Universität Augsburg

Institut für Mathematik

# Advanced Statistical Methods for Prognostic Biomarkers and Disease Incidence Models

von

**Stefan Schiele**

**Dissertation**

zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
(Doctor rerum naturalium, Dr. rer. nat.)

eingereicht an der Mathematisch-Naturwissenschaftlich-Technischen Fakultät
der Universität Augsburg

Augsburg, November 2022

| | |
|---|---|
| Erstgutachter: | Prof. Dr. Gernot Müller,<br>University of Augsburg |
| Zweitgutachter: | Prof. Ph.D. Donna Ankerst,<br>Technical University of Munich |
| Tag der mündlichen Prüfung: | 27.02.2023 |

# Acknowledgement

First of all, I would like to thank Prof. Gernot Müller for the possibility to be part of his research group and for his support on the realization of this thesis. His feedback provided a lot of ideas to this thesis.

Further, I would like to thank my colleagues at the chair of Computational Statistics and Data Analysis for their helpful ideas and our collaboration on many projects.

I would also like to thank all physicians from the University Hospital Augsburg with whom I collaborated during the last years. I have always enjoyed the work and the possibility to apply theoretical concepts in medical applications.

Finally, without my family and their encouragement this thesis would not have been possible. I would like to thank them for their support in every situation.

# Abstract

Due to their prognostic value, biomarkers can support physicians in making the appropriate choice of therapy for a patient. In this thesis, several advanced statistical methods and machine learning algorithms were considered and applied to projects in collaboration with departments of the University Hospital Augsburg. A machine learning algorithm capturing hidden structures in binary immunohistologically stained images of colon cancer was developed to identify patients with a high risk of occurrence of distant metastases. Further, generalized linear models were used to estimate the probability of the need for a permanent shunt in patients after an aneurysmatic subarachnoid hemorrhage. Patients with oligometastatic colon cancer were stratified by a score developed using approaches from survival analysis to investigate which groups might benefit from surgical removal of metastases with prolonged overall survival.

Another important point is the selection of suitable statistical models dependent on the structure of the data. We found that a linear regression may only be suited with a transformation of the response variable in the context of association of a COVID-19 infection with lymphocyte subsets. In addition, modeling the course of daily reported new COVID-19 cases is a relevant task and requires suitable statistical models. We compared non-seasonal and seasonal ARIMA models and examined the performance of different log-linear autoregressive Poisson models. To add more structure and enable theoretical prognosis for the further course depending on non-pharmaceutical interventions, we fitted a Bayesian SEIR model with several change points and set the determined change points in context with the distribution of variants of the virus.

# Zusammenfassung

Biomarker können Ärtze durch ihren prognostischen Wert bei der Auswahl geeigneter Therapieoptionen unterstützen. In dieser Arbeit wurden mehrere fortgeschrittene statistische Methoden sowie Algorithmen des maschinellen Lernens eingeführt und in Zusammenarbeit mit verschiedenen Abteilungen des Universitätsklinikums Augsburg angewendet. Mit Hilfe eines Algorithmus des maschinellen Lernens, der versteckte Strukturen in binären, immunhistologisch gefärbten Bildern von Darmkrebstumoren feststellen kann, wurden Patienten mit einem hohen Risiko für auftretende Fernmetastasen identifiziert. Ebenso wurden Generalisierte Lineare Modelle verwendet, um eine Vorhersage der Wahrscheinlichkeit für eine dauerhafte Shunt-Anlegung nach einer aneurysmatischen Subarachnoidalblutung zu treffen. Patienten mit oligometastastischen Darmkrebs wurden mittels eines Scores, der anhand von Methoden der Survival Analysis entwickelt wurde, stratifiziert, um eine Gruppe zu identifizieren, die von einer operativen Entfernung der Metastasen durch ein langes Gesamtüberleben profitieren kann.

Ein weiterer wichtiger Punkt bei der Datenanalyse ist die geeignete Auswahl der statistischen Methode abhängig von der Datenstruktur. Es konnten am Beispiel der Assoziation einer Coronainfektion mit der Anzahl von Lymphozytensubpopulationen gezeigt werden, dass eine Transformation der Zielvariable notwendig sein kann, um die Voraussetzungen der linearen Regression zu erfüllen. Die Modellierung der Anzahl an täglichen Neuinfektionen stellt eine relevante Aufgabe dar und benötigt passende statistische Modelle. Ein non-seasonal und ein seasonal ARIMA-Model wurden ebenso wie mehrere log-linearen autoregressiven Poisson-Modellen verglichen. Zusätzlich wurde ein weiterer Modellierungsansatz untersucht, der die biologischen Mechanismen stärker einbezieht und eine theoretische Prognose für den weiteren Verlauf unter verschiedenen Szenarien ermöglicht. Der Verlauf wurde mittels eines bayesschen SEIR Modell mit mehreren Wendepunkten an die Daten angepasst. Die gefundenen Wendepunkte wurden in Kontext der Verteilung der Virusvarianten analysiert.

# Contents


**1 Introduction** **1**

1.1 Context . . . . . 1

1.2 Structure of the thesis . . . . . 2

**2 Selected Theoretical Concepts of Statistical Modeling** **4**

2.1 Generalized linear models . . . . . 4

2.1.1 Linear regression . . . . . 4

2.1.2 Generalized linear models . . . . . 8

2.2 Basic concepts of survival analysis . . . . . 12

2.2.1 Censoring . . . . . 12

2.2.2 Notation . . . . . 13

2.2.3 Estimation of the survival function . . . . . 14

2.2.4 Cox proportional hazards model . . . . . 16

**3 Statistical Approaches for Predicting Survival and Metastasis in Colon Cancer Patients using Machine Learning** **18**

3.1 Introduction . . . . . 18

3.2 Theoretical concepts: machine learning . . . . . 22

3.3 Machine learning for prognosis of overall survival . . . . . 24

3.3.1 Data and statistical approaches . . . . . 24

3.3.2 Results . . . . . 28

3.3.3 Discussion . . . . . 32

3.4 Machine learning for a prognosis of metastasis-free survival . . . . . 33

3.4.1 Data and statistical approaches . . . . . 33

3.4.2 Results . . . . . 37

3.4.3 Discussion . . . . . 44

**4 Development of Scores for Medical Research with Generalized Linear Regression Models and Methods from Survival Analysis** **46**

4.1 Development of a score for prediction of shunt risk for patients after an aSAH with generalized linear regression models . . . . . 46

4.1.1 Biological background . . . 47
4.1.2 Data . . . 47
4.1.3 Statistical approaches . . . 50
4.1.4 Results . . . 52
4.1.5 Discussion . . . 59
4.2 Development of a score for stratification of patients according to their survival using methods from survival analysis . . . 60
4.2.1 Biological background . . . 60
4.2.2 Data . . . 61
4.2.3 Statistical approaches . . . 61
4.2.4 Development of score . . . 63
4.2.5 Validation of a prognostic score . . . 69
4.2.6 Discussion . . . 74

**5 Modified Linear Regression Models for Associations between Lymphocytes and COVID-19 76**
5.1 Introduction . . . 76
5.2 Data . . . 78
5.3 Statistical approaches . . . 78
5.4 Results . . . 80
5.5 Discussion . . . 85

**6 Statistical Models for the Incidence of COVID-19 in Germany 87**
6.1 Comparison of modeling approaches for incidence of COVID-19 . . . 87
6.1.1 Introduction and background . . . 87
6.1.2 Data: incidence . . . 90
6.1.3 Statistical approaches . . . 90
6.1.4 Non-seasonal and seasonal ARIMA models . . . 94
6.1.5 Models for incidence without smoothing . . . 96
6.1.6 Log-linear autoregressive Poisson model . . . 103
6.1.7 Discussion . . . 109
6.2 Semi-mechanistic SEIR model with change points . . . 112
6.2.1 Background of interventions against COVID-19 . . . 112
6.2.2 Statistical approaches . . . 114
6.2.3 Modifying SEIR system with multiple change points . . . 117
6.2.4 Discussion . . . 122

**7 Further Studies 128**
7.1 Autopsies of COVID-19 patients . . . 128
7.2 New biomarker for gastric and colon cancer . . . 128

7.3 Lymphocyte subsets in patients with colorectal carcinoma . . . . . 129
7.4 Comparison of surgery techniques for parotidectomy . . . . . 130
7.5 Accuracy of ultrasound-guided core needle biopsy . . . . . 130
7.6 Survival analysis for parotid gland . . . . . 131
7.7 VR-based relaxation for enhancement of perioperative well-being . . . 131

**8 Summary** **133**

**Appendices** **145**

**A Software Code** **146**
A.1 Software code used for chapter 3 - training of CNN for images of tumor sections . . . . . 146
A.2 Software code used in chapter 6 - Bayesian SEIR model with change points . . . . . 148

# Chapter 1

# Introduction

## 1.1 Context

Statistical methods are an essential component of the analysis of medical data and various approaches have been developed over the last several years. Improved quality as well as higher availability of data have increased the possibilities for data analysis. A complex structured biological or medical research hypothesis must be evaluated with an appropriate statistical method and thus requires intensive cooperation between statisticians and physicians.

Biomarkers can support the diagnosis of diseases and facilitate the prognosis of the further course of a patient. The concept of personalized medicine is rapidly rising based on the detection of biomarkers in a wide range of fields. Values of blood parameters and scores based on sociodemographic characteristics of patients or disease-specific metrics from a tumor are only a portion of the variables on which a biomarker can be based.

An excellent example is the Framingham Risk Score, which identifies risk factors and combines them into a score for the estimation of 10-year cardiovascular risk (Wilson et al. (1998)). With this score, physicians can identify patients at high risk and advise them regarding modifications of their lifestyle or treat them with preventive drugs.

Another option to develop a biomarker is machine learning. Machine learning algorithms have received increased interest due to their ability to accommodate highly complex structures. Hidden features that are associated with the further course of a patient can be detected in images by convolutional neural networks (CNNs) and hence provide a prognosis based on an image, without previous manual detection of

relevant features. Skrede et al. (2020) have presented a CNN that was trained on H&E stained histological images of patients with colon cancer. They were able to classify patients based on the output of their CNN. Improved technical possibilities will enable more frequent use of machine learning to detect hidden structures in data and images.

The importance of prediction of the future course of a disease can be seen not only in the prognosis of patients with cancer but also in the behavior of the virus SARS-CoV-2. Its onset had a significant influence on all aspects of daily life around the world. Due to a high mortality rate, a higher transmission rate compared to previously known viruses, and severe infections leading to long-term medical issues, interventions needed to be implemented to protect elderly people and those with underlying health problems. Models forecasting the future course of the incidence of COVID-19 can help to estimate how the strength of interventions might affect the future number of COVID-19 cases.

## 1.2 Structure of the thesis

This thesis provides an overview of different theoretical statistical concepts for biomarkers and prognosis that can be applied in medical projects. We begin with mathematical and statistical background to introduce relevant concepts of statistical modeling and machine learning. We then define linear and generalized linear models (GLMs) due to their wide range of applications in several settings. One type of analysis that is not covered by GLMs relates to studies concerning the survival of patients, as censoring is often present and has to be taken into account. All methods that are only used in a single chapter are presented in the statistical approaches of that particular chapter.

Each chapter follows a similar structure. We provide the biological background of the application and a detailed description of the statistical methods that were used. Afterwards, we present the application in medical research and discuss the results.

Chapter 3 addresses an application of machine learning algorithms for patients with colon cancer and describes the development of a tool that can support treatment decisions by estimating the risk of a short time until death and the occurrence of metastases. When recurrence of the tumor is detected early, the therapy of a patient can be adapted and hence improved.
The machine learning algorithm used stained histological images from a selected

region of the tumor, binarized them into black-and-white images and assigned each patient to either a high- or a low-risk group dependent on the predicted risk obtained from the CNN.

Two medical studies in which scores were introduced for the prediction of the further course of a disease are presented in chapter 4. The first study considered the necessity of a permanent shunt for patients with a hemorrhage located in their brain. Clinical variables regarding the hemorrhage, the health conditions at admission to the hospital, and measurements during the first days after admission were combined to a score. Significant variables were determined in a generalized linear model with a logit link function, and different weightings for relevant variables were compared. In the second project, a score for overall and disease-free survival was developed for patients with colon cancer and distant metastases at the time of diagnosis. We performed a univariable preselection and a multivariable Cox proportional hazard regression to determine risk factors for shorter overall survival. The score was validated in patients from an independent cohort and compared to another score. We aimed to identify a group of patients that could benefit from surgical resection of the metastases.

Chapter 5 considers the influence that a COVID-19 infection has on the counts of lymphocytes. Measurements for multiple subsets of lymphocytes were collected from healthy individuals and people infected with COVID-19. A univariable linear regression was performed to determine differences between healthy and infected individuals. Because age and gender might have an impact on lymphocyte counts, we included both factors in a multivariable linear regression model with the severity of COVID-19 infection to adjust for them. All lymphocyte counts were logarithmically transformed to ensure that the residuals are approximately normally distributed.

Chapter 6 compares different modeling approaches for daily incidence of COVID-19. We fitted non-seasonal and seasonal ARIMA models and different log-linear autoregressive Poisson models. Furthermore, we fitted a compartment-based model with change points to investigate whether a model with a mechanical structure would be suited. We also examined whether change points might be explainable by changes in variants of interest or changes in the severity of non-pharmaceutical interventions.

In chapter 7, we present other projects that were processed during statistical consulting for the medical faculty of Augsburg University. Chapter 8 provides a summary of results of this thesis.

# Chapter 2

# Selected Theoretical Concepts of Statistical Modeling

This chapter introduces the theoretical background of several mathematical approaches for modeling that are needed in the following chapters. Methods that are used only in a single chapter are defined in that chapter. The definition of GLMs is mainly based on Vittinghoff et al. (2006) and Dunn and Smyth (2018), and the fundamentals of survival analysis are based on Kleinbaum and Klein (2012) and Kalbfleisch and Prentice (2011).

## 2.1 Generalized linear models

Regression models are commonly used for statistical data analysis. In this section, we present fundamental, theoretical concepts of linear regression and extend them to generalized linear regression models. We provide definitions and details of the estimation of parameters.

### 2.1.1 Linear regression

A linear regression aims to model a random variable $Y$ denoted as **response variable** with a set of $k$ known variables $x_1, \ldots, x_k$ which are called **explanatory variables** with $n$ data points. Our model consists of a systematic and a random component. We expect the relationship between the explanatory variables and the response variable for every data point to be linear except a normally distributed error term.

For the linear regression model, the following assumptions are necessary:

**Definition 2.1.1** (Assumptions of linear models)**.**

1. **Linearity:** For each individual $i$ with $i = 1, \ldots, n$ the relationship between explanatory variables $x_{i1}, \ldots, x_{ik}$ and response variable $Y_i$ has the form

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$$

   We call $\beta_i$ the **coefficient** for the $i-$th explanatory variable and $\epsilon_i$ should be a random variable that satisfies $\mathbb{E}[\epsilon_i] = 0$. $\epsilon_i$ is also denoted as the error term.

2. **Independence:** All random variables $\epsilon_i$ are independent.

3. **Variance homogeneity:** All random variables $\epsilon_i$ have constant variance $\sigma^2$.

4. **Normality:** The random variables $\epsilon_i$ are normally distributed.

From Definition 2.1.1 it follows directly that $\epsilon_i$ are independent and identically distributed with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

To facilitate the notation, we introduce a vector notation of the linear regression. We define a **design matrix** $X \in \mathbb{R}^{n \times (k+1)}$ where the first column corresponds to the intercept and all other columns correspond to the explanatory variables such that each data point in the regression is represented by one row. The design matrix has the following form:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ & & \vdots & & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}$$

Furthermore we define

$$\begin{aligned} \boldsymbol{Y} &= (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n \\ \boldsymbol{\epsilon} &= (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^n \\ \boldsymbol{\beta} &= (\beta_1, \ldots, \beta_k)^T \in \mathbb{R}^{(k+1)}. \end{aligned}$$

With $I_n$ indicating the $n \times n$ identity matrix, the linear model can be written as

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{with} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 I_n) \tag{2.1}$$

We are now interested in a parameter estimation of the coefficients and the variance of the error term of our linear regression model. In the following, we call the obtained estimates $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \ldots, \hat{\beta}_k\right)^T$. We use the maximum likelihood approach to estimate $\hat{\beta}$. By the assumptions in Definition 2.1.1 we know that our error term is normally distributed with mean 0 and variance $\sigma^2$ and that our independent variables are known for each patient. Hence, the likelihood of $(\boldsymbol{\beta}, \sigma)$ given $\boldsymbol{y}$ as the vector of observed values of the dependent variable can be expressed as

$$L(\boldsymbol{\beta}, \sigma|\boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2\right) \tag{2.2}$$

and the corresponding log likelihood as

$$g(\boldsymbol{\beta}, \sigma|\boldsymbol{y}) = \frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2. \tag{2.3}$$

For the maximum likelihood estimate we need to derive the log likelihood given in equation (2.3) with respect to $\boldsymbol{\beta}$. This leads to an optimum when $\boldsymbol{\beta}$ fulfills the **normal equation**

$$X^T X\boldsymbol{\beta} = X^T\boldsymbol{y}.$$

When our matrix X is of full rank, we can obtain an estimate for $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}} = (X^T X)^{-1}X^T\boldsymbol{y}$. With a similar calculation we get $\hat{\sigma^2} = \frac{1}{n}\|\boldsymbol{r}\|^2$, where $\boldsymbol{r}$ is called **raw residual vector** and defined as:

$$\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}}, \quad \text{where} \quad \hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}.$$

Different models can be compared via the Akaike Information Criterion (AIC). AIC measures how well the dependent variable is fitted by the model and penalizes the number of parameters. Without the penalty, the value could not decrease even if more unnecessary variables are integrated as independent variables.

**Definition 2.1.2** (Akaike Information Criterion (AIC))**.** The Akaike Information Criterion (AIC) for a model with $k$ parameters is defined as

$$\text{AIC} = -2l(\bullet) + 2 \cdot k \text{ ,where } l() \text{ is the loglikelihood function.}$$

Besides point estimates of coefficients, we are also interested which of the independent variables have a significant influence on the dependent variable. In the special case of a hypothesis test for a single coefficient, the null hypothesis can be written as

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0 \quad \text{for a fixed } j \in 0, \ldots, k.$$

We now need the distribution of $\hat{\boldsymbol{\beta}} = (X^T X)^{-1}X^T\boldsymbol{y}$. It is known that $\boldsymbol{y} \sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2)$

from the definition of our linear model. Since $\hat{\boldsymbol{\beta}}$ is only a transformed normally distributed random variable, it is also normally distributed with

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = (X^T X)^{-1} X^T \mathbb{E}[\boldsymbol{y}] = (X^T X)^{-1} X^T X \boldsymbol{\beta} = \boldsymbol{\beta}$$

and

$$\mathbb{V}ar[\hat{\boldsymbol{\beta}}] = (X^T X)^{-1} X^T \mathbb{V}ar[\boldsymbol{y}]((X^T X)^{-1} X^T)^T = \sigma^2 (X^T X)^{-1}.$$

This implies that $\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2 \left((X^T X)^{-1}\right)_{jj}\right)$. As $\sigma^2$ is in general unknown, we can define

$$T_j = \frac{\hat{\beta}_j}{\hat{se}(\beta_j)}$$

where $\hat{se}(\beta_j)$ is estimated as $s\sqrt{((X^T X)^{-1})_{jj}}$ with $s^2$ being the unbiased estimator of $\sigma^2$:

$$s^2 = \frac{1}{n-k-1} \|\boldsymbol{r}\|^2.$$

Under the null hypothesis, it holds that $T_j \sim t_{n-k-1}$ which can now be used for a test of the hypothesis and the calculation of confidence intervals. Furthermore, we can perform a backwards selection by removing the parameter from the model with the highest p value until all coefficients are significant.

Whether the model assumptions hold, can be checked by plotting the residuals. For the linearity assumption a plot of the residuals against the explanatory variables should show no pattern. A present pattern would indicate that the linear relationship is not appropriate for an explanatory variable.
The homogeneity assumption implies that the variance of the error term $\hat{\epsilon}_i$ is independent of the expected value. Hence, we can plot the residuals against the fitted values to see if a pattern is present.
For the normality assumption, we standardize the residuals and firstly calculate their variance. It holds that:

$$\begin{aligned}\mathbb{V}ar[\boldsymbol{r}] &= \mathbb{V}ar[\boldsymbol{y} - \hat{\boldsymbol{y}}] = \mathbb{V}ar[(I_n - X(X^T X)^{-1} X^T)\boldsymbol{y}] \\ &= (I_n - X(X^T X)^{-1} X^T)^T \sigma^2 (I_n - X(X^T X)^{-1} X^T) = \sigma^2 (I_n - X(X^T X)^{-1} X^T).\end{aligned}$$

We define **standardized residuals** to ensure that residuals should have mean 0 and variance 1.

**Definition 2.1.3** (Standardized Residuals)**.**
The standardized residuals are defined as:

$$r_{j,st} := \frac{r_j}{\sqrt{1 - (X(X^T X)^{-1} X^T)_{jj}}\, s}.$$

### 2.1.2 Generalized linear models

In comparison to linear regression models, we are not restricted to a linear relationship between the explanatory and the response variable and the structure of the error term is more flexible in GLMs. This allows our response variable to account for binary and count variables. In this section, we define the class of GLMs and provide details about parameter estimation and hypothesis testing. We also introduce more details to the logistic and poisson regression.

The notation is identical to the previous section with a response variable $Y_i$ and a vector of explanatory variables $\boldsymbol{x_i} = (1, x_{i1}, \ldots, x_{ik})$. We can expand the concept that we have seen for the linear regression and define a GLM based on three components:

**Definition 2.1.4** (Components of generalized linear model)**.**

1. **Random Component:** All responses $Y_i$ are independent and distributed according to a probability density function or a probability mass function from the **exponential family** with parameter $\theta$ and $\phi > 0$ given by

$$f(y|\theta, \phi) = \exp\left(\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right). \tag{2.4}$$

   $\phi$ is called the **dispersion parameter** and $\theta$ is called the **canonical parameter**. The functions $a, b$ and $c$ are known.

2. **Systematic Component:** We define the **linear predictor** as

$$\eta_i(\boldsymbol{\beta}) \coloneqq \boldsymbol{x_i}^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.$$

3. **Parametric Link Component:** The relationship between the linear predictor and the mean $\mu_i$ of $Y_i$ is set by the **link function**:

$$g(\mu_i) = \eta_i(\boldsymbol{\beta}) = \boldsymbol{x_i}^T\boldsymbol{\beta}.$$

Many known distributions like the normal distribution, binary distribution or the Poisson distribution belong to the exponential family. For a normal distribution $\mathcal{N}(\mu_i, \sigma^2)$, we set $\theta = \mu_i$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \theta^2/2$, and $c(y_i, \phi) = -\frac{1}{2}\log(2\pi\phi) - \frac{y_i^2}{2\phi}$ to prove the property. This shows that the linear regression model is only a special case of a GLM.

Definition 2.4 enables us for a closed form of the moments of a distribution from the

exponential family. The expectation and the variance of Y satisfy:

$$\mathbb{E}[Y] = b'(\theta)$$
$$\mathbb{V}ar[Y] = b''(\theta)a(\phi).$$

Link functions have no restrictions but are recommended to be monotone and real valued functions of the mean $\mu$. Some distributions, like the Poisson distribution, require a transformation. A family of transformations that fulfill these conditions is the Box-Cox transformation.

**Definition 2.1.5** (Box-Cox transformation)**.**
A Box-Cox transformation $f_\alpha$ is defined for a real valued $\alpha$ as:

$$f_\alpha(\mu) := \begin{cases} \frac{\mu^\alpha - 1}{\alpha} & \alpha \neq 0 \\ \log(\mu) & \alpha = 0. \end{cases}$$

The coefficients are estimated with a maximum likelihood estimation which will be explained in detail. We assume that $\boldsymbol{Y}$ follows a generalized linear model and that $\boldsymbol{x_i}$ is the explanatory variable of the $i-$th response $Y_i$. Further, we define the inverse mean function $h(\cdot)$ as the inverse of $b'(\cdot)$. Since $\mu_i = b'(\theta_i)$, it holds that $h(\mu_i) = \theta_i$.

We consider the log likelihood function for MLE:

**Definition 2.1.6** (Log likelihood in GLM)**.**
For observed data $\boldsymbol{y}$ of $\boldsymbol{Y}$ the **log likelihood** is defined by

$$l(\boldsymbol{\beta}, \phi|\boldsymbol{y}) := \sum_{i=1}^{n} l_i(\mu_i, \phi|y_i) = \sum_{i=1}^{n} \left( \frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right).$$

The derivatives of the log likelihood are given by

$$\begin{aligned} \frac{\partial l_i}{\partial \beta_j} &= \frac{\partial l_i}{\partial \mu_i} \frac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\ &= \frac{y_i - \mu_i}{b''(\theta_i)a(\phi)} \frac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i} x_{ij} \\ &= \frac{W_i(y_i - \mu_i)}{a(\phi)} \frac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i} x_{ij}, \quad \text{with} \quad W_i = \left( \frac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i} \right)^2 / b''(\theta_i). \end{aligned}$$

Because the scaling has no impact on the estimation, we use the unscaled score equations. These equations are non-linear in $\boldsymbol{\beta}$ and have in general no closed form of solution but need to be solved with an iterative algorithm. One possibility is the iterative weighted least squares (IWLS) algorithm. In every step of the IWLS algorithm, based on the current linear predictor and the current fitted means, an

intermediate dependent variable and intermediate weights are computed. The intermediate dependent variable is regressed on the explanatory variables with the intermediate weights, and the value of the coefficient vector is updated until it reaches convergence.
Shao has proven that the iterative, numerical solution leads to a maximum likelihood estimate that is consistent and asymptotically normal under regularity conditions. For further details we refer to Shao (2003). This asymptotic normality can be used for the computation of confidence intervals and hypothesis testing.

The two following subsections provide more details regarding the logistic regression and the Poisson regression as specific regression models that we applied in projects.

### Logistic regression

In many applications, the response variable is not continuous but rather is binary, for example whether a patient had a recurrence or whether a patient survived. Linear regression models can not map the structure of responses. However, the exponential family contains a suited distribution for binary responses. We use a similar notation as above except from some small modifications. $Y_i$ can either be 0 or 1 and we define the probability of an event as

$$p(\boldsymbol{x_i}) \coloneqq \mathbb{P}(Y_i = 1 | \boldsymbol{X_i} = \boldsymbol{x_i}).$$

Of the many different approaches to modeling this probability, only the logistic regression is presented here. We assume that the probability can be expressed as the logit-transformed linear predictor, which is given by

$$p(\boldsymbol{x_i}) = \mathbb{P}(Y_i = 1 | \boldsymbol{X_i} = \boldsymbol{x_i}) = \frac{\exp\left(\boldsymbol{x_i}^T \boldsymbol{\beta}\right)}{1 + \exp\left(\boldsymbol{x_i}^T \boldsymbol{\beta}\right)}.$$

This model fulfills all requirements of a GLM. The Bernoulli distribution belongs to the exponential family, there is a systematic component, and the parametric link is provided by the inverse of the logit function $g(\mu) \coloneqq \log\left(\frac{\mu}{1-\mu}\right)$.

For interpretation of the model, we evaluate the ratio of the probability of an event to the probability of no event.

**Definition 2.1.7** (Odds of event)**.**
The odds of event in a binomial regression model for $X = x$ is defined as

$$o(x) = \frac{p(x)}{1 - p(x)}. \tag{2.5}$$

We can describe a dependency between the response variable and the explanatory variable by the **odds ratio**.

**Definition 2.1.8** (Odds Ratio)**.**
The odds ratio (OR) for a binomial regression with a binary explanatory variable $X$ is defined as

$$OR = \frac{o(1)}{o(0)}.$$

We can now reformulate this equation:

$$\begin{aligned}\log OR &= \log\left(\frac{o(1)}{o(0)}\right) = \log\left(\frac{\frac{p(1)}{1-p(1)}}{\frac{p(0)}{1-p(0)}}\right) = \log\left(\frac{p(1)}{1-p(1)}\right) - \log\left(\frac{p(0)}{1-p(0)}\right) \\ &= \log\left(\exp\left(\beta_0 + \beta_1\right)\right) - \log\left(\exp\left(\beta_0\right)\right) = \beta_1.\end{aligned}$$

An odds ratio OR greater than 1 means that an observation with $x = 1$ has a $OR$-times higher odds for an event than when $x = 0$. Similar calculations deliver odds ratios when variables with multiple categories are included in the model. The odds ratio is then computed against a reference category for each level. In the case of a multiple logistic regression, the OR is calculated by fixing all other parameters to their reference.

### Poisson regression

Since count variables are often modeled by a Poisson distribution, a Poisson regression is suited when the response variables $Y_i$ are counts. An example is the number of new registered COVID-19 cases each day which might be associated to the case numbers of previous days.

We can prove that the Poisson regression belongs to the exponential family. We can rewrite the probability mass function:

$$\mathbb{P}(Y_i = y_i) = \exp\left(-\mu_i\right)\frac{\mu_i^{y_i}}{y_i!} = \exp y_i \log\left(\mu_i\right) - \mu_i - \log y_i!.$$

Now we chose $\theta_i = \log\left(\mu_i\right)$, $b(\theta_i) = \exp\left(\theta_i\right)$, $c(y_i, \phi) = -\log\left(y_i!\right)$. In addition, we can set $\phi = 1$ and $a(\phi) = 1$ and have proven the claim. Further, we select $g(\mu_i) = \log\left(\boldsymbol{x_i}^T\boldsymbol{\beta}\right)$ as link function. Thus, the Poisson regression is a GLM. The

logarithmic link function enables the coefficients of the GLM to be interpreted as multiplicative factors. Considering the systematic component we get:

$$\begin{aligned}\mu_i &= \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) \\ &= \exp(\beta_0) + \exp(\beta_1)^{x_{i1}} + \cdots + \exp(\beta_k)^{x_{ik}}.\end{aligned}$$

If we increase $x_{ij}$ by one unit $\mu_i$ is increased by a factor of $\exp(\beta_j)$. $\beta_j > 0$ leads to an increased $\mu_i$ and vice versa.

## 2.2 Basic concepts of survival analysis

In addition to GLMs, survival analysis is a central concept for medical data analysis. This is because time-to-event data are commonly used in research, for example for validation of a drug or a therapy after a cancer diagnosis, influencing the survival. In this section, fundamentals of survival analysis such as censoring and estimation of survival times are introduced.
Survival analysis aims to analyze the time from a starting point such as a diagnosis until a specified event occurs. This event can be death but can also be an adverse reaction to a drug or the time until a drug takes effect. We consider here only one possible event and no competing risk, which needs adapted methods for analysis.

### 2.2.1 Censoring

**Censoring** is a problem that occurs in many studies and reflects the condition that the real survival time of some participants is unknown. There are three main reasons that censoring occurs: no event until the end of the study, loss of follow-up, and withdrawal from the study.
The first reason can arise if time until recurrence after a cancer therapy is of interest. Patients who do not have a recurrence at the end of the study have an unknown recurrence-free survival time because it is only known that the patient was recurrence-free at least for the duration of study. In the other two scenarios, patients either have not attended their control appointments or have withdrawn from the study due to medical or personal reasons.

The following three different types of censoring can be distinguished:

- **Right-censored:** A survival time is right-censored when the true survival time is greater than or equal to the observed survival time. This is the most

common type of censoring and can be caused by the reasons described above.

- **Left-censored:** A survival time is left-censored when the true survival time is less than or equal to the observed survival time. For example, if a person has been tested positive for COVID-19, the exact time from infection to the positive test is not known.
- **Interval-censored:** A survival time is interval-censored when the true survival time is within a known time interval. If the above mentioned person had been tested at several time points, the interval in which the infection occurred could be determined.

Commonly, three assumptions for censoring in a survival analysis are distinguished, as follows:

- **Independent censoring:** Independent censoring is essential for survival analysis and states that within any subgroup subjects who are censored at time $t$ should have the same failure rate as subjects in the subgroup who remained at risk.
- **Random censoring:** Random censoring means that the failure rate for subjects who are censored should be equal to the failure rate of subjects who are not censored.
- **Non-informative censoring:** Non-informative censoring means that the distribution of the survival time $T$ has no influence on the censoring time $C$, and vice versa.

### 2.2.2 Notation

In this section, we present the general notation of survival analysis. For every subject $i$ ($i = 1, \ldots, n$) in our study, we denote the random variable for the **true survival time** as $T_i$, the random variable for the **censoring time** as $C_i$, and a random variable for the status of each patient as $D_i$. Because the patients are independent, it can be assumed that all $T_i$ and $C_i$ are independent. The **observed survival time** is denoted as $\tilde{T}_i$ and define as

$$\tilde{T}_i = \min\{T_i, C_i\}.$$

The status $D_i$ can be concluded from $T_i$ and $C_i$ as $D_i = I(T_i \leq C_i)$ with $I(\cdot)$ the indicator function. $D_i = 1$ indicates that an event was observed and $D_i = 0$ that the individual was censored because no event occurred during the study period or the person was lost of follow-up.

We define the **survivor function** $S(t)$ as $S(t) \coloneqq \mathbb{P}(T > t)$ such that $S$ is the probability that the survival of an individual is longer than t. An accurate estimation of our survival function is important to receive information for the study. The survival function is non-increasing and fulfills $S(0) = 1$ and $\lim_{t \to \infty} S(t) = 0$.

Furthermore, we define the **hazard function** $\lambda(t)$ as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

The hazard function $\lambda(t)$ can be interpreted as the current rate of occurrence of an event at time $t$ under the condition that a person survived until time $t$. The hazard function is always non-negative and is not bounded from above. Whereas the survival function directly describes the survival, the hazard function is used for modeling because specific parametric and non-parametric forms can be identified.

Despite this differences, survival and hazard function are connected. With the knowledge of one function the other one can be derived with the equations:

$$S(t) = \exp\left(-\int_0^t \lambda(u)\mathrm{d}u\right)$$
$$\lambda(t) = -\frac{\mathrm{d}S(t)/\mathrm{d}t}{S(t)}.$$

For example, if we know that the hazard function is constant with $\lambda(t) = \lambda^*$ then we can compute the survival function as $S(t) = \exp\left(-\int_0^t \lambda(u)\mathrm{d}u\right) = \exp\left(-\lambda^* t\right)$.

For time points without an observed event, the hazard function is 0. To overcome this issue, we consider the **cumulative hazard function** $\Lambda(t)$ and define it as the integral over the hazard function:

$$\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s.$$

The cumulative hazard function is directly connected to the survival function as $S(t) = \exp\left(-\Lambda(t)\right)$ and can be estimated by the Nelson-Aalen estimator.

### 2.2.3 Estimation of the survival function

The survival function can be estimated with either a parametric or a non-parametric approach. Both have advantages and disadvantages. A parametric approach assumes that the survival function and hence the hazard function have a parametric

distribution. It can be seen from the above description that a constant hazard results in an exponentially declining survival function. Other often used distributions are a Weibull distribution or a lognormal distribution. However, it is essential to choose a suitable distribution for the survival function, which can be difficult for applications involving real projects. A more detailed overview for parametric survival functions is provided by Kleinbaum and Klein (2012).

An alternative is the non-parametric estimation of the survival function by a Kaplan-Meier curve (KM curve). A KM curve is a non-increasing step function starting with a survival probability of 1 at time 0. The formula for the computation of the KM curve includes ordered failure times and a product of conditional probabilities. Let $t_{(1)}, \ldots, t_{(N)}$ indicate the set of distinct failure times. For each failure time $t_{(i)}$, the conditional probability of a survival time greater than $t_{(i)}$ can be computed given that the probability that the individual is in the risk set at time $t_{(i)}$.

**Definition 2.2.1** (Kaplan-Meier estimator)**.** Let $t_{(1)}, \ldots, t_{(N)}$ be the distinct and ordered failure times in a dataset. Further, we denote $d_{(i)}$ as the number of events at time $t_{(i)}$ and $n_{(i)}$ as the individuals at risk at $t_{(i)}$ (no event and no censoring up to time $t_{(i)}$). Then, the survival function can be estimated with the Kaplan-Meier estimator as

$$\hat{S}(t) = \prod_{i:t_{(i)} \leq t} \left(1 - \frac{d_i}{n_i}\right).$$

To test whether the KM curves are different, a log-rank test can be performed. The log-rank test uses differences between observed and expected numbers of failures to obtain the value of a statistic that is approximately chi-squared distributed. With this statistic, the p value for the null hypothesis that both curves are equal can be calculated.

Harrell Jr et al. (1996) have suggested the **c-index** as a measure that can be used to validate the discrimination of survival curves and the predictive information of a model. The c-index or concordance index is defined as the fraction of usable pairs of patients for whom the outcome and the prediction are concordant, which is satisfied if patient A is predicted to survive longer than B and actually survived longer than B. A usable pair means that either both patients have died at different times or patient A died and patient B survived longer than patient A. In all other cases, the pair of patients cannot be compared. If the predicted survival times are equal, then only 1/2 is added to the count of concordance. The c-index has a range between 0 and 1 with 1 indicating perfect discrimination.

### 2.2.4 Cox proportional hazards model

The Cox proportional hazards model (Cox PH model) expresses the hazard at time $t$ for an individual depending on the explanatory variables. No distribution of data is required compared to parametric models, but an underlying distribution can be well approximated. If the real distribution is not known, the Cox PH model is preferred as it provides reliable results. Denoting the explanatory variables as $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ with $\boldsymbol{X} \in \mathbb{R}^p$ and the vector of coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, the hazard function can be expressed as

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t) \exp\left(\sum_{i=1}^{p} \beta_i X_i\right). \tag{2.6}$$

According to equation (2.6), the hazard at time $t$ is split into a **baseline hazard** $\lambda_0(t)$ which is independent of explanatory variables and a multiplicative term depending on the explanatory variables but not on time $t$. Hence, the Cox PH model is a semi-parametric model and the ratio of hazards between two groups remains constant over time if $\boldsymbol{X}$ is time-independent. This justifies the requirement of proportional hazards.

If two individuals are identical except an one unit change for variable $X_k, k \leq p$, then the **hazard ratio** is:

$$\frac{\lambda(t|\boldsymbol{X})}{\lambda(t|\boldsymbol{X^*})} = \frac{\lambda_0(t) \exp\left(\sum_{i=1}^{p} \beta_i X_i\right)}{\lambda_0(t) \exp\left(\sum_{i=1}^{p} \beta_i X_i^*\right)} = \exp\left(\sum_{i=1}^{p} \beta_i (X_i - X_i^*)\right) = \exp\left(\beta_k\right).$$

The coefficient $\beta_k$ can thus be interpreted as the change in the logarithm of the hazard which is constant over time.

