dauer von 36.8 Monaten durchgeführt. Die Patientenbeteiligung hieran betrug 80.7 %, 73.2 %, und 68.2 % nach jeweils zwei, drei und fünf Jahren. Therapieversagen wurde definiert durch einen Eckardt Score größer als 3 oder die Durchführung einer erneuten Therapie.

Mehrfach imputierte multivariate logistische Regressionsmodelle für die Vorhersage eines Therapieversagens nach zwei, drei und fünf Jahren wurden berechnet. Sie beinhalten frühere Therapien, das Geschlecht, das Alter, den IRP sowie den Achalasietyp als Prädiktoren. Im Zwei-Jahres-Modell konnten ein höheres Alter als Schutzfaktor (OR je 10 Jahre: 0.66, $p < 0.001$) sowie der Achalasietyp III als Risikofaktor (OR: 2.5, $p = 0.048$) identifiziert werden. Übereinstimmend zeigte sich ein höheres Alter auch im Drei-Jahres-Modell (OR je 10 Jahre: 0.72, $p = 0.002$) sowie im Fünf-Jahres-Modell (OR je 10 Jahre: 0.74, $p = 0.022$) als Schutzfaktor. Zusätzlich wurde ein mehrfach imputiertes multivariates Cox-Regressionsmodell basierend auf derselben Auswahl an Parametern berechnet. Es identifizierte frühere Therapien als einen Risikofaktor (HR: 1.5, $p = 0.031$). Als Schutzfaktoren zeigten sich hier abermals ein höheres Alter (HR je 10 Jahre: 0.81, $p < 0.001$) sowie abschließend auch ein höherer IRP (HR je 5 mmHg: 0.90, $p = 0.038$). Alle Modelle wurden sorgfältig validiert. Ihre Anpassungsgüte wurde jeweils für hoch befunden.

Diese Studie liefert starke Evidenz für die Annahme, dass das Risiko eines Therapieversagens nach POEM für vortherapierte Patienten signifikant höher ist als für therapienaive Patienten. Es wird ein Literaturvergleich durchgeführt. Implikationen und Einschränkungen werden diskutiert.

6 References

Akaike, H. (ed), (1973) Information Theory and an Extension of the Maximum Likelihood Principle. Budapest, Hungary: Akadémiai Kiadó.

Akdis, M., Palomares, O., van de Veen, W., van Splunter, M. & Akdis, C. A. (2012) Th17 and Th22 Cells: A Confusion of Antimicrobial Response with Tissue Inflammation Versus Protection. J. Allergy Clin. Immunol., 129(6): 1438-1449.

Altman, D. G. (2009) Missing Outcomes in Randomized Trials: Addressing the Dilemma. Open Med., 3(2): e51-53.

Arber, N., Grossman, A., Lurie, B., Hoffman, M., Rubinstein, A., Lilos, P., Rozen, P. & Gilat, T. (1993) Epidemiology of Achalasia in Central Israel. Rarity of Esophageal Cancer. Dig. Dis. Sci., 38(10): 1920-1925.

Armstrong, D., Bennett, J. R., Blum, A. L., Dent, J., De Dombal, F. T., Galmiche, J. P., Lundell, L., Margulies, M., Richter, J. E., Spechler, S. J., Tytgat, G. N. & Wallin, L. (1996) The Endoscopic Assessment of Esophagitis: A Progress Report on Observer Agreement. Gastroenterology, 111(1): 85-92.

Bloomston, M., Nields, W. & Rosemurgy, A. S. (2003) Symptoms and Antireflux Medication Use Following Laparoscopic Nissen Fundoplication: Outcome at 1 and 4 Years. JSLS, 7(3): 211-218.

Booy, J. D., Takata, J., Tomlinson, G. & Urbach, D. R. (2012) The Prevalence of Autoimmune Disease in Patients with Esophageal Achalasia. Dis. Esophagus, 25(3): 209-213.

Bortolotti, M., Mari, C., Lopilato, C., Porrazzo, G. & Miglioli, M. (2000) Effects of Sildenafil on Esophageal Motility of Patients with Idiopathic Achalasia. Gastroenterology, 118(2): 253-257.

Bramhall, S. R. & Mourad, M. M. (2019) Wrap Choice During Fundoplication. World J. Gastroenterol., 25(48): 6876-6879.

van Buuren, S. (2018) Flexible Imputation of Missing Data, 2nd edition. CRC/Chapman & Hall, FL: Boca Raton.

Cariati, M., Chiarello, M. M., Cannistra, M., Lerose, M. A. & Brisinda, G. (2019) Gastrointestinal Uses of Botulinum Toxin., Handb. Exp. Pharmacol. Berlin, Heidelberg: Springer.