We do not estimate the coefficients with a full likelihood but use a partial likelihood function based on the order of observed events. The conditional probability of observing a specific failure is computed, conditioned on the current risk set and that one event occurs.

Like above, we use the distinct ordered uncensored failure times $t_{(1)}, \ldots, t_{(N)}$ and define a risk set $R(t) = \{k : t_k \geq t\}$ containing individuals that are at risk at time

$t$. The likelihood is provided by

$$\begin{aligned}\mathcal{L} &= \prod_{i=1}^{N} \mathbb{P}(\text{Subject } i \text{ has an event at } t_{(i)} | \text{one event occurs in } R(t_{(i)})) \\ &= \prod_{i=1}^{N} \frac{\lambda(t_{(i)}|\boldsymbol{X}_{(i)})}{\sum_{k \in R(t_{(i)})} \lambda(t_{(i)}|\boldsymbol{X}_{(k)})} = \prod_{i=1}^{N} \frac{\lambda_0(t_{(i)}) \exp\left(\boldsymbol{\beta}^T \boldsymbol{X}_{(i)}\right)}{\sum_{k \in R(t_{(i)})} \lambda_0(t_{(i)}) \exp\left(\boldsymbol{\beta}^T \boldsymbol{X}_{(k)}\right)} \\ &= \prod_{i=1}^{N} \frac{\exp\left(\boldsymbol{\beta}^T \boldsymbol{X}_{(i)}\right)}{\sum_{k \in R(t_{(i)})} \exp\left(\boldsymbol{\beta}^T \boldsymbol{X}_{(k)}\right)}.\end{aligned}$$

The estimation is independent of $\lambda_0(t)$ and it is sufficient to consider the exponential part of equation (2.6) for the coefficient estimation. An iterative Newton-Raphson algorithm can be performed to detect a maximum of the partial log likelihood, and the variance of the coefficients is obtained via the inverse observed information matrix at the maximum partial likelihood estimate. Further details can be found in Kalbfleisch and Prentice (2011) and Kleinbaum and Klein (2012).

# Chapter 3

# Statistical Approaches for Predicting Survival and Metastasis in Colon Cancer Patients using Machine Learning

**Keywords:** machine learning, VGG-net, InceptionResNetV2, prognostic biomarker, risk prediction, survival analysis, colon cancer

## 3.1 Introduction

Colon cancer is a highly relevant field of research because of its prevalence, its substantial mortality risk and the probability for an onset of distant metastases in the liver or lung. The German Robert-Koch Institute has stated that in 2016, every eighth cancer patient suffered from colon cancer and that approximately 60,000 incidents as well as approximately 25,000 deaths were recorded in connection with colon cancer. Although the possibilities for treatment and therapy have been improved in recent years, predictions of further course of the disease and patient' survival are still difficult. Nevertheless, they are necessary for suitable choices of therapy.

Currently, the classification of colon cancer and hence the stratification of patients with regard to prognostic estimates is mainly based on the Union Internationale Contre le Cancer (UICC) stage. This tool combines information regarding the local extent of the tumor and the occurrence of regional or distant metastases (cf. Brierley et al. (2017)). Thus far, most tissue samples have been treated in the same manner whereby the tissue sample is first stained for example with H&E, which is one of the most commonly applied histological stains and consists of hematoxylin

and eosin. Hematoxylin highlights the cell nuclei with a blue color whereas eosin is responsible for staining the cytoplasm (cf. Fischer et al. (2008)). The resulting sample is then evaluated based on tumor-node-metastasis (TNM) staging. In this way, three important properties are investigated by the pathologists: the size of the tumor and whether nearby tissue is invaded, the status of nearby nodes and the prevalence of metastasis. This information results in a UICC stage, which has been shown to be suited for the classification of the tumor in many cases. Most patients characterized as a stage II case have a good prognosis, whereas for stage III cases additional therapy is often needed and recommended.

Despite its usefulness in general, the UICC staging also exhibits weaknesses in daily routine. The assessment of TNM staging can be biased by inter-pathologist and intra-pathologist variability. Specifically, the grading of the same tissue sample may differ between two independently concerned pathologists, which might result in contradictory statements about optimal therapy and survival chances. In comparison, intra-pathologist variability involves different staging from the same pathologist if the same tissue sample is graded twice. Both problems persist in many pathological grading tasks and remain an impediment in diagnosis. High reproducibility and a comparable standard across different institutes would be preferable.

Another problem is the time-consuming task of grading a tumor manually. Here, a computer-driven approach could facilitate the work of pathologists and could help to improve the quality of prediction by incorporating hidden structures and small parts of the tumor that cannot be detected by humans.

Although UICC staging is still seen as a start-of-the-art method to classify tumors, there are cases in which predictions and the real course of the disease do not coincide. To improve the stratification of patients and support physicians in their decisions, reliable biomarkers that can easily be integrated into the daily routine are urgently needed.

In recent years, several promising biomarkers were found, for example based on the microsatellite stability status and cells from the immune system. However, only a few of these have been established in practical use. Tumor budding and the tumor-stroma-ratio (TSR) are examples of recently developed biomarkers that have obtained increased attention (cf. Lugli et al. (2017), Huijbers et al. (2013), and Mesker et al. (2007)). TSR is defined as the ratio between the area of invasive neoplastic cells and the surrounding nonneoplastic tissue and is often estimated based on of the H&E-stained slides. The estimation process can also be accomplished by

advanced methods to increase the quality of the estimate, as shown by Geessink et al. (2019) and Martin et al. (2020). A threshold of 50% is commonly used for separation of two groups. One fundamental idea of the TSR is to integrate not only TNM staging but also information regarding the tumor structure into the prediction. This approach can be further extended if methods of artificial intelligence are included in the process to identify complex prognostically meaningful patterns that have not been detected by humans thus far.

Machine learning algorithms are a modern approach for the classification of images in different research fields (cf. Goodfellow et al. (2016)). The basic approach is based on predefined features, for example tumor proportions, extracted from images. Rather than manual review of images, algorithms can find useful properties of a tumor. These features can be combined with known data from the patient such as age or gender and used as input for regression models chosen in dependence of the target variable. Logistic regression modeling of a binary outcome or Cox Regression modeling of survival probability are common choices. Besides classic statistical methods, machine learning approaches are also suitable for the classification of the input variables. Typical examples are support vector machines, neural networks, and random forests.

In recent years, convolutional neural networks (CNNs) and machine learning algorithms have opened new and wider possibilities for image analysis (cf. Krizhevsky et al. (2012)). Instead of relying on predefined features for a specific medical task, the algorithm itself finds representative, complex, hidden structures in a labeled training set of images. Several studies have shown that CNNs can outperform other algorithms in many different applications of medical imaging such as classification of cell types and grading of tissue samples. In recent years, an increasing number of algorithms and architectures have been made available for an improved performance. CNNs have learned to quantify the TSR in tumor slides (cf. K. Zhao et al. (2020)) and are also able to determine tumor budding (cf. Weis et al. (2018)).

Several studies have investigated the application of a CNN for predicting a patient's prognosis and have proved the prognostic value of an artificial intelligence approach. Kather et al. (2019) have presented a CNN that can distinguish between different tissue types in H&E-stained colon cancer tissue and can decompose a given image into its constituent parts. The fractions of the tissue types are used to build a weighted sum, which leads to a calculated prognostic score that could improve state-of-the-art methods. Jiang et al. (2020) have followed a similar approach in their work.

Instead of segmentation of the image, Bychkov et al. (2018) have fine-tuned a pre-trained CNN to identify hidden structures in H&E-stained colon cancer tissue images in a so called feature vector and have investigated the impact of different machine learning approaches on the performance of the model to predict the 5-year survival of patients by this feature vector. Their algorithm generated appropriate results for classification, and their predicted survival was a useful predictor in survival analysis. Skrede et al. (2020) followed the same approach and fitted a model directly onto image data, without a previous classification of tissue and computation of tissue proportions.

We refer to Pacal et al. (2020) for a comprehensive review of machine learning in colon cancer. Despite great progress in research for deep learning methods for the classification of medical images, algorithms predicting the further course of a patient's disease based on histological tumor images are rare.

The aim of our work was to investigate and compare different statistical approaches for the prediction of survival of colon cancer patients. Therefore, we trained multiple logistic regression models and derived a convolutional neural network-based approach for binary images to classify colon cancer patients according to their 5-year overall-survival. We selected a pre-trained convolutional neural network called VGG-net to extract hidden features of binary images and a neural network structure for the classification layers for the model. The model's architecture is based on VGG-net which is commonly used in the context of images and consists of several convolutional and pooling layers (cf. Simonyan and Zisserman (2014)).
In the second part of this chapter, we used the occurrence of distant metastases instead of overall survival as our primary endpoint and expanded the structure of the algorithm through further measures to avoid overfitting. As in the first approach, labeled pure black-and-white histological images were the input for the training such that the algorithm could not include any morphological features in the prediction. We examined if a machine learning algorithm could predict the occurrence of distant metastases, and compared the prediction with established criteria.

## 3.2 Theoretical concepts: machine learning

This section is mainly based on Goodfellow et al. (2016).

### Neural networks

Several machine learning algorithms are based on neural networks (NN) due to the ability of these constructs to model complex structured data better than other approaches. A NN consists of different types of layers. In an initial step, an input layer provides data that should model the target variable to an inner layer of the model. These layers consist of many neurons that form hidden layers and that are mostly not connected within the same layer. Finally, a probability distribution or a prediction at the output layer is obtained.

If we denote a neuron as $a_{l,n}$, where $l$ is an index for the layer and $n$ is an index for the neurons within a layer, then the value of $a_{l+1,n}$ can be computed by a weighted sum $w_{l,n}$ of the neurons of the previous layer as

$$w_{l+1,m} \coloneqq \sum_{i=1}^{n} \omega_{l,n,m} a_{l,n}.$$

In our notation $\omega_{l+1,m}$ represents a weight which needs to be optimized during the training process.

Subsequently, the weighted sum is transformed with an activation function $g(x)$ such that we obtain:

$$a_{l+1,m} = g(w_{l+1,m}) = g(\sum_{i=1}^{n} \omega_{l,n,m} a_{l,n}).$$

Two common choices for the activation function are the rectified linear unit (ReLU) and the softmax function.
The ReLU for an input $x$ is defined as:

$$f(x) = \max(0, x).$$

The softmax function can normalize the output vector of length $L$ of a NN such that its entries sum up to 1 and lie between 0 and 1. Hence, we can interpret this vector as the probabilities that a patient belongs to a certain class. The softmax for

an input vector $\mathbf{x} = (x_1, x_2, ..., x_L)$ is defined as

$$(f(\mathbf{x}))_j = \frac{e^{x_j}}{\sum_{i=1}^{L} e^{x_i}}.$$

## Convolutional neural networks (CNN)

Convolutional neural networks (CNNs) represent a state-of-the-art method in image analysis. Instead of manually defined features, more complex structures can be captured. A CNN is composed of a series of convolutions, pooling and further operations.

Convolutions are discrete operators and can be defined by a discrete function $F : \mathbb{Z}^2 \to \mathbb{R}$ and a filter $k$ of size $(2r + 1)^2$. This filter can be represented by a mapping $k : \Omega_r \to \mathbb{R}$, where $\Omega_r = [-r, r]^2$. Combining the discrete function and the filter, results in an equation for the convolution which is given by

$$(F \circ k)(p) \coloneqq \sum_{s+t=p} F(s)k(t).$$

## Pooling layer

Pooling layers are necessary to reduce the information within a network so that important features can be focused. Although some information is lost, the entire network benefits of pooling layers because the number of parameters as well as calculations can be reduced. Hence, the training process can be accelerated. Furthermore, the risk of overfitting can be lowered. Overfitting is a problem often faced in machine learning algorithms in which the generality of the model is lost because the number of parameters is too high.

There are different options for a pooling layer. In a max-pooling layer, the image is split into quadratic tiles with fixed and constant side lengths $a$ such that every pixel belongs to one tile. All pixels in each tile are replaced by one pixel whose value is assigned as the maximum value of the previous tile. A regular choice for the value of $a$ is 2, but higher values are also possible.
Another pooling layer for CNN is an average pooling layer. Rather than the maximum of a tile, the average of the tile is assigned as the value of the new pixel in this option.

## 3.3 Machine learning for prognosis of overall survival

This section was based on a collaboration between the pathology of Augsburg University Hospital (Bruno Märkl, Benedikt Martin) and the Institute of Mathematics (Gernot Müller, Stefan Schiele). Pathologists prepared the tumor images and mathematicians processed them, trained the model and fitted statistical models.

### 3.3.1 Data and statistical approaches

#### Case collective

For this study we included patients with colon adenocarcinomas of no special type that were assigned by the pathology of the University Hospital Augsburg as pT3/pT4 without metastasis at the time of diagnosis. The term pT3 specifies the severity of the tumor and is composed of three components. The first letter, p, stands for "pathological" and expresses that the classification is based on a pathological examination. The second letter indicates that the extent of the primary tumor is concerned. Finally, the number indicates the severity of the extent. pT3 means that the tumor has grown through the inner parts of the tumor and into the subserosa. In a pT4 case the tumor has grown through all layers of the colon (cf. Greene et al. (2006)).

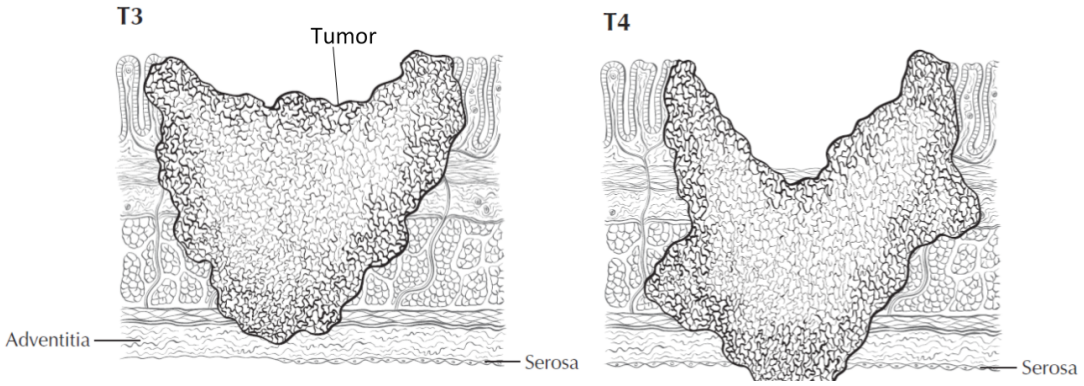


Figure 3.1: Example of tumors assigned with severe staging pT3 (left) and pT4 (right). On the left, the tumor has grown into the subserosa or adventitia. On the right, the tumor has passed through the serosa and all layers of the colon. This image has been adapted from Greene et al. (2006).

The sample had been used in another study by Martin et al. (2020) at the University Hospital Augsburg, but here we restricted the sample further such that the patients had to be less than or equal to 70 years old and their survival status five years after

diagnosis had to be known. For all cases, clinical-pathological information as well as a histopathological image of the invasive front of the tumor were available. All patients underwent surgery at the University Hospital Augsburg. Follow-up data were provided by the Tumor Data Management of the University Hospital Augsburg.

## Preprocessing of images

The following procedure was applied by the pathologists to preprocess the images and is similar to that of Martin et al. (2020). The entire H&E slide was viewed, and the best-fitting region that contained no artifacts of blood vessels, necrosis or other special type was selected. In the next step, a rectangular region with a field size of 3.58 $mm^2$ was extracted from the entire slide, containing tumor cells at all borders of the image field. The selected regions were digitized with a computer connected camera attached to the microscope. All images were immunohistochemically stained with cytokeratin (cytokeratin AE1/AE3) in order to highlight tumor tissue.

In further steps, the obtained image was processed with the open-source image software ImageJ (Version 1.48 v) (cf. Abràmoff et al. (2004) and Rasband et al. (1997)). Tumor-containing tissue reacted immunohistochemically and was marked in brown. The differentiation between tumor and stroma could be accomplished via binary coding. After translation of the image into a binary color, holes were filled and the images were reviewed by a pathologist. If necessary, the resulting image was manually improved by filling gaps that had not been closed by the software algorithm. This image was used as input for our machine learning model.

For further analysis, the images were measured and the tumor proportion of the image was calculated as the sum of all tumor areas divided by the area of the whole image.

## Feature extraction

The binary images were resized by a factor of 3 to $840 \times 680$ pixels to improve the learning performance of the algorithm. The images were split into a training and a validation set (80%/20%). The training images were then input to a pretrained VGG-neural network with removed classification layers to extract important features of the histological images. These features where used in a fully connected neural network with two hidden layers and an additional SoftMax activation to obtain the associated class probabilities. We trained only the classification part of the model and used binary cross entropy as the loss function and the Adam optimizer with a step size of $\tau = 10^{-3}$ for the optimization process.

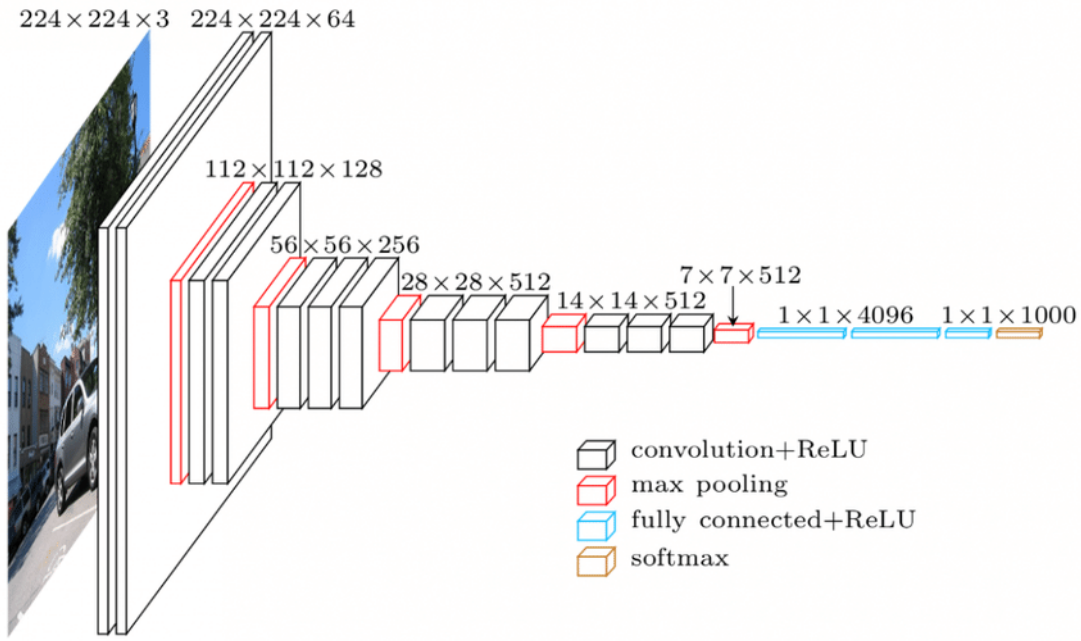


Figure 3.2: Structure of the VGG-net which is a convolutional neural network consisting of multiple convolutions and max pooling layers to reduce the size of the image and simultaneously increase the channel size. This figure is an adapted version of a corresponding figure in Loukadakis et al. (2018).

## VGG-Neural network

The architecture of our neural network was a VGG neural network. VGG is often used for the classification of images and was used for example in the ImageNet competition. It consists of multiple convolution layers of $3 \times 3$ filter with a stride of 2. Each convolution operator is followed by a ReLu activation. Between the convolution blocks a maxpool layer of $2 \times 2$ filter and stride 2 is executed, reducing the size of the image and extracting hidden features of the original image. For the entire architecture, the number of channels is increased. We removed the fully connected part at the end of the network to use the extracted features.

## Transfer learning

The weights of this neural network have been trained on the images of the ImageNet challenge, which is one of best known challenges concerning the classification of images. Other studies have proven that neural networks trained on this dataset can be used for other tasks, especially in medical and digital pathology. In the medical context, in which datasets from a large number of patients are very difficult to obtain, training from scratch or fine-tuning of the parameters is generally impossible. Further, we were not focused on a classification of the tissue as malignant or benign, in which case every pixel could be annotated and small patches of the entire slide

could be used to generate many images for training the model. Instead, only the 5-year survival outcome of the patient was available. As this outcome could not be referred to a special part of the entire slide, generating patches was impossible for this task.

## Variable description

Several predefined variables were extracted from the tissue images. For each measured variable and for each patient, the median, mean and the standard deviation were computed. These values could be used in a logistic regression model to analyze the impact of measurements of the tumor. According to Zdilla et al. (2016), we evaluated the roundness, the circularity, the solidity and the aspect ratio (AR) of every tumor region as explained in the following:

- **Roundness:** Roundness measures how similar the tumor region is to a circle by using the major axis of the best fitting ellipse of the tumor region and is calculated as:
$$4 \cdot \frac{Area}{\pi \cdot [\text{Major axis}]^2}$$
- **Circularity:** Similar to the roundness, circularity measures the degree of similarity of the tumor region to a perfect circle, but this time by using the perimeter of the best fitting ellipse of the tumor region. The circularity is calculated as:
$$4 \cdot \pi \frac{Area}{[\text{Perimeter}]^2}$$
- **Solidity:** Solidity of the tumor area describes how convex or concave the tumor area is. Its value ranges between 0 (high concave) and 1 (absolutely convex).
$$\frac{Area}{\text{Convex Area}}$$
- **Aspect Ratio:** Aspect Ratio is computed as the ratio of the major and the minor axis of the best fit ellipse
$$\frac{\text{Major axis}}{\text{Minor axis}}$$

## Statistical analysis

All statistical analysis was performed using the statistics software R 4.1.0. We analyzed clinical parameters of patients descriptively and divided the patients into two groups dependent on their 5-year survival. Patients with unknown status were

removed. We performed univariate logistic regression models with 5-year survival as response variable.

### 3.3.2 Results

#### Sample description

We enrolled 69 patients in our analysis. 51 (73.9%) survived longer than 5 years after diagnosis. The mean age was 60.0 years, and only 21.7% were less than or equal to 50 years old. The tumor was graded as pT4 in 14.5% and pN-positive in 44.9% of all cases. For the categorization of the tumor-stroma-ratio (TSR), we used previously stated cutoffs determined by the Institute of Pathology in Augsburg according to Martin et al. (2020). They identified three groups: low tumor proportion ($\leq 15\%$), medium tumor proportion (15% to $< 54\%$), and high tumor proportion ($\geq 54\%$). With this partitioning, nearly 15% of the patients had a high tumor ratio and 15% had a low tumor ratio (Table 3.1).

Building two groups dependent on the patient's 5-year-survival, we found differences between the group of survivors versus non-survivors. Non-survivors had a higher mean age (62.2 vs. 59.3 years) and a tumor that had been graded as more severe, with a higher fraction of pT4 (22.2% vs. 11.8%), a higher fraction of positive pN (55.6% vs. 41.2%), a higher fraction of a high grading (50% vs. 27.5%), and more often a low or high tumor-stroma-ratio (Table 3.2).
We further compared the two groups concerning their measurements of properties concerning the tumor shape. These were calculated for all tumor areas of a patient and its individual mean, median, and standard deviation over all tumor areas. The results of all patients were visualized via boxplots (Figure 3.3). We found only minor differences between both groups.

#### Generalized linear regression models

As a first step, we analyzed whether a generalized linear regression model as a classical statistical approach could predict the probability of survival of a patient if clinical data were available. For each variable, a separate univariate logistic regression model was fitted for all patients. The response variable was 5-year survival.
We found no significant association between the survival and tumor-stroma-ratio, age, sex, pT, pN and L status. It is conceivable that age has no influence because of the restriction in the sample. Furthermore, the clinical parameters pT and pN showed no association, but the grading dividing low and high showed a trend that was not significant.

Table 3.1: Patient characteristics of the study sample

| Variable | | N | % |
|---|---|---|---|
| 5 yr-Survival | | | |
| No | | 18 | 26.1 |
| Yes | | 51 | 73.9 |
| Age | mean: 60.0 ± 9.4 | | |
| $\leq 50$ | | 15 | 21.7 |
| $> 50$ to $\leq 60$ | | 12 | 17.4 |
| $> 60$ to $\leq 70$ | | 42 | 60.9 |
| pT | | | |
| 3 | | 59 | 85.5 |
| 4 | | 10 | 14.5 |
| pN | | | |
| positive | | 31 | 44.9 |
| negative | | 38 | 55.1 |
| Grading | | | |
| low | | 46 | 66.7 |
| high | | 23 | 33.3 |
| MSI | | | |
| positive | | 8 | 11.6 |
| negative | | 61 | 88.4 |
| V | | | |
| 0 | | 60 | 87.0 |
| 1 | | 9 | 13.0 |
| L | | | |
| 0 | | 56 | 81.2 |
| 1 | | 13 | 18.8 |
| Tumor Ratio | | | |
| $\leq 0.15$ | | 11 | 15.9 |
| $> 0.15$ to $< 0.54$ | | 48 | 69.6 |
| $\geq 0.54$ | | 10 | 14.5 |

Table 3.2: Comparison of patient characteristics stratified by their 5yr-survival

| | 5 yr-Surv n = 51 | | 5 yr-Surv no n = 18 | |
|---|---|---|---|---|
| Variable | N | % | N | % |
| Age | mean: 59.3 ± 9.6 | | 62.2 ± 9.0 | |
| $\leq 50$ | 13 | 25.5 | 2 | 11.1 |
| $> 50$ to $\leq 60$ | 10 | 19.6 | 2 | 11.1 |
| $> 60$ to $\leq 70$ | 28 | 54.9 | 14 | 77.8 |
| pT | | | | |
| 3 | 45 | 88.2 | 14 | 77.8 |
| 4 | 6 | 11.8 | 4 | 22.2 |
| pN | | | | |
| positive | 21 | 41.2 | 10 | 55.6 |
| negative | 30 | 58.8 | 8 | 44.4 |
| Grading | | | | |
| low | 37 | 72.5 | 9 | 50.0 |
| high | 14 | 27.5 | 9 | 50.0 |
| MSI | | | | |
| positive | 8 | 15.7 | 0 | 0.0 |
| negative | 43 | 84.3 | 18 | 100.0 |
| V | | | | |
| 0 | 47 | 92.2 | 13 | 72.2 |
| 1 | 4 | 7.8 | 5 | 27.8 |
| L | | | | |
| 0 | 43 | 84.3 | 13 | 72.2 |
| 1 | 8 | 15.7 | 5 | 27.8 |
| Tumor Ratio | | | | |
| $\leq 0.15$ | 7 | 13.7 | 4 | 22.2 |
| 0.15 to $< 0.54$ | 38 | 74.5 | 10 | 55.6 |
| $\geq 0.54$ | 6 | 11.8 | 4 | 22.2 |

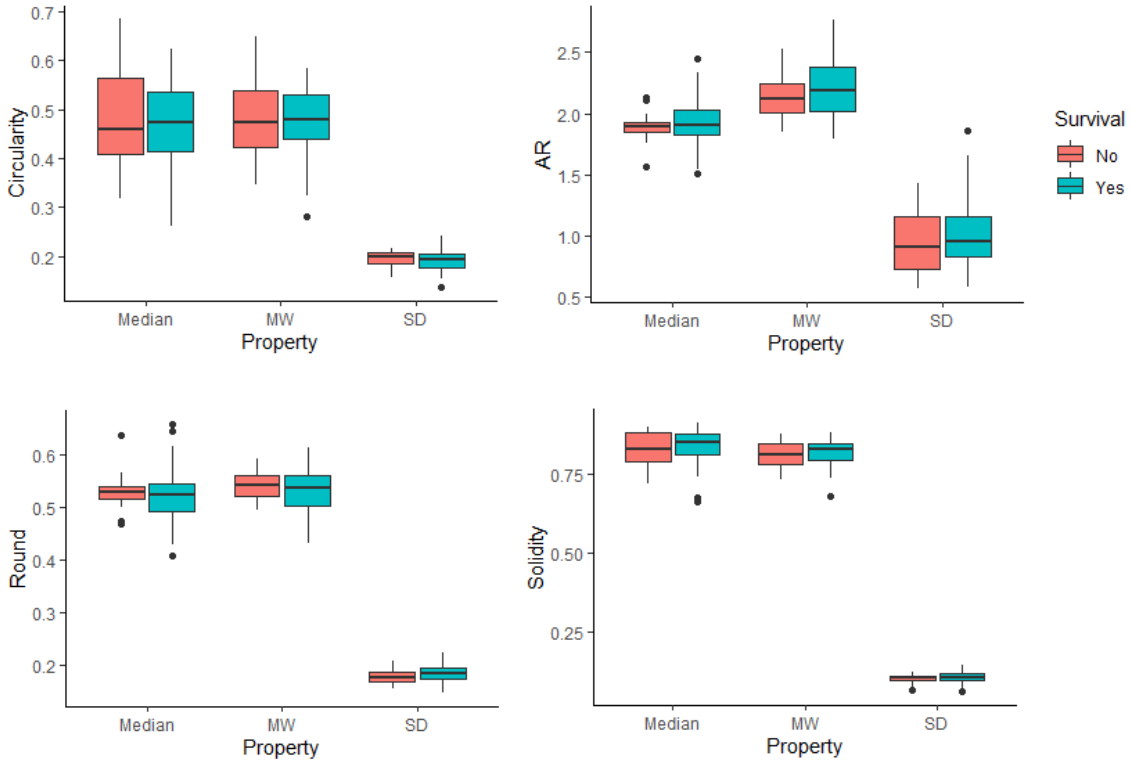


Figure 3.3: Boxplots for the distribution of mean, median and standard deviation of four measurements (circularity, AR, roundness, solidity) of patients tumor areas stratified by their 5-year-survival

From the inspected parameters, the V status and the existence of metastases were associated with a lower chance of survival five years after diagnosis.

We repeated the same analysis for the measurements from the tumor (roundness, circularity, aspect ratio, and solidity). For each property, we used the mean, the median and the standard deviation as covariates such that twelve univariate models were computed.
There was no association between the 5-year survival and the measurements of the tumor, with the exception of the mean of the aspect ratio. The higher the aspect ratio, the lower was the odds of survival, with an odds ratio of 0.76 for every 0.1 increase of the aspect ratio.

Finally, we performed a multiple logistic regression model by incorporating all variables with a p-value $< 0.1$ in one logistic regression model. These variables were grading of the tumor, V status, metastasis, and the mean of the aspect ratio. We performed a backward elimination based on the Akaike Information Criterion (AIC) to remove variables. In this process, only the V status was left in the final model. Another model without metastasis was also computed, as metastases occur only on follow-up rather than at diagnosis. Again, only grading and the mean of the aspect

Table 3.3: Factors associated with the 5-year survival from a multiple logistic regression with AIC-based backward elimination (without metastases)

| Variable | OR | CI | p-value |
|---|---|---|---|
| Grading high(ref: low) | 0.21 | [0.04-0.89] | $< 0.001$ |
| Aspect ratio mean (continuous with on unit=0.1) | 0.73 | [0.62-0.84] | 0.04 |

ratio were left in the model. We found that high grading and a higher mean of the aspect ratio of tumor areas were associated with a lower odds of survival at least 5 years after diagnosis.

### CNN

To investigate whether we could make a prediction based only on images, we developed a CNN approach using the images of the patient to capture more complex structures. We extracted the features with a pre-trained VGG network and only fine-tuned the classification layer. The patients were divided into two groups (training and validation).

Different parameters were used to fine-tune the classification layer of the network. Finally, a model with satisfactory performance on the training and validation set was trained. In both, an accuracy of approximately 85% was reached. When we investigated the performance of the model using an independent test set, we reached an accuracy of only 56%.

## 3.3.3 Discussion

We investigated the ability of different approaches to classify patients regarding their 5-year survival probability. In general, prediction was more difficult for our sample because of the restrictions on the age and the pT status. To ensure that the analysis of overall survival was less influenced by side-effects such as other diseases, we defined an age of 70 years or younger as an inclusion criterion. Furthermore, all patients in the sample had been diagnosed with a pT3 or pT4 cancer, leading to some similarities within the patients as compared to a sample containing all four subtypes.
A large difference between the models concerning interpretability could be observed. Whereas the classical statistical model provides pathologists with an insight on the decision process, the machine learning models have more of a black box character. This might be an impediment for wider usage of machine learning algorithms for all applications.

In spite of restrictions on age, we could see that the prediction of overall survival based only on histological images remains difficult due to other causes of death aside from colon cancer. These effects could not be taken into account.

## 3.4 Machine learning for a prognosis of metastasis-free survival

Because of restrictions in the overall survival and a low rate of cancer specific deaths we restricted the analysis on occurrence of distant metastases. This project was based on a collaboration between the pathology of Augsburg University Hospital (Bruno Märkl, Benedikt Martin) and the Institute of Mathematics (Prof. Gernot Müller, Tobias Arndt, Stefan Schiele). Pathologists prepared the tumor images, and the mathematicians processed those images, trained the model and fitted statistical models. Tobias Arndt and I worked on the machine learning algorithm. Tobias Arndt provided the main part of the model training, whereas I focused on the statistical analysis of the data. The developed methodology of the algorithm as well as results of the model performance for the classification of patients with colon cancer resulted in the publication Schiele et al. (2021).

### 3.4.1 Data and statistical approaches

#### Case collectives

We investigated our new hypothesis on a larger sample than the first one. Restrictions on age were not necessary in this case because the occurrence of distant metastasis could clearly be related to the primary tumor.

Both case collectives consisted of locally advanced colon adenocarcinomas of no special type, pT3/4, N±, M0, and R0 that were treated in the University Hospital Augsburg. For the training cohort (n = 163), we included patients whose surgery had been performed between 2012 and 2016 and the occurrence of distant metastases or documented metastasis-free survival of at least five years. The validation set fulfilled the same inclusion criteria and consisted of 128 patients (surgery between January 2002 and December 2011). Follow-up data for all cases were provided by the Tumor Data Management of the University Hospital Augsburg and complemented with data from patient files. The patients were treated in accordance with valid guidelines at that time.

## Sample preparation

Sample preparation was performed in exactly the same manner as described above. In this case, we further investigated whether the automatically selected threshold of the software could influence the resulting images. A sensitivity analysis showed that deviations of the threshold could introduce noise into the images, impeding the classification. The problem was that either pixels of the background had been classified incorrectly as tumor due to a lower threshold or parts of the tumor had been assigned to the background due to a higher threshold. Overall, the software algorithm performed well.

## Architecture of machine learning algorithm

The neural network described in the following section was similar to the one shown above, but extended to an additional preparation of the images before the network and layers to avoid overfitting. All implementations were performed in Python 3.6.9 using the Keras framework supplied by the TensorFlow 2.3.1 platform and trained using a Nvidia Tesla V100 GPU.

Our model was based on the InceptionResNetV2 (cf. Szegedy et al. (2017)) and was not trained from scratch but only the initial convolution layer and the fully connected layers at the end while the parameters of the InceptionResNetV2 were maintained. The model was optimized for 300 epochs with batches of 21 by the RMSprop with a learning rate of 0.0005. We chose the categorical cross entropy as loss function. After every epoch the model was validated using test data and the best performing model was finally selected. Tobias Arndt was responsible for the implementation and training of the algorithm.

## Feature extraction

As previously, all images were downscaled by a factor of three to $680 \times 840$ pixels to ensure a satisfactory performance during training. We normalized the images to obtain binary images with pixel values between 0 and 1. The images of the training and testing sets were split (80%/20%). With a convolution layer consisting of three $20 \times 20$ filters with a stride of three, padding as well as a hyperbolic tangent activation function, the images were reduced to a size of $216 \times 287$ with three channels. In comparison to above, we extracted the features with a pretrained InceptionResNetV2 network. The weights of the network had been previously trained on the ImageNet challenge. Afterwards, the resulting 1536 feature maps with a size of $5 \times 7$ were pooled with GlobalAveragePooling with a stride of two. The final output

of our neural network was obtained by two fully connected layers with Relu activation functions, containing 256 nodes in the first layer and 64 nodes in the second layer, and a fully connected output layer containing two nodes with a SoftMaxactivation function. This algorithm provided a predicted probability for the occurrence of metastasis and the absence of metastasis to classify the images.

## Measures against overfitting

Overfitting can occur during the training of a neural network, especially when the dataset contains a low number of samples. The algorithm then loses the ability to generalize the prediction from the training sample because data points have been fitted too closely during the training. Hence, the algorithm must be adapted to ensure that the model can provide satisfactory predictions in general and not only for the training dataset. During the training, we presented the same images multiple times each time with small geometric changes (rotation, shifting, mirroring) as described by Shorten and Khoshgoftaar (2019). The prediction of the model should be identical because the structure had been maintained under all modifications. We generated altered images with random augmentations in each training epoch based on the ImageDataGenerator implemented in Keras to reduce overfitting.

We set a range of possible parameters for each transformation from which the algorithm uniformly sampled. The images were rotated between $-15$ and 15 degrees, shifted between $-10$ and 10 percent in width and height, and sheared in the interval of $[0, 1]$ degrees. If present, voids were filled by reflecting the image to obtain the right format. We chose random rectangular sections of the image and substituted them with uniformly distributed and smoothed noise to create additional modifications and avoid overfitting (cf. Zhong et al. (2020)). Examples of augmentations in the images are provided in Figure 3.4.

Another important tool to reduce the risk of overfitting is dropout. During the training phase in each step, a fraction of layers was chosen and set to 0. This reduced the number of parameters adapted in this step and could thus assist in improving the training of the model. We selected a dropout of 10% for the output of the InceptionResNetV2, 20% for the first fully connected layer, and 10% for the second layer.

## Statistical analysis

All statistical analyses were conducted using R 4.1.0. The performance of the model was validated on an external dataset of patients that was not incorporated during

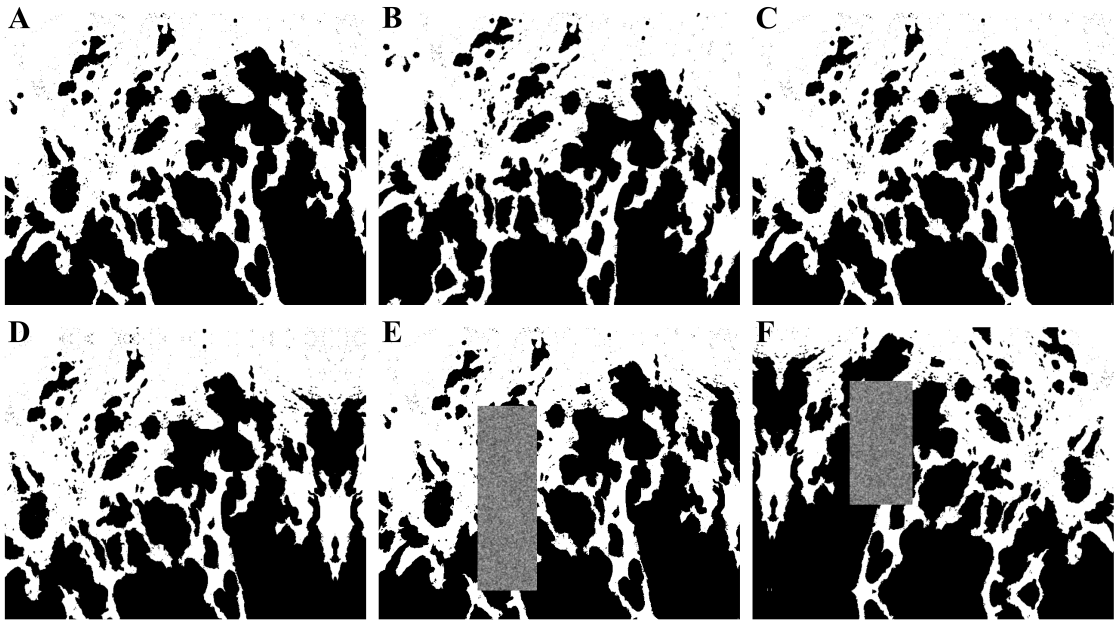


Figure 3.4: Different augmentations of the original image (A) through rotation (B), shearing (C), shifting (D), random erasing (E), and a combination of all methods (F). Adapted from Schiele et al. (2021)

the training process. We aimed to split the patients into two groups according to the trained model. A histological image was provided to the trained model and the outcome of the model, namely the probability for the occurrence of metastasis, was used for further analysis. We decided to use 0.5 as a cutoff value because patients above this value were more likely to experience a recurrence than to be metastasis-free (high-risk group). All other patients were assigned to the low-risk group. In the following text, the model as well as the prediction of the risk group is termed Binary ImaGe Colon Metastasis classifier (BIg-CoMet) reflecting that we used a black-and-white image of the tumor section.

The sample was characterized by counts and percentages for categorized variables and mean as well as standard deviation for continuous variables. The high-risk and the low-risk group were compared with a t-test or a Wilcoxon-Mann-Whitney test for continuous variables and a chi-squared test or a Fisher's exact test for categorical variables to identify differences in age, sex, or clinicopathological parameters.

Our main focus was on the occurrence of metastasis after diagnosis, and we performed a survival analysis with time until metastasis. We computed Kaplan-Meier curves for both risk groups and compared them with a log-rank test. The aim was to identify a significant separation between both groups, which would indicate that BIg-CoMet had provided accurate results and was capable of the prediction.

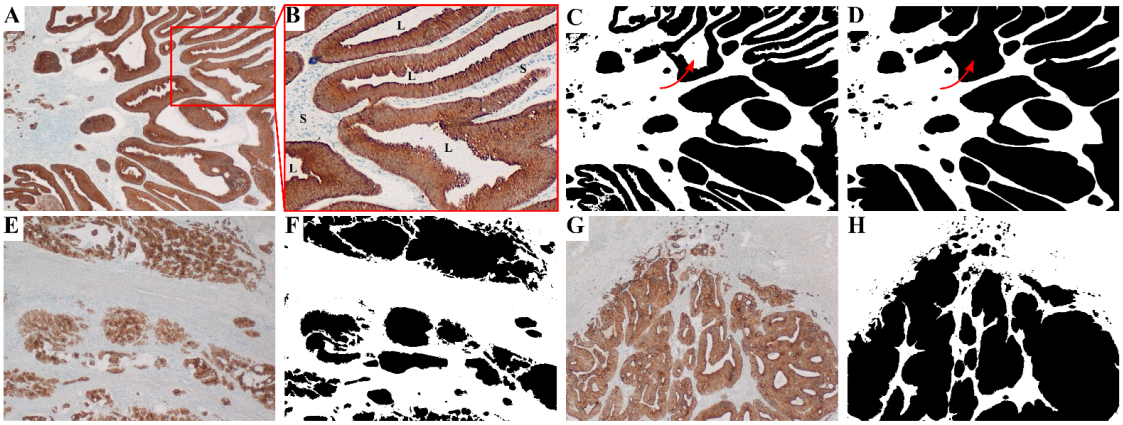


Figure 3.5: Different steps of the image preprocessing; (A) shows the selected stained region of the tissue sample and (B) a zoomed section where the difference between lumen (L) and stroma (S) is visible as lumen remains white whereas stroma is stained with blue dots. All stained images are translated into binary images and lumen is tried to be closed with ImageJ (C). The remaining lumen, which is marked with an arrow in (C) and (D) needs to be treated manually. (E) shows the slide of a patient without metastasis and the preprocessed image (F) predicted as low-risk with a risk of distant metastasis of 9.6%. (G) shows the slide of a patient with metastasis and the preprocessed image (H) which was predicted as high-risk with a predicted risk of distant metastasis of 85.5% by BIg-CoMet. Adapted from Schiele et al. (2021)

In the last step, we adjusted our stratification against several clinical parameters in a Cox proportional hazard regression. Due to the size of the sample, we previously fitted univariate Cox proportional hazard regression models for metastasis-free survival for each clinicopathological parameter and included only parameters with a p-value $< 0.3$ in the multivariable model. We computed the hazard ratio and the corresponding 95% confidence intervals as well as the p-value. Schoenfeld residuals were checked for the model to ensure that the proportional hazards assumption had been fulfilled. In a subgroup analysis, we repeated the above analysis for the subgroups of the UICC staging.

### 3.4.2 Results

#### Characteristics of the validation and the training sample

The training sample (TS) consisted of 163 patients, and the validation sample (VS) contained 128 patients. An overview of both groups is provided in Table 3.4. The patients were nearly the same age in both groups with a mean age of 69 years, and were mostly male (TS: 58%, VS: 61%). In nearly one of three cases, metastases had occurred during the follow-up period. The training sample had a median follow-up period of 5.2 years and the validation sample a slightly longer follow-up period of

5.8 years.

The clinicopathological characteristics of both samples were comparable. In the training sample, 113 cases (69%) had been classified as pT3, 78(48%) had a positive nodal status, and in 76 cases (59%) the tumor was located on the right side. The validation set had a higher fraction of pT3 (85%) and a lower number of patients with positive nodal status (41%) but nearly the same number of tumors on the right side (59%). During follow-up 43 patients in the training set and 53 patients in the validation set had died, with approximately one third of deaths attributed to the tumor.

## Comparison of the risk groups

All 128 patients in the validation collective received a predicted risk of occurring metastases during follow-up by BIg-CoMet and were classified as either low-risk (59%) or high-risk (41%). In the low-risk group, metastasis was detected in 10 of 76 patients (13%), whereas 31 out of 52 patients (41%) developed a metastasis in the high-risk group. Of note, the fraction of patients with metastasis in the high-risk group was similar to the corresponding fraction in the training sample.

We calculated several performance indicators to judge the performance of BIg-CoMet in terms of classification. The proportion of correctly classified patients (accuracy) was 75.8% (95% CI: 67.4–82.9%). The specificity was 75.9% (95% CI: 66.9 − 84.9%) and the sensitivity was 75.6% (95% CI: 62.5–88.8%). This result indicates that 75% of patients who had developed metastasis were classified as a high-risk patient and 75% of patients without the occurrence of a metastasis were assigned to the low risk group.
The positive predictive value of 59.6% (95% CI: 49.5–69.0%) can be interpreted as the fraction of patients in the high-risk group who had developed a metastasis, whereas the negative predictive value of 86.8% (95% CI: 79.2–92.0%) means that 86.8% of patients with low-risk had remained without a metastasis.

The two risk groups of BIg-CoMet were compared for all characteristics (Table 3.5). No differences in age, sex, and many clinicopathological characteristics were observed between the low-risk group and the high risk group. The high risk group showed a higher fraction of deceased patients (56% vs. 32%, p=0.011) and a higher proportion of patients with metastasis (60% vs. 13%, $p < 0.001$). Furthermore, the tumor proportions were lower in the medium category and higher at the tails in the high-risk group.

Table 3.4: Patient characteristics in training and validation set

| Variable | Validation set (n=128) | Training set (n = 163) |
|---|---|---|
| **Sociodemographic factors** | | |
| Age, mean (SD), y | 69 (12) | 69 (11) |
| Sex, n (%) | | |
| Female | 50 (39) | 68 (42) |
| Male | 78 (61) | 95 (58) |
| Follow-up duration, median, years | 5.8 | 5.2 |
| **Clinicopathological characteristics** | | |
| Tumor stadium, n (%) | | |
| pT3 | 109 (85) | 113 (69) |
| pT4 | 19 (15) | 50 (31) |
| Nodal status, n (%) | | |
| Negative | 75 (59) | 85 (52) |
| Positive | 53 (41) | 78 (48) |
| Mean lymph node harvest (n) | 21 (11) | 43 (20) |
| Positive lymph nodes (n) | 1.2 (2.3) | 2.0 (3.5) |
| UICC, n (%) | | |
| II | 75 (59) | 85 (52) |
| III | 53 (41) | 78 (48) |
| Grading, n (%) | | |
| Low grade | 76 (59) | 138 (85) |
| High grade | 52 (41) | 25 (15) |
| Vascular invasion, n (%) | | |
| Negative | 114 (89) | 140 (86) |
| Positive | 14 (11) | 23 (14) |
| Lymphovascular invasion, n (%) | | |
| Negative | 104 (81) | 122 (75) |
| Positive | 24 (19) | 41 (25) |
| Tumor budding, n (%) | | |
| Bd 1 | 103 (80) | 104 (64) |
| Bd 2 | 15 (12) | 36 (22) |
| Bd 3 | 10 (8) | 23 (14) |
| Location of tumor, n (%) | | |
| Right | 76 (59) | 91 (56) |
| Left | 52 (41) | 72 (44) |
| Microsatellite status, n (%) | | |
| MSS | 115 (90) | 137 (85) |
| MSI | 13 (10) | 24 (15) |
| Died, n (%) | | |
| Yes | 53 (41) | 43 (26) |
| No | 75 (59) | 120 (74) |
| Died of tumor, n (%) | | |
| Yes | 21 (16) | 18 (11) |
| No | 107 (84) | 145 (89) |
| Distant Metastasis, n (%) | | |
| Yes | 41 (32) | 64 (39) |
| No | 87 (68) | 99 (61) |
| Tumor proportion, mean (SD) | 0.358 (0.184) | 0.507 (0.111) |
| Tumor proportion, n (%) | | |
| Low | 21 (16) | 0 (0) |
| Medium | 80 (63) | 105 (65) |
| High | 27 (21) | 57 (35) |
| Adjuvant Chemotherapy, n (%) | | |
| Yes | 66 (52) | 69 (42) |
| No | 62 (48) | 94 (58) |

Table 3.5: Comparison of patient characteristics in the validation set stratified by the predicted risk group from BIg-CoMet

| Variable | BIg-CoMet Low risk (n=76) | BIg-CoMet High risk (n=52) | p-value |
|---|---|---|---|
| **Sociodemographic factors** | | | |
| Age, mean (SD), y | 69 (12) | 69 (12) | 0.753 |
| Sex, n (%) | | | 0.764 |
| Female | 31 (41) | 19 (37) | |
| Male | 45 (59) | 33 (64) | |
| Follow-up duration, median, years | 5.9 | 5.5 | 0.780 |
| **Clinicopathological characteristics** | | | |
| Tumor stadium, n (%) | | | 0.056 |
| pT3 | 69 (91) | 40 (77) | |
| pT4 | 7 (9) | 12 (23) | |
| Nodal status, n (%) | | | 0.278 |
| Negative | 48 (63) | 27 (52) | |
| Positive | 28 (37) | 25 (48) | |
| Mean lymph node harvest (n) | 20 (9) | 21 (12) | 0.911 |
| Positive lymph nodes (n) | 0.9 (1.7) | 1.6 (2.9) | 0.110 |
| UICC, n (%) | | | 0.278 |
| II | 48 (63) | 27 (52) | |
| III | 28 (37) | 25 (48) | |
| Grading, n (%) | | | 0.819 |
| Low grade | 44 (58) | 32 (62) | |
| High grade | 32 (42) | 20 (38) | |
| Vascular invasion, n (%) | | | 1.000 |
| Negative | 68 (89) | 46 (88) | |
| Positive | 8 (11) | 6 (12) | |
| Lymphovascular invasion, n (%) | | | 1.000 |
| Negative | 62 (82) | 42 (81) | |
| Positive | 14 (18) | 10 (19) | |
| Tumor budding, n (%) | | | 0.328 |
| Bd 1 | 64 (84) | 39 (75) | |
| Bd 2 | 8 (11) | 7 (13) | |
| Bd 3 | 4 (5) | 6 (12) | |
| Location of tumor, n (%) | | | 0.614 |
| Right | 47 (62) | 29 (56) | |
| Left | 29 (38) | 23 (44) | |
| Microsatellite status, n (%) | | | 0.896 |
| MSS | 69 (91) | 46 (88) | |
| MSI | 7 (9) | 6 (12) | |
| Died, n (%) | | | 0.011 |
| Yes | 24 (32) | 29 (56) | |
| No | 52 (68) | 23 (44) | |
| Died of tumor, n (%) | | | 0.004 |
| Yes | 6 (8) | 15 (29) | |
| No | 70 (92) | 37 (71) | |
| Distant Metastasis, n (%) | | | <0.001 |
| Yes | 10 (13) | 31 (60) | |
| No | 66 (87) | 21 (40) | |
| Tumor proportion, mean (SD) | 0.376 (0.171) | 0.326 (0.199) | 0.143 |
| Tumor proportion, n (%) | | | 0.002 |
| Low | 6 (8) | 15 (29) | |
| Medium | 56 (74) | 24 (46) | |
| High | 14 (18) | 13 (25) | |
| Adjuvant Chemotherapy, n (%) | | | 0.431 |
| Yes | 37 (49) | 29 (56) | |
| No | 39 (51) | 23 (44) | |

## Prognostic analysis of BIg-CoMet

In addition to a comparison between the risk groups, we were interested in the prognostic capabilities of BIg-CoMet and performed an survival analysis of metastasis-free survival. BIg-CoMet demonstrated a high capability of prognosis, as the Kaplan-Meier curves of the low- and high-risk groups were clearly separated (Figure 3.6A; log-rank-test: $p < 0.001$). We further performed a univariable Cox regression. The risk group classification was a significant factor, with a hazard ratio of 6.9 (95% CI: 3.4–14.2, $p < 0.001$).

In comparison, we conducted the same analysis for the UICC staging that is commonly used as an important tool to classify patients into risk groups. The Kaplan-Meier curve showed less separation between both groups than in the BIg-CoMet setting (Figure 3.6B).

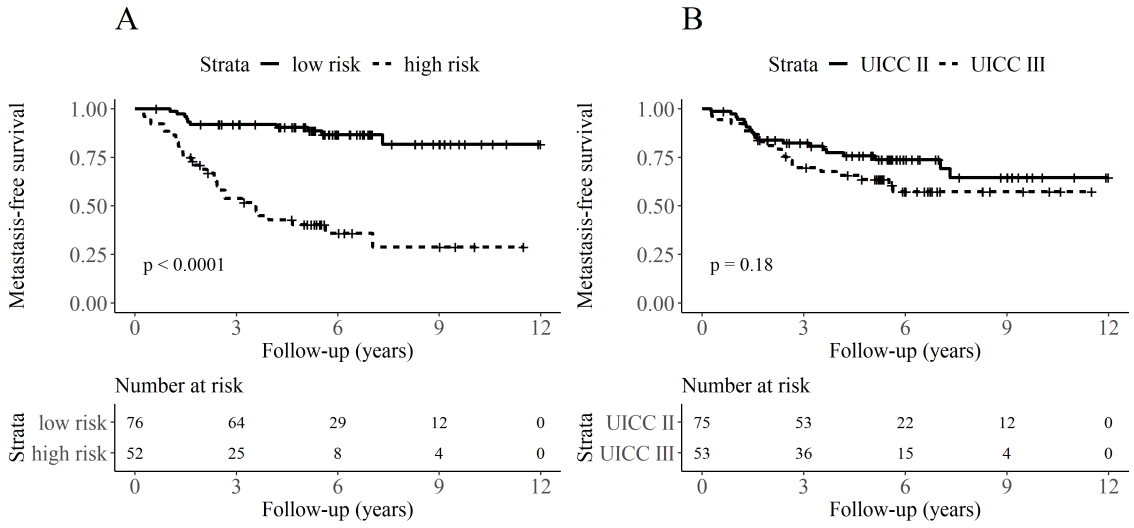


Figure 3.6: (a) Kaplan-Meier curves for occurrence of metastasis based on the classification of BIg-CoMet ($p < 0.0001$); (b) Kaplan-Meier curves for occurrence of metastasis based on the different UICC stages ($p = 0.18$). Adapted from Schiele et al. (2021)

We investigated whether BIg-CoMet remains a prognostic factor even if adjusted for other clinicopathological variables. In an initial step, we performed a univariable Cox regression for each of the variables of interest. Our BIg-CoMet risk group, age, tumor proportion, tumor budding, and tumor staging had a p-value $< 0.05$ (Table 3.6). We further considered all variables with a p-value $< 0.3$ and enhanced the set of parameters with sex, nodal status, lymphovascular invasion status, location of the tumor and microsatellite status.

A higher risk for the occurrence of metastasis was associated with the high-risk group classified by BIg-CoMet (HR = 5.4, 95% CI: 2.5–11.7, p-value $< 0.001$) and pT4 tumor staging (HR = 2.6, 95% CI: 1.1–6.0, p-value = 0.029). Patients with medium

Table 3.6: Univariate Cox PH regression for occurrence of metastasis.

| Variable | p-value |
|---|---|
| **Sociodemographic factors** | |
| Age (continuous) | $< 0.001$ |
| Sex (ref.: female) | 0.287 |
| **Clinicopathological factors** | |
| Big-CoMet risk group (ref.: low) | $< 0.001$ |
| Tumor proportion (ref.: High) | 0.003 |
| Tumor stadium (ref.: pT3) | 0.001 |
| Nodal status (ref.: negative) | 0.183 |
| Lymphovascular invasion (ref.: negative) | 0.059 |
| Tumor budding | 0.014 |
| Location of tumor (ref: right side) | 0.079 |
| Microsatellite status (ref.: MSS) | 0.094 |
| Mean lymph node harvest (n) | 0.373 |
| Grading | 0.722 |
| Vascular invasion | 0.315 |
| Adjuvant Chemotherapy | 0.750 |

tumor proportion had a lower risk for metastasis compared to patients with high tumor proportion (HR = 0.4, 95% CI: 0.2–0.99, p-value = 0.047). Microsatellite instable tumors showed a non-significant trend toward a lower risk for metastasis (p= 0.076) (Table 3.7).

## Subgroup analysis for UICC

We performed an additional subgroup analysis, in which both UICC subgroups were considered separately. We computed sensitivity as well as specificity and performed a survival analysis similar to the metastasis-free survival analysis as above.

BIg-CoMet performed well in the UICC II group, with an area under the curve of 0.76, a sensitivity of 55.0% and a specificity of 70.9%. Although the sensitivity was relatively low, patients without a metastasis were mainly correctly classified.
For UICC II, the risk groups differed significantly according to metastasis-free survival (log-rank-test, p = 0.016) and the risk group assignment was shown to be a prognostic predictor in a univariable Cox regression with a hazard ratio of 2.9 (95%-CI: $1.2 - 7.0$, p= 0.021).

BIg-CoMet demonstrated better ability to stratify patients correctly for patients with a UICC III classified tumor with a high area under the curve (AUC = 0.927) and a sensitivity of 95.2% as well as a specificity of 84.4%. This result suggests that BIg-CoMet may be especially helpful for UICC III tumors. Only few patients had

Table 3.7: Multivariable Cox PH regression for occurrence of metastasis.

| Variable | HR (95% CI) | p-value |
|---|---|---|
| **Sociodemographic factors** | | |
| Age (continuous) | 1.01 (0.98–1.04) | 0.592 |
| Sex (ref.: female) | 1.2(0.6–2.6) | 0.626 |
| **Clinicopathological factors** | | |
| Big-CoMet risk group (ref.: low) | 5.4(2.5–11.7) | $< 0.001$ |
| Tumor proportion (ref.: High) | | |
| Medium | 0.4(0.2–0.99) | 0.047 |
| Low | 0.7(0.3–1.7) | 0.410 |
| Tumor stadium (ref.: pT3) | 2.6(1.1–6.0) | 0.029 |
| Nodal status (ref.: negative) | 0.9(0.5–1.8) | 0.838 |
| Lymphovascular invasion (ref.: negative) | 1.3(0.6–3.2) | 0.517 |
| Tumor budding | 1.6(0.96–2.7) | 0.069 |
| Location of tumor (ref: right side) | 1.5(0.7–3.2) | 0.245 |
| Microsatellite status (ref.: MSS) | 0.2(0.02–1.2) | 0.076 |

developed metastasis during the follow-up period in the low-risk group. However, of patients with a high predicted risk, 80% had developed a metastasis.
Metastasis-free survival differed between both risk groups (log-rank-test, $p< 0.001$). BIg-CoMet was a prognostic predictor in the UICC III subgroup, although the results must be interpreted carefully due to the small sample size (HR = 45.2, 95% CI 6.0–340.8, $p < 0.001$)(Figure 3.7).

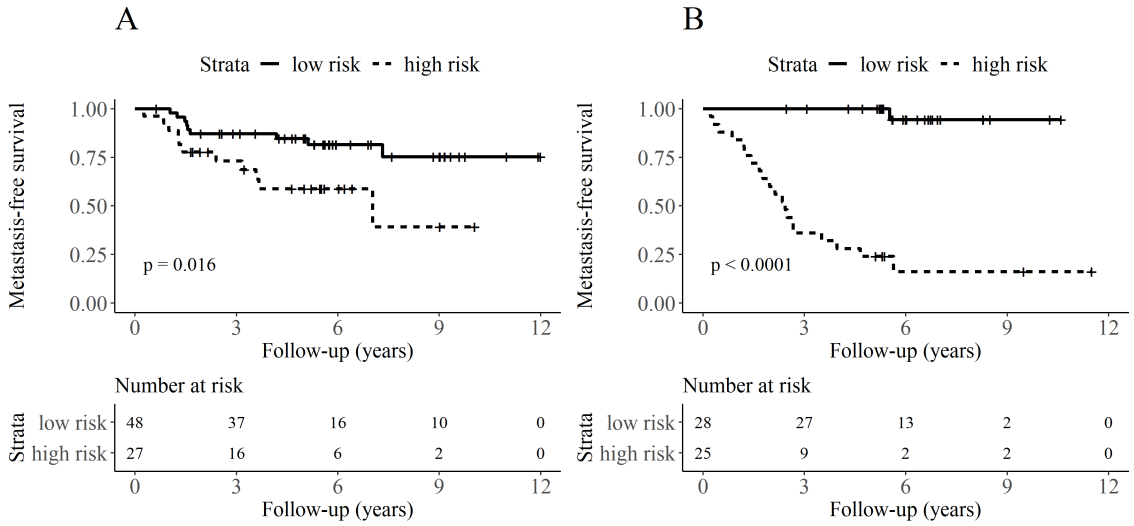


Figure 3.7: (a) Kaplan-Meier curves for occurrence of metastasis based on the classification of BIg-CoMet for UICC II cases ($p = 0.016$); (b) Kaplan-Meier curves for occurrence of metastasis based on the classification of BIg-CoMet for UICC III cases ($p < 0.001$). Adapted from Schiele et al. (2021)

### 3.4.3 Discussion

In this study, we have presented an application of modern statistical methods, especially artificial intelligence, to stratify patients with locally advanced colon cancer according to their risk of occurrence of metastasis. In comparison to many others applications, we aimed to provide a prognosis for the further course of the disease instead of a prediction of whether an illness is currently present. The decision of our algorithm, BIg-CoMet, was based on a selected and preprocessed region of a histological slide of the tumor area, which could be easily accessed, and the predicted probability for metastasis of a trained deep learning network for the image.

With this combination, we found a low-risk and a high-risk group, which are well separated based on metastasis-free survival. The accuracy of the classification was better for BIg-CoMet than for other established criteria such as UICC staging. Furthermore, even after adjustment for other clinically important factors, our risk group prediction remained an independent risk factor in a multivariable Cox regression. This quality of BIg-CoMet could be confirmed for both UICC subgroups.

To date, only a few studies by Bychkov et al. (2018), Kather et al. (2019), and Skrede et al. (2020) have investigated models based on deep learning for prognosis of metastasis occurrence in patients with colon cancer. All of these studies used H&E stained images, in contrast to our immunohistochemical staining.
Kather et al. (2019) performed a tissue classification of images in which they divided each image into several small tiles and predicted the type of tissue with a trained CNN. The proportions of different tissue types were then summarized in a score that could be used for prognosis. Another approach trained a deep learning model on the whole slide image without any prior tissue classification. Skrede et al. (2020) and Bychkov et al. 2018 trained their algorithm with large sample sizes of 828 and 280, respectively, and presented results indicating that their risk classification is a promising risk factor for metastasis-free survival (HR= 3.04, 95% CI: 2.07–4.47 and HR= 2.3, 95% CI: 1.79–3.03). BIg-CoMet could also be shown as a significant predictor, with a hazard ratio of 5.4 (95% CI: 2.5–11.7) in a smaller sample size of 163 patients but using binary images containing less information than an H&E image.

Our findings suggest that the structure of the tumor contains a sufficiently high amount of information for appropriate stratification. Because only black-and-white histological tumor images were used for the training, other components such as the nuclear configuration of the tumor cells or presence of tumor-infiltrating lymphocytes could not influence the algorithm. In our view, reducing the information

during the training of an algorithm is beneficial to focus on certain facets associated with a poor prognosis, such as the tumor architecture.

If algorithms such as BIg-CoMet could be implemented in daily routine, physicians would be supported in their decision making and the subsequent therapy decisions could be improved as follow-up intervals of patients in the high risk group could be shortened or further medical interventions could be recommended. The subgroup analysis suggests that in addition to already established criteria, BIg-CoMet can be an additional tool to further improve the classification of patients.

One limitation of our approach is that we can reduce the field of attention of the algorithm but still do not know which exact feature or structure might have led to the classification. One widely discussed issue is whether the difficult interpretation of such algorithms presents an obstacle to further dissemination, particularly in health science and how this problem might be solved (cf. F. Wang et al. (2020) and Castelvecchi (2016)). Applications that are not traceable by physicians may suffer from a lack of acceptance despite their strong performance and benefit for daily routine.
Furthermore, it must be recognized that all patients had been treated at the same center and their histological images had been prepared by the same pathologist. A wider application might introduce interobserver variability into the process. However, H&E staining is also not standardized and exhibits variability among different laboratories.

To address these issues, it will be necessary to validate our findings independently on another sample and also to investigate the performance in a prospective study to further implement BIg-CoMet in clinical routines. An important question is whether BIg-CoMet can be applied to other cancer entities, for example gastric cancer, without changing the structure of the algorithm or whether the type of staining must be changed to produce an accurate prognosis for the further course of a patient.

BIg-CoMet has demonstrated that a deep learning algorithm based on binary histological tumor images can stratify patients with colon cancer with regard to their risk of occurrence of metastasis. Further validation using other samples is needed to provide insight regarding important structures in the image and ensure its wider application.

# Chapter 4

# Development of Scores for Medical Research with Generalized Linear Regression Models and Methods from Survival Analysis

**Keywords:** prognostic score, generalized linear models, optimal cutoff determination, survival analysis, risk groups, aneurysmatic subarachnoid hemorrhage, oligometastatic colon cancer

## 4.1 Development of a score for prediction of shunt risk for patients after an aSAH with generalized linear regression models

One major task of statistics in medicine concerns the development of appropriate predictive models and scores by using measured patient characteristics. These predictions can be the survival time of a patient, the patient's risk for a disease or an information useful for the treatment of the patient. Dependent on the task, there is a high number of options for statistical models. Generalized linear models and particularly logistic regressions are often the basis of statistical models. As an example, in this chapter, we introduce the theory of the development of a score by a logistic regression and applied this in the context of patients with an aneurysmatic subarachnoid hemorrhage (aSAH), which is a special kind of cerebral hemorrhage. For patients suffering from an aSAH, a statistical model should help to predict the risk of a shunt implantation. Further, we built a score based on predictors from the derived model to ensure an easy and traceable usage in hospitals. The appli-

cation arose from a collaboration with Bastian Stemmer from the Department of Neurosurgery at the University Hospital Augsburg.

### 4.1.1 Biological background

Around 5% of all strokes are caused by an aSAH, according to Bederson et al. (2009). In nearly 80% of cases, the initial event of an aSAH is the rupture of an aneurysm in the basal arteries of the brain. Consequently, the intracranial pressure increases immediately and leads to a wide range of symptoms from sudden severe headache and nausea to unconsciousness in critical cases (cf. Spendel (2008)).

Because of the high mortality rate of 50% for patients suffering from an aSAH, an immediate treatment is important to decrease the risk for possible consequential damages. One possible complication is the build-up of a hydrocephalus that occurs in around 25% of all patients within a few days after the aSAH. A hydrocephalus is a collection of cerebrospinal fluid that cannot be dissolved and hence leads to an increased pressure in the brain. For treatment, a temporary shunt is implanted to avoid further neurological damages as explained by Hasan et al. (1989). If the ability to transport the liquor in the brain is not improved during the follow-up period, a permanent shunt must be implanted. However, this implant reduces the life quality of a patient and should therefore only be done if it is absolutely necessary. To support physicians in their decisions, a statistical model is developed to predict the risk for a permanent shunt implantation.

### 4.1.2 Data

For this study, patients who had suffered from an aSAH and underwent an endovascular treatment between January 2010 and July 2015 at the Department of Neurosurgery of the University Hospital Augsburg were included. Furthermore, patients had to have received a cerebrospinal fluid drainage via an external ventricular drainage within three days after a beginning treatment, and patients who were not alive 12 months after aSAH were removed from the sample. The outcome was the defined as whether a patient obtained a shunt or not within 12 months after their diagnosis to avoid patients who received a shunt not at the hospital but at a later point in time. For each patient, demographical variables (age, sex), clinical variables (shunt, different scores concerning the health status), radiological variables (measurements of the ventricle extracted from CT imaging of the brain), and volume of cerebrospinal fluid loss for each day were available and are described below in detail.

## Duty of shunt implantation

The duty of a shunt was modeled as a binary variable, where 0 indicated the absence of an implanted shunt within 12 months after an aSAH and 1 indicated that the course of the disease necessitated a permanent shunt either during hospitalization or during the first year after an aSAH. This decision was made according to medical guidelines.

## Further variables

### Hunt and Hess grading

The Hunt and Hess-Grading is used to judge the severity of an aSAH and the perioperative mortality according to patients symptoms (cf. Hunt and Hess (1968)). The scale can be divided into five grades between 1 (best health status) and 5 (worst health status)(Table 4.1).

Table 4.1: Criteria of the five grades of the Hunt and Hess-Grading

| Category | Criteria |
|---|---|
| Grade I | Asymptomatic, or minimal headache and slight nuchal rigidity |
| Grade II | Moderate to severe headache, nuchal rigidity, no neurological deficit other than cranial nerve palsy |
| Grade III | Drowsiness, confusion, or mild focal deficit |
| Grade IV | Stupor, moderate to severe hemiparesis, possibly early decerebrate rigidity and vegetative disturbances |
| Grade V | Deep coma, decerebrate rigidity, moribund appearance |

### Glasgow Coma Scale

The Glasgow Coma Scale is used to assess a person's consciousness according to three items: motor response, verbal response, and eye response. The items are scored according to Table 4.2, and the sum of these items builds the Glasgow Coma Scale with a range between 3, negative status, and 15, positive status (Teasdale and Jennett (1974)).

### Vasospasms

Especially in the context of aSAH, there is an increased risk for a Vasospasm, which is an arterial spasm that leads to vasoconstriction. As a result, tissue ischemia and, hence, tissue death can occur. The variable was binary encoded.

Table 4.2: Criteria of the Glasgow Coma Scale

| Score | Eye Response | Verbal Response | Motor Response |
|---|---|---|---|
| 1 point | Does not open eyes | Makes no sounds | Makes no movements |
| 2 points | Opens eyes in response to pain | Makes sounds | Extension to painful stimuli (decerebrate response) |
| 3 points | Opens eyes in response to voice | Words | Abnormal flexion to painful stimuli (decorticate response) |
| 4 points | Opens eyes spontaneously | Confused, disoriented | Flexion / Withdrawal to painful stimuli |
| 5 points | | Oriented, converses normally | Localizes to painful stimuli |
| 6 points | | | Obeys commands |

### mRS at admission

Rankin (1957) introduced the modified Rankin Scale (mRS) that can be used to determine the severity of a patient's disability after a stroke or another neurological disease. Low values indicate no or a few symptoms for a disability, whereas values of 4 to 5 are assigned for (moderately) severe disabilities. The highest possible rating, 6, is equal to the death of the person.

### Fisher Grading Scale

The Fisher Grading Scale judges the amount of subarachnoid hemorrhage on CT images. The scale possesses four different levels with Grade 1 indicating the lowest amount and Grade 4 indicating the most severe level (cf. Fisher et al. (1980)).

### Size and region of aneurysm

For every patient, the size and the region of the aneurysm were determined. The aneurysm reasoning the SAH is mainly located in four regions: the anterior cerebral artery (ACA), the middle cerebral artery (MCA), the internal carotid artery (ICA), and the basilar artery or vertebrobasilar We included the location to examine whether this variable was associated with the outcome.

### Measurements resulting from CT imaging

Based on a CT image, relevant properties of ventricles were measured and labeled with letters A to F (Figure 4.1). For example, C measured the width of the third ventricle.
We further calculated indices with the taken measurements for further analysis.:

- Evans ratio = $\frac{A}{E}$

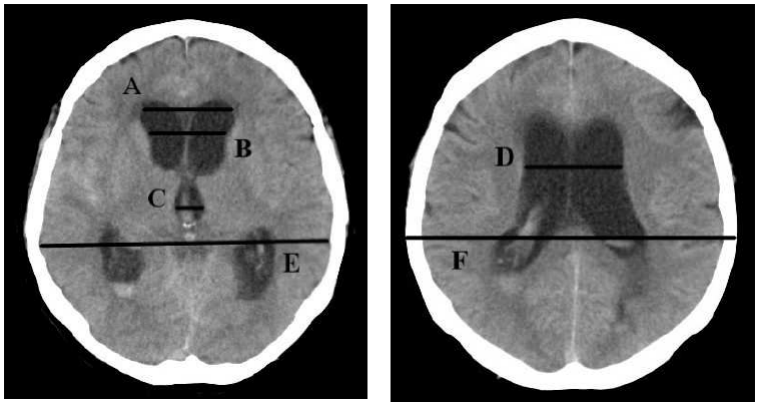


Figure 4.1: CT image of the brain with annotations for the measurements A-F. This image was taken at the University Hospital Augsburg and is adapted from Stemmer (2019)

- Third-ventricle-index = $\frac{C}{E}$
- Cella-media-index = $\frac{D}{F}$
- Ventricular score = $[(A + B + C + D)/E]$

### 4.1.3 Statistical approaches

All statistical analysis were conducted with the free-available software R (version 4.1.1). In the following subsections, the different steps of the analysis are described in detail.

#### Descriptive statistics and comparison of patients dependent on the duty of shunt

To describe the characteristics of the sample, we calculated the counts and percentage of categorized variables and means as well as standard deviations for continuous variables. We split the sample into two groups dependent on the implantation of a permanent shunt and investigated whether there are differences between these two groups by conducting statistical tests. For the comparison of a categorized variable, we used chi-squared tests and Fisher's exact tests. Further, in case of continuous variables, we chose t-tests if the data was normally distributed or Wilcoxon-Mann-Whitney tests.

#### Cutoff determination

Because the developed score should be widely and easily applicable at hospitals, we decided to dichotomize continuous variables to facilitate the computation and

increase its comprehensibility for not only the doctors but also the patients. Like in many other tasks, the acceptance of a score is improved if it can be explained in a few simple words to patients. To find optimal cutoff values for all variables, we developed Algorithm 1.

First, we defined a suitable discrete parameter space $\Theta_{x_i} = \{\theta_0, \theta_1, ...\}$ for each

**Algorithm 1:** Optimal cutoff determination

**Result:** Determine of the optimal cutoff-value
Initialization of a suitable parameter space $\Theta_{x_i}$
**for** $\theta \in \Theta_{x_i}$ **do**
  Define a binary variable $x_\theta$
  Fit a univariate logistic regression and calculate $\text{AIC}_\theta$
**end**
**return** *Select* $\theta^*$ *as* $argmin_{\theta \in \Theta} AIC_\theta$

variable $x_i$ according to the range of $x_i$. Then, for every $\theta$ in $\Theta_{x_i}$, we created a binary variable $x_\theta$ based on the concerned variable $x_i$:

$$x_{i,\theta} = \begin{cases} 0 & \text{for } x \leq \theta \\ 1 & \text{for } x > \theta \end{cases}$$

With the calculated $x_{i,\theta}$ as an independent variable, we fitted a univariate logistic regression model for the duty of shunt implantation. The different models and ,hence, cutoff values were evaluated based on the Akaike Information Criterion (AIC). As a lower AIC value represents a better cutoff, we finally assigned the cutoff value as those with the lowest AIC. For this selected model, we reported the odds ratio (OR), the 95% confidence interval (CI), and the p-value of the independent variable.

## Generalized linear model

In a next step, we determined significant predictors within a multiple generalized linear model (GLM) using the binary variables with the determined cutoffs as independent variables. As the output (shunt implantation) is binary, a logistic regression model was suitable. Because of the small sample size, we pre-selected candidates from the whole set of variables. Our choice was reasoned by the AIC of the univariate logistic model such that we took those variables whose univariate model had the lowest AIC. To remove variables without significant association from the multivariable model, we performed an AIC backwards-stepwise selection. We reported the results of the final model by ORs with 95%-CI and p-values. Further, we visualized the ORs with a forest plot.