Chuah, S. K., Wu, K. L., Hu, T. H., Tai, W. C. & Changchien, C. S. (2010) Endoscope-Guided Pneumatic Dilation for Treatment of Esophageal Achalasia. World J. Gastroenterol., 16(4): 411-417.

Cox, D. R. (1972) Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological), 34(2): 187-220.

DeCarlo, L. T. (1997) On the Meaning and Use of Kurtosis. Psychol. Methods, 2(997): 292-307.

Eckardt, V. F., Aignherr, C. & Bernhard, G. (1992) Predictors of Outcome in Patients with Achalasia Treated by Pneumatic Dilation. Gastroenterology, 103(6): 1732-1738.

Engstrom, C., Lonroth, H., Mardani, J. & Lundell, L. (2007) An Anterior or Posterior Approach to Partial Fundoplication? Long-Term Results of a Randomized Trial. World J. Surg., 31(6): 1221-1225; discussion 1226-1227.

Fagerland, M. W. & Hosmer, D. W. (2013) A Goodness-of-Fit Test for the Proportional Odds Regression Model. Stat. Med., 32(13): 2235-2249.

Farrukh, A., DeCaestecker, J. & Mayberry, J. F. (2008) An Epidemiological Study of Achalasia among the South Asian Population of Leicester, 1986–2005. Dysphagia, 23(2): 161-164.

Feng, X., Linghu, E., Chai, N. & Ding, H. (2018) New Endoscopic Classification of Esophageal Mucosa in Achalasia: A Predictor for Submucosal Fibrosis. Saudi J. Gastroenterol., 24(2): 122-128.

Field, A. P., Miles, J. & Field, Z. (2012) Discovering Statistics Using R, 1st edition. SAGE Publications.

Fisher, R. A. (1922) On the Interpretation of X 2 from Contingency Tables, and the Calculation of P. J. Royal Stat. Soc., 85(1): 87-94.

Furuzawa-Carballeda, J., Aguilar-Leon, D., Gamboa-Dominguez, A., Valdovinos, M. A., Nunez-Alvarez, C., Martin-del-Campo, L. A., Enriquez, A. B., Coss-Adame, E., Svarch, A. E., Flores-Najera, A., Villa-Banos, A., Ceballos, J. C. & Torres-Villalobos, G. (2015) Achalasia--an Autoimmune Inflammatory Disease: A Cross-Sectional Study. J Immunol Res, 2015: Online publication. doi: 10.1155/2015/729217.

Gelfond, M., Rozen, P., Keren, S. & Gilat, T. (1981) Effect of Nitrates on Los Pressure in Achalasia: A Potential Therapeutic Aid. Gut, 22(4): 312-318.

Gennaro, N., Portale, G., Gallo, C., Rocchietto, S., Caruso, V., Costantini, M., Salvador, R., Ruol, A. & Zaninotto, G. (2011) Esophageal Achalasia in the Veneto Region: Epidemiology and Treatment. J. Gastrointest. Surg., 15(3): 423-428.

Ghosh, S. K., Janiak, P., Schwizer, W., Hebbard, G. S. & Brasseur, J. G. (2006) Physiology of the Esophageal Pressure Transition Zone: Separate Contraction Waves above and Below. Am. J. Physiol. Gastrointest. Liver Physiol., 290(3): G568-576.

Ghosh, S. K., Pandolfino, J. E., Rice, J., Clarke, J. O., Kwiatek, M. & Kahrilas, P. J. (2007) Impaired Deglutitive Egj Relaxation in Clinical Esophageal Manometry: A Quantitative Analysis of 400 Patients and 75 Controls. Am. J. Physiol. Gastrointest. Liver Physiol., 293(4): G878-885.

van Ginkel, J. R., Linting, M., Rippe, R. C. A. & van der Voort, A. (2020) Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data. J. Pers. Assess., 102(3): 297-308.

Gockel, I., Becker, J., Wouters, M. M., Niebisch, S., Gockel, H. R., Hess, T., Ramonet, D., Zimmermann, J., Vigo, A. G., Trynka, G., de Leon, A. R., de la Serna, J. P., Urcelay, E., Kumar, V., Franke, L., Westra, H. J., Drescher, D., Kneist, W., Marquardt, J. U., Galle, P. R., Mattheisen, M., Annese, V., Latiano, A., Fumagalli, U., Laghi, L., Cuomo, R., Sarnelli, G., Muller, M., Eckardt, A. J., Tack, J., Hoffmann, P., Herms, S., Mangold, E., Heilmann, S., Kiesslich, R., von Rahden, B. H., Allescher, H. D., Schulz, H. G., Wijmenga, C., Heneka, M. T., Lang, H., Hopfner, K. P., Nothen, M. M., Boeckxstaens, G. E., de Bakker, P. I., Knapp, M. & Schumacher, J. (2014) Common Variants in the HLA-DQ Region Confer Susceptibility to Idiopathic Achalasia. Nat. Genet., 46(8): 901-904.