Furthermore, we repeated Algorithm 1, but this time with a high-dimensional parameter space of all $k$ parameters of the final model, $\hat{\Theta} = \Theta_{x_1} \times \Theta_{x_2} \times ... \times \Theta_{x_k}$, to check whether our choice of cutoffs could be improved. For every $\theta \in \hat{\Theta}$, we fitted a logistic regression model based on the binary variables $x_{1,\theta_1}, x_{2,\theta_2}, ..., x_{k,\theta_k}$ and compared the resulting AIC. The optimal parameter $\theta^*$ was chosen from the model with the lowest AIC.

### Development and evaluation of a score

The final model built the origin for developing a score. For every binary variable $x_{i,\theta_i^*}$, we decided whether a value of 0 or a value of 1 is awarded with one point based on the OR. If the variable had an $OR > 1$ in the multiple model then $x_{i,\theta_i^*} = 1$ showed a worse prognosis and hence one point was added to the score. In case of OR $< 1$ we took the other decision.

With this procedure, we obtained three different scores that were computed for every patient:

1. $s_{dis}$, which is a discrete score like the one described above,
2. $s_{pro}$, which is based on the predicted probability of the model,
3. $s_{wdis}$, which is a discrete score, but this time the variables are weighted by their OR in the multivariable model.

The performance of classification between the three scores was compared by their area under the ROC curve (AUC).

## 4.1.4 Results

The sample of patients satisfying the inclusion criteria consisted of 91 patients of whom 64% (n = 58) underwent a permanent shunt implantation. The average age of patients in this sample was $55.3 \pm 14.0$ years of age. 35.2% were assigned a Hunt and Hess grading of $> 3$, which is associated with a severe status, and 38.5% obtained a score of 9 or less on the Glasgow Coma Scale, showing an impaired consciousness. In nearly half of all treated patients (45.1%), a severe disability was diagnosed at admission. Moreover, a vasospasm was present in 46.2% and an intraventricular hemorrhage was found in 68.1%.
During the first seven days of treatment the daily volume of liquor drainage of all patients was measured and listed. Their individual mean of the first seven days was calculated. On average, the computed mean of volume was $229.5ml \pm 76.9ml$.

Furthermore, the aneurysm of the patients had a mean size of $8.0mm \pm 2.7mm$ and the third ventricle was wider than 6 mm for 73.6%. On average, the determined ventricular score was $0.71 \pm 0.13$.

## Comparison between the two groups

After a description of the whole sample, we divided the patients into two groups dependent on their need for a permanent shunt implantation. We found differences between both groups, especially for clinical parameters and CT measurements from the aneurysm. There were no differences between the groups concerning the intraventricular hemorrhage, but the group without the implantation of a permanent shunt was on average younger (51.5 years vs. 57.4 years, $p < 0.001$ and had a wealthier health status at admission with a lower fraction on the modified Rankin Scale $> 4$ (18.2% vs. 60.3%, $p < 0.001$)), a lower fraction of GCS $\leq 9$ (15.2% vs. 51.7%, $p < 0.001$), and a lower fraction of HH $> 3$ (15.2% vs. 65.9%, $p = 0.005$). Moreover, the fraction of patients for whom a vasospasm was diagnosed was lower than for patients with a shunt (21.2% vs. 60.3%).

Patients with a shunt had a mean daily volume of liquor drainage of 243.7 ml during the first week, whereas without a shunt only 204.4 ml was measured. In a comparison, we found differences between the groups for parameters of the aneurysm. The third ventricle was wider (mean: 8.6 mm vs. 6.9 mm; p=0.003) and the ventricular score was higher (mean: 0.74 vs. 0.65; p=0.003) with a shunt implantation, but no differences in the location of the aneurysm could be observed (Table 4.4).

### Determination of optimal cutoff values

According to Algorithm 1, we started the procedure by choosing an appropriate parameter grid $\Theta_{x_i}$ for each variable $x_i$. Then, we followed the steps described above to obtain the best suited cutoff for the parameter. In case of GCS, we made an exception and took not the optimal but a value with an AIC close to the optimal. This is reasoned by its common use as the threshold for GCS and the habit of physicians in their daily routine. A change could lead to reduced acceptance for the developed score. Because there is hardly a difference between both parameters with respect to the AIC, our approach was acceptable.

We performed univariate logistic regression models to investigate associations between several parameters and the implantation of a permanent shunt. Sex was not associated with a shunt. However, greater odds for implantation were associated with a higher age (OR: 5.88[2.20–16.91]), a higher HH grading (OR: 4.88[1.76 –

Table 4.3: Characteristics of the sample

| Variable | N | % |
|---|---|---|
| **Shunt implantation** | | |
| Yes | 58 | 64.0 |
| No | 33 | 36.0 |
| **Sociodemographic factors** | | |
| Age | 55.2 ± 14.0 | |
| ≤ 45 years | 24 | 26.4 |
| > 45 years | 67 | 73.6 |
| Sex | | |
| Female | 63 | 69.2 |
| Male | 28 | 30.8 |
| **Clinical data** | | |
| Hunt & hesse grading | | |
| ≤ 3 | 59 | 64.8 |
| > 3 | 32 | 35.2 |
| Fisher | | |
| ≤ 3 | 28 | 30.8 |
| > 3 | 63 | 69.2 |
| GCS | | |
| ≤ 9 | 35 | 38.5 |
| > 9 | 56 | 61.5 |
| Modified Rankin Scale at admission | | |
| ≤ 4 | 50 | 54.9 |
| > 4 | 41 | 45.1 |
| Vasospasm | | |
| No | 49 | 53.8 |
| Yes | 42 | 46.2 |
| Liquor (CSF)-Drainage [Mean over first 7 days] | 229.5 ± 76.9 | |
| ≤ 180 ml | 23 | 25.3 |
| > 180 ml | 68 | 74.7 |
| Intraventricular Hemorrhage | | |
| No | 29 | 31.9 |
| Yes | 62 | 68.1 |
| **Aneurysm** | | |
| Size | 7.1 ± 3.6 | |
| Region | | |
| MCA | 9 | 9.9 |
| ACA | 37 | 40.7 |
| ICA | 24 | 26.4 |
| Vertebrobasilar | 21 | 23.1 |
| **Measurements from CT imaging** | | |
| Width of third ventricle | 8.0 ± 2.7 | |
| ≤ 6 mm | 24 | 26.4 |
| > 6 mm | 67 | 73.6 |
| Ventricle Score | 0.71 ± 0.13 | |
| ≤ 0.6 | 17 | 18.7 |
| > 0.6 | 74 | 81.3 |
| Evans Index | 0.28 ± 0.04 | |
| Third-Ventricle-Index | 0.06 ± 0.02 | |
| Cella-Media-Index | 0.19 ± 0.04 | |

Table 4.4: Comparison of patient characteristics stratified by the duty of a permanent shunt implantation

| | Shunt n = 58 | | No Shunt n = 33 | | |
|---|---|---|---|---|---|
| Variable | N | % | N | % | p-value |
| **Sociodemographic factors** | | | | | |
| Age | 57.4 ± 13.0 | | 51.5 ± 15.0 | | |
| ≤ 45 years | 8 | 13.8 | 16 | 48.5 | < 0.001 |
| > 45 years | 50 | 86.2 | 17 | 51.5 | |
| Sex | | | | | |
| Female | 38 | 65.5 | 25 | 75.8 | 0.435 |
| Male | 20 | 34.5 | 8 | 24.2 | |
| **Clinical data** | | | | | |
| Hunt & hesse grading | | | | | |
| ≤ 3 | 31 | 34.1 | 28 | 84.8 | 0.005 |
| > 3 | 27 | 65.9 | 5 | 15.2 | |
| Fisher | | | | | |
| ≤ 3 | 16 | 27.6 | 12 | 36.4 | 0.261 |
| > 3 | 42 | 72.4 | 21 | 63.6 | |
| GCS | | | | | |
| ≤ 9 | 30 | 51.7 | 5 | 15.2 | < 0.001 |
| > 9 | 28 | 48.3 | 28 | 84.8 | |
| Modified Rankin Scale at admission | | | | | |
| ≤ 4 | 23 | 39.7 | 27 | 81.8 | < 0.001 |
| > 4 | 35 | 60.3 | 6 | 18.2 | |
| Vasospasm | | | | | |
| No | 23 | 39.7 | 26 | 78.8 | < 0.001 |
| Yes | 35 | 60.3 | 7 | 21.2 | |
| Liquor (CSF)-Drainage [Mean over first 7 days] | 243.7 ± 80.7 | | 204.4 ± 63.4 | | |
| ≤ 180 ml | 8 | 13.8 | 15 | 45.5 | 0.002 |
| > 180 ml | 50 | 86.2 | 18 | 54.5 | |
| Intraventricular Hemorrhage | | | | | |
| No | 17 | 29.3 | 12 | 36.4 | 0.645 |
| Yes | 41 | 70.7 | 21 | 63.6 | |
| **Aneurysm** | | | | | |
| Size | 7.5 ± 4.0 | | 6.4 ± 2.5 | | 0.383 |
| Region | | | | | |
| MCA | 4 | 6.9 | 5 | 15.2 | 0.096 |
| ACA | 25 | 43.1 | 12 | 36.4 | |
| ICA | 12 | 20.7 | 12 | 36.4 | |
| Vertebrobasilar | 17 | 29.3 | 4 | 12.1 | |
| **Measurements from CT imaging** | | | | | |
| Width of third ventricle | 8.6 ± 2.8 | | 6.9 ± 2.1 | | 0.003 |
| ≤ 6 mm | 8 | 13.8 | 16 | 48.5 | < 0.001 |
| > 6 mm | 50 | 86.2 | 17 | 51.5 | |
| Ventricle Score | 0.74 ± 0.13 | | 0.65 ± 0.11 | | 0.003 |
| ≤ 0.6 | 4 | 6.9 | 13 | 39.4 | < 0.001 |
| > 0.6 | 54 | 93.1 | 20 | 60.6 | |
| Evans Index | 0.29 ± 0.04 | | 0.26 ± 0.04 | | 0.002 |
| Third-Ventricle-Index | 0.07 ± 0.02 | | 0.05 ± 0.02 | | 0.003 |
| Cella-Media-Index | 0.20 ± 0.04 | | 0.18 ± 0.03 | | 0.003 |

Table 4.5: Determination of optimal cutoff values for continuous parameter and univariable logistic regression for the duty of a permanent shunt

| Variable | Optimal cutoff | AIC | OR [CI] | p-value |
|---|---|---|---|---|
| **Sociodemographic factors** | | | | |
| Age | $> 45$ | 110.45 | $5.88[2.20 - 16.91]$ | $< 0.001$ |
| Sex | | 122.10 | | n.s. |
| **Clinical data** | | | | |
| Hunt & hesse grading | $> 3$ | 113.38 | $4.88[1.76 - 15.94]$ | 0.004 |
| Fisher | $> 3$ | 122.44 | | n.s. |
| GCS | $\leq 9$ | 110.34 | $6.00[2.17 - 19.62]$ | 0.001 |
| Modified Rankin Scale at admission | $> 4$ | 107.13 | $6.85[2.58 - 20.74]$ | $< 0.001$ |
| Vasospasm | | 109.59 | $5.65[2.20 - 16.11]$ | $< 0.001$ |
| Liquor (CSF)-Drainage [Mean over first 7 days] | $> 180$ | 112.32 | 5.21[1.93–14.97] | 0.001 |
| Intraventricular Hemorrhage | | 122.72 | | n.s. |
| **Aneurysm** | | | | |
| Size | $> 10$ | 113.50 | $8.35[1.53 - 155.86]$ | 0.047 |
| Width of third ventricle | $> 6$ | 113.35 | 4.52[1.75–12.22] | 0.002 |
| Evans Index | $> 0.29$ | 115.19 | 4.24[1.53–13.88] | 0.009 |
| Ventricle Score | $> 0.6$ | 108.91 | 8.77[2.75–34.14] | $< 0.001$ |

15.94]), a lower GCS (OR: $6.00[2.17 - 19.62]$), a severe disability (mRS $> 4$) (OR: 6.85[2.58–20.74]), and a higher daily mean liquor drainage during the first seven days (OR: 5.21[1.93–14.97]). On the other hand, there was no association with the Fisher score or the intraventricular hemorrhage.

Finally, measurements of the tumor and CT-based imaging were explored. From the analysis, we found that the more severe the aSAH is, the higher the odds for a shunt implantation can be estimated. In detail, the following parameters showed significant associations with an increased odds of a permanent shunt: a greater size of the aneurysm (OR: $8.35[1.53 - 155.86]$), a wider third ventricle (OR: 4.52[1.75–12.22]), a higher Evans index (OR: 4.24[1.53–13.88]), and a higher ventricular score (OR: 8.77[2.75–34.14]).

## Multiple model

The score was developed by fitting a multiple model. Although a higher sample size would be necessary to obtain more precise results, the multivariable model contained eight variables with the lowest AIC. In our example, this was smaller than 113.50. Of note, according to this approach, the selection was not focused on a special group of parameters but composed of variables distributed across parameter categories such as sociodemographic parameter (age), clinical data (HH grading, GCS, mRS, vasospasm, and mean daily volume of liquor drainage during the first seven days after implantation of a temporary shunt), and measurements of the aneurysm (width of the third ventricle and the ventricular score).

We removed non-significant variables via an AIC based stepwise backward selection that resulted in five remaining variables. Risk factors for a shunt implantation were a higher age (OR: 8.10 [2.04–38.44]), a smaller GCS (OR: 8.88 [1.81–74.13]), a vasospasm (OR: 4.43 [1.32 − 16.60]), a higher mean volume of liquor drainage during the first seven days (OR: 7.88 [2.03–36.49]), and a higher ventricular score (OR: 15.23 [2.76–133.41]) (Figure 4.2).

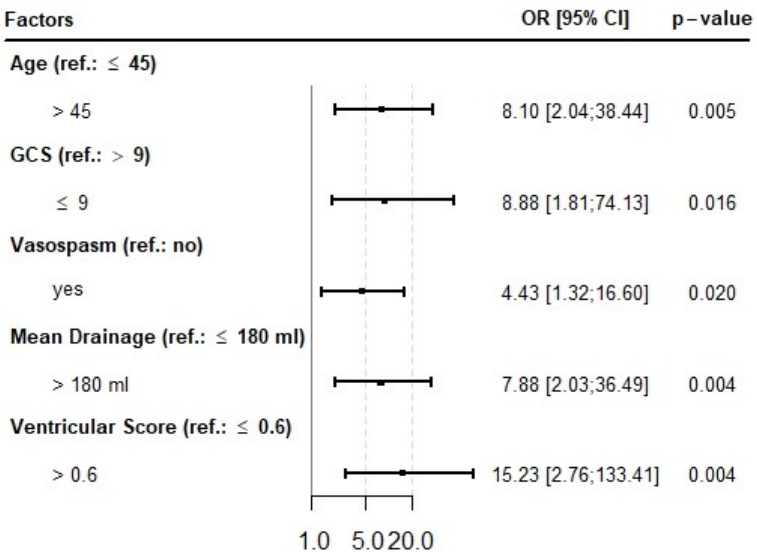


Figure 4.2: Forestplot of all variables remaining in the multiple logistic regression model for the duty of a permanent shunt after selection

Furthermore, we repeated the analysis of suitable cutoffs again but simultaneously for all parameters remaining in the final model. $\hat{\Theta} = \Theta_{x_1} \times \Theta_{x_2} \times \Theta_{x_3} \times \Theta_{x_4} \times \Theta_{x_5}$ was defined as the parameter space for Algorithm 1. We obtained the same result like the combination of the univariate cases, which underlines that the choice of thresholds was suitable for the parameters.

### Score and evaluation

Based on the fitted multiple model, we experienced three different approaches for calculating a risk prediction score for shunt implantation.

1. $s_{dis}$: For every of the five dichotomized risk factors (age, GCS, vasospasm, mean daily liquor, and ventricular score) one point was awarded if the risk factor was present for a patient and its sum was assigned as the final score. The score has a range of 0 to 5.

2. $s_{pro}$: We took the outcome of the fitted logistic regression model as score. This was possible since the model predicts the probability for a shunt implantation

Table 4.6: Weighting of the factors of the score in three different variants

| Variable | Score $s_{dis}$ | Coefficient for $s_{pro}$ | Score $s_{wdis}$ |
|---|---|---|---|
| Age $> 45$ | 1 point | 2.09 | 2 points |
| Vasospasm | 1 point | 1.49 | 1 point |
| GCS $\leq 9$ | 1 point | 2.18 | 2 points |
| Mean of daily volume of liquor $> 180$ | 1 point | 2.06 | 2 points |
| Ventricular score $> 0.6$ | 1 point | 2.72 | 4 points |
| Range of the score | $0-5$ | $0-1$ | $0-11$ |

based on independent variables, so all predictions stayed in the interval $[0, 1]$. Because of the use of binary predictors, the outcomes from this model are also discrete but with more differentiation.

3. $s_{wdis}$: We created a discrete score that is a combination of $s_{dis}$ and $s_{pro}$. Every of the five dichotomized risk factors received a score, but in comparison to $s_{dis}$, not only one point is awarded, but this time the score is dependent on the OR from the multiple model. We divided the OR by 4 and rounded the result, so we obtained one point for the presence of a vasospasm, two points for age above the cutoff, GCS below the cutoff, mean of the daily volume of liquor drainage above the cutoff, and four points for a ventricular score above the cutoff.

To evaluate the described scores, we calculated the AUC. We found that all three scores demonstrated an accurate predictive power with an AUC over 0.90 and hardly differed from each other (Table 4.7). This was a strong indication that all three scores are useful for clinical prediction. As we are interested in a score with a simple

Table 4.7: Comparison of the scores with respect to the AUC

| Type of score | AUC | 95 %-Confidence Interval |
|---|---|---|
| Score $s_{dis}$ | 0.90 | [0.84,0.96] |
| Score $s_{pro}$ | 0.91 | [0.85,0.96] |
| Score $s_{wdis}$ | 0.90 | [0.85,0.95] |

calculation and a tractable explanation for the patient, we decided to use the $s_{dis}$ as the final score. Its performance with respect to AUC was on the same level as the other tested scores, but its interpretability was higher.

To show that the selected score is suitable to predict the risk, we computed the fraction of patients with a shunt implantation for every level of the score. With a score of 0 or 1, hardly any patient needed the implantation of a permanent shunt. However, this risk increased with an increased score, as patients with a score of 2

received a shunt in 30% of cases and patients with a score of 3 in 62.5%. Moreover, a score greater than or equal to 4 resulted in an implantation in nearly all cases.

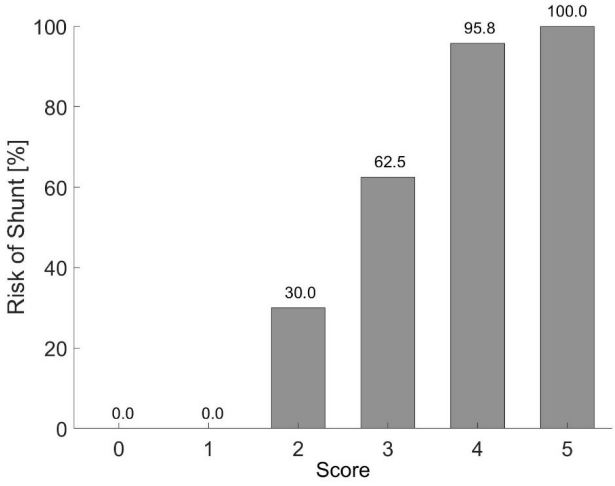


Figure 4.3: Fraction of patients receiving a permanent shunt for every level of the final score.

### 4.1.5 Discussion

In this section, we presented a score to predict the risk of shunt implantation for patients suffering from an aSAH. To incorporate continuous variables in the score, we determined optimal cutoff values for those with respect to the AIC. We identified risk factors in a multiple logistic regression model via AIC-based backward selection. Finally, we included the age, the GCS, the presence of a vasospasm, the mean drainage, and the ventricular score in the score.

We compared three different approaches for the calculation of a score: assigning one point to every identified risk factor, assigning each patient the predicted probability from the logistic regression model, and assigning a weight proportional to the computed OR to each variable. We found that, in our setting, there were hardly any differences and decided to score each risk factor with one point. Patients with a score of 4 or 5 received a shunt in nearly 100%, whereas the percentage of patients with a shunt was low for patients with a low score.

Moreover, this study has some limitations. The low case number was a restriction for our study. Due to a small sample size, we could not split the sample into a validation and exploration set, which denies an evaluation of the developed score for a separate data set. Hence, the score should be tested further for an independent set of patients to prove its predictive power and enable a general use.

## 4.2 Development of a score for stratification of patients according to their survival using methods from survival analysis

This project was based on a cooperation between the III. Medical Clinic at the University Hospital Augsburg (Giuliano Velazquez and Rainer Claus) and the Institute of Mathematics (Gernot Müller and Stefan Schiele). The method that is described on the following pages was applied to develop a score in Velázquez et al. (2022).

### 4.2.1 Biological background

Patients who have been diagnosed with colorectal cancer are likely to develop metastases in the lung or liver, which often causes death. If multiple and widespread metastases are already present at the time of diagnosis, the therapy becomes particularly complicated, and a surgical resection is often impossible. Patients with low-stage liver metastases, who are called oligometastatic, might benefit from resection (cf. Tomlinson et al. (2007)). Although this state is not uniformly defined, it is characterized as an intermediate stage between the localized and the widespread form of the disease, in which only a few metastases are present. This definition was introduced by Weichselbaum and Hellman (2011).

Whether performing surgery on an oligometastatic patient is desirable depends on their chances of survival after the surgery. When the primary cancer is treated and all metastases can be detected and ablated, a long disease-free period is possible. However, if the survival time is short, a medical intervention might not be justified because of the high physical burden of surgery. Therefore, in this study we attempted to find a score that can help physicians to assess overall survival time and disease-free survival time preoperatively. A newly introduced score is employed to identify subgroups which are likely to benefit from a surgical resection.

Malik et al. (2007) and Fong et al. (1999) found several variables that exert a significant influence on the overall survival of patients. They found that the nodal status of the primary tumor, the number of metastases that are detected in a patient at diagnosis, the size of the tumor, and inflammatory response to the tumor (IRT), among others, are important risk factors for overall and disease-free survival. Despite detecting patients who might benefit from a surgery in a satisfactory manner, those factors are still not reliable enough for clinical application, as explained by Schreckenbach et al. (2015).

Our sample consists of patients from several hospitals who had undergone a surgical resection of liver metastases from colorectal cancer. We investigated which variables are associated with overall survival in course of a survival analysis. The set of variables includes previously identified risk factors as well as new variables such as the sidedness of the tumor. We used the variables that we identified as relevant to develop a score. This score stratifies the patients into subgroups, and we validated it on an independent set of patients that was not considered during training.

### 4.2.2 Data

A total of 512 patients with metastatic colorectal cancer who had been treated for a primary tumor and had undergone surgical resection of *de novo* liver metastases between January 2006 and December 2016 at 16 different hospitals were enrolled in this study. The patients had mainly been treated at the University Hospital Augsburg (n=92), at the University Hospital Regensburg (n=186), and at the Katharinen Hospital Stuttgart (n=44) as well as at 13 smaller hospitals. Patient data from these smaller hospitals were documented by the Center of Tumor Registry at the University of Regensburg. Patients with extrahepatic metastases were excluded from the study.

Tumors in the ascending colon and in the colon transversum were defined as right-sided and tumors in the beginning of the left colon flexure were defined as left-sided. Furthermore, patients with a CRP level of $\geq 1 \frac{mg}{dl}$ were considered positive for IRT. Physicians collected all medical data. Informed consent was obtained from all patients. All analyses were performed according to the terms of the declaration of Helsinki.

### 4.2.3 Statistical approaches

The 512 patients were split into a training set and a validation set to ensure that the score could be validated on an independent group of patients. The applicability of an appropriate score should not depend on the hospital at which the patients in the training set are treated. We built the training set ($n = 282$) from the data of the University of Augsburg and the Tumor Registry, whereas the validation set contains 230 patients from the Regensburg and Stuttgart hospitals.

The two endpoints were disease-free survival (DFS) and overall survival (OS). DFS is the period from the date of the surgery to the recurrence of disease. If there had been no recurrence at the most recent follow-up, the patient would be censored at

this point for the analysis. OS is defined similarly, but it terminates with the death of the patient.

Differences between the training and the validation group were tested statistically with chi-squared test for categorical variables and with a t-test for continuous variables. All tests were two sided and with a significance level of 5%. We used the statistical computing program R (version 4.0.2.).
We performed the following steps, which we describe in detail on the following pages, in order to develop the score:

1. Searching for an optimal cutoff for continuous variables,
2. Developing univariable regression models for OS and DFS
3. Fitting a multivariable model and assigning a score to risk factors
4. Validating the score

**Search for optimal cutoff for continuous variables**

Because age must be translated into a discrete variable to be included in a score, we used the training set to find an optimal cutoff value. We selected a range of possible thresholds and performed a simple univariable Cox regression for OS for every split. The best-performing threshold was chosen for further analysis.

**Univariable regression model for OS and DFS**

Since only relevant variables should be included in the multivariable model, we fitted a separate univariable proportional hazards model for DFS and OS for every variable. We only included variables in the next step if the p-value from the univariable regression was below 0.15.

**Multivariable model and assigning a score to risk factors**

We performed a multivariable Cox proportional hazard regression with the remaining variables. A backward selection based on the p-value was chosen to determine the final model. We verified the assumption of proportional hazards via Schoenfeld residuals. We computed hazard ratios and 95% confidence intervals for both models. We computed the quotient of hazard ratios and the smallest hazard ratio of the model, rounded these values and assigned each risk factor its corresponding points. The sum of present risk factors was calculated for each patient.

#### Validation of the score

The validation set was stratified according to risk scores. DFS and OS were estimated from Kaplan-Meier curves and compared via log-rank tests. Median OS and DFS were calculated with confidence interval for each subgroup. We compared our predicted score with the score from Malik et al. (2007).

### 4.2.4 Development of score

A total of 512 patients, 68.9% of whom were male, were enrolled in the study. The median age at the time of liver surgery was 66 years. The median number of liver metastases was two, and 267 patients (52.1%) had more than one liver metastasis. Among the patients, 59.6% had synchronous disease, 64.3% had a positive nodal status, and 22.3% had an IRT. The tumor was located on the right side in 133 cases (26.0%) and had positive resection margins in 53 cases (10.4%)(Table 4.8). The median follow-up period was 81.2 months, and the median OS and DFS were 60.4 months (95%-CI $52.2-68.5$ months) and 17.0 months (95%-CI $14.3-19.8$ months), respectively.
Patients were split into a training set (TS) and a validation set (VS), and the two groups were compared (Table 4.9). The two groups were not different in terms of gender (fraction of males: TS:67.4%; VS:70.9%), IRT status (TS:23.0%; VS:21.3%), primary tumor side (right side: TS:23.8%; VS:28.7%), and nodal-positive tumors (TS:63.5%; VS: 65.2%). The patients in the training set were older than those in the validation set (median age: TS:68 years; VS:65 years; $p< 0.001$), and the training set contained a higher proportion of patients over the age of 72 (TS:30.5%; VS:22.2%; $p= 0.044$). The validation cohort had a slightly higher proportion of patients with multiple metastases (TS:48.2%; VS:57.0%; $p= 0.041$). The median follow-up period was 83.2 months for the TS and 70.3 months for the VS.

#### Cutoffs and univariable regression

In an initial step, we inquired whether continuous variables and categorical variables with multiple groups can be dichotomized to facilitate the calculation of a score. The number of liver metastases and age at the time of surgery were relevant to our study.

In our sample, only a few patients had four or more metastases. Therefore, we bundled those patients into one group. We compared different patient stratifications, which were based on the number of metastases, and found that all yielded almost identical results. When we only used two groups (one metastasis and multiple metastases), it emerged that a higher number of metastases leads to a higher risk of recurrence (HR= 1.5) or death (HR= 2.1). When we split the group with

Table 4.8: Patient characteristics in the whole sample

| Variable | Total (%) (n=512) |
|---|---|
| **Sociodemographic factors** | |
| Sex | |
| Female | 159 (31.1 %) |
| Male | 353 (68.9 %) |
| Median age at time of surgery (range) | 66y (27-89) |
| Age at time of surgery | |
| $<$ 72 years | 375 (73.2 %) |
| $\geq$ 72 years | 137 (26.8 %) |
| Inflammatory response to tumor (IRT) | |
| No IRT | 398 (77.7 %) |
| IRT | 114 (22.3 %) |
| Primary tumor side | |
| Left | 379 (74.0 %) |
| Right | 133 (26.0 %) |
| Median number of liver metastases | 2 (1-14) |
| Solitary vs multiple liver metastases | |
| Solitary | 242 (47.3 %) |
| Multiple | 267 (52.1 %) |
| Missing data | 3 (0.6 %) |
| Node positive primary tumor | |
| Negative | 168 (32.8 %) |
| Positive | 329 (64.3 %) |
| Missing data | 15 (2.9 %) |
| Synchronous vs metachronous disease | |
| Metachronous | 207 (40.4 %) |
| Synchronous | 305 (59.6 %) |
| KRAS | |
| Wildtyp | 136 (26.6 %) |
| Mutated | 68 (13.3 %) |
| Missing data | 308 (60.2 %) |
| Resection margin status | |
| R0 | 411 (80.3 %) |
| R1 | 53 (10.4 %) |
| Missing data | 48 (9.4 %) |

Table 4.9: Patient characteristics compared between training and validation set

| Variable | Training (n=282) | Validation (n=230) | p-value |
|---|---|---|---|
| **Sociodemographic factors** | | | |
| Sex | | | 0.451 |
| Female | 92 (32.6 %) | 67 (29.1 %) | |
| Male | 190 (67.4 %) | 163 (70.9 %) | |
| Median age at time of surgery (range) | 68y (31-89) | 65y (27-88) | <0.001 |
| Age at time of surgery | | | 0.044 |
| $<$ 72 years | 196 (69.5 %) | 179 (77.8 %) | |
| $\geq$ 72 years | 86 (30.5 %) | 51 (22.2 %) | |
| Inflammatory response to tumor (IRT) | | | 0.715 |
| No IRT | 217 (77.0 %) | 181 (78.7 %) | |
| IRT | 65 (23.0 %) | 49 (21.3 %) | |
| Primary tumor side | | | 0.244 |
| Left | 215 (76.2 %) | 164 (71.3 %) | |
| Right | 67 (23.8 %) | 66 (28.7 %) | |
| Median number of liver metastases | 1 (1-9) | 2 (1-14) | 0.009 |
| Solitary vs multiple liver metastases | | | 0.041 |
| Solitary | 146 (51.8 %) | 96 (41.7 %) | |
| Multiple | 136 (48.2 %) | 131 (57.0 %) | |
| Missing data | 0 (0.0 %) | 3 (1.3 %) | |
| Node positive primary tumor | | | 0.816 |
| Negative | 94 (33.3 %) | 74 (32.2 %) | |
| Positive | 179 (63.5 %) | 150 (65.2 %) | |
| Missing data | 9 (3.2 %) | 6 (2.6 %) | |
| Synchronous vs metachronous disease | | | 0.525 |
| Metachronous | 110 (39.0 %) | 97 (42.2 %) | |
| Synchronous | 172 (61.0 %) | 133 (57.8 %) | |
| KRAS | | | 0.875 |
| Wildtyp | 91 (32.3 %) | 45 (19.6 %) | |
| Mutated | 44 (15.6 %) | 24 (10.4 %) | |
| Missing data | 147 (52.1 %) | 161 (70.0 %) | |
| Resection margin status | | | 0.336 |
| R0 | 227 (80.5 %) | 184 (80.0 %) | |
| R1 | 25 (8.9 %) | 28 (12.2 %) | |
| Missing data | 30 (10.6 %) | 18 (7.8 %) | |

Table 4.10: Univariable analysis of overall survival and disease-free survival

| Variable | p-value from log-rank test for OS | p-value from log-rank test for DFS |
|---|---|---|
| Male sex | 0.225 | 0.016 |
| Age at time of surgery (> 72 years) | <0.001 | 0.400 |
| Inflammatory response to tumor (IRT) | <0.001 | <0.001 |
| Right-sided primary tumor | 0.015 | 0.016 |
| Solitary vs Multiple Metastases | <0.001 | 0.005 |
| Node positive primary tumor | 0.021 | 0.149 |
| Synchronous disease | 0.014 | 0.143 |
| Resection margin status (R1) | 0.082 | 0.003 |
| KRAS-mutated | 0.055 | 0.016 |

multiple metastases into a subgroup of patients with two metastases and a subgroup of patients with more than two metastases, the two hazard ratios remained almost identical. Therefore, we only differentiated between patients with a single metastasis and patients with multiple metastases.

In order to categorize patients by age at surgery, we performed a univariable regression model for OS and DFS and selected the cutoff with the lowest p-value. For DFS, no model yielded a statistically significant p-value. For OS, we reached optimal separation by differentiating between patients with an age $\leq 72$ and patients with an age $> 72$ years, with a p-value of less than 0.001 and a HR of 1.7.

After a description of the training and validation sample and preparations of variables, we performed a simple univariable Cox proportional hazard regression for every important variable with DFS and OS as endpoints (Table 4.10).

The factors that were associated with OS were age at surgery ($p < 0.001$), presence of IRT ($p < 0.001$), primary tumor side ($p = 0.015$), presence of more than one metastasis ($p < 0.001$), node status ($p = 0.021$) and synchronous disease ($p = 0.014$). The variables that emerged to be associated with DFS were sex ($p = 0.016$), presence of IRT ($p < 0.001$), primary tumor side ($p = 0.016$), presence of more than one metastasis ($p = 0.005$), resection margin status ($p = 0.003$), and KRAS-mutation ($p = 0.016$). Due a high proportion of missing values for KRAS and resection margin status, we excluded them from further analysis and did not take them into account in the multivariable model. Further studies of their roles must be conducted in the future. Variables with p-value $p < 0.15$ were considered in the multivariable model.

Due to their high importance for the characterization of a tumor, primary tumor side and IRT-status were investigated further. The median OS for patients with left-sided primary tumors was 62.0 months (95%-CI: $50.5 - 75.5$ months). It was 40.4 months (95%-CI: $30.7 - 64.6$ months) for patients with right-sided primary

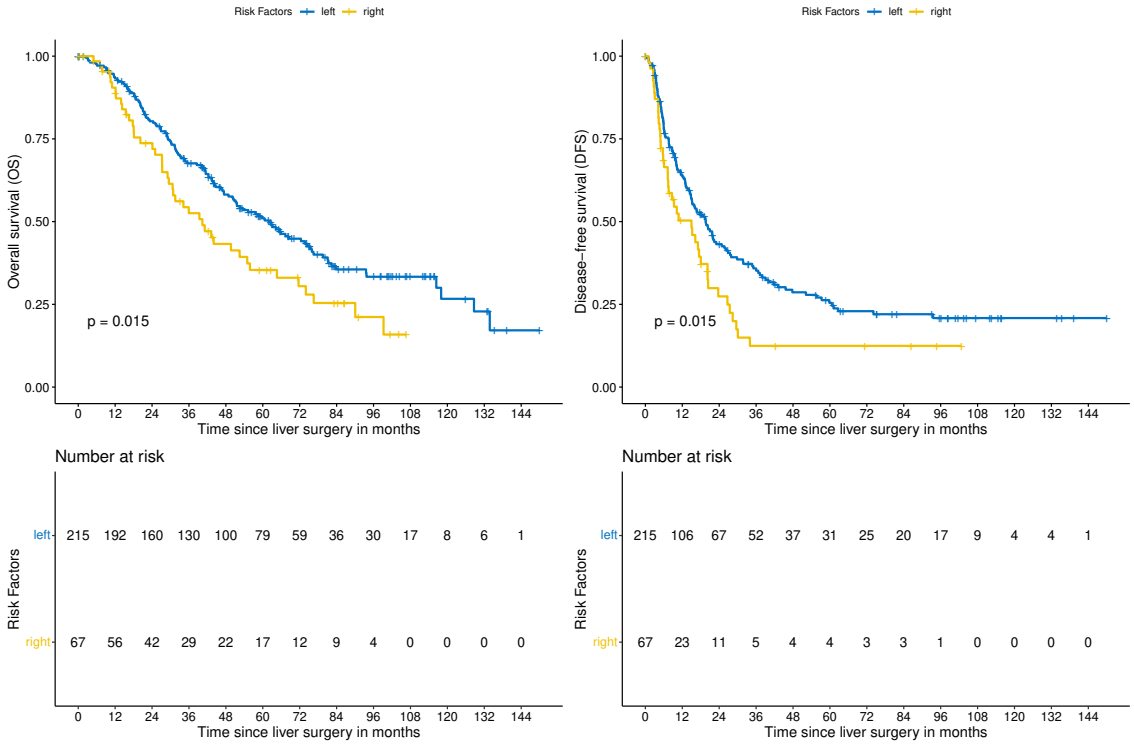


Figure 4.4: Kaplan-Meier curves for OS and DFS in the training sample stratified by the tumor sidedness of the patient (right side in yellow and left side in blue).

tumors. OS differed between the two groups (p= 0.015, log-rank-test).
Patients whose primary tumor was located on the left side had a median DFS of 19.6 months, with a 95% confidence interval from 15.1 to 26.3 months. In comparison, patients with a right-sided tumor had a median DFS of 15.0 months (95%-CI:7.4 – 20.2 months). Once more, there were differences between the two groups (p= 0.015, log-rank-test) (Figure 4.4. Median OS and DFS differed between patients with and without IRT. Patients with a positive IRT status survived for a shorter period, with a median survival of 30.9 months (95%-CI: 24.1-44.0) compared to a median survival of 68.4 months (95%-CI: 58.5 – 79.5 months) without IRT ($p < 0.001$, log-rank-test). A positive IRT status was also associated with shorter DFS, with a median of 11.5 months (95%-CI: 8.6 – 15.2), whereas patients with a negative IRT status had a median DFS of 20.3 months (95%-CI: 16.7 – 26.9 months). This difference was statistically significant in a log-rank-test with a p-value $< 0.001$ (Figure 4.5.

### Multivariable analyses

Based on the results of the univariable analysis, we conducted a multivariable analyses with backward selection in order to identify the variables for our score. A lower OS was associated with the presence of IRT (HR= 1.92; 95%-CI: 1.35 – 2.75, $p < 0.001$), a right-side primary tumor (HR= 1.63; 95%-CI: 1.14 – 2.34, $p = 0.008$),

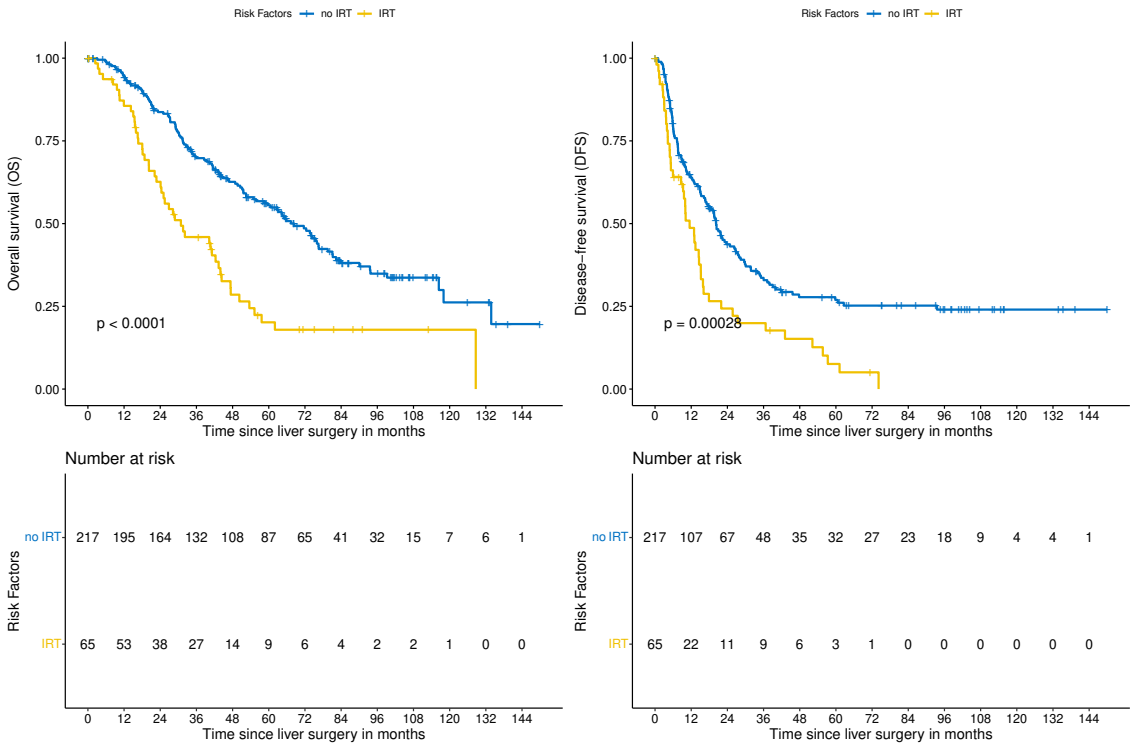


Figure 4.5: Kaplan-Meier curves for OS and DFS in the training sample stratified by presence of inflammatory response to tumor (IRT) (IRT in yellow and no IRT in blue).

multiple metastases (HR= 1.75; 95%-CI: $1.27 - 2.42$, $p < 0.001$), a node-positive primary tumor (HR= 1.49; 95%-CI: $1.05 - 2.13$, $p = 0.026$), and age at surgery $> 72$ years (HR= 1.74; 95%-CI: $1.24 - 2.44$, $p = 0.001$).

As far as DFS is concerned, a higher risk was associated with the presence of IRT (HR= 1.74; 95%-CI: $1.23 - 2.47$, $p = 0.002$), a right-sided primary tumor (HR= 1.56; 95%-CI: $1.09 - 2.21$, $p = 0.014$), multiple metastases (HR= 1.46; 95%-CI: $1.07 - 1.98$, $p = 0.016$), and male sex (HR= 1.44; 95%-CI: $1.03 - 2.03$, $p = 0.035$) (Table 4.11). When developing the score, we mainly focused on OS as an important indicator of the benefits that can accrue to a patient as a result of surgery. Three variables

Table 4.11: Multivariable Cox PH regression for overall survival and disease-free survival

| Variable | Overall Survival<br>p-value Hazard ratio (CI 95%) | Disease-free Survival<br>p-value Hazard ratio (CI 95%) |
|---|---|---|
| Inflammatory response to tumor | <0.001 1.92 (1.35-2.75) | 0.002 1.74 (1.23-2.47) |
| Right-sided primary tumor | 0.008 1.63 (1.14-2.34) | 0.014 1.56 (1.09-2.21) |
| Solitary vs multiple liver metastases | <0.001 1.75 (1.27-2.42) | 0.016 1.46 (1.07-1.98) |
| Node positive primary tumor | 0.026 1.49 (1.05-2.13) | — — |
| Age at time of therapy (> 72y) | 0.001 1.74 (1.24-2.44) | — — |
| Male sex | — — | 0.035 1.44 (1.03-2.03) |

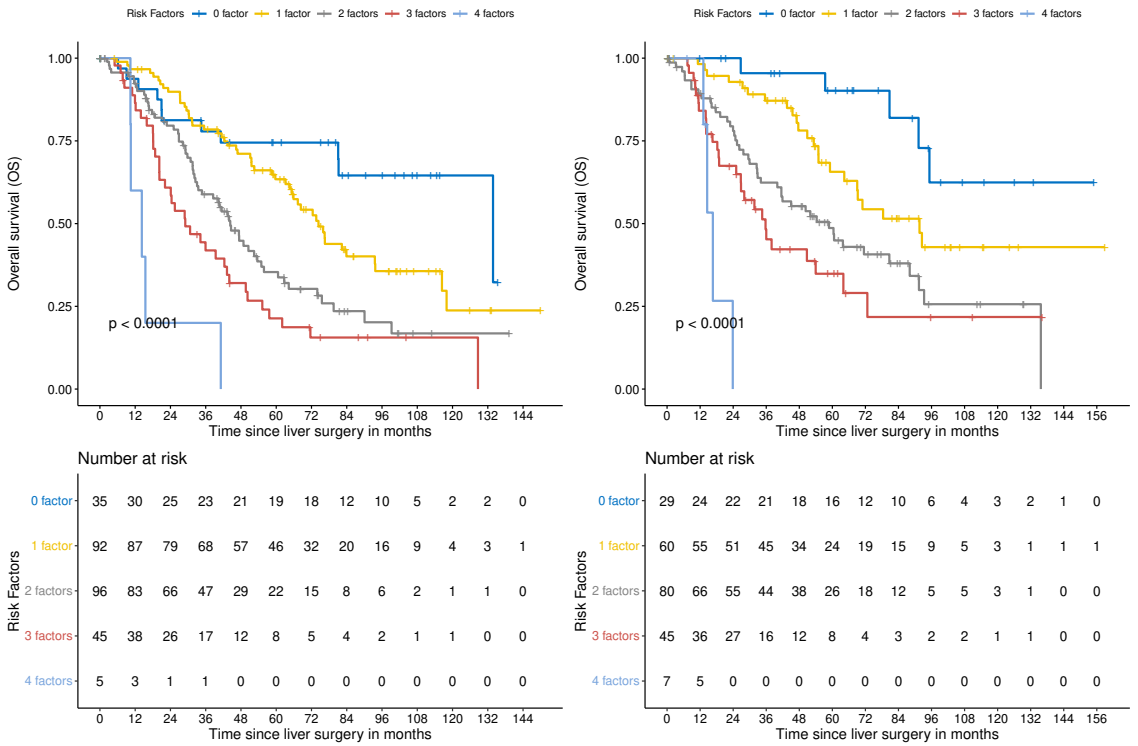


Figure 4.6: Kaplan-Meier curves of OS of all five risk groups for the training sample (left) and the validation sample (right).

that were significantly associated with OS remained in the final model of DFS. Hence, IRT status, with a HR of 1.92, sidedness of the primary tumor, with a HR of 1.63, and number of metastases, with a HR of 1.75 were included in the score. Furthermore, we decided that node status should be included in the final score as an important characteristic of the tumor. Age at surgery is not part of our prognostic score because it is not a cancer-specific risk factor. Adjusting for age differences is helpful in our model but not suitable for clinical decision-making.
Due to the comparable HRs of the four selected variables, which are between 1.49 and 1.92, each variable was assigned one point when the risk factor was present in a patient. This enables the score to be computed in a simple manner. Patients were stratified into five different risk groups according to the number of risk factors that they exhibited, which ranged between 0 and 4.

### 4.2.5 Validation of a prognostic score

When validating our score, we focused primarily on OS. However, we conducted the same analysis for DFS, and present the two together. The OS and DFS of all five groups was displayed with Kaplan-Meier curves, which showed satisfactory separation between the risk groups ($p < 0.001$) (Figure 4.6,4.7).
Importantly, the group with no risk factors exhibited a median OS of 133.8 months

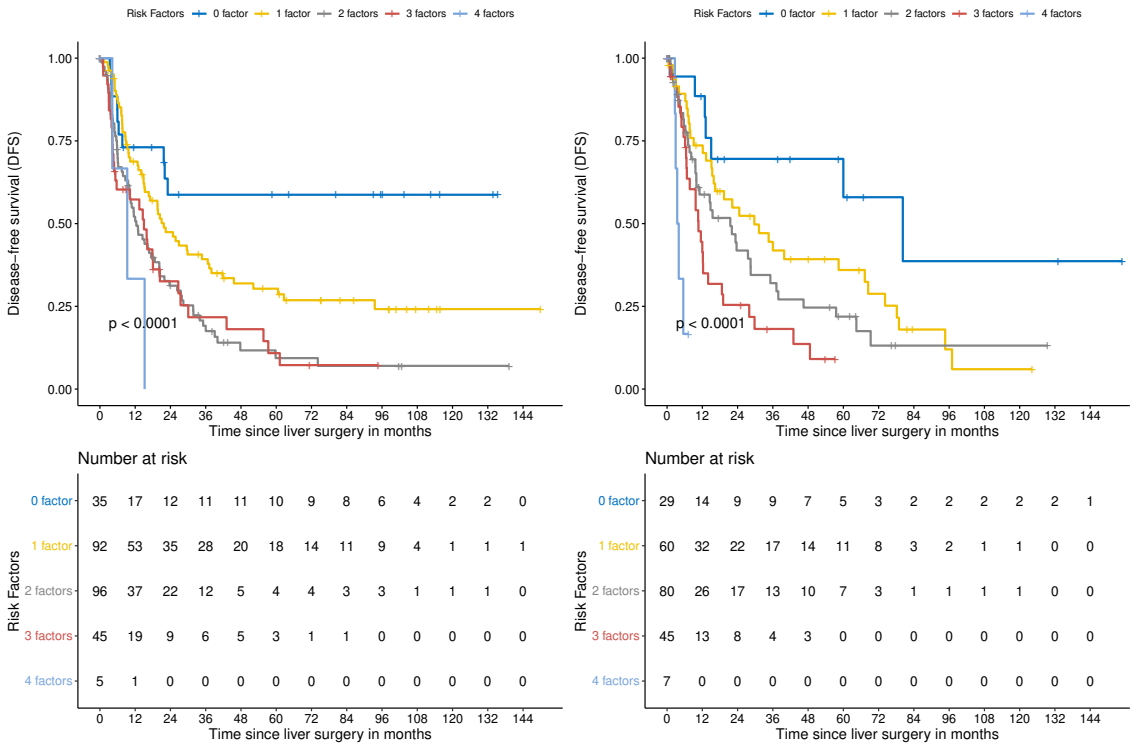


Figure 4.7: Kaplan-Meier curves of DFS of all five risk groups for the training sample (left) and the validation sample (right).

Table 4.12: Median OS of risk groups stratified according to prognostic score

| Risk Group | Number of patients | Median OS (months) Training (CI 95%) | Validation (CI 95%) |
|---|---|---|---|
| 0 risk factors | 35/29 | 133.8 (81.2-nr) | Not reached (95.2-nr) |
| 1 risk factor | 92/60 | 74.4 (65.3-93.7) | 91.6 (69.0-nr) |
| 2 risk factors | 96/80 | 44.4 (34.7-54.9) | 58.8 (41.5-91.4) |
| 3 risk factors | 45/45 | 29.0 (22.1-44.0) | 35.7 (26.8-72.7) |
| 4 risk factors | 5/7 | 14.3 (10.5-nr) | 16.6 (14.6-nr) |

(95%-CI: $81.2 - nr$) in the training sample, which is longer than that of the other groups. The median value for DFS was not reached in the training sample. Patients with three risk factors had a median OS of 29.0 months (95%-CI: $22.1 - 44.0$), and patients with four risk factors, who were few in number, had a median OS of 14.3 months.

The results for the validation cohort were similar. Patients with no risk factors did not reach the median survival time. Patients with three or four risk factors had a median OS of 35.7 months (95%-CI: $26.8 - 72.7$ months) and 16.6 months (95%-CI: $14.6 - nr$), respectively (Table 4.12).

We repeated the analysis for DFS and found that our score can distinguish between patients appropriately. Patients without risk factors exhibited the highest median survival or did not reached the median. In the validation group, the median DFS

Table 4.13: Median DFS of risk groups stratified according to prognostic score

| Risk Group | Number of patients | Median DFS (months) Training (CI 95%) | Validation (CI 95%) |
|---|---|---|---|
| 0 risk factors | 35/29 | Not reached (22.1-nr) | 80.2 (60.0-nr) |
| 1 risk factor | 92/60 | 21.7 (15.3-37.2) | 29.7 (15.9-68.4) |
| 2 risk factors | 96/80 | 12.4 (10.1-20.2) | 21.5 (10.0-35.2) |
| 3 risk factors | 45/45 | 15.0 (5.3-26.7) | 10.7 (6.7-18.7) |
| 4 risk factors | 5/7 | 9.3 (4.2-nr) | 3.7 (2.9-nr) |

was 80.2 months (95%-CI: $60.0 - nr$).
Patients with three or four risk factors had a shorter disease-free survival. When three risk factors were present, patients had a disease-free period of 15.0 months (95%-CI: $5.3 - 26.7$ months) in the training sample and of 10.7 months (95%-CI: $6.7 - 18.7$ months) in the validation sample (Table 4.13).
The group of patients with four risk factors that was used for the purposes of this model was small ($n = 5$ for the training cohort and $n = 7$ for the validation cohort). Due to the similarities between patients with two and three risk factors, we decided to combine the three risk groups and classified patients according to the presence of risk factors: "low-risk" = no risk factors, "intermediate-risk" = one risk factor, "high-risk" = 2, 3, or 4 risk factors.

The Kaplan-Meier analyses of OS and DFS demonstrated relevant stratification between the risk groups in the training sample, which could be reproduced in the validation sample (Table 4.14). In the training sample, low-risk patients had a median OS of 133.8 months (95%-CI: $81.2 - nr$ months), whereas patients in the intermediate and high-risk groups had median survival times of 74.4 months (95%-CI: $65.3 - 93.7$ months) and 40.4 months (95%-CI: $31.8 - 47.3$ months), respectively.
The median OS for the low-risk group in the validation sample was not reached. For the intermediate-risk group, the median OS was 91.6 months (95%-CI: $69.0 - nr$ months) and for the high-risk group the median OS was 41.9 months (95%-CI: $32.9 - 63.8$ months). The score distinguished the patients regarding OS in the two samples ($p < 0.001$)(Figure 4.8).
The capacity of our score to stratify patients could also be shown for DFS (Table 4.15). The low-risk group did not reach the median DFS (95%-CI: $22.0 - nr$ months) in the training sample. In comparison, patients with one risk factor had a lower median DFS of 21.7 months (95%-CI: $15.3 - 37.2$ months) and patients with more than one risk factor had the lowest median DFS (13.0 months; 95%-CI: $10.2 - 17.4$ months).
The similar results that we obtained from the analysis of the validation cohort

Table 4.14: Median OS of low, intermediate, and high risk group for the training sample (left) and the validation sample (right).

| Risk Group | Number of patients | Median OS (months) Training (CI 95%) | Validation (CI 95%) |
|---|---|---|---|
| Low risk (0 risk factors) | 35/29 | 133.8 (81.2-nr) | Not reached (95.2-nr) |
| Intermediate risk (1 risk factor) | 92/60 | 74.4 (65.3 -93.7) | 91.6 (69.0-nr) |
| High risk (2 - 4 risk factors) | 146/132 | 40.4 (31.8-47.3) | 41.9 (32.9-63.8) |

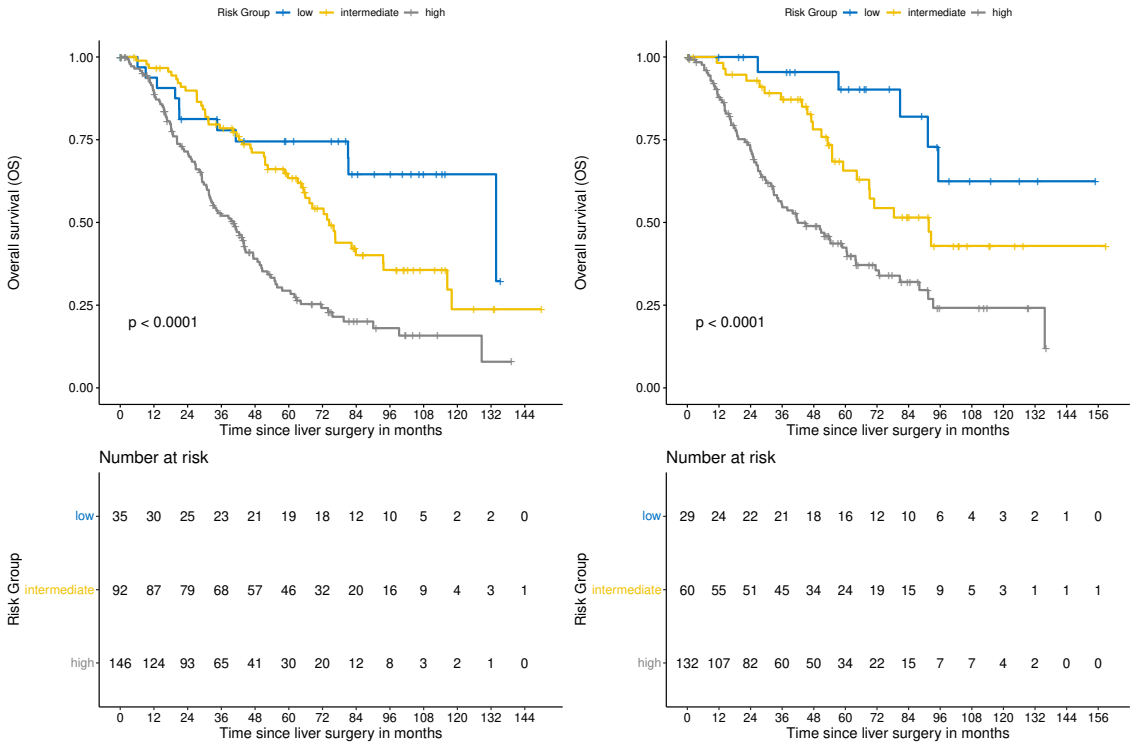


Figure 4.8: Kaplan-Meier curves of OS of low, intermediate, and high risk group for the training sample (left) and the validation sample (right).

Table 4.15: Median DFS of low, intermediate, and high risk group for the training sample (left) and the validation sample (right).

| Risk Group | Number of patients | Median DFS (months) Training (CI 95%) | Validation (CI 95%) |
|---|---|---|---|
| Low risk (0 risk factors) | 35/29 | Not reached (22.0-nr) | 80.2 (60.0-nr) |
| Intermediate risk (1 risk factor) | 92/60 | 21.7 (15.3-37.2) | 29.7 (15.9-68.4) |
| High risk (2 - 4 risk factors) | 146/132 | 13.0 (10.2-17.4) | 12.0 (9.7-21.5) |

highlight the performance of our score (median DFS: low-risk group: 80.2 months (95%-CI: $60.0 - nr$ months); intermediate-risk group: 29.7 months (95%-CI: $15.9 - 68.4$ months); high-risk group: 12.0 months (95%-CI: $9.7 - 21.5$ months). The three risk groups differed significantly ($p < 0.001$, log-rank-test) (Figure 4.9).

Beyond the stratification capability of our score, we also examined its performance by reference to sensitivity, specificity and area under the curve (AUC). AUC is a measure of the capacity of a score to discriminate between two groups. Patients with censored survival within 12 months after diagnosis were excluded from the analysis to guarantee that the follow-up period would be sufficient for classification. Furthermore, the values of the score had to be dichotomized for a calculation of sensitivity and specificity.

We obtained an AUC of 0.652 for OS in the validation cohort. If one supposes that patients with a score of 2 or more are likely to die, our score exhibits a sensitivity of 73.1% and a specificity of 53.1%. For DFS, we obtained an AUC of 0.670 for the validation cohort. If one supposes that patients with a score of 2 or more are likely to have a recurrence, our score exhibits a sensitivity of 63.4% and a specificity of 59.4%. Remarkably, with a cutoff of 1, sensitivity is high (OS: 95.2%, DFS: 93.8%), while specificity remains close to 20% (OS: 19.4%, DFS: 28.1%). This finding shows that our low-risk group contained a certain fraction of patients with satisfactory courses of illness and only a few with dissatisfactory courses of illness.

Finally, we compared the performance of our score to that of Malik et al. (2007). They used IRT status and the number of metastases and assigned one point for a positive IRT status and one point for having more than eight metastases. Since the patients in our sample rarely had more than eight metastases, only a few patients had a score of 2.

We computed the c-index for OS and DFS in the validation sample for both scores. The c-index is a commonly used measure that describes how adequately groups can be separated by a score. It ranges between 0 and 1, with 1 denoting a perfect sepa-

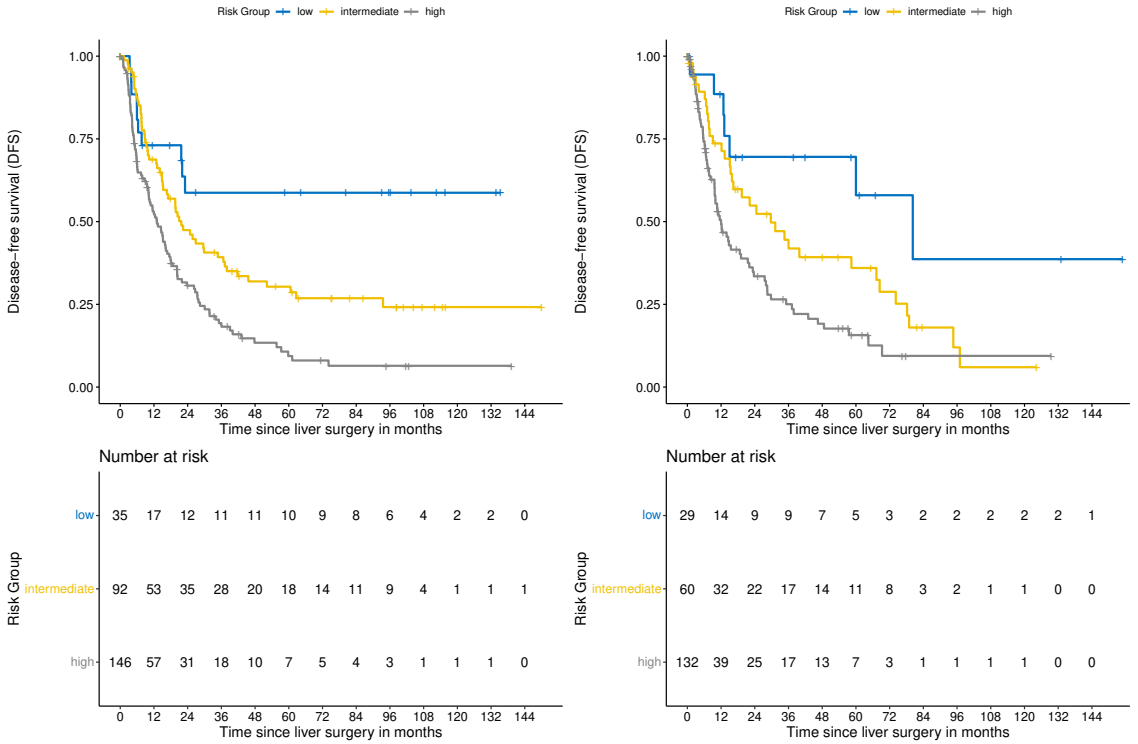


Figure 4.9: Kaplan-Meier curves of DFS of low, intermediate, and high risk group for the training sample (left) and the validation sample (right).

ration.

For OS our score, which is based on five risk groups, achieved a c-index of 0.676. In comparison, the score of Malik et al. (2007) had a c-index of 0.616. For DFS our score had a c-index of 0.629 and performed better than the score of Malik et al. (2007) (c-index 0.572).

### 4.2.6 Discussion

This study concerns patients with colon cancer whose disease has produced metastases, which complicate the treatment. Our aim was to identify a score that is based on relevant variables that can support clinical decisions about surgical resections of metastases. We included the preoperative available parameters in our final score and tested its performance with predictions about the OS and DFS.

Our score evaluates for each patient the number of risk factors that are present in each patient. Positive IRT status, tumor located on the right side, multiple metastases, and node-positive primary tumor could be identified as risk factors. Patients without any of these four risk factors exhibited a longer OS and DFS than those in other risk groups. This particular group might benefit from surgery, and

thus physicians should considered patient in this group for a treatment.

Intermediate-risk patients, who exhibited one risk factor, had a shorter OS and DFS than patients without risk factors. However, they may still benefit from surgical resection.
In contrast, patients in the high-risk group showed short OS and, moreover, DFS. The presence of at least two risk factors is indicative of an illness that would be difficult to address surgically. In many cases, multiple metastases have already occurred, which makes curing the disease more difficult. Although it is not exceedingly likely that the lives of such patients would be prolonged by surgery, it is questionable whether they can be deprived of its potential benefits. One solution would be to treat high-risk patients with additional, perioperative chemotherapy to increase their chances of survival.

We also compared our score to that of Malik et al. (2007), whereby one point is allocated to a patient for the presence of IRT and one point is allocated to them if they have a high number of metastases. Our score accounts for the sidedness of the tumor and its nodal status as additional risk factors. We showed that our score outperforms the score of Malik et al. (2007) and that the addition of two risk factors improves the stratification. Our score could identify a high-risk group, which is not possible with the score of Malik et al. (2007), at least for our sample.

A limitation of our study was the high number of missing data for KRAS and the resection margin so that they could not be incorporated into the score. Further studies should examine the importance of those for prognostic scores. In addition, the present study is based on retrospective data, which is also a limitation. A prospective study should be conducted to evaluate the quality of our score. The evaluation was conducted with an independent sample here. Although comparisons with other established score are important, we could only refer to Malik et al. (2007) because many of the parameters that would be necessary to compute other scores were not available in our retrospective sample.

Our score is based on variables which are easily accessible for every patient. We established the stratification capacity of the score by using an independent validation sample, which comprises data from multiple hospitals. Therefore, our score has the potential to be implemented widely in clinical practice and to identify patients who can benefit from a surgical resection of metastases.

# Chapter 5

# Modified Linear Regression Models for Associations between Lymphocytes and COVID-19

This project was based on a collaboration between the III. Medical Clinic of Augsburg University Hospital (Andreas Rank and Phillip Löhr) and the chair of Computational Statistics and Data Analysis (Gernot Müller, Stefan Schiele, and Tobias Arndt). The application of the methodology that was developed resulted in a publication (Löhr et al. (2021)).

**Keywords:** multivariable linear regression, transformed response variable, distribution of residuals, COVID-19

## 5.1 Introduction

Coronaviruses in general are highly transmissible and are present in many animals. They are known to have non-severe effects on humans, normally causing only common colds and infections of the respiratory tract. The risks mainly concern elderly individuals and those with comorbidities. However, especially in the last decades, two novel coronaviruses, namely severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV), have exhibited increased morbidity and case-fatality ratios that are significantly higher than those of previous coronaviruses. In 2019, a new and even more dangerous human pathogenic coronavirus, SARS-CoV-2, emerged in Wuhan, China, and spread quickly in several countries due to its high transmission rate (cf. N. Zhu et al. (2020)). For a more profound introduction to the biological characteristics of SARS-CoV-2, we refer to Hu et al. (2021) and Felsenstein et al. (2020).

SARS-CoV-2 often causes an acute respiratory tract infection that is called coronavirus disease 2019 (COVID-19). Many cases of SARS-CoV-2 are either asymptomatic, which means that a person is infected but does not develop any symptoms, or see the infected suffer from a moderate infection without needing to be hospitalized. However, some individuals, in particular patients with comorbidities, need further treatment for more severe symptoms such as acute respiratory distress syndrome (ARDS) (cf. Schaller et al. (2020)). Several studies, such as W. Huang et al. (2020) and G. Chen et al. (2020), have shown that severe COVID-19 infection is related to lower lymphocyte counts than those found in mildly affected patients.

Lymphocytes are a central part of the human immune system (cf. Murphy and Weaver (2016)). The immune system is necessary for protection against harmful and dangerous substances. It detects many of them, which enables an infection to be avoided. The immune system relies on two connected systems. The first defense mechanism is innate immunity, which is present since birth and can initiate inflammatory responses from specific cells, such as NK cells, macrophages, and dendritic cells, causing typical symptoms such as fever. The second mechanism is adaptive immunity, which develops as a suitable response to pathogens. Specific cells handle those pathogens and present antigens to initiate an immune response.

T and B lymphocytes are the main adaptive immunity cells and perform different tasks during an immune response. Two types of T cells are involved, namely helper T cells (CD4+ cells) and cytotoxic T cells (CD8+ cells). The helper T cell is activated when a pathogen is extra cellular and recognized by a B cell. The helper T cell interacts with a B cell and stimulates the B cell to produce antibodies that immobilize the pathogen so that the innate immunity can destroy it. In contrary, a cytotoxic T cell can recognize a cell that has already been infected through specific antigens and then destroy it. Both cell types proliferate into short-living effector cells and long-living memory cells. Effector cells are involved in a current immune response. Memory cells create an immunological memory in order to prevent repeated infection by inducing more rapid reactions.

The correlation of age or sex with lymphocyte counts is important. Many studies have compared lymphocyte counts across different groups, but have not adjusted their analyses for other influencing variables such as age. Age needs to be considered because it is mainly older patients who suffer severely. Older individuals have a lower number of naïve T cells and an increased number of memory cells. Sex can also have an influence on the count of lymphocytes (cf. Kverneland et al. (2016)).

Therefore, pronounced differences in age and sex might bias comparisons between those who are infected with COVID-19 and healthy individuals.

On the following pages, we only present the results for the main types of lymphocytes (total lymphocytes, total T cells (CD3+ cells), helper T cells (CD4+ cells), cytotoxic T cells (CD8+ cells), natural killer cells (NK cells), and total B cells (CD19+ cells). We analyzed the influence of a COVID-19 infection on the count of lymphocytes after adjustment for age and sex.

## 5.2 Data

This study covers patients with a COVID-19 infection that was confirmed by a positive PCR test between April and October 2020. Patients for whom the onset of COVID-19 symptoms occurred more than 28 days prior to the date of the observation, pregnant individuals, and patients with severe comorbidities of the immune system, such as malignancies or autoimmune disorders, were excluded. The study was conducted in line with the declaration of Helsinki. Signed informed consent was obtained from all patients, and the research was approved by the internal ethics committee.

A total of 50 healthy individuals were included in our study to provide benchmark counts of lymphocytes in the absence of COVID-19 infection. Their counts of lymphocytes were collected and analyzed before the first onset of SARS-CoV-2 in order to avoid hidden infections within the group.
The patients who were infected with SARS-CoV-2 were divided into two groups depending on the severity of their infections. Following the World Health Organization (WHO), we classified SARS-CoV-2 infections as either moderate (uncomplicated upper respiratory tract infection or pneumonia without supplemental oxygen) or severe (pneumonia with additional oxygen, ARDS, sepsis, or septic shock).
Every blood sample was analyzed with flow cytometry in order to measure total lymphocyte counts as well as several subsets.

## 5.3 Statistical approaches

We presented the characteristics of the sample with counts and percentages. Counts of every lymphocyte subset were expressed as median, first quartile, and third quartile and comparisons are drawn between healthy individuals and patients with moderate or severe COVID-19 infection by using Wilcoxon-Mann-Whitney tests. We ran univariable linear regression models for each subset of lymphocytes in order to

examine associations with age and sex, and we examined differences in cell counts over 10-year steps. A multivariable linear regression model was fitted to each subset of lymphocytes, with severity of COVID-19 as explanatory variable and with adjustments for age and sex. The variance inflation factor (VIF) was calculated for each model to ensure that the explanatory variables are not collinear.

We also inspected the residuals of each regression model. A necessary requirement for a linear regression is that the residuals follow a normal distribution. The residuals in our study did not, for the most part, approximate a normal distribution, in particular in the tails. Therefore, we modified the counts of lymphocytes by applying a logarithmic transformation (cf. Bland and D. G. Altman (1996) and Keene (1995)). A logarithmic transformation is a special case of a Box-Cox transformation. Instead of

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon$$

we model

$$log(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon. \tag{5.1}$$

Here, $Y_i$ is the count of a subset of lymphocytes for the $i$-th patient, $\beta_0$ is the intercept, and $\beta_k$ is the coefficient of $x_{ik}$. In our case, $x_{i1}$ is the age of patient $i$, $x_{i2}$ is the sex of patient $i$, and $x_{i3}, x_{i4}$ are dummy-coded for the severity of COVID-19. $x_{i3}$ takes a value of 1, when the infection is moderate, and $x_{i4}$ takes a value of 1, when an infection is severe. If both take values of 0, the patient is in the healthy control group. $\epsilon$ is approximately normally distributed.

The transformation of our dependent variable means that, the coefficients must either be interpreted on the log-scale or transformed into the original one. It should be noted that the means on the log-scale do not coincide with the arithmetic mean but the geometric mean after an inverse transformation, because $\exp(\frac{1}{n}\sum_{i=1}^{n}\log(y_i) = \exp(\log(\prod_{i=1}^{n} {y_i}^{\frac{1}{n}})) = \prod_{i=1}^{n} {y_i}^{\frac{1}{n}}$.
On the original scale, an exponentiation of (5.1) leads to

$$Y_i = \exp(\beta_0)\exp(\beta_1)^{x_{i1}}\exp(\beta_2)^{x_{i2}}\exp(\beta_3)^{x_{i3}}\exp(\beta_4)^{x_{i4}} * \exp(\epsilon) \tag{5.2}$$

For every unit increase in age, our count of lymphocytes is multiplied by a factor $\exp(\beta_1)$. In comparison to the reference group, moderate and severe COVID-19 infections are linked with a multiplication of the counts of lymphocytes by $\exp(\beta_3)$ and $\exp(\beta_4)$, respectively.

Beyond the afore mentioned condition of normally distributed residuals, another

Table 5.1: Median count of lymphocytes and subsets stratified by the severity of COVID-19 infection

| | Healthy controls (n = 50) | Moderate COVID-19 (n = 11) | p-value | Severe COVID-19 (n = 22) | p-value |
|---|---|---|---|---|---|
| Total lymphocytes | 1884 (1439–2288) | 1120 (867–1390) | 0.002 | 730 (463–1043) | $< 0.001$ |
| Total T cells | 1175 (839–1675) | 879 (576–971) | 0.017 | 380 (263–547) | $< 0.001$ |
| Cytotoxic T cells cells | 292 (209–488) | 219 (118–354) | 0.058 | 112 (71–179) | $< 0.001$ |
| T helper cells | 782 (554–993) | 523 (388–554) | 0.016 | 209 (113–277) | $< 0.001$ |
| Natural killer cells | 226 (143–300) | 127 (85–197) | 0.017 | 81 (60–166) | $< 0.001$ |
| Total B cells | 211 (149–274) | 59 (43–87) | $< 0.001$ | 71 (42–161) | $< 0.001$ |

justification of the use of log-transformation in our study is that we had to arrive at a biologically meaningful interpretation. Applying the obtained estimate of the coefficient of age in a linear regression without transformation, would entail adding or subtracting a certain amount of lymphocyte counts. In the worst case, this could lead to a subgroup with a negative count of lymphocytes, which would have no biological interpretation. A logarithmic transformation that is combined with an inverse transformation guarantees that the count of lymphocytes is always multiplied so that it remains positive.

## 5.4 Results

The median age of the 83 participants was 54 years, ranging between 17 and 94 years. A total of 50 patients were allocated to the healthy control group, and 33 had suffered COVID-19 infections, of which 33% were classified as moderate and 66% were classified as severe. There were 29 female and 54 male participants.

Those with COVID-19 infections were older than the healthy individuals (median age: 71 years vs. 43 years, $p < 0.001$). The proportion of female participants was similar in both groups. On average, the patients had developed COVID-19 specific symptoms four days before lymphocytes were measured.