Goodman, L. A. & Kruskal, W. H. (1954) Measures of Association for Cross Classifications. J. Am. Stat. Assoc., 49(268): 732-764.

Goodman, L. A. & Kruskal, W. H. (1959) Measures of Association for Cross Classifications. II: Further Discussion and References. J. Am. Stat. Assoc., 54(285): 123-163.

Goodman, L. A. & Kruskal, W. H. (1963) Measures of Association for Cross Classifications III: Approximate Sampling Theory. J. Am. Stat. Assoc., 58(302): 310-364.

Goodman, L. A. & Kruskal, W. H. (1972) Measures of Association for Cross Classifications, IV: Simplification of Asymptotic Variances. J. Am. Stat. Assoc., 67(338): 415-421.

Gosset, W. S. (1908) The Probable Error of a Mean. Biometrika, 6(1): 1-25.

Graham, J. W., Olchowski, A. E. & Gilreath, T. D. (2007) How Many Imputations Are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. Prev Sci, 8(3): 206-213.

Greene, C. L., Chang, E. J., Oh, D. S., Worrell, S. G., Hagen, J. A. & DeMeester, S. R. (2015) High Resolution Manometry Sub-Classification of Achalasia: Does It Really Matter? Does Achalasia Sub-Classification Matter? Surg. Endosc., 29(6): 1363-1367.

Hanna, A. N., Datta, J., Ginzberg, S., Dasher, K., Ginsberg, G. G. & Dempsey, D. T. (2018) Laparoscopic Heller Myotomy Vs Per Oral Endoscopic Myotomy: Patient-Reported Outcomes at a Single Institution. J. Am. Coll. Surg., 226(4): 465-472.

Harrell, F. E. (2016) Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2nd edition. Springer.

Harrell, F. E., Jr., Lee, K. L. & Mark, D. B. (1996) Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. Stat. Med., 15(4): 361-387.

Hawkins, D. M. (1981) A New Test for Multivariate Normality and Homoscedasticity. Technometrics, 23(1): 105-110.

Heller, E. (1913) Extramucöse Cardioplastie beim Chronischen Cardiospasmus mit Dilatation des Oesophagus. Mitt. Gren. Med. Chir., 27: 141-194.

Hemmert, G. A. J., Schons, L. M., Wieseke, J. & Schimmelpfennig, H. (2016) Log-Likelihood-Based Pseudo-R2 in Logistic Regression. Sociol. Methods Res., 47(3): 507-531.

Herbella, F. A. & Patti, M. G. (2015) Laparoscopic Heller Myotomy and Fundoplication in Patients with End-Stage Achalasia. World J. Surg., 39(7): 1631-1633.

von Hippel, P. T. (2007) Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data. Sociol. Methodol., 37(1): 83-117.

Ho, K. Y., Tay, H. H. & Kang, J. Y. (1999) A Prospective Study of the Clinical Features, Manometric Findings, Incidence and Prevalence of Achalasia in Singapore. J. Gastroenterol. Hepatol., 14(8): 791-795.

van Hoeij, F. B. & Bredenoord, A. J. (2016) Clinical Application of Esophageal High-Resolution Manometry in the Diagnosis of Esophageal Motility Disorders. J. Neurogastroenterol. Motil., 22(1): 6-13.

Hong, Y. S., Min, Y. W. & Rhee, P. L. (2016) Two Distinct Types of Hypercontractile Esophagus: Classic and Spastic Jackhammer. Gut Liver, 10(5): 859-863.

Hosmer, D. W. & Lemesbow, S. (1980) Goodness of Fit Tests for the Multiple Logistic Regression Model. Commun. Stat. - Theory and Methods, 9(10): 1043-1069.

Huffmann, L. C., Pandalai, P. K., Boulton, B. J., James, L., Starnes, S. L., Reed, M. F., Howington, J. A. & Nussbaum, M. S. (2007) Robotic Heller Myotomy: A Safe Operation with Higher Postoperative Quality-of-Life Indices. Surgery, 142(4): 613-620.

Hunter, J. G., Trus, T. L., Branum, G. D. & Waring, J. P. (1997) Laparoscopic Heller Myotomy and Fundoplication for Achalasia. Ann. Surg., 225(6).