The healthy individuals had a median count of $1884/\mu l$, whereas patients with a moderate or severe COVID infection had much lower counts (moderate: $1120/\mu l$; severe: $730/\mu l$). In general, patients with a moderate or severe COVID-19 infection had lower lymphocyte counts in nearly all subsets. In particular, patients with a severe infection had significantly lower amounts of lymphocytes (all $p < 0.001$)(Table 5.1).
The distribution of each subset of lymphocytes is displayed in Figure 5.1.
We performed univariable regressions for log-transformed counts of lymphocytes in

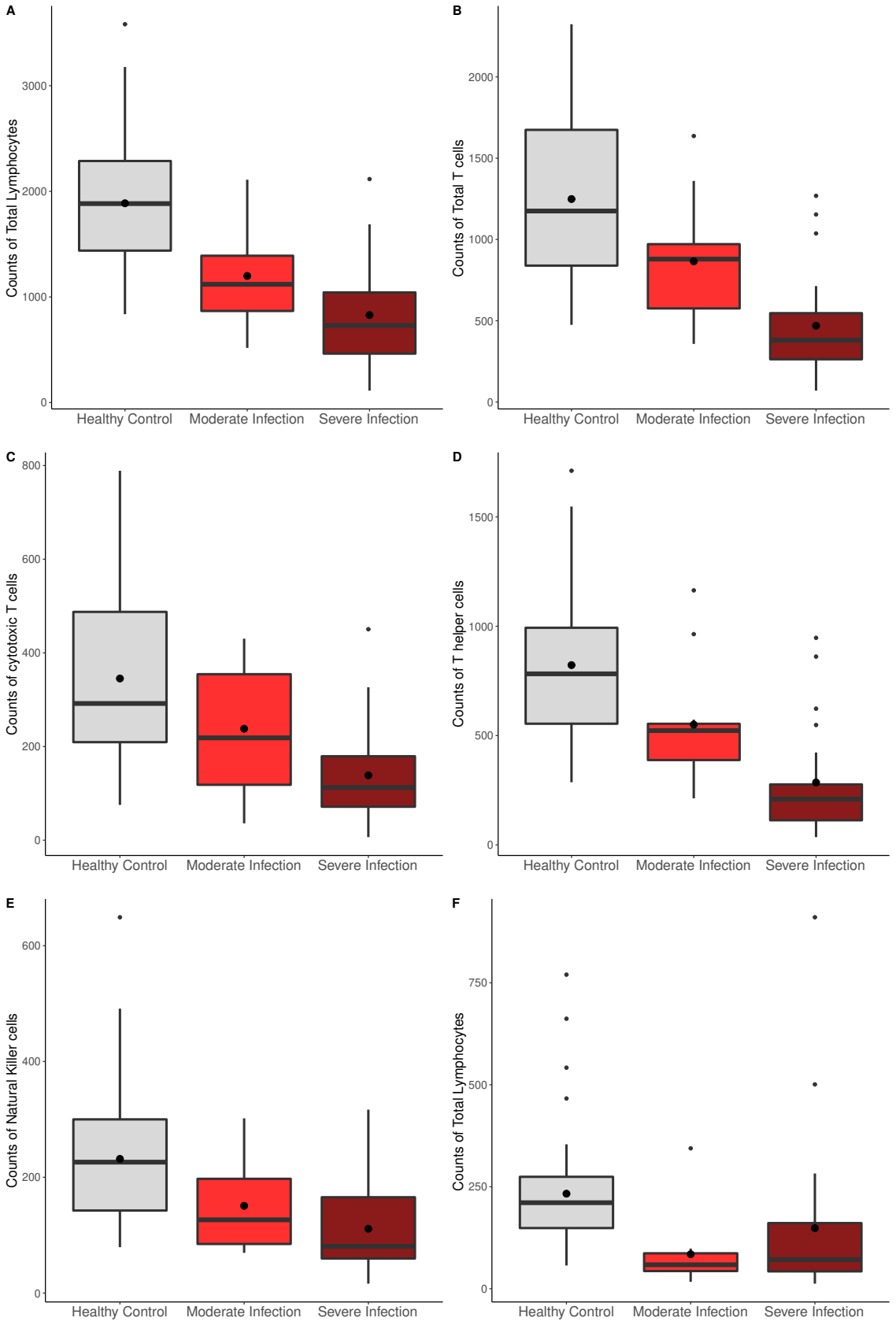


Figure 5.1: Boxplots of considered subsets of lymphocytes for healthy, moderate infected, and severe infected individuals: (A) Total lymphocytes, (B) Total T cells, (C) Cytotoxic T cells, (D) T helper cells, (E) Natural killer cells, and (F) Total B cells

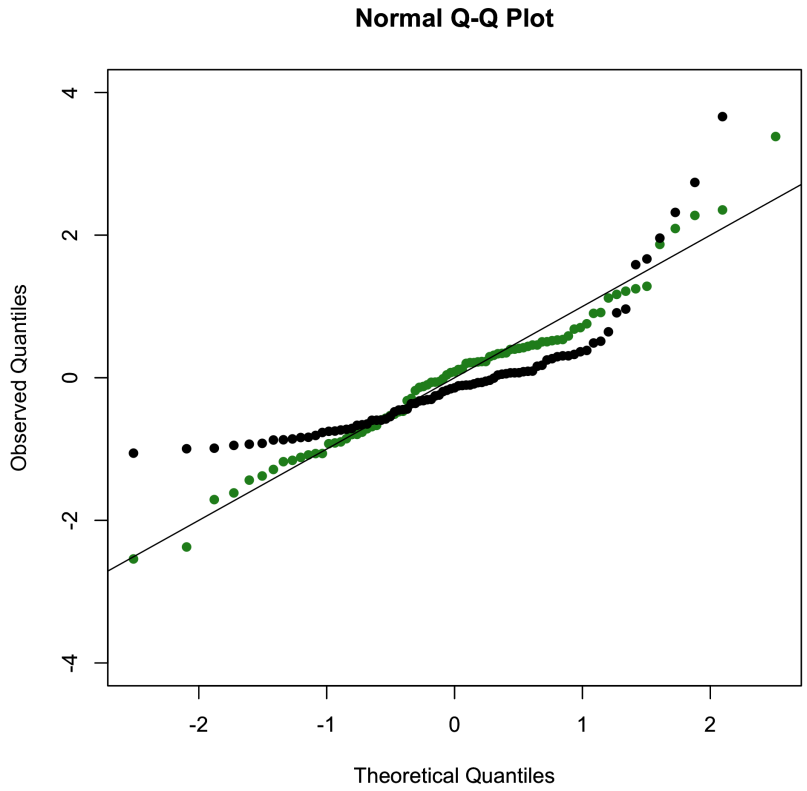


Figure 5.2: Comparison of the distribution of residuals without a transformation (black) and with a logarithmic transformation (green) of the response variable

the entire cohort, with age as the only explanatory variable, and found a significant reduction of the geometric mean for all subsets of lymphocytes in older patients. The reduction in total lymphocytes was 12.5% for every 10 years, and total T cells were 14.1% lower for every 10 years. We observed a similar trend for cytotoxic T cells (−19.1% per 10 years), helper T cells (−14.8% per 10 years), natural killer cells (−13.6% per 10 years), and B cells (−14.4% per 10 years).

### Multivariable analyses

In the next step, we examined the association between moderate or severe COVID-19 infections and the lymphocyte counts in a multivariable linear regression, including sex and age. All measured counts of subsets were logarithmically transformed, and the coefficient of age was examined over 10-year steps. In all models, the VIF was small enough to assume the absence of collinearity. The resulting coefficients are presented as multiplicative factors after inverse transformation.

Figure 5.2 displays the standardized residuals from a linear regression model with and without a transformation of the total count of B cells as response variable. The

Table 5.2: Multivariable linear regression of logarithmic transformed counts of lymphocytes under adjustment for gender and age

| Subset of Lymphocytes | Variables | Multiplicative Coefficient (95%-CI) | p-value |
|---|---|---|---|
| Total lymphocytes | moderate COVID-19 | 0.640 (0.463 – 0.885) | 0.008 |
| | severe COVID-19 | 0.427 (0.318 – 0.574) | < 0.001 |
| | gender | 1.176 (0.935 – 1.478) | 0.163 |
| | age per 10 years | 0.958 (0.899 – 1.020) | 0.174 |
| Total T cells | moderate COVID-19 | 0.723 (0.519 – 1.006) | 0.054 |
| | severe COVID-19 | 0.395 (0.292 – 0.534) | < 0.001 |
| | gender | 1.291 (1.022 – 1.632) | 0.033 |
| | age per 10 years | 0.941 (0.883 – 1.004) | 0.064 |
| Cytotoxic T cells | moderate COVID-19 | 0.738 (0.470 – 1.160) | 0.185 |
| | severe COVID-19 | 0.472 (0.312 – 0.714) | 0.001 |
| | gender | 1.121 (0.814 – 1.544) | 0.478 |
| | age per 10 years | 0.876 (0.803 – 0.957) | 0.004 |
| helper T cells | moderate COVID-19 | 0.686 (0.466 – 1.011) | 0.057 |
| | severe COVID-19 | 0.320 (0.225 – 0.457) | < 0.001 |
| | gender | 1.363 (1.036 – 1.793) | 0.027 |
| | age per 10 years | 0.952 (0.883 – 1.027) | 0.202 |
| Natural killer cells | moderate COVID-19 | 0.707 (0.470 – 1.062) | 0.094 |
| | severe COVID-19 | 0.499 (0.344 – 0.723) | < 0.001 |
| | gender | 1.151 (0.863 – 1.535) | 0.334 |
| | age per 10 years | 0.929 (0.859 – 1.006) | 0.068 |
| Total B cells | moderate COVID-19 | 0.335 (0.201 – 0.556) | < 0.001 |
| | severe COVID-19 | 0.503 (0.316 – 0.801) | 0.004 |
| | gender | 1.156 (0.807 – 1.657) | 0.425 |
| | age per 10 years | 0.930 (0.842 – 1.026) | 0.147 |

black dots that show the standardized residuals of the model without transformation do not approximate a normal distribution. However, transformation provides a superior approximation of a normal distribution.

All subsets of lymphocytes had lower counts in patients with severe COVID-19 infection, even after adjustment (Figure 5.3 and Table 5.2).

Total lymphocyte count was 32.0% lower (95%-CI:11.5 – 53.7%, $p = 0.008$) in patients with moderate infection and 57.3% lower (95%-CI:42.6 – 68.2%, $p < 0.001$) in patients with severe infection. Age and sex were not associated with total lymphocyte count.

Total T cell populations were lower, but not significant, when the COVID infection was classified as moderate ($p = 0.054$). However, a severe infection was associated with a reduction in T cells of 60.5% (95%-CI:46.6 – 70.8%, $p < 0.001$). Furthermore, female sex was associated with a higher count of T cells than male sex (increase of 29.1% (95%-CI:2.2 – 63.2%, $p = 0.033$).

In comparison with healthy individuals, patients with severe COVID-19 infection

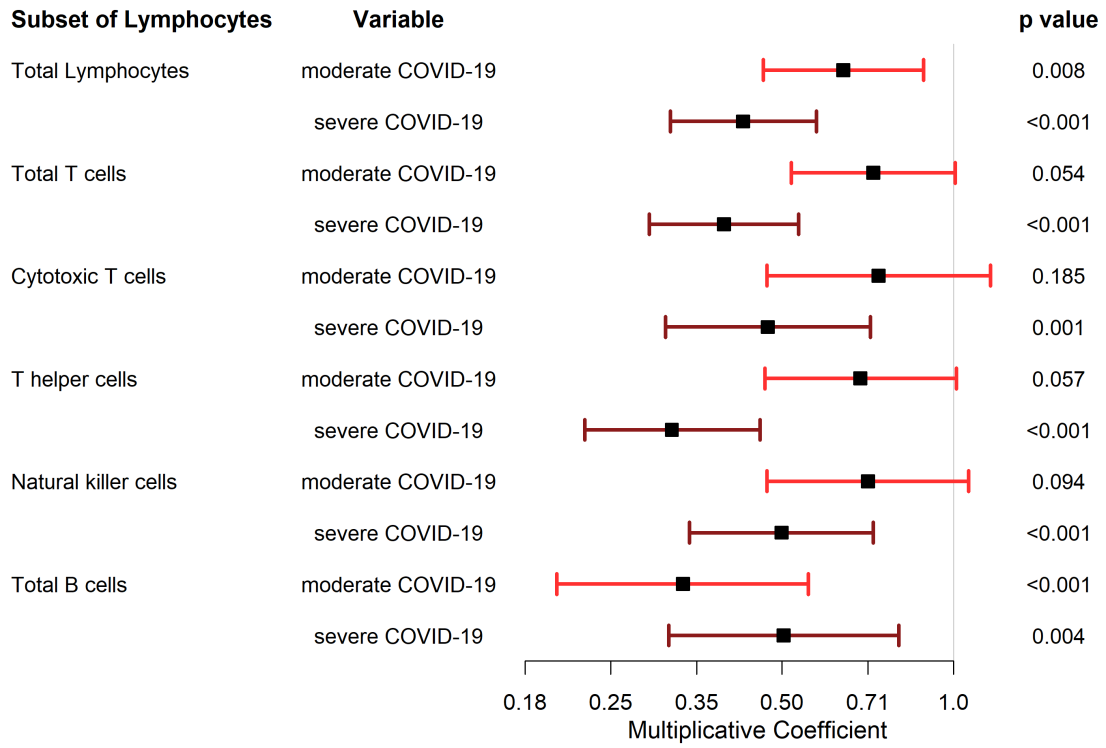


Figure 5.3: Transformed coefficients of moderate/severe COVID-19 infection in a linear multivariable regression for lymphocytes subsets with adjustment for gender and age. For each subset the multiplicative coefficient is depicted with 95% confidence interval.

had 52.8% (95%-CI:28.6 − 68.8%, $p = 0.001$) fewer cytotoxic T cells. Patients with a moderate infection had no reduced counts. This was the only subset in which the age was significantly associated with a decrease. We found reductions that were not significant in all other subsets.
Helper T cells behaved similarly to T cells as a whole, and the same patterns could be observed: counts were 68.0% (95%-CI:54.3 − 77.5%, $p < 0.001$) lower in patients with severe COVID-19 infection and 36.3% (95%-CI:3.6 − 79.3%, $p = 0.027$) higher among women.
A severe COVID-19 infection was also associated with a 50.1% (95%-CI:27.7−65.6%, $p < 0.001$) decrease in the number of natural killer cells and a 49.7% (95%-CI:19.9 − 68.4%, $p = 0.004$) decrease in the number of total B cells .

## 5.5 Discussion

We examined the influence of a COVID-19 infection on several subsets of lymphocytes by measuring counts in infected and healthy individuals. A univariable analysis showed that the counts of nearly all subsets are lower in patients with a COVID-19 infection, relative to healthy controls. If the infection is classified as severe, the reduction is even higher than in the case of a moderate one. These results could also be confirmed by a multivariable linear regression on logarithmically transformed lymphocyte counts.

It may seem contradictory that an infection leads to a lower lymphocyte count because infections generally cause an activation of the human immune system. However, a reduction in lymphocytes is evidence, among others, in Diao et al. (2020). Furthermore, after recovery from a mild infection, the original counts of lymphocytes can be restored (cf. Rank et al. (2021)). The exact reason has not been identified yet, but defense against the virus in several parts of the body might require a high number of T cells, which affects the observed behavior. Further research is needed to investigate this hypothesis.

We decided to apply a logarithmic transformation to the dependent variable for two reasons. Firstly, the conditions of the linear regression model are such that normally distributed error is required. Thus, the residuals have to be transformed if they do not approximate a normal distribution. Using B cells as an example, we showed that additional transformation can improve the distribution of the residuals.

Secondly, we think that our choice is justified by the biological nature of our problem. It is more realistic to assume that the effect of age is multiplicative than to

assume that it is additive. Negative predicted values for certain age ranges were thus avoided.

Transformations are often criticized for using data in a manner that does not ensure that a normal distribution is approximated accurately or for being susceptible to misinterpretation (cf. Feng et al. (2013)). However, we have shown the benefits of adopting the approach in question for our particular regression model.
Xiao et al. (2011) and Changyong et al. (2014) inquired whether non-linear regression models or general estimation equations can outperform linear regressions on transformed data and yield more reliable results because the interpretation of results complicates without an inverse transformation. Although a non-linear model might not be as dependent on parametric assumptions, it is questionable whether interpretation is facilitated. Furthermore, overfitting is a possibility in applications with few samples.

The small number of participants with a COVID infection is a limitation of our study. We could only include 11 individuals with moderate symptoms. The large confidence intervals may have resulted from this limitation. Even with this low number of patients, we could still observe a decrease in the counts of nearly all subsets of lymphocytes. In particular, patients with severe COVID-19 had significantly lower counts for all subsets.

In summary, patients with COVID-19 infections have lower total lymphocyte, T cell, B cell, and natural killer cell counts. This effect is still present when adjustments are made for age and sex. Due to some associations of age and sex with lymphocytes, the two factors must be considered in estimates of the influence of COVID-19 on the immune system.

# Chapter 6

# Statistical Models for the Incidence of COVID-19 in Germany

We have shown in chapter 5 that a COVID-19 infection is associated with a reduction of lymphocytes and thus affects the human immune response. In this chapter, we focus on different approaches to model the time series of daily reported new infections and compare their performance. We first use a non-mechanistic approach without additional information and investigate whether the model assumptions are fulfilled for each model. Subsequently, we employ a mechanistic approach based on a structured compartment model.

**Keywords:** ARIMA, log-linear autoregressive Poisson model, compartment model, change points, Bayesian analysis, non-pharmaceutical interventions, COVID-19

## 6.1 Comparison of modeling approaches for incidence of COVID-19

### 6.1.1 Introduction and background

General information concerning SARS-CoV-2 and a COVID-19 infection were already covered in Chapter 5. In this chapter we discuss the course of incidence in more detail. The first infection in Germany was detected at the end of January 2020. In the early phase of the pandemic, a lack of tests and limited information about the virus impeded a detailed registration of infections. The first wave of infections with exponentially increasing case numbers of new infections happened in March

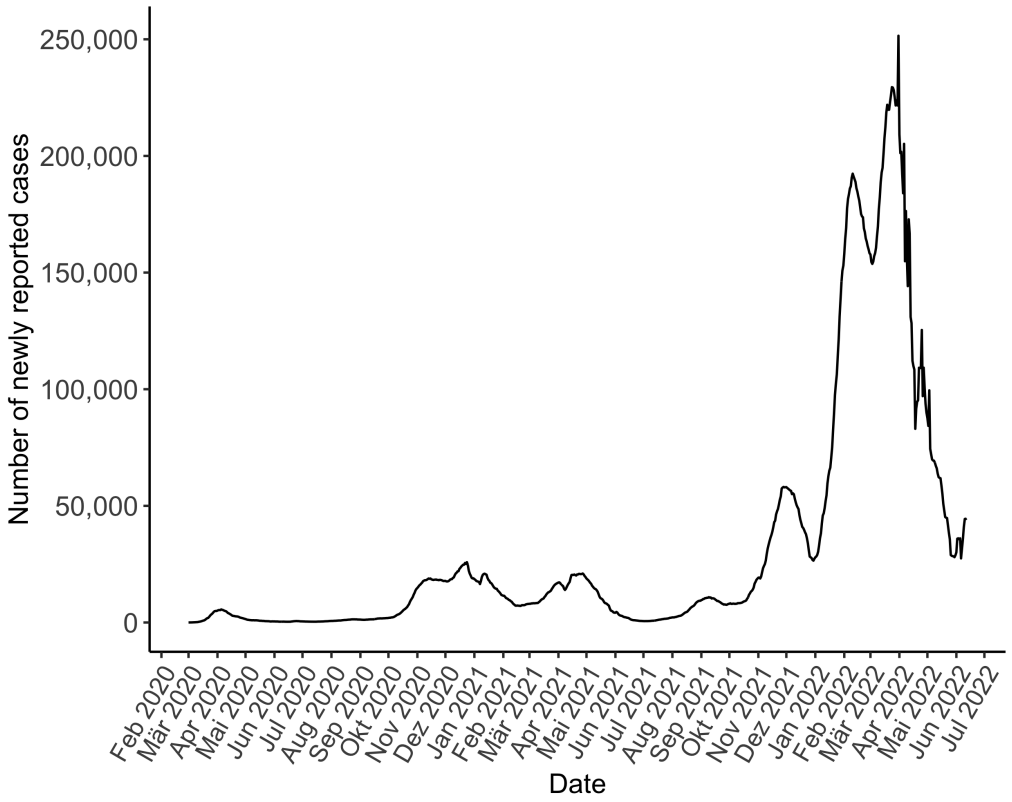


Figure 6.1: Number of daily reported new cases of COVID-19 (smoothed)

2020. Since then, several non-pharmaceutical interventions have been implemented in order to avoid the risk of a collapsing medical system and to protect elderly people as well as people with preconditioned illnesses.

Interventions have to be tailored to the actual situation and the danger that COVID-19 pose. Thus, they were reduced or completely lifted when the number of new infections lowered. Measures must be adapted to the situation because all of them influence the economy and society. Therefore, they should only be implemented if needed. Variants of the virus in combination with less restrictive interventions led to multiple waves with high incidence, hospital admissions, and deaths (Figure 6.1 and Figure 6.2). After the initial wave in March 2020, further peaks were registered between October 2020 and January 2021 (Second Wave) as well as between February 2021 and June 2021 (Third Wave) as described by Salzberger et al. (2021).
Models of daily COVID-19 incidence are important for forecasting the future developments, judging the expansion in the population, and reacting immediately when case numbers are rising in order to prevent high number of infections that might exceed the capacity limit of the German health system. In particular, when no vaccination was available and the dominant virus variant had a high fatality, non-pharmaceutical interventions needed to be well timed.

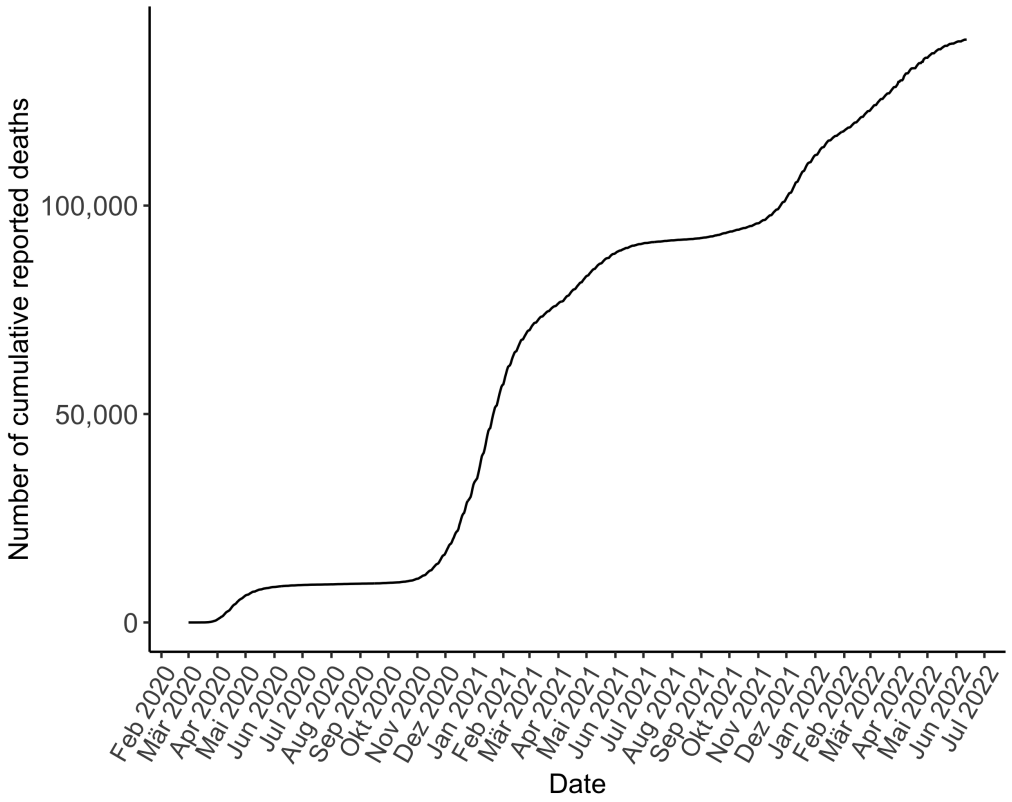


Figure 6.2: Cumulative number of reported deaths related to COVID-19

Several articles have focused on statistical methods for modeling daily reported new COVID-19 infections. Because the data is collected on a daily basis, methods from time series analysis are well suited. For example, Benvenuto et al. (2020) applied an Autoregressive Integrated Moving Average (ARIMA) model to predict COVID-19 case numbers. Barrıa-Sandoval et al. (2021) compared a wide range of different techniques including an ARIMA model, a Poisson process with a linear trend, a GLARMA model, and an adaptation of the Holt-Winters method. The ARIMA model showed the best performance for predicting case numbers.

In contrast to these techniques, Agosto and Giudici (2020) presented a log-linear autoregressive Poisson model of daily new observed cases which combined a short-term component that referred to the previous day and a long-term component that referred to the predicted value of the previous day. Therefore, their prediction indirectly incorporated data from all past days.

In this section, we develop and compare several models of COVID-19 incidence following the approach of a log-linear autoregressive Poisson model. Furthermore, we fit non-seasonal and seasonal ARIMA models and evaluate the performance dif-

ferences. Many published articles make no statement whether model assumptions could be verified in their model, and focus only on the quality of their prediction. However, it is important for an accurate model that all assumptions are fulfilled. We examine the residual structure of all models in detail.

### 6.1.2 Data: incidence

For our analysis, we used data from the period between July 28, 2020 and May 09, 2022. All data were obtained from a data repository from Our World in Data which collects data from several source such as the Johns Hopkins University and provides daily updates on new cases and deaths. It also reports on a stringency index (Oxford COVID-19 Government Response Tracker (OxCGRT)) and the number of vaccinations (cf. Ritchie et al. (2020)). The distribution of the variants of interest in Germany was obtained from the website of the Robert Koch Institute (RKI). Because the information is only published on a weekly basis, we assigned to all days of the week the same value.

Reported case numbers exhibit high variance because not all cases are registered on weekends. We did not include a multiplicative factor to adjust for this effect but used a smooth version of the case number in some applications. Instead of the reported number of cases we used the average number of cases during the last 7 days and rounded them to an integer.

### 6.1.3 Statistical approaches

#### ARIMA model

An Autoregressive integrated moving average model (ARIMA) is used often for time series analysis and can be applied to many situations. We assume that the current value of a variable $x_t$ at time $t$ depends on the previous values of that variable $(x_{t-1}, x_{t-2}, \dots)$. We give here only a short introduction and refer to Shumway et al. (2000) and Brockwell and Davis (2002) for details.
An ARIMA model is composed of three main parts: an autoregressive model, a moving average model, and an integrated model. For the definition of an autoregressive model, we introduce a stationary process.

**Definition 6.1.1** (Stationary Process)**.**
A time series $x_t$ is called stationary if

1. the mean value $\mu_t = \mathbb{E}[x_t]$ is constant over time and

2. the autocovariance function $\gamma(s,t) = \mathbb{E}[(x_s - \mu_s)(x_t - \mu_t)]$ depends on $s$ and $t$ only through their difference $|s-t|$.

In the following definitions of the autoregressive model and the moving average model, we assume that the mean of $x_t$ is 0. If the mean is unequal 0, the mean is subtracted to obtain a time series with mean 0. We define an autoregressive model of order $p$ ($AR(p)$) as

$$\phi(B)x_t = w_t,$$

where $x_t$ is stationary, $w_t$ is white noise. Furthermore,

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots \phi_p B^p,$$

with $B$ defined as the backward shift operator ($B^j X_t = X_{t-j}$ , $j \in \mathbb{N}$) and $\phi_1, \dots, \phi_p$ are coefficients that need to be estimated. This model includes $p$ previous values into the model of $x_t$ and adds a white noise.

Another class of models is a moving average model of order $q$ ($MA(q)$) that is defined as

$$x_t = \theta(B)w_t,$$

where $w_t$ is white noise. Furthermore, we define

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots \theta_q B^q,$$

with $B$ as backward shift operator and $\theta_1, \dots, \theta_q$ are coefficients. Instead of a linear combination of history values of $x_t$, the previous white noise values are weighted and summed up for $x_t$.

These two models are often combined into an $ARMA(p,q)$ **model** that is defined as

$$\phi(B)x_t = \theta(B)w_t,$$

where $\beta_p \neq 0, \theta_q \neq 0$ and $w_t$ is white noise. The parameters $p$ and $q$ are called autoregressive and moving average orders, respectively. The full model can be written as

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \dots + \theta_q w_{t-q}.$$

So far, we have assumed that the distribution of $x_t$ is stationary. If instead non stationary trend is included, we can consider an integrated model which combines an $ARMA(p,q)$ process and the differencing of $x_t$. The differencing removes linear

or higher order as well as non stationary trends from the time series so that we obtain a stationary time series on which we can apply an $ARMA$ model.

**Definition 6.1.2** (ARIMA(p,d,q) model)**.**
A process $x_t$ is called ARIMA(p,d,q) if
$\Delta^d x_t \coloneqq (1-B)^d x_t$ is a ARMA(p,q) model. $B$ is the backward shift operator.

The ARIMA model can be further expanded by adding an additional component to eliminate seasonal effects. Seasonal effects can be reduced with additional differencing of the series $x_t$.

**Definition 6.1.3** (ARIMA$(p,d,q)\times(P,D,Q)_s$ model)**.**
A process $x_t$ is called ARIMA $(p,d,q)\times(P,D,Q)_s$ if the differenced series $\Delta^d(1-B^s)^D x_t = (1-B)^d(1-B^s)^D x_t$ is an ARMA process defined by

$$\begin{aligned}\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D x_t &= \theta(B)\Theta(B^s)w_t,\\ \phi(B) &= 1-\phi_1 B-\phi_2 B^2-\dots\phi_p B^p,\\ \Phi(B^s) &= 1-\Phi_1 B^s-\Phi_2 B^{2s}-\dots\Phi_P B^{Ps},\\ \theta(B) &= 1-\theta_1 B-\theta_2 B^2-\dots\theta_p B^p,\\ \Theta(B^s) &= 1-\Theta_1 B^s-\Theta_2 B^{2s}-\dots\Theta_p B^{Qs},\end{aligned}$$

where $B$ is the backward shift operator and $w_t$ is white noise.

### Generalized linear autoregressive models

We focus on modified versions of a linear and a log-linear Poisson GLM which was introduced by Fokianos and Tjøstheim (2011). We used the R package *tscount* that was developed by Liboschik et al. (2017) for the implementation.
For a log-linear autoregressive Poisson model we assume that $x_t$ is approximately Poisson distributed with a parameter $\lambda_t$ being the expected mean of the current incidence from day $t$. This estimation is based on $r$ previous values $x_{t-t_1},\dots,x_{t-t_r}$ as well as on $s$ previous expected mean values $\lambda_{t-t_1},\dots,\lambda_{t-t_s}$. Since previous expected values depend on previous case numbers, our prediction accounts for all reported case numbers indirectly.
Furthermore, we apply a logarithmic transformation to the previous values and add 1 for the Poisson regression, with a log-link function such that all counts of reported new cases are on the same scale and days with 0 new cases can be included. All in

all, a log-linear autoregressive Poisson model can be written as:

$$x_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t) \tag{6.1}$$

$$\log(\lambda_t) = \alpha_0 + \sum_{i=1}^{r} (\alpha_i \log(1 + x_{t-t_i})) + \sum_{j=1}^{s} \left(\beta_j \log(\lambda_{t-t_j})\right), \tag{6.2}$$

where $\mathcal{F}_{t-1}$ denotes the $\sigma$-field generated by $\{x_0, \ldots, x_{t-1}\}$ and $x_t \in \mathbb{N}, \omega \in \mathbb{R}$, $\alpha_i \in \mathbb{R}, \beta_j \in \mathbb{R}$.

Our model incorporates an intercept term ($\alpha_0$), terms that reflect the historical observed values multiplied by a coefficient ($\alpha_1, \ldots, \alpha_r$), and terms that reflect long-term behavior via the expected historical values multiplied by a coefficient ($\beta_1, \ldots, \beta_s$). Because we have to ensure that our model has a stationary and ergodic solution, the parameters of the model need to be constrained. They must fulfill:

$$|\beta_1|, \ldots, |\beta_s|, |\alpha_1|, \ldots, |\alpha_r| < 1$$
$$\left|\sum_{k=1}^{s} \beta_k + \sum_{l=1}^{r} \alpha_l\right| < 1.$$

We investigated two modifications in this thesis. First, we considered the use of the identity function as link function instead of a logarithmic function. This change results in a model:

$$x_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t)$$
$$\lambda_t = \alpha_0 + \sum_{i=1}^{r} (\alpha_i x_{t-t_i}) + \sum_{j=1}^{s} \left(\beta_j \lambda_{t-t_j}\right),$$

where $\mathcal{F}_{t-1}$ denotes the $\sigma$-field generated by $\{x_0, \ldots, x_{t-1}\}$ and $x_t \in \mathbb{N}, \omega \in \mathbb{R}$, $\alpha_i \in \mathbb{R}, \beta_j \in \mathbb{R}$.

Compared to the log-linear autoregressive Poisson model, the identity link function needs further constraints to guarantee that all estimated means are positive. The logarithmic link function yields positivity automatically. The parameters in the setting of the identity link function must fulfill:

$$\beta_0 > 0$$
$$\beta_1, \ldots, \beta_s, \alpha_1, \ldots, \alpha_r \geq 0$$
$$\sum_{k=1}^{s} \beta_k + \sum_{l=1}^{r} \alpha_l < 1.$$

A disadvantage of the Poisson distribution assumption is that the estimated mean is equal to the estimated variance. This could be problematic for applications where overdispersion or underdispersion is present. A negative binomial distribution possesses more flexibility, and can thus support the modeling of overdispersion. We used a parametrization of the negative binomial distribution in terms of estimated mean and a dispersion parameter $\phi \in (0, \infty)$. The variance of $x_t$ conditional $\mathcal{F}_{t-1}$ can be written as $\lambda_t + \lambda_t^2/\phi$ instead of $\lambda_t$.

To account for external influences, we modified equation (6.2) to incorporate additional covariates. We selected the stringency of NPIs and the distribution of variants at each time point because the transmission rate is dependent on variants. A shift in the distribution also influences COVID-19 incidence. We focused only on the main variants for the model (Wildtype, Alpha variant, Delta variant, Omicron BA.1 variant, Omicron BA.2 variant, and Omicron BA.5 variant). The relative frequency $z_{i,t}$ for every variant $i$ at every time point $t$ lies in the range of 0 and 1 and sums up to 1 over all variants at each time point. The stringency of NPIs is rated on a scale between 0 and 1 at each time point and denoted by $u_t$. With this definition, the second part of our model can be written as:

$$\log(\lambda_t) = \alpha_0 + \sum_{i=1}^{r} (\alpha_i \log(1 + x_{t-t_i})) + \sum_{j=1}^{s} \left(\beta_j \log(\lambda_{t-t_j})\right) + \sum_{i=1}^{5} \eta_i z_{i,t} + \nu u_t. \quad (6.3)$$

**Measurements for model evaluation**

The performance of these models were mainly evaluated bases on the mean absolute error (MAE) and the mean absolute percentage error (MAPE). Both provide a measure how well the data approximate the observed data.
When the observed incidences are denoted by $C_t$, the estimated incidences are denoted by $\hat{C}_t$, and $n$ data points are considered, we can define the both measures

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^{n} |C_t - \hat{C}_t|, \quad (6.4)$$

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \frac{|C_t - \hat{C}_t|}{C_t}. \quad (6.5)$$

### 6.1.4 Non-seasonal and seasonal ARIMA models

First, we examined the autocorrelation function (ACF) of the smoothed COVID-19 incidence and the ACF of the smoothed COVID-19 incidence after a first order differencing (Figure 6.3). The smoothed incidence values were highly correlated because we used the average of the last seven days instead of the observed value. We used the method *auto.arima* from the package *forecast* in R to detect the optimal

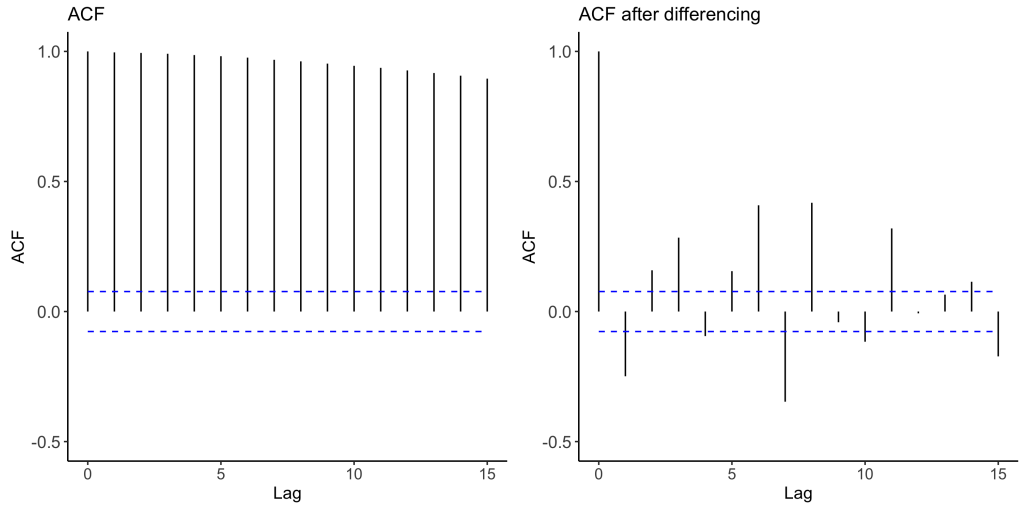


Figure 6.3: ACF of the COVID-19 smoothed incidence (left) and of the smoothed COVID-19 indicence after differencing (right)

ARIMA model regarding AIC. The algorithm identified $p = 5$, $d = 1$, $q = 2$ and no seasonal component as optimal parameters. The model yielded a MAE of 1413 and a MAPE of 4.25%. Figure 6.4 shows the course of fitted values in comparison to that of the observed ones. Although the result was a relatively close fit, the

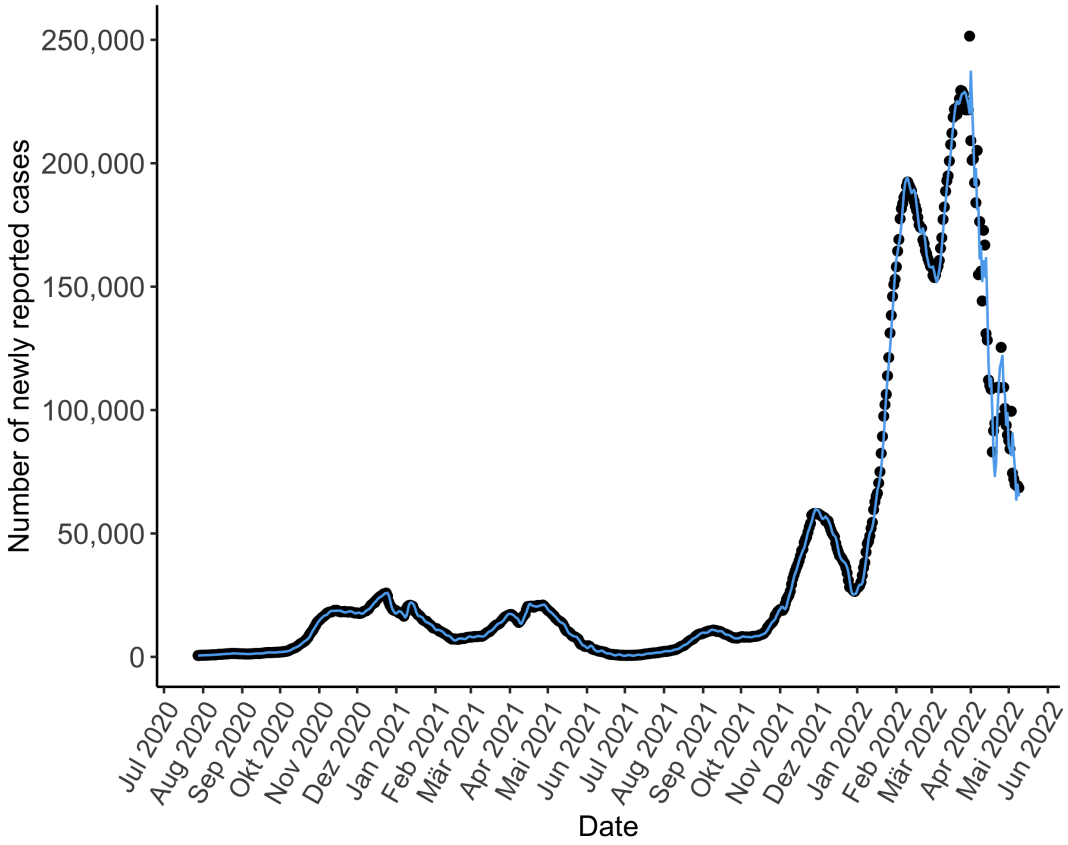


Figure 6.4: Fitted values of the optimal ARIMA model (blue line) compared with the actual reported numbers (black)

distribution of the residuals could not be approximated by a normal distribution (Figure 6.5). Problems arose from the number of infected individuals in different COVID-19 waves. Whereas during the initial waves, the number of daily reported cases was below 50,000, the numbers increased in 2022. This behavior was also present in the residuals, with small values around 0 at the beginning and higher values at the end. Furthermore, in the selected model there was still a correlation between the residuals. These problems with the correct distribution of residuals

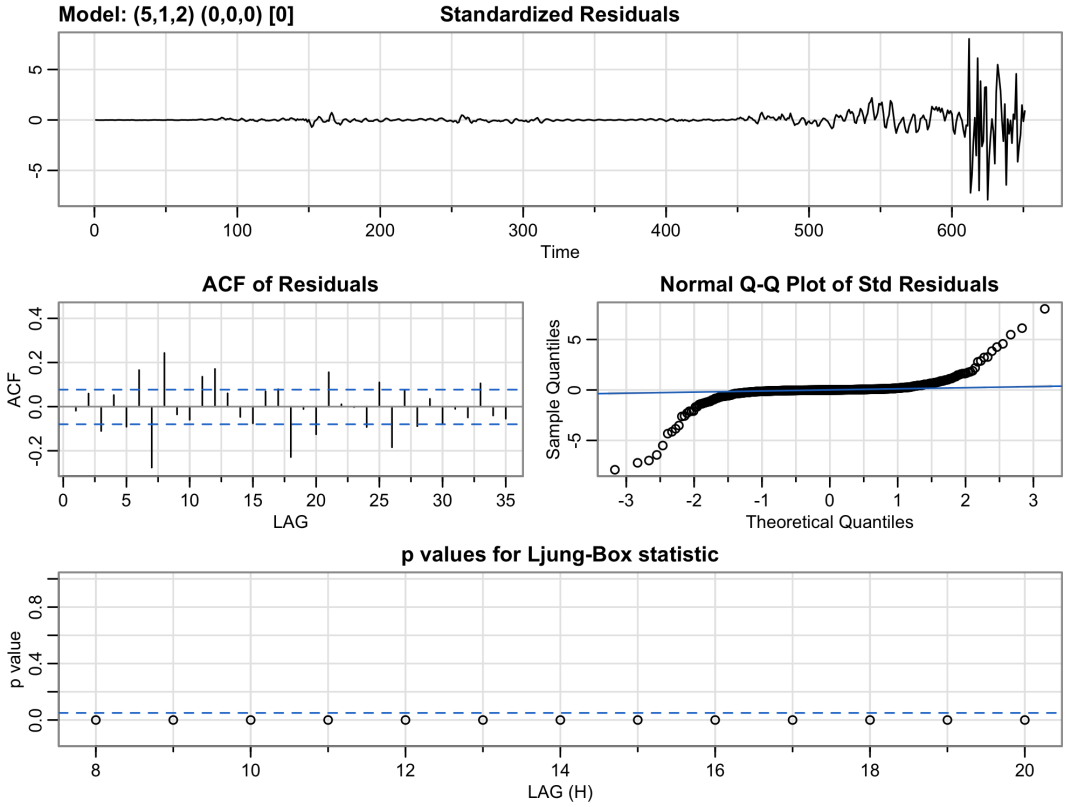


Figure 6.5: Residuals of fitted ARIMA Model over all data points

made the model unusable. A high correlation induced by the smoothing of the curve through the replacement of the reported case number by the mean of the last 7 days was another problem. To improve the distribution of residuals and reduce the correlation, we analyzed every wave of COVID-19 separately and used the daily incidence without a smoothing.

### 6.1.5 Models for incidence without smoothing

The ACF plot of the incidence without smoothing emphasised that the correlation between the actual and previous values is smaller than the correlation with smoothing (Figure 6.6). We found the highest correlation at lag 7 which is an indicator for a seasonal component in our model. The same structure is also present when a first order differencing is applied. However, the correlations are lower after differencing.

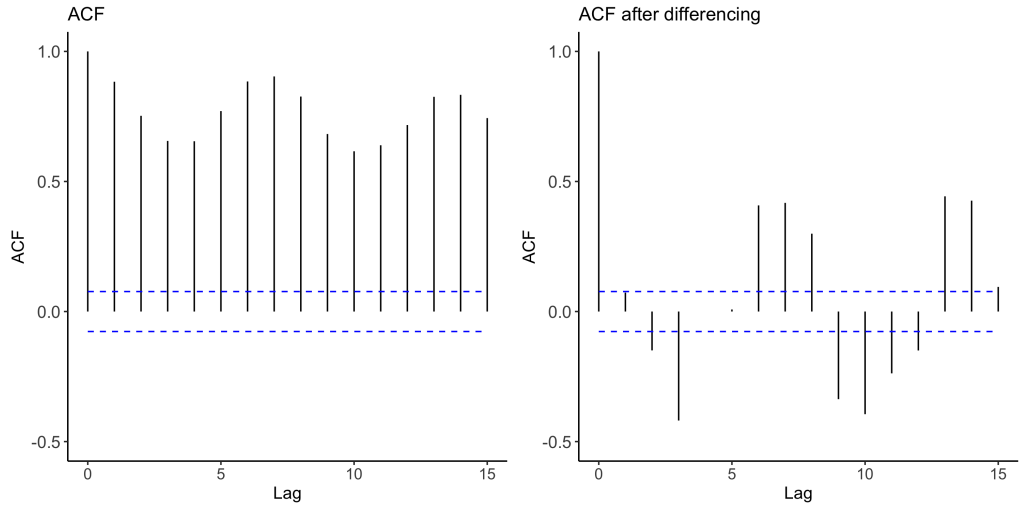


Figure 6.6: ACF of the COVID-19 incidence (left) and of the COVID-19 indicence after differencing (right)

Although a seasonal component is reasonable from the ACF plot, we fitted for each COVID-19 wave a non-seasonal and one seasonal ARIMA model. We used *auto.arima* to find the optimal ARIMA model regarding AIC. The aim of this section was also to investigate whether model assumptions are fulfilled for different models because often only the results of prediction are presented without any information about model assumptions.

### Second COVID wave

For the second COVID wave (09/28/2020 until 02/28/2021), the optimal parameters were chosen as ARIMA(3,1,2) without a seasonal component. This resulted in the following measures: AIC = 2847.7, MAE = 1848.8, and MAPE = 15.3. Figure 6.7 shows the analysis of the residuals. Although the residuals approximated the normal distribution well except outliers, the middle part of the analyzed time period showed high standardized residuals. Furthermore, this model still contains a certain degree of correlation between residuals, in particular at lag 7. This correlation could be reduced with an additional seasonal component (Figure 6.8). The optimal model according to *auto.arima* had parameters ARIMA$(1,1,1)\times(0,1,2)_7$. This model outperformed the non-seasonal model in all measurements in the dataset. The model had an AIC of 2708.8, a MAE of 1533.5, and a MAPE of 11.6.
We also accounted for the high variability in the variance and applied a logarithmic transformation to the daily incidence numbers and selected the best seasonal ARIMA model. The analysis of the residuals of this ARIMA$(0,1,1)\times(0,1,2)_7$ is shown in Figure 6.9. The variance of this transformed model is more homogeneous than without a transformation. Another benefit of the transformation is that the

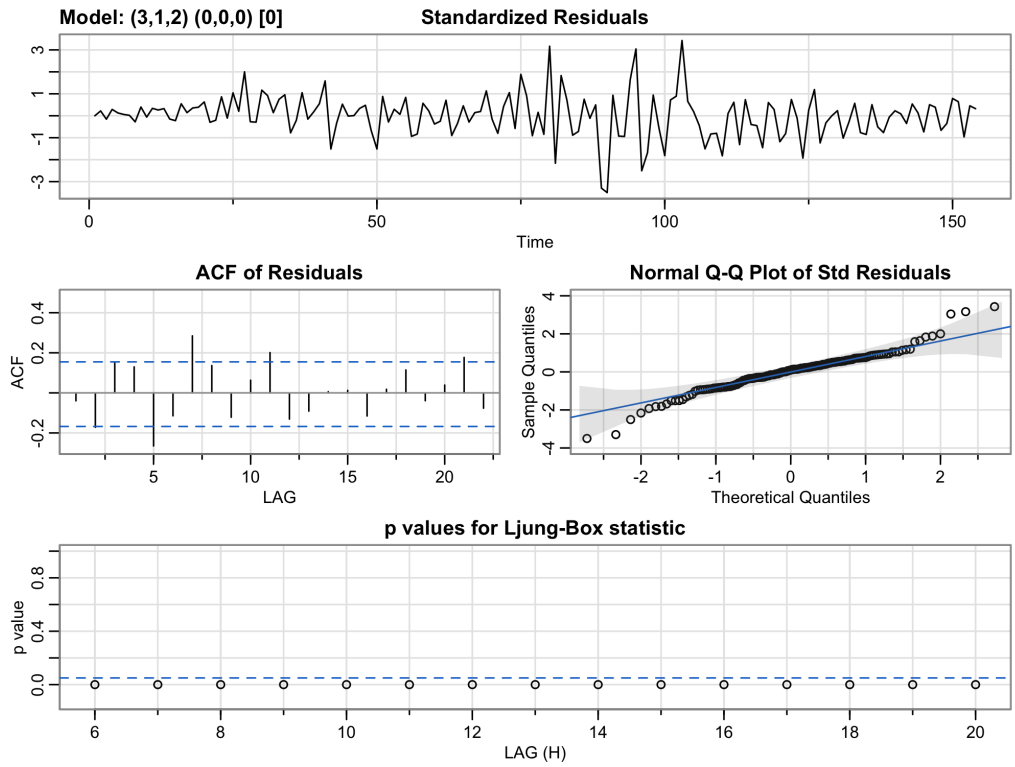


Figure 6.7: Analysis of residuals for non-seasonal ARIMA model for second wave

correlation between residuals could be reduced and that the residuals are closer to a normal distribution. The fit of the model was also most accurate, with a MAE of 1498.9 and a MAPE of 10.4. This means a slightly better result than the model without transformation (Figure 6.10).

### Third COVID wave

The third COVID-19 wave lasted from 03/01/2021 until 06/13/2021. The optimal parameter for the non-seasonal model was ARIMA(2,1,2). Whereas the residuals approximated the normal distribution well, residuals were correlated at several lags and the incidence could be modeled accurately. The model had an AIC of 1949.7, a MAE of 2110.8 and a MAPE of 23.7.

We further fitted an ARIMA model with a seasonal component, and obtain the optimal model as ARIMA$(2,1,1) \times (0,1,1)_7$. The performance could be improved compared to the previous model. The model had a lower AIC (1773.8), a lower MAE (1427.9), and a reduced MAPE of 18.6. The residuals approximated the normal distribution well and only slight correlation between the residuals could be observed (Figure 6.11). High residuals could be detected in the middle of the analyzed time period and at point around 90 days since the beginning of the wave. The residual showed some heterogeneous variance, in particular in the middle of the wave.

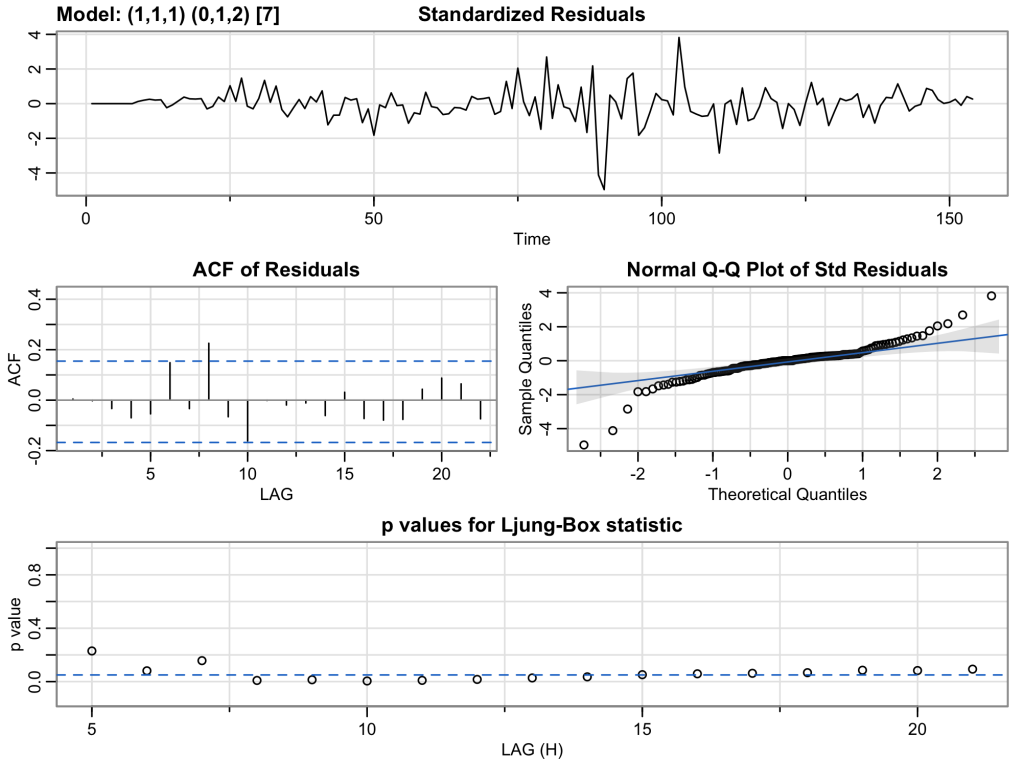


Figure 6.8: Analysis of residuals for seasonal ARIMA model for second wave.

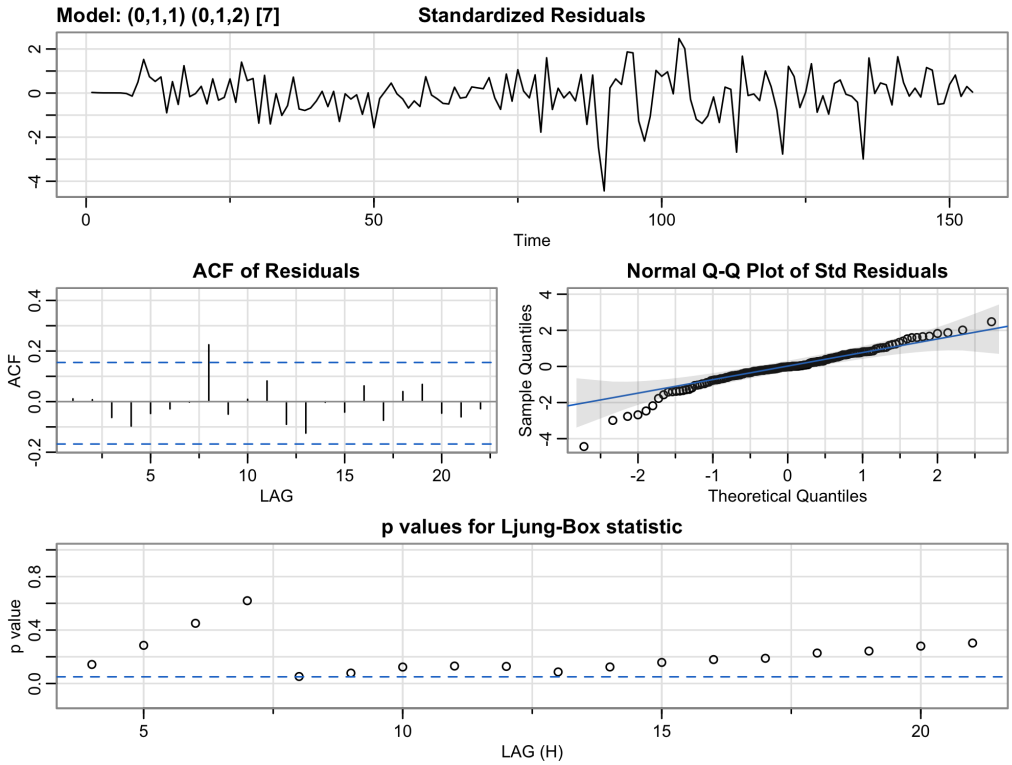


Figure 6.9: Analysis of residuals for seasonal log-transformed ARIMA model for second wave.

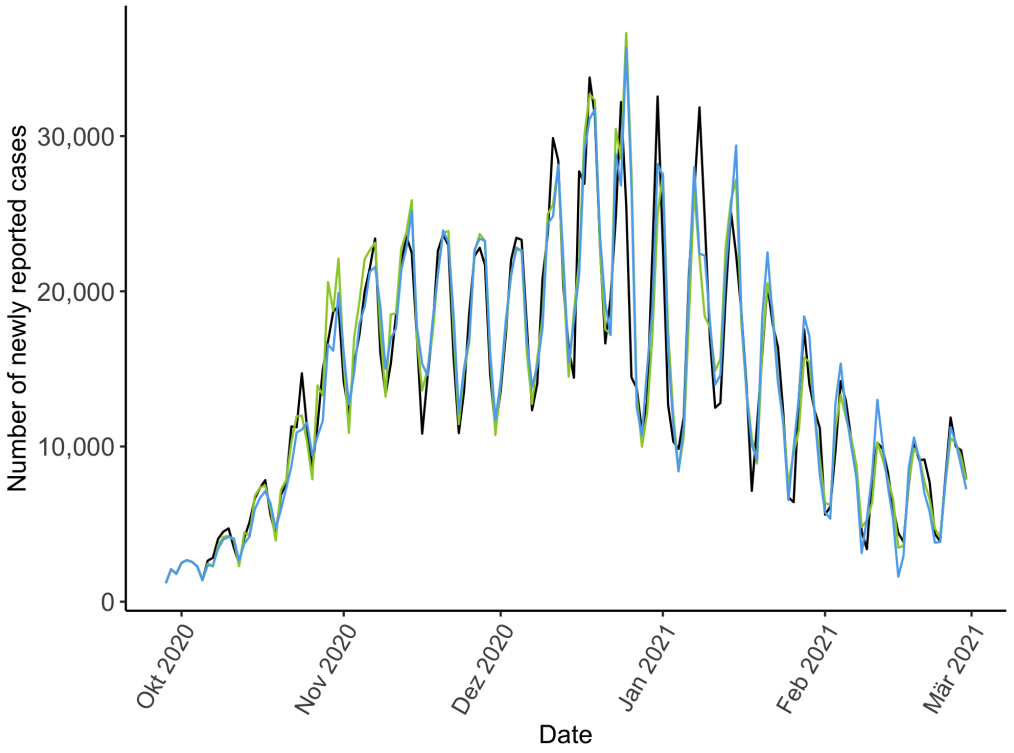


Figure 6.10: Fitted COVID-19 incidence of a seasonal ARIMA model without transformation (blue) and with logarithmic transformation (green) compared to the observed values (black) in the second wave.

We applied a logarithmic transformation to the daily incidence of COVID-19, and fitted a seasonal ARIMA model. Figure 6.12 shows the analysis of the residuals for the optimal ARIMA model (ARIMA$(3,1,0) \times (0,1,1)_7$). The variance of the residuals was more homogeneous than before and the correlation between the residuals could be further reduced. The p-values for the Ljung-Box statistic indicated that the residuals are independent. The distribution of the standardized residuals was closer to a normal distribution. Compared to a model without transformation, the model with transformation improved the MAE (1427.9 to 1346.6) and MAPE (18.6 to 12.4). The fitted value of both models are visualized in Figure 6.13.

### Fourth COVID-19 wave

We repeated a similar analysis for the fourth COVID-19 wave that lasted from 08/02/2021 until 12/26/2021. We selected ARIMA(2,1,2) as the best model without a seasonal component regarding AIC. Similar problems like for the second and third wave arose. This model could not approximate the data adequately. AIC was 2886.9, MAE was 3178.2, and MAPE was 20.7. The residuals approximated the normal distribution relatively well but the residuals showed correlation, especially at lag 7. This indicated that a seasonal component could improve the model.

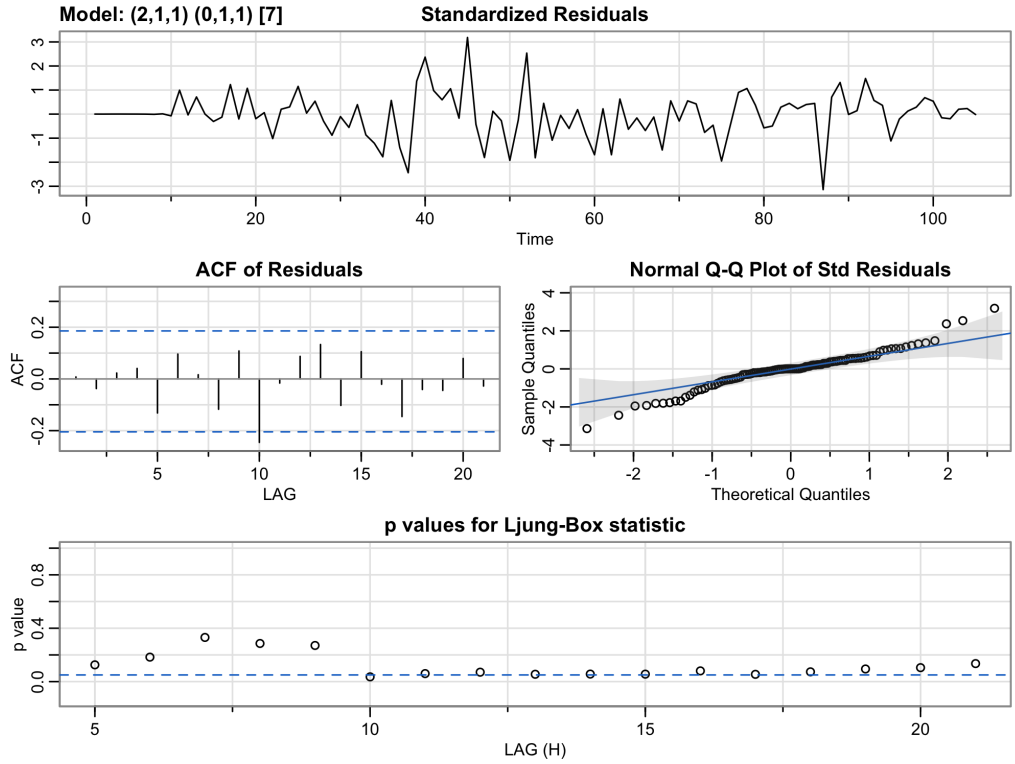


Figure 6.11: Analysis of residuals for seasonal ARIMA model for third wave.

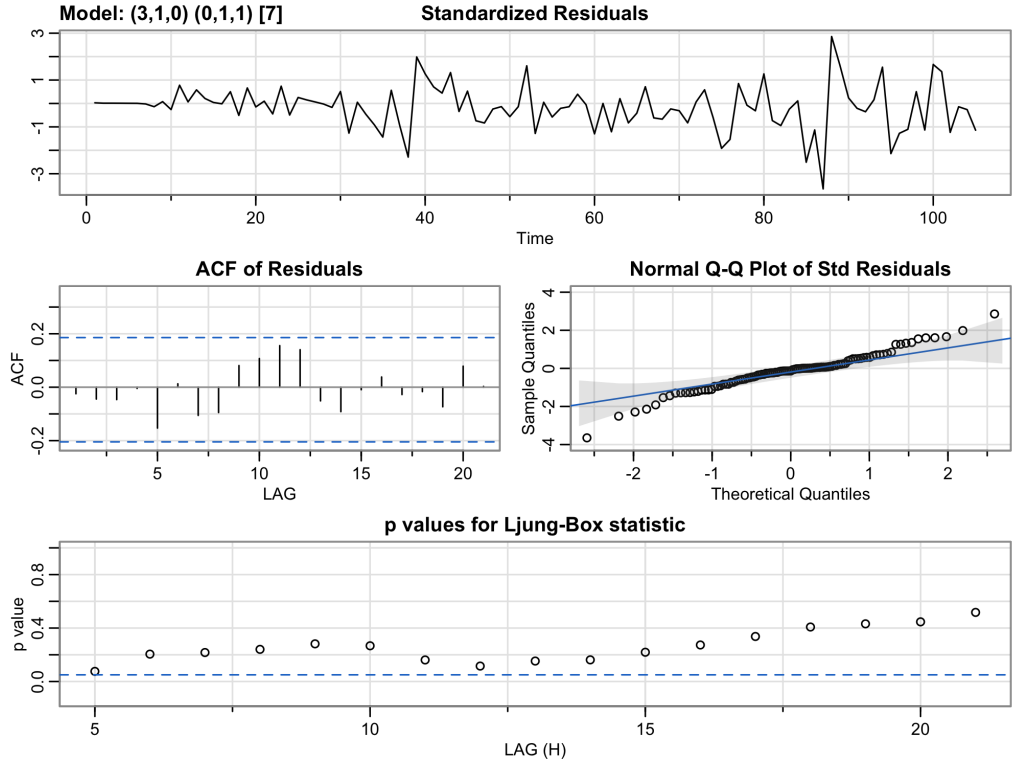


Figure 6.12: Analysis of residuals for seasonal log-transformed ARIMA model for third wave.

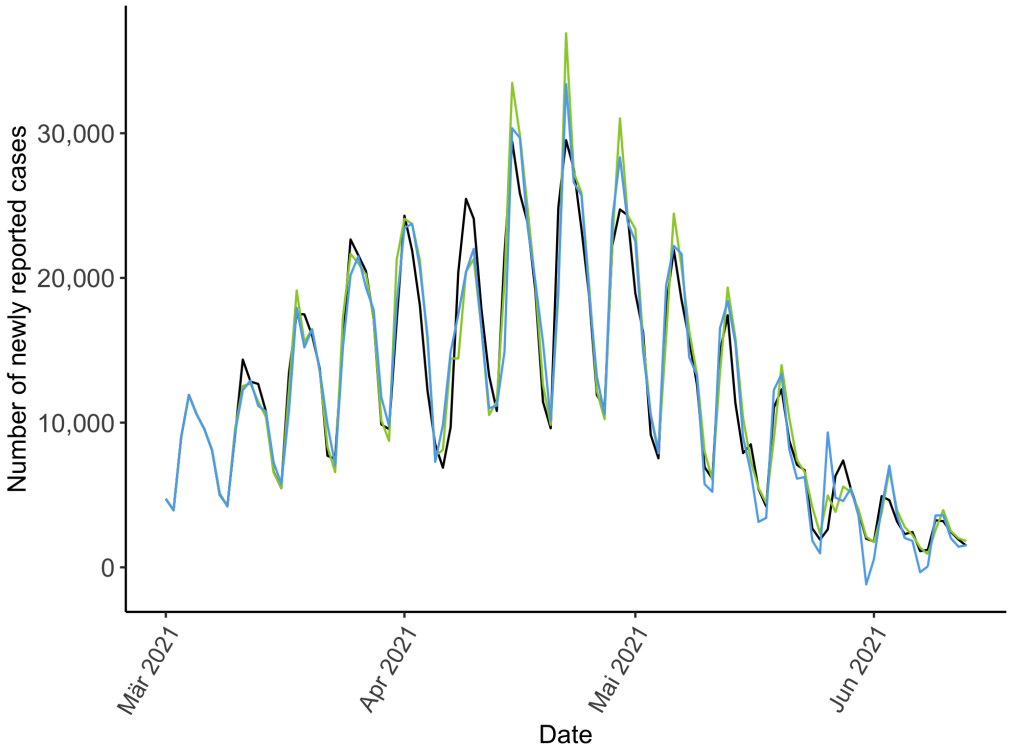


Figure 6.13: Fitted COVID-19 incidence of a seasonal ARIMA model without transformation (blue) and with logarithmic transformation (green) compared to the observed values (black) in the third wave.

An ARIMA model with seasonal component (ARIMA$(1,1,1) \times (1,1,0)_7$) had a better fit to the observed daily incidence. AIC was 2685.6, MAE was 2254.9, and MAPE was 11.4. Despite this fit, several problems were present in the analysis of residuals (Figure 6.14). The distribution of the residuals was not gaussian and a slight correlation between residuals was present. Moreover, the residuals showed a higher variance at the second half of the time period compared to the variance at the first half.

We addressed this issue with a logarithmic transformation and obtained an (ARIMA$(0,1,1) \times (1,1,0)_7$) model. This model had the highest benefit from a transformation of the incidence among all COVID-19 waves. The variance of residuals was homogeneous over the whole interval, and the residuals were less correlated. In addition, the residuals could be well approximated by a normal distribution. The model fit was comparable to the other model without transformation, with a MAE of 2306.5 and a MAPE of 10.3.

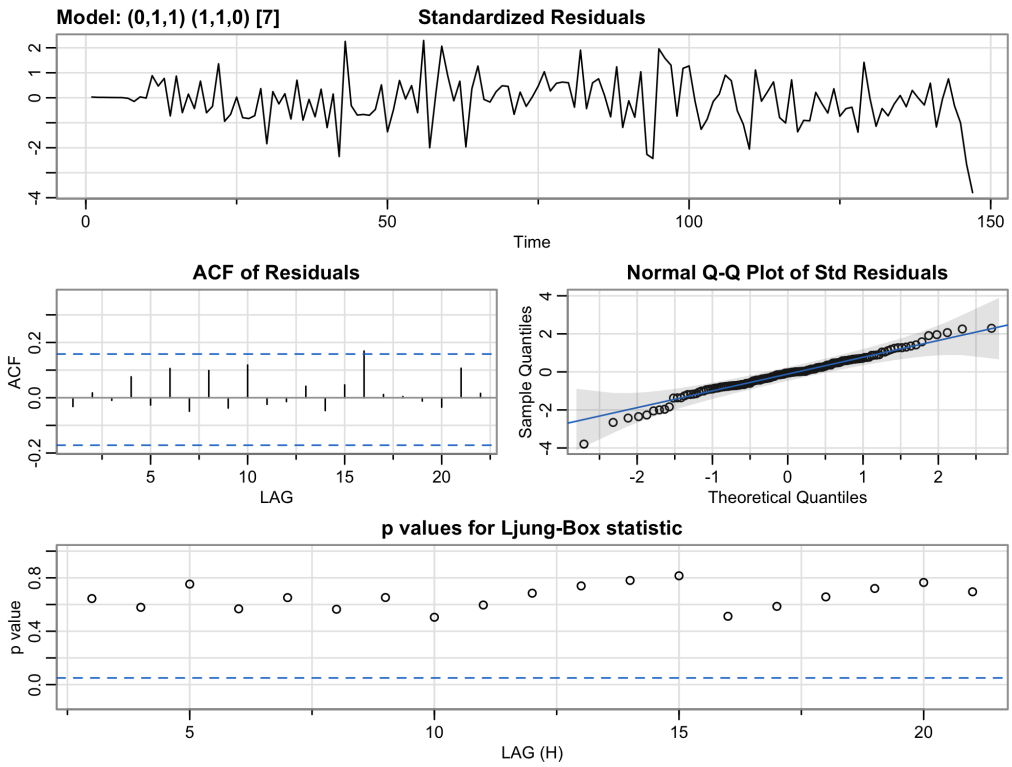


Figure 6.14: Analysis of residuals for seasonal log-transformed ARIMA model for fourth wave.

### 6.1.6 Log-linear autoregressive Poisson model

We developed and examined a log-linear autoregressive Poisson model and adaptions for each wave and compared them in this section.

#### Second COVID-19 wave

We used the second COVID-19 wave to examine the general performance of different settings to model the incidence of COVID-19. Many articles such as Agosto and Giudici (2020) build their model based on the observed value and the estimated mean from the previous day. However, the analysis was not performed on data from Germany. We assumed that our model is defined as

$$x_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t)$$
$$\log(\lambda_t) = \alpha_0 + \alpha_1 \log(1 + x_{t-1}) + \beta_1 \log(\lambda_{t-1}).$$

The ACF plot of the residuals indicated that there was a weekly correlation in the data that needed further observed values to be integrated (Figure 6.16). We expanded the model to include the observed value one week ago and could reduce the correlation between the residuals. The identity link function is an alternative to the logarithmic link function for the Poisson model. We examined whether the

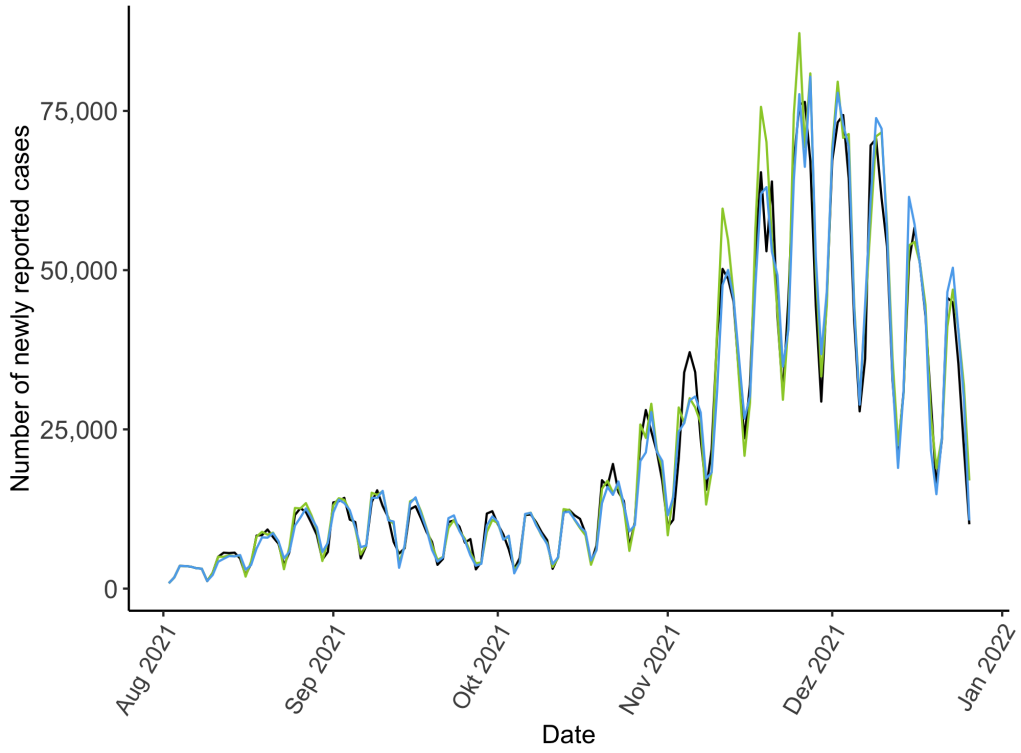


Figure 6.15: Fitted COVID-19 incidence of a seasonal ARIMA model without transformation (blue) and with logarithmic transformation (green) compared to the observed values (black) in the fourth wave.

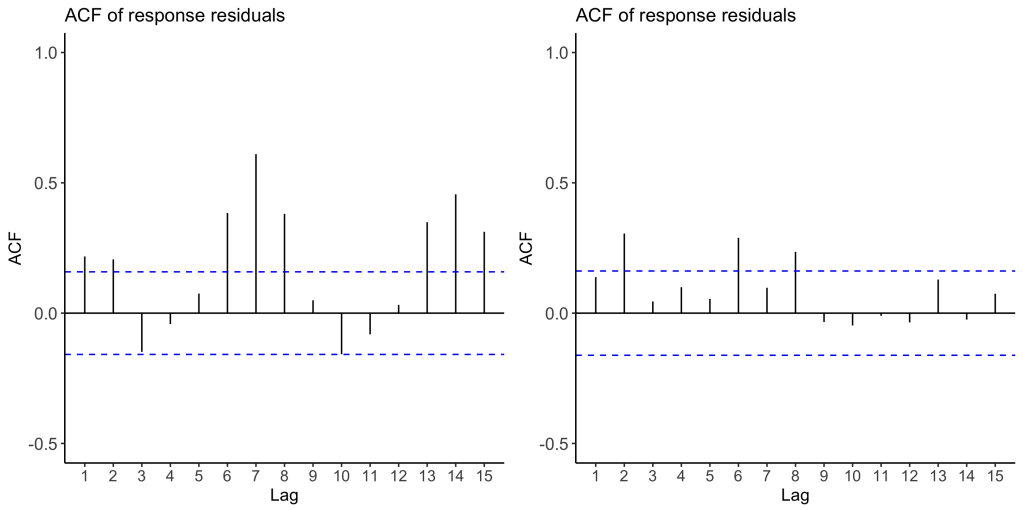


Figure 6.16: ACF plot of a log-linear autoregressive Poisson model with information from prior day only (left) and in addition with the observed value one week before (right).

performance improved by using the identity link function. The coefficients of all parameters need to be positive to ensure a stable solution in this model. This condition caused problems with the fit in the application and could not be used for modeling (Figure 6.17). Hence, we relied on the logarithmic link function for

further models. In addition to previous requirements, it is also important that the

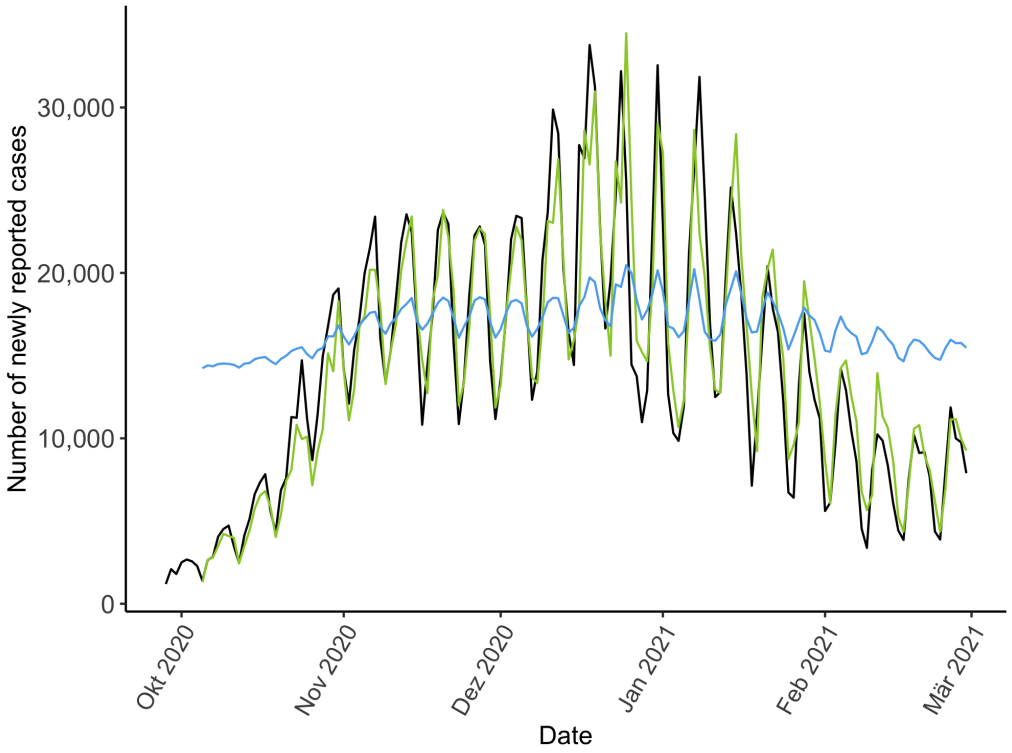


Figure 6.17: Fitted COVID-19 incidence of a log-linear autoregressive Poisson model with identity link function (blue) and with logarithmic link function (green) compared to the observed values (black) in the second COVID-19 wave

predictive performance of a model is correct. We define $P_t(y) = \mathbb{P}(Y_t \leq y|\mathcal{F}_{t-1})$ the cumulative density function of a predictive distribution. We can judge the predictive distribution by examining whether the probability integral transform (PIT) follows a uniform distribution. Czado et al. (2009) defined the PIT value for count data for the observed value $y_t$ and the predictive distribution $P_t(y)$ by

$$F_t(u|y) = \begin{cases} 0, & \text{for } u \leq P_t(y-1) \\ \frac{u-P_t(y-1)}{P_t(y)-P_t(y-1)}, & \text{for } P_t(y-1) < u < P_t(y) \\ 1, & \text{for } u \geq P_t(y). \end{cases} \tag{6.6}$$

Furthermore, the mean PIT is defined by

$$\bar{F}(u) = \frac{1}{n}\sum_{t=1}^{n} F_t(u|y_t),\ 0 \leq u \leq 1. \tag{6.7}$$

Czado et al. (2009) suggest to plot a histogram with 10 bins to check that the mean PIT is uniformly distributed. If the model is well suited, a histogram of PIT should consist of equal high bars. In contrast, a U-shape indicates dispersion.