Inoue, H., Minami, H., Kobayashi, Y., Sato, Y., Kaga, M., Suzuki, M., Satodate, H., Odaka, N., Itoh, H. & Kudo, S. (2010) Peroral Endoscopic Myotomy (POEM) for Esophageal Achalasia. Endoscopy, 42(4): 265-271.

Inoue, H., Sato, H., Ikeda, H., Onimaru, M., Sato, C., Minami, H., Yokomichi, H., Kobayashi, Y., Grimes, K. L. & Kudo, S. E. (2015) Per-Oral Endoscopic Myotomy: A Series of 500 Patients. J. Am. Coll. Surg., 221(2): 256-264.

Inoue, H., Shiwaku, H., Iwakiri, K., Onimaru, M., Kobayashi, Y., Minami, H., Sato, H., Kitano, S., Iwakiri, R., Omura, N., Murakami, K., Fukami, N., Fujimoto, K. & Tajiri, H. (2018) Clinical Practice Guidelines for Peroral Endoscopic Myotomy. Dig. Endosc., 30(5): 563-579.

Inoue, H., Ueno, A., Shimamura, Y., Manolakis, A., Sharma, A., Kono, S., Nishimoto, M., Sumi, K., Ikeda, H., Goda, K., Onimaru, M., Yamaguchi, N. & Itoh, H. (2019) Peroral Endoscopic Myotomy and Fundoplication: A Novel NOTES Procedure. Endoscopy, 51(2): 161-164.

Jamshidian, M. & Jalal, S. (2010) Tests of Homoscedasticity, Normality, and Missing Completely at Random for Incomplete Multivariate Data. Psychometrika, 75(4): 649-674.

Jamshidian, M., Jalal, S. & Jansen, C. (2014) MissMech: An R Package for Testing Homoscedasticity, Multivariate Normality, and Missing Completely at Random (MCAR). J. Stat. Softw., 1(6).

Jones, E. L., Meara, M. P., Pittman, M. R., Hazey, J. W. & Perry, K. A. (2016) Prior Treatment Does Not Influence the Performance or Early Outcome of Per-Oral Endoscopic Myotomy for Achalasia. Surg. Endosc., 30(4): 1282-1286.

Kahrilas, P. J., Bredenoord, A. J., Fox, M., Gyawali, C. P., Roman, S., Smout, A. J., Pandolfino, J. E. & International High Resolution Manometry Working, G. (2015) The Chicago Classification of Esophageal Motility Disorders, V3.0. Neurogastroenterol. Motil., 27(2): 160-174.

Kahrilas, P. J. & Pandolfino, J. E. (2017) Treatments for Achalasia in 2017: How to Choose among Them. Curr. Opin. Gastroenterol., 33(4): 270-276.

Kendall, M. G. (1938) A New Measure of Rank Correlation. Biometrika, 30(1-2): 81-93.

King, G. (1986) How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science. Am. J. Political Sci., 30: 666–687.

Korn, E. L. & Simon, R. (1990) Measures of Explained Variation for Survival Data. Stat. Med., 9(5): 487-503.

Kruskal, W. H. (1958) Ordinal Measures of Association. J. Am. Stat. Assoc., 53(284): 814-861.

Kumar, S., Choi, S. S. & Gupta, S. K. (2020) Eosinophilic Esophagitis: Current Status and Future Directions. Pediatr. Res.: Online publication ahead of print. doi: 10.1038/s41390-41020-40770-41394.

Kumbhari, V., Tieu, A. H., Onimaru, M., El Zein, M. H., Teitelbaum, E. N., Ujiki, M. B., Gitelis, M. E., Modayil, R. J., Hungness, E. S., Stavropoulos, S. N., Shiwaku, H., Kunda, R., Chiu, P., Saxena, P., Messallam, A. A., Inoue, H. & Khashab, M. A. (2015) Peroral Endoscopic Myotomy (POEM) Vs Laparoscopic Heller Myotomy (LHM) for the Treatment of Type III Achalasia in 75 Patients: A Multicenter Comparative Study. Endosc Int Open, 3(3): e195-201.

Li, Q. L., Wu, Q. N., Zhang, X. C., Xu, M. D., Zhang, W., Chen, S. Y., Zhong, Y. S., Zhang, Y. Q., Chen, W. F., Qin, W. Z., Hu, J. W., Cai, M. Y., Yao, L. Q. & Zhou, P. H. (2018) Outcomes of Per-Oral Endoscopic Myotomy for Treatment of Esophageal Achalasia with a Median Follow-up of 49 Months. Gastrointest. Endosc., 87(6): 1405-1412 e1403.

Ling, T., Guo, H. & Zou, X. (2014) Effect of Peroral Endoscopic Myotomy in Achalasia Patients with Failure of Prior Pneumatic Dilation: A Prospective Case-Control Study. J. Gastroenterol. Hepatol., 29(8): 1609-1613.

Little, R. J. A. (1988) A Test of Missing Completely at Random for Multivariate Data with Missing Values. J. Am. Stat. Assoc., 83(404): 1198-1202.

Liu, Z., Wang, Y., Fang, Y., Huang, Y., Yang, H., Ren, X., Xu, M., Chen, S., Chen, W., Zhong, Y., Zhang, Y., Qin, W., Hu, J., Cai, M., Yao, L., Li, Q. & Zhou, P. (2020) Short-Term Safety and Efficacy of Peroral Endoscopic Myotomy for the Treatment of Achalasia in Children. J. Gastroenterol., 55(2): 159-168.

Liu, Z. Q., Li, Q. L., Chen, W. F., Zhang, X. C., Wu, Q. N., Cai, M. Y., Qin, W. Z., Hu, J. W., Zhang, Y. Q., Xu, M. D., Yao, L. Q. & Zhou, P. H. (2019) The Effect of Prior Treatment on Clinical Outcomes in Patients with Achalasia Undergoing Peroral Endoscopic Myotomy. Endoscopy, 51(4): 307-316.

Louie, B. E., Schneider, A. M., Schembre, D. B. & Aye, R. W. (2017) Impact of Prior Interventions on Outcomes During Per Oral Endoscopic Myotomy. Surg. Endosc., 31(4): 1841-1848.

Mantel, N. (1966) Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration. Cancer Chemother. Rep., 50(3): 163-170.

Martins, R. K., Ribeiro, I. B., DTH, D. E. M., Hathorn, K. E., Bernardo, W. M. & EGH, D. E. M. (2020) Peroral (Poem) or Surgical Myotomy for the Treatment of Achalasia: A Systematic Review and Meta-Analysis. Arq. Gastroenterol., 57(1): 79-86.

Mayberry, J. F. & Atkinson, M. (1985) Studies of Incidence and Prevalence of Achalasia in the Nottingham Area. Q. J. Med., 56(220): 451-456.

Mikaeli, J., Bishehsari, F., Montazeri, G., Yaghoobi, M. & Malekzadeh, R. (2004) Pneumatic Balloon Dilatation in Achalasia: A Prospective Comparison of Safety and Efficacy with Different Balloon Diameters. Aliment. Pharmacol. Ther., 20(4): 431-436.

Minami, H., Inoue, H., Haji, A., Isomoto, H., Urabe, S., Hashiguchi, K., Matsushima, K., Akazawa, Y., Yamaguchi, N., Ohnita, K., Takeshima, F. & Nakao, K. (2015) Per-Oral Endoscopic Myotomy: Emerging Indications and Evolving Techniques. Dig. Endosc., 27(2): 175-181.

Miwa, H., Yokoyama, T., Hori, K., Sakagami, T., Oshima, T., Tomita, T., Fujiwara, Y., Saita, H., Itou, T., Ogawa, H., Nakamura, Y., Kishi, K., Murayama, Y., Hayashi, E., Kobayashi, K., Tano, N., Matsushita, K., Kawamoto, H., Sawada, Y., Ohkawa, A., Arai, E., Nagao, K., Hamamoto, N., Sugiyasu, Y., Sugimoto, K., Hara, H., Tanimura, M., Honda, Y., Isozaki, K., Noda, S., Kubota, S. & Himeno, S.

(2008) Interobserver Agreement in Endoscopic Evaluation of Reflux Esophagitis Using a Modified Los Angeles Classification Incorporating Grades N and M: A Validation Study in a Cohort of Japanese Endoscopists. Dis. Esophagus, 21(4): 355-363.

Moonen, A., Annese, V., Belmans, A., Bredenoord, A. J., Bruley des Varannes, S., Costantini, M., Dousset, B., Elizalde, J. I., Fumagalli, U., Gaudric, M., Merla, A., Smout, A. J., Tack, J., Zaninotto, G., Busch, O. R. & Boeckxstaens, G. E. (2016) Long-Term Results of the European Achalasia Trial: A Multicentre Randomised Controlled Trial Comparing Pneumatic Dilation Versus Laparoscopic Heller Myotomy. Gut, 65(5): 732-739.

Mukaka, M., White, S. A., Terlouw, D. J., Mwapasa, V., Kalilani-Phiri, L. & Faragher, E. B. (2016) Is Using Multiple Imputation Better Than Complete Case Analysis for Estimating a Prevalence (Risk) Difference in Randomized Controlled Trials When Binary Outcome Observations Are Missing? Trials, 17: 341.

Nabi, Z., Ramchandani, M., Chavan, R., Tandan, M., Kalapala, R., Darisetty, S., Lakhtakia, S., Rao, G. V. & Reddy, D. N. (2018) Peroral Endoscopic Myotomy in Treatment-Naive Achalasia Patients Versus Prior Treatment Failure Cases. Endoscopy, 50(4): 358-370.

Nabi, Z., Ramchandani, M., Darisetty, S., Kotla, R. & Reddy, D. N. (2019) Impact of Prior Treatment on Long-Term Outcome of Peroral Endoscopic Myotomy in Pediatric Achalasia. J. Pediatr. Surg.: Online publication ahead of print. doi: 10.1016/j.jpedsurg.2019.1007.1010.

Nabi, Z., Ramchandani, M., Kotla, R., Tandan, M., Goud, R., Darisetty, S., Rao, G. V. & Reddy, D. N. (2020) Gastroesophageal Reflux Disease after Peroral Endoscopic Myotomy Is Unpredictable, but Responsive to Proton Pump Inhibitor Therapy: A Large, Single-Center Study. Endoscopy: Online publication ahead of print. doi: 10.1055/a-1133-4354.

Nabi, Z., Reddy, D. N. & Ramchandani, M. (2017) Severe Submucosal Fibrosis – the “Achilles’ Heel” of Peroral Endoscopic Myotomy. Endoscopy, 49(11): 1116.

Nagelkerke, N. J. D. (1991) A Note on a General Definition of the Coefficient of Determination. Biometrika, 78(3): 691-692.

Orenstein, S. B., Raigani, S., Wu, Y. V., Pauli, E. M., Phillips, M. S., Ponsky, J. L. & Marks, J. M. (2015) Peroral Endoscopic Myotomy (POEM) Leads to Similar Results in Patients with and without Prior Endoscopic or Surgical Therapy. Surg. Endosc., 29(5): 1064-1070.

Orth, W. (2010) The Predictive Accuracy of Credit Ratings: Measurement and Statistical Inference. University of Cologne Statistics and Econometrics Discussion Paper, 2/10.

Pandolfino, J. E., Fox, M. R., Bredenoord, A. J. & Kahrilas, P. J. (2009) High-Resolution Manometry in Clinical Practice: Utilizing Pressure Topography to Classify Oesophageal Motility Abnormalities. Neurogastroenterol. Motil., 21(8): 796-806.

Pandolfino, J. E. & Gawron, A. J. (2015) Achalasia: A Systematic Review. JAMA, 313(18): 1841-1852.

Pearson, K. (1905) Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson. A Rejoinder. Biometrika, 4(1-2): 169-212.

Peto, R. & Peto, J. (1972) Asymptotically Efficient Rank Invariant Test Procedures. J. Royal Stat. Soc., 135(2): 185-207.

Podboy, A. J., Hwang, J. H., Rivas, H., Azagury, D., Hawn, M., Lau, J., Kamal, A., Friedland, S., Triadafilopoulos, G., Zikos, T. & Clarke, J. O. (2020) Long-Term Outcomes of Per-Oral Endoscopic Myotomy Compared to Laparoscopic Heller Myotomy for Achalasia: A Single-Center Experience. Surg. Endosc.: Online publication ahead of print. doi: 10.1007/s00464-00020-07450-00466.

Pressman, A. & Behar, J. (2017) Etiology and Pathogenesis of Idiopathic Achalasia. J. Clin. Gastroenterol., 51(3): 195-202.

Richardson, W. S., Willis, G. W. & Smith, J. W. (2003) Evaluation of Scar Formation after Botulinum Toxin Injection or Forced Balloon Dilation to the Lower Esophageal Sphincter. Surg. Endosc., 17(5): 696-698.

Rohof, W. O., Salvador, R., Annese, V., Bruley des Varannes, S., Chaussade, S., Costantini, M., Elizalde, J. I., Gaudric, M., Smout, A. J., Tack, J., Busch, O. R., Zaninotto, G. & Boeckxstaens, G. E. (2013) Outcomes of Treatment for Achalasia Depend on Manometric Subtype. Gastroenterology, 144(4): 718-725; quiz e713-714.

Rösch, T., Repici, A. & Boeckxstaens, G. (2017) Will Reflux Kill POEM? Endoscopy, 49(7): 625-628.

Rubin, D. B. (1976) Inference and Missing Data. Biometrika, 63(3): 581-592.

Rubin, D. B. (1987) Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.

Sadowski, D. C., Ackah, F., Jiang, B. & Svenson, L. W. (2010) Achalasia: Incidence, Prevalence and Survival. A Population-Based Study. Neurogastroenterol. Motil., 22(9): e256-261.

Samo, S., Carlson, D. A., Gregory, D. L., Gawel, S. H., Pandolfino, J. E. & Kahrilas, P. J. (2017) Incidence and Prevalence of Achalasia in Central Chicago, 2004-2014, since the Widespread Use of High-Resolution Manometry. Clin. Gastroenterol. Hepatol., 15(3): 366-373.

Samo, S. & Qayed, E. (2019) Esophagogastric Junction Outflow Obstruction: Where Are We Now in Diagnosis and Management? World J. Gastroenterol., 25(4): 411-417.

Sanaka, M. R., Parikh, M. P., Subramanium, S., Thota, P. N., Gupta, N. M., Lopez, R., Gabbard, S., Murthy, S. & Raja, S. (2020) Obesity Does Not Impact Outcomes or Rates of Gastroesophageal Reflux after Peroral Endoscopic Myotomy in Achalasia. J. Clin. Gastroenterol., 54(4).

Sanaka, M. R., Thota, P. N., Parikh, M. P., Hayat, U., Gupta, N. M., Gabbard, S., Lopez, R., Murthy, S. & Raja, S. (2019) Peroral Endoscopic Myotomy Leads to Higher Rates of Abnormal Esophageal Acid Exposure Than Laparoscopic Heller Myotomy in Achalasia. Surg. Endosc., 33(7): 2284-2292.

Schlottmann, F., Herbella, F., Allaix, M. E. & Patti, M. G. (2018) Modern Management of Esophageal Achalasia: From Pathophysiology to Treatment. Curr. Probl. Surg., 55(1): 10-37.

Schlottmann, F., Herbella, F. A. & Patti, M. G. (2017) Understanding the Chicago Classification: From Tracings to Patients. J. Neurogastroenterol. Motil., 23(4): 487-494.

Schoenfeld, D. (1980) Chi-Squared Goodness-of-Fit Tests for the Proportional Hazards Regression Model. Biometrika, 67(1): 145-153.

Scholz, F. W. & Stephens, M. A. (1987) K-Sample Anderson-Darling Tests. J. Am. Stat. Assoc., 82(399): 918-924.

Schwarz, G. (1978) Estimating the Dimension of a Model. Ann. Stat., 6(2): 461-464.

Shapiro, S. S. & Wilk, M. B. (1965) An Analysis of Variance Test for Normality (Complete Samples). Biometrika, 52(3/4): 591-611.

Sharata, A., Kurian, A. A., Dunst, C. M., Bhayani, N. H., Reavis, K. M. & Swanstrom, L. L. (2013) Peroral Endoscopic Myotomy (POEM) Is Safe and Effective in the Setting of Prior Endoscopic Intervention. J. Gastrointest. Surg., 17(7): 1188-1192.

Shiwaku, H., Inoue, H., Nimura, S., Yamashita, K., Ohmiya, T., Takeno, S., Sasaki, T. & Yamashita, Y. (2016a) Histological Findings of Divided Muscle after Peroral Endoscopic Myotomy. Ann Gastroenterol, 29(1): 94-95.

Shiwaku, H., Inoue, H., Onimaru, M., Minami, H., Sato, H., Sato, C., Tanaka, S., Ogawa, R. & Okushima, N. (2020) Multicenter Collaborative Retrospective Evaluation of Peroral Endoscopic Myotomy for Esophageal Achalasia: Analysis of Data from More Than 1300 Patients at Eight Facilities in Japan. Surg. Endosc., 34(1): 464-468.

Shiwaku, H., Inoue, H., Yamashita, K., Ohmiya, T., Beppu, R., Nakashima, R., Takeno, S., Sasaki, T., Nimura, S. & Yamashita, Y. (2016b) Peroral Endoscopic Myotomy for Esophageal Achalasia: Outcomes of the First over 100 Patients with Short-Term Follow-Up. Surg. Endosc., 30(11): 4817-4826.

Short, T. P. & Thomas, E. (1992) An Overview of the Role of Calcium Antagonists in the Treatment of Achalasia and Diffuse Oesophageal Spasm. Drugs, 43(2): 177-184.

Sodikoff, J. B., Lo, A. A., Shetuni, B. B., Kahrilas, P. J., Yang, G. Y. & Pandolfino, J. E. (2016) Histopathologic Patterns among Achalasia Subtypes. Neurogastroenterol. Motil., 28(1): 139-145.

Somers, R. H. (1962) A New Asymmetric Measure of Association for Ordinal Variables. Am. Sociol. Rev., 27(6): 799-811.

Spechler, S. J., Konda, V. & Souza, R. (2018) Can Eosinophilic Esophagitis Cause Achalasia and Other Esophageal Motility Disorders? Am. J. Gastroenterol., 113(11): 1594-1599.

Srivastava, M. S. & Dolatabadi, M. (2009) Multiple Imputation and Other Resampling Schemes for Imputing Missing Observations. J. Multivar. Anal., 100(9): 1919-1937.

Tebaibia, A., Boudjella, M. A., Boutarene, D., Benmediouni, F., Brahimi, H. & Oumnia, N. (2016) Incidence, Clinical Features and Para-Clinical Findings of Achalasia in Algeria: Experience of 25 Years. World J. Gastroenterol., 22(38): 8615-8623.

Tukey, J. W. (1977) Exploratory Data Analysis. Reading, Mass.: Addison-Wesley Pub. Co.

Tyberg, A., Choi, A., Gaidhane, M. & Kahaleh, M. (2018) Transoral Incisional Fundoplication for Reflux after Peroral Endoscopic Myotomy: A Crucial Addition to Our Arsenal. Endosc Int Open, 6(5): E549-E552.

United Nations (2019) World Population Prospects 2019: Volume II: Demographic Profiles.

Wu, Q. N., Xu, X. Y., Zhang, X. C., Xu, M. D., Zhang, Y. Q., Chen, W. F., Cai, M. Y., Qin, W. Z., Hu, J. W., Yao, L. Q., Li, Q. L. & Zhou, P. H. (2017) Submucosal Fibrosis in Achalasia Patients Is a Rare Cause of Aborted Peroral Endoscopic Myotomy Procedures. Endoscopy, 49(8): 736-744.

Yeniova, A. O., Yoo, I. K., Jeong, E. & Cho, J. Y. (2020) Comparison of Peroral Endoscopic Myotomy between De-Novo Achalasia and Achalasia with Prior Treatment. Surg. Endosc.: Online publication ahead of print. doi:10.1007/s00464-00020-07380-00463.

Zárate, N., Mearin, F., Gil-Vernet, J. M., Camarasa, F. & Malagelada, J. R. (1999) Achalasia and Down's Syndrome: Coincidental Association or Something Else? Am. J. Gastroenterol., 94(6): 1674-1677.

Zheng, Z., Zhao, C., Su, S., Fan, X., Zhao, W., Wang, B., Jin, H., Zhang, L., Wang, T. & Wang, B. (2019) Peroral Endoscopic Myotomy Versus Pneumatic Dilation - Result from a Retrospective Study with 1-Year Follow-Up. Z. Gastroenterol., 57(3): 304-311.

Zou, B. C., Zhang, L., Qin, B., Wang, S. H., Cheng, Y. & Zhao, H. L. (2020) Effects of Peroral Endoscopic Myotomy on Esophageal Function in the Treatment of Achalasia. Surg. Innov.: Online publication ahead of print. doi:10.1177/1553350620913133.

7 Acknowledgments

I would like to express my gratitude to my primary adviser, Prof. Dr. Thomas Rösch, and to my secondary adviser, Dr. Yuki Werner, for the guidance and advice they provided throughout my research. The knowledge they shared with me will reverberate in my future work.

Furthermore, I owe thanks to Dr. Maren Vens and Gerhard Schön of the Institute of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Germany. Their statistical guidance helped me to develop and refine the early methodology of this thesis.

Last but not least, I want to commemorate the invention of sugar-free energy drinks, without which I would indubitably have succumbed to diabetes during the writing of this thesis.

8 Curriculum Vitae

Der Lebenslauf wurde aus datenschutzrechtlichen Gründen entfernt.
This chapter is omitted for reasons of privacy.

9 Declaration of Academic Honesty

Eidesstattliche Erklärung:

Ich versichere ausdrücklich, dass ich die Arbeit selbständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die aus den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen einzeln nach Ausgabe (Auflage und Jahr des Erscheinens), Band und Seite des benutzten Werkes kenntlich gemacht habe.

Ferner versichere ich, dass ich die Dissertation bisher nicht einem Fachvertreter an einer anderen Hochschule zur Überprüfung vorgelegt oder mich anderweitig um Zulassung zur Promotion beworben habe.

Ich erkläre mich einverstanden, dass meine Dissertation vom Dekanat der Medizinischen Fakultät mit einer gängigen Software zur Erkennung von Plagiaten überprüft werden kann.

Unterschrift: .....