Because the PIT distribution showed high tails and a strong U-shape, the predictive distribution needed further parameters for dispersion. We fitted a model with the same parameters but a negative binomial function and found that the distribution of PIT was closer to a uniform distribution than the distribution for a Poission distribution function (Figure 6.18). We further explored if the estimated mean seven

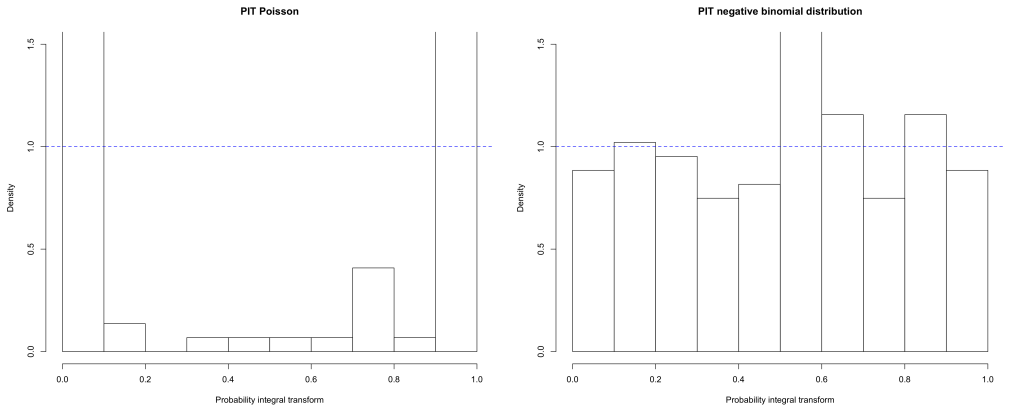


Figure 6.18: Comparison of PIT distribution between a log-linear autoregressive model with Poisson distribution (left) and with negative binomial distribution (right)

days prior can improve the estimation and the analysis of the residuals. This showed only little effect and was neglected. Because the daily COVID-19 incidence is influence by non-pharmaceutical interventions and the current distribution of COVID-19 variants, we included the stringency of interventions and the variants of interest as internal factors in the model. During the second wave only the wildtype and the Alpha variante was present. Both expansions showed an improvement for the model compared to other settings (Table 6.1). We drop the first observed values from the calculation to avoid errors through the first estimates because they cannot be estimated with the model. Thus, a comparison with the ARIMA models based on this measures would be biased. Of note, the point estimation of incidence is independent of the chosen distribution function. The distribution has only influence on confidence intervals.

Considering the residuals (Figure 6.19), the distribution of Pearson residuals was approximately normal and the PIT of the model was uniformly distributed. The model still had some correlation in the residuals that persisted even after adding more observed values. Here should be noted that many articles do not present the correlation structure of their investigated models but only rely on the fitted values and their prediction so that a comparison with other articles is not available.

Table 6.1: Comparison of log-linear autoregressive Poisson models for the second wave

| Model Setting | MAPE | ME | MAE |
|---|---|---|---|
| $x_1, \lambda_1$ | 27.7 | -15.2 | 2990 |
| $x_1, x_7, \lambda_1$ | 14.5 | -0.04 | 1932 |
| $x_1, x_7, \lambda_1, \lambda_7$ | 14.7 | -5.5 | 1927 |
| $x_1, x_7, \lambda_1$ + stringency index | 13.5 | -7.0 | 1797 |
| $x_1, x_7, \lambda_1$ + SI + VOI | 13.1 | -8.9 | 1780 |

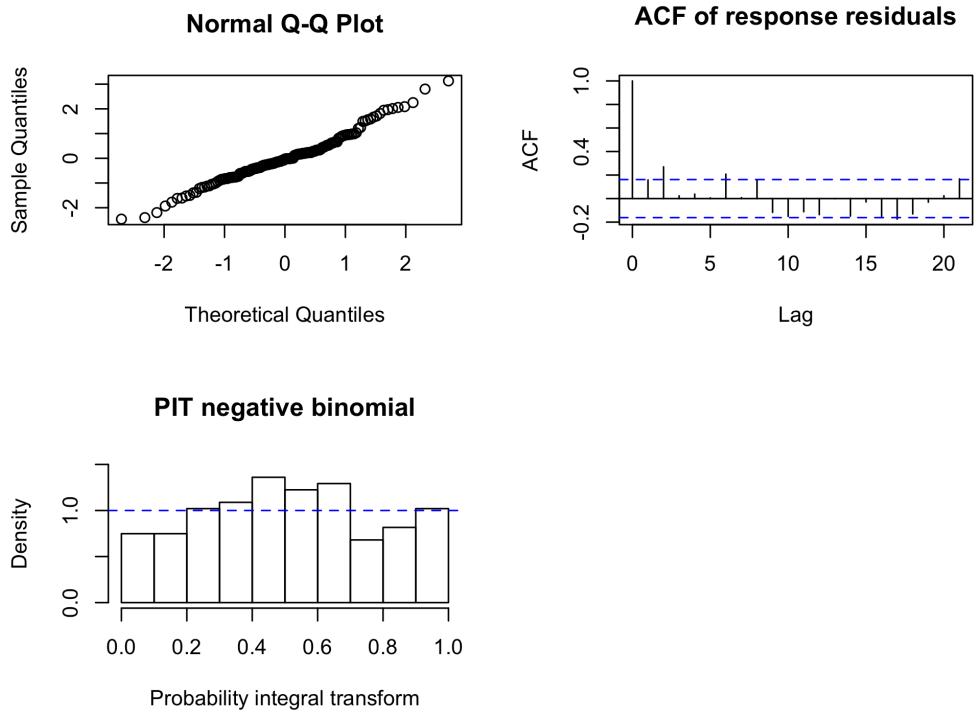


Figure 6.19: Analysis of the residuals of the log-linear autoregressive negative binomial model with stringency and distribution of variants; shown is the QQ plot of the pearson residuals, the ACF plot of the response residuals and a histogram of the PIT.

### Third COVID-19 wave

We compared different models for the third COVID-19 wave, and found similar results like for the second COVID-19 wave (Table 6.2). The stringency of interventions did not improve the model, whereas the VOIs (here the Alpha and Delta variante) led to a reduction in MAE and MAPE. Because the model had high correlation of residuals, we included the observed value of the incidence two days ago. The model performance on the data set remained equal, with a slight increase in the MAPE from 19.5 for the model without $x_2$ to 20.3. However, the correlation between the residuals could be reduced (Figure 6.21). The PIT distribution was close to a uni-

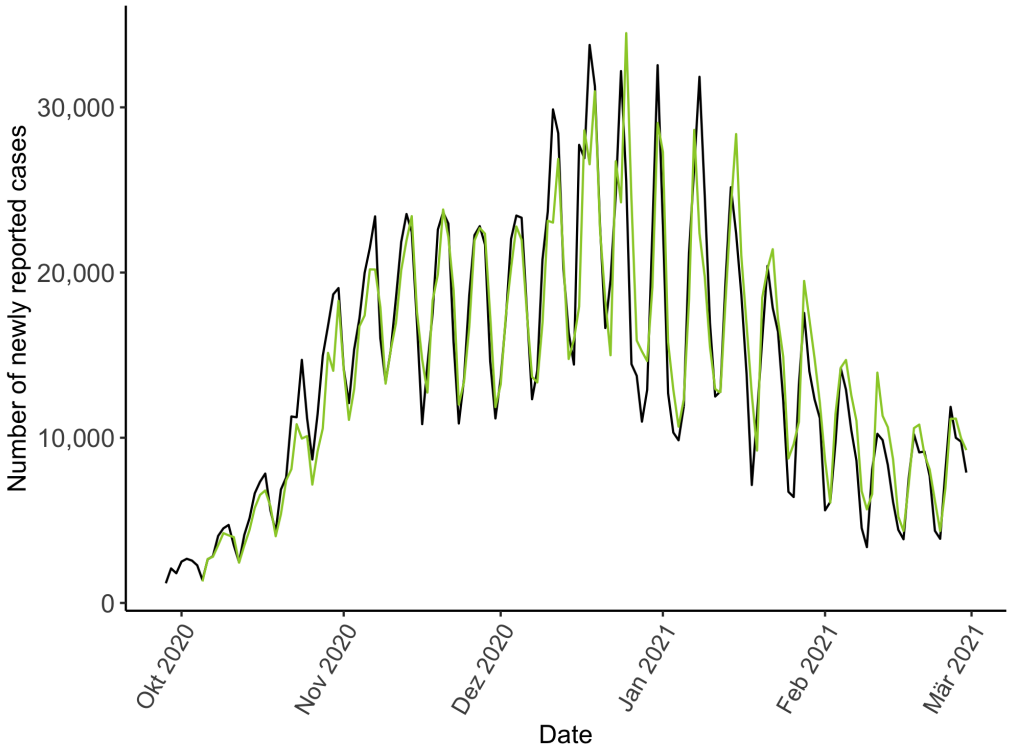


Figure 6.20: Fitted COVID-19 incidence of a log-linear autoregressive negative binomial model with logarithmic link function and included stringency and distribution of variants, compared to the observed values (black) in the second COVID-19 wave.

Table 6.2: Comparison of log-linear autoregressive Poisson models for third wave

| Model Setting | MAPE | ME | MAE |
|---|---|---|---|
| $x_1, x_7, \lambda_1$ | 20.1 | 70.7 | 1807 |
| $x_1, x_7, \lambda_1$ + stringency index | 26.8 | 84.0 | 2127 |
| $x_1, x_7, \lambda_1$ + SI + VOI | 19.5 | 0.3 | 1738 |
| $x_1, x_2, x_7, \lambda_1$ + SI + VOI | 20.3 | -16.9 | 1741 |

form distribution and the pearson residuals had an accurate approximation to a normal distribution in a QQ plot. The fitted model was close to the observed data (Figure 6.22).

### Fourth COVID-19 wave

We fitted different models for the fourth COVID-19 wave (Table 6.3). When we fitted similar models like for the second and third COVID-19 wave, the residuals of all these models were highly correlated, even for higher lags. In order to reduce this correlation, we included the observed incidence two weeks prior in our model. This improved the performance of the model and reduced the correlation (Figure 6.23. Although also other observed values and estimated means were included, the correlation of the residuals could not be further improved. The PIT distribution was

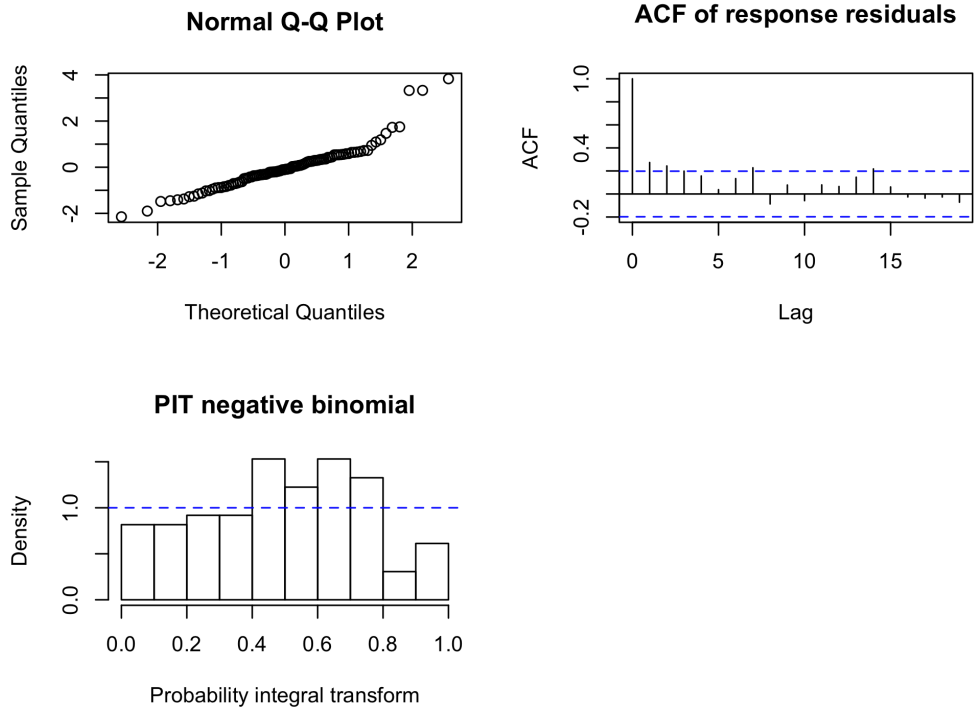


Figure 6.21: Analysis of the residuals of the log-linear autoregressive negative binomial model with stringency and distribution of variants in the third COVID-19 wave; shown is the QQ plot of the pearson residuals, the ACF plot of the response residuals and a histogram of the PIT.

Table 6.3: Comparison of log-linear autoregressive Poisson models for fourth wave

| Model Setting | MAPE | ME | MAE |
|---|---|---|---|
| $x_1, x_7, \lambda_1$ | 16.8 | 17.1 | 3373 |
| $x_1, x_7, \lambda_1$ + stringency index | 16.9 | -53.2 | 3282 |
| $x_1, x_7, \lambda_1$ + SI + VOI | 18.2 | 30.0 | 3366 |
| $x_1, x_7, x_{14}, \lambda_1$ | 15.9 | -0.1 | 3117 |

close to a uniform distribution and the pearson residuals could be approximated by a normal distribution. The fitted model was close to the observed data (Figure 6.24).

### 6.1.7 Discussion

In this section, we compared non-seasonal and seasonal ARIMA models for each COVID-19 wave and analyzed their residuals. The seasonal ARIMA model had a better fit than a non-seasonal ARIMA model, in particular with a logarithmic transformation that stabilizes the variance of the time series. Some articles have tried to fit ARIMA models to predict COVID-19 cases. In many cases, this was done during the first COVID-19 wave only. Although model assumptions should be considered

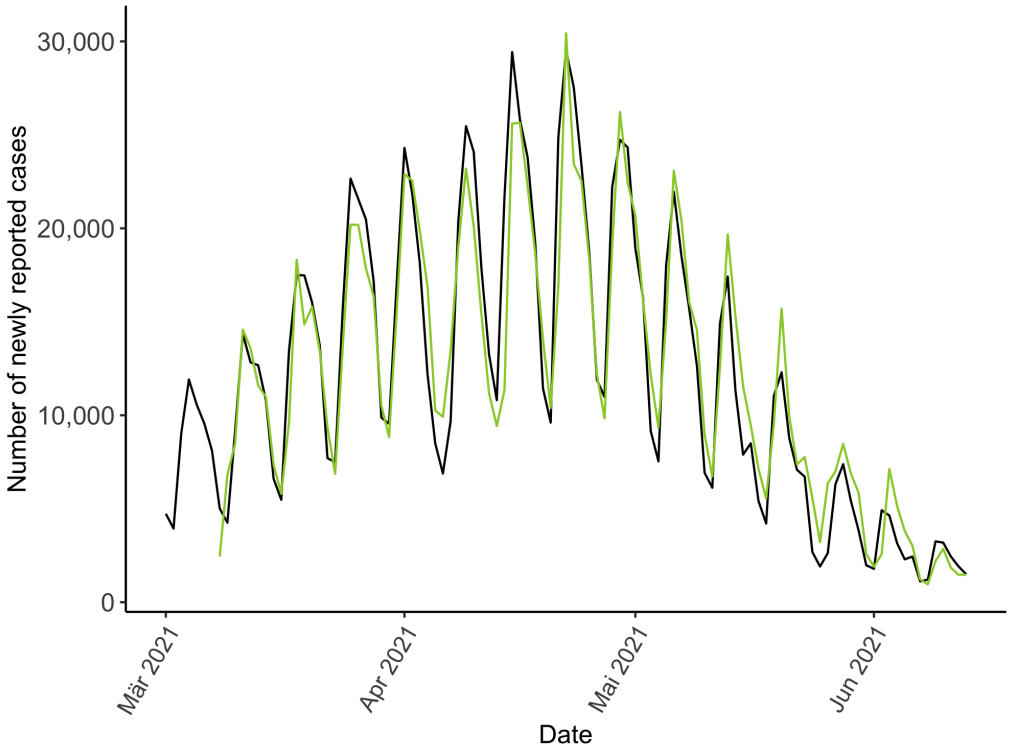


Figure 6.22: Fitted COVID-19 incidence of a log-linear autoregressive negative binomial model with logarithmic link function and included stringency and distribution of variants, compared to the observed values (black) in the third COVID-19 wave.

during the development of a model, many articles present no analysis whether model assumptions could be verified in their model. If a model does not fulfill the assumptions the model is not adequate and cannot be used for forecasting.
As future work it might improve the seasonal ARIMA model to include external covariates such as temperature.

We also examined a log-linear autoregressive Poisson model with several adaptions for COVID-19 incidence. We have shown that for the modeling of COVID-19 incidence a logarithmic link is important because negatively correlated parameters can be included. Furthermore, the negative binomial distribution improved the distribution of the PIT, and thus the predictive distribution, through a higher flexibility in dispersion than the Poisson distribution.

Agosto and Giudici (2020) also used log-linear autoregressive Poisson regressions. They trained their model based on the previous day and the estimated mean of the previous day and showed promising results. However, in this analysis, seasonal effects remained in the model such that further previous observed values were needed in the model.

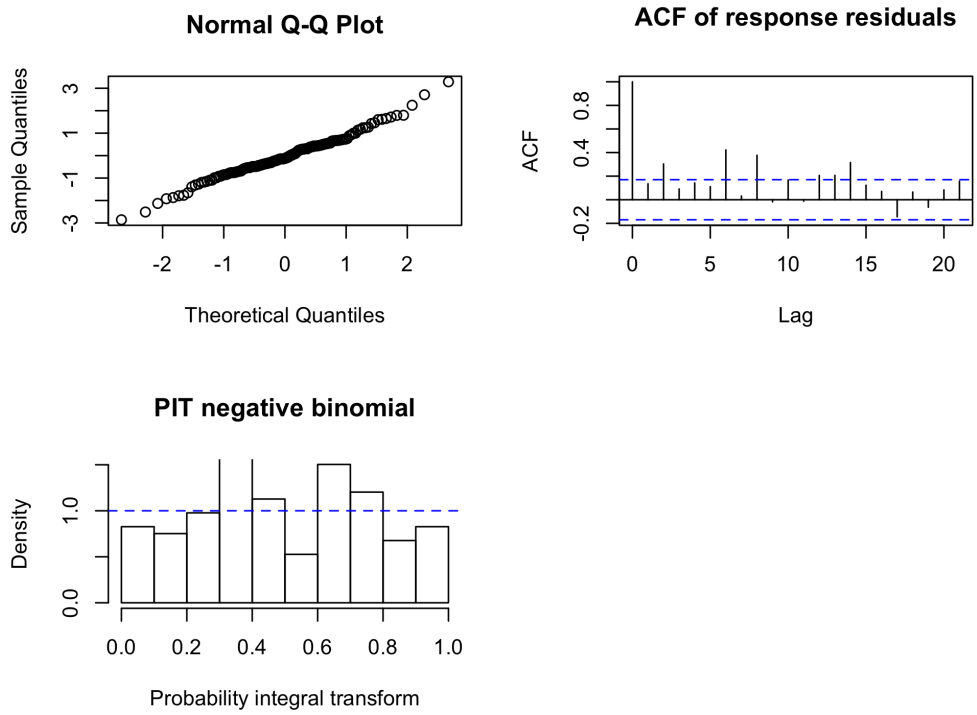


Figure 6.23: Analysis of the residuals of the log-linear autoregressive negative binomial model in the fourth COVID-19 wave; shown is the QQ plot of the pearson residuals, the ACF plot of the response residuals and a histogram of the PIT.

A prediction with several days ahead is complex because the development of daily incidence has a high variability and external influences affect the course. We included the relative frequency of variants in the model and improved the performance of the fit. This approach enables a prediction of the further course some days in advance since the distribution of variants remains relatively stable. Models including a time component as external covariate were also introduced by Agosto and Giudici (2020).

Liboschik et al. (2017) underlined in their work the possibility to include different types of interventions in the model. They distinguished between an outlier at only one point, a decreasing effect of an intervention, and a constant shift through external effects. A closer fit to the data might be reached in further work by applying interventions, in particular to capture outliers in the dataset.

One major limitation is the quality of the data. In particular, at the start of the pandemic, a high number of unreported asymptomatic cases arose which caused a bias of case numbers. Furthermore, cases are often reported lately such that the number might increase after publishing. Preprocessing the data and using them to nowcast the actual number of cases could be an interesting aspect. Studies have shown that this could be achieved either with a method using time-series like Alaimo Di Loro

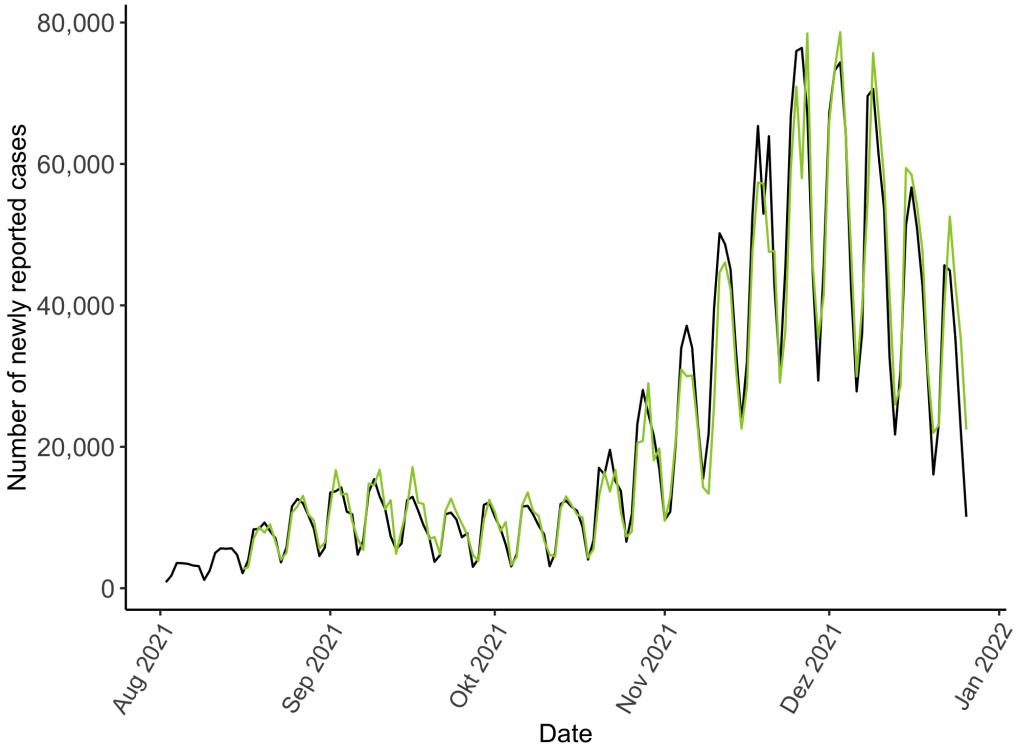


Figure 6.24: Fitted COVID-19 incidence of a log-linear autoregressive negative binomial model with logarithmic link function, compared to the observed values (black) in the fourth COVID-19 wave.

et al. (2021) or with Bayesian hierarchical models like Günther et al. (2021).

## 6.2 Semi-mechanistic SEIR model with change points

All of the models that were presented on the preceding pages relied only on case numbers but did not focus on any biological meaningful interpretation. Due to the importance of this, another type of model was now investigated.

### 6.2.1 Background of interventions against COVID-19

At the beginning of the pandemic in 2020 no vaccine was available and other interventions were implemented in order to reduce the spread of the virus and avoid a high number of deaths. Different statistical models were employed to generate prognoses about the further course of the epidemic and the impact of non-pharmaceutical interventions (NPIs). In the first period a lack of data as well as the systematic under-reporting of cases impeded the estimation of parameters. Several statistical approaches were introduced with the aim of estimating the main parameters of the pandemic and of simulating the course of the pandemic to investigate which mea-

sures mitigate the spread of the virus. At present, more and better-quality data is available due to reliable tests. Several NPIs have been introduced. An index that summarizes many interventions is provided by the Oxford COVID-19 Government Response Tracker (OxCGRT). It covers the current interventions in Germany on a daily record basis.

Several parameters other than NPIs have influenced the course of the pandemic and the characteristics of the virus. Like many viruses, SARS-CoV-2 tends to mutate to improve its conditions and to increase its transmission rate.

In December 2020, the European Medicines Agency approved the first vaccines against COVID-19 which could be shown to deliver a significant protection against severe courses. The first vaccines were released from Pfizer, Moderna, and Jansen. An vaccine from AstraZeneca became available later. During time, the vaccine from Pfizer was mainly used because the vaccine from Jansen and AstraZeneca had potential security issues. Due to the low amount of available vaccines, the elderly and individuals with severe preexisting conditions were prioritized initially. After the administration of a low number of doses at the beginning of 2021, a peak of daily vaccinations could be reached in July 2021. Two doses were needed at first. After the wildtype mutated, three doses became necessary for full immunization. For a detailed overview of COVID-19 vaccines, we refer to Desson et al. (2022).

A large number of statistical approaches were employed to estimate important characteristics from small samples and to provide an outlook on the development of the incidence of COVID-19. The SIR model was the starting point for several models. It is used often for epidemic outbreaks. A mechanical model with the underlying structure of the classic SIR model is suitable when a new virus is in its initial, exponential growth. Then, the inclusion of biological knowledge can help to reduce overfitting compared to other models with more degrees of freedom. Multiple COVID-19 waves made it impossible to focus exclusively on the classical SIR model without further expansions.

Unlike several other authors, Dehning et al. addressed this problem and introduced a novel version of the SIR that relies on a Bayesian framework combined with Markov Chain Monte Carlo (MCMC) sampling. Furthermore, they introduced change points to capture changing properties of the virus and the pandemic. Multiple change points can thus be used to model the reaction of the government, say at the beginning of the pandemic. They can also be employed to capture changes in the transmission rate of the virus.

Bayesian analysis is beneficial due to multiple reasons. One can use priors for the parameters to incorporate prior knowledge about them, which is relevant mainly to the later phases of the pandemic when more information about the virus is available. Prior knowledge is important for the forecast and the estimation of the parameters. Another advantage is that Bayesian inference is not bounded by a density assumption but complex models can be implement more easily.

Xu and Tang (2021) expanded the SIR system with a further compartment and showed that multiple change points can enable the modeling of a long-term period with multiple waves. They also added vaccination to the SEIR model because of its impact on susceptible individuals.

In this section we expand the approach of Dehning et al. (2020) and Xu and Tang (2021) by fitting a SEIR model with multiple change points over a long period. Furthermore, we inspect whether changes that we identify can be explained by the variants of interest.

### 6.2.2 Statistical approaches

#### Estimation with Bayesian MCMC

Bayesian inference is applied for the estimation of parameters in many situations and for nearly all types of models. For the theoretical aspects, we refer to Gelman et al. (1995). The method are intended to combine the likelihood of observed data and knowledge about the parameters (in the form of priors) in order to obtain a posterior distribution of the parameters. All considerations are based on the Bayes Theorem which yields the following:

$$\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{X}, M) = \frac{\mathbb{P}(\boldsymbol{X}|\boldsymbol{\theta}, M)\mathbb{P}(\boldsymbol{\theta})}{\mathbb{P}(\boldsymbol{X})} \propto \mathbb{P}(\boldsymbol{X}|\boldsymbol{\theta}, M)\mathbb{P}(\boldsymbol{\theta}).$$

$\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{X}, M)$ is the posterior distribution of parameter vector, $\mathbb{P}(\boldsymbol{X}|\boldsymbol{\theta}, M)$ is denoted as the likelihood of the observed data $C_t$, and $\mathbb{P}(\boldsymbol{\theta})$ is denoted as prior of the parameter vector.

A major concern in Bayesian inference is that the posterior distribution may not be expressed in an analytical form, thus, we cannot draw samples directly but need an algorithm to sample from the posterior distribution. For posterior inference of our parameter vector, we drew random samples from the conditional distribution of the

given number of real cases at a specific day $t$. In the present case, we obtain

$$\mathbb{P}(\boldsymbol{\theta}|C_t) \propto \mathbb{P}(C_t|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta}).$$

**NUTS algorithm**

In this section, we introduce the No-U-Turn Sampler (NUTS) algorithm that is capable to overcome these sampling issues and which was introduced by Hoffman, Gelman, et al. (2014). It is a special case of the Markov Chain Monte Carlo (MCMC) algorithm and creates a Markov chain in which the probability of $\boldsymbol{\theta}$ at time $t+1$ depends only on the value at $t$ and not on other historical values such that

$$\mathbb{P}(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_t) = \mathbb{P}(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t).$$

Many applications of the MCMC rely on the Metropolis-Hastings algorithm to sample from the posterior distribution and to generate a sequence whose distribution converges against the target. Every new generated sampled value of the posterior distribution is accepted with a ratio:

$$\mathbb{P}(\text{accept}(\boldsymbol{\theta}_t)) = \min\left(1, \frac{\mathbb{P}(\boldsymbol{\theta}_t|C_t)\mathbb{P}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})}{\mathbb{P}(\boldsymbol{\theta}_{t-1}|C_t)\mathbb{P}(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t)}\right)$$

The Hamiltonian Monte Carlo algorithm (HMC) can outperform the commonly used symmetric random walk in the generation of new values of the posterior distribution. According to Neal et al. (2011), it can often solve the problem of a low acceptance rate by providing a more directed search in the parameter space. The concept is based on Hamiltonian dynamics. We denote the parameter space as $\boldsymbol{\theta}_i$ and introduce an auxiliary variable $r_i$ in every step. We use the current gradient to guide the next state of the Markov chain to a value with a high probability of acceptance. We define the Hamiltonian $H(\boldsymbol{\theta}, \boldsymbol{r}) = U(\boldsymbol{\theta}) + K(\boldsymbol{r}) = \log p(\boldsymbol{\theta}, \boldsymbol{r}) + K(\boldsymbol{r})$, where $U(\boldsymbol{\theta})$ is the negative log-likelihood of the posterior distribution and $K$ is an additional function that is needed for the direction that is specified by a Gaussian kernel with a covariance matrix $\Sigma$ and defined as:

$$K(r) = \frac{\boldsymbol{r}^T \Sigma^{-1} \boldsymbol{r}}{2}.$$

HMC uses a leapfrog integrator that performs a half step in the direction of $\boldsymbol{r}$, a full step with stepsize $\epsilon$ in the direction of $\boldsymbol{\theta}$ and another half step in the direction of $\boldsymbol{r}$. Then, the newly obtained parameter vector is accepted with a probability

$$\mathbb{P}(\text{accept}(\boldsymbol{\theta}^*)) = \min\left(1, \frac{\mathbb{P}(\boldsymbol{\theta}^*|C_t, H(\boldsymbol{\theta}, \boldsymbol{r}))}{\mathbb{P}(\boldsymbol{\theta}|C_t, H(\boldsymbol{\theta}, \boldsymbol{r}))}\right).$$

Selecting suitable parameters is essential for the performance of the algorithm. The NUTS algorithm is an extension of the HMC whereby the discretization stepsize $\epsilon$ and the trajectory length $L$ are tuned automatically. The stepsize $\epsilon$ is optimized during an initial burn-in phase by setting a target mean acceptance probability, and $L$ is chosen by adding steps iteratively until the trajectory begins to retrace itself. For further details, we refer to Hoffman, Gelman, et al. (2014).

### Compartment models

#### SIR model

A classical SIR model is a compartment model which consists of three different compartments. Susceptible individuals (S) can be infected by others and have not been infected so far. Infected individuals (I) have been in contact with the virus and can transmit the virus to other individuals. Recovered individuals (R) have experienced the virus and are immune to reinfection. We assume that the disease infects susceptible individuals with a rate $\beta$. Therefore, $\beta\frac{SI}{N}$ individuals are moved from the S to the I compartment at every time point. Furthermore, the recovery rate is denoted by $\gamma$.

The compartment model that describes the spread of the disease can be formulated with ordinary differential equations:

$$\begin{aligned} \frac{dS}{dt} &= -\beta\frac{SI}{N} \\ \frac{dI}{dt} &= \beta\frac{SI}{N} - \gamma I \\ \frac{dR}{dt} &= \gamma I. \end{aligned} \tag{6.8}$$

#### SEIR model

Beyond the classical SIR model, further compartments can be included to generate a four state Markov chain model. In the case of COVID-19, patients can be exposed to the virus, but are not infective because of a latent period of time. Therefore, exposed individuals should be considered alongside susceptible, infected, and recovered ones.

When a susceptible individual is in contact with an infected one, it becomes an exposed individual and is only infectious after an incubation period. This extension

can be modelled by the following equations:

$$\begin{aligned} \frac{dS}{dt} &= -\beta \frac{SI}{N} \\ \frac{dE}{dt} &= \beta \frac{SI}{N} - \sigma E \\ \frac{dI}{dt} &= \sigma E - \gamma I \\ \frac{dR}{dt} &= \gamma I. \end{aligned} \tag{6.9}$$

$\beta$ is the transmission rate, $\gamma$ the recovery rate, $\sigma$ is the incubation rate.

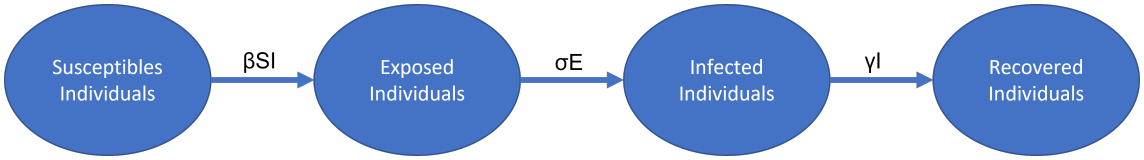


Figure 6.25: Compartments of a SEIR model with transition rate

### 6.2.3 Modifying SEIR system with multiple change points

#### State-of-the-Art

Our modeling approach follows an article from Dehning et al. (2020) which is explained here in more detail. They started with a basic SIR model and introduced multiple change points. Because measurements are taken on a daily basis, we can reformulate system (6.8) on a discrete time scale, such that we obtain the following system, where $S_t$ denotes the number of susceptible individuals at day $t$:

$$\begin{aligned} S_{t+1} - S_t &= -\beta \frac{S_t I_t}{N} \\ I_{t+1} - I_t &= \beta \frac{S_t I_t}{N} - \gamma I_t \\ R_{t+1} - R_t &= \gamma I_t. \end{aligned} \tag{6.10}$$

It should be noted that the size of the population $N$ is set to be constant because we ignore the number of deaths and newborn offspring.

Based on equations (6.10) we can define the number of newly infected persons by

$$I_t^{new} := \beta \frac{S_t I_t}{N}.$$

Dehning et al. (2020) propose incorporating a time delay between the time when a new infection occurred and when it is reported. They defined $C_t := I_{t-D}^{new}$ as the

number of newly reported cases on day $t$.

One reason for the introduction of change points into the model was to investigate the effect of non-pharmaceutical interventions. At the beginning of the epidemic this was reasonable because only a few measures were implemented such that their impact could be more easily tracked. Furthermore, only the wildtype virus was present and no mutations to the virus had occurred. During the further course, a direct association between measures and case numbers is more difficult to be estimated. A SIR model with potential change points can include a changed spreading rate of the virus via governmental interventions or mutations of the virus over a period of time. Thus, a higher number of change points is necessary for a long-term modeling of the number of newly reported cases.

Dehning et al. (2020) introduced a weekly modulation due to the unevenly distributed number of cases. At weekends, low case numbers are reported because the registration of the infection occurs on the following days. They propose to multiply the delayed number of new infections by a factor $(1 - f(t))$, such that we obtain

$$\begin{aligned} C_t &= I_{t-D}^{new}(1 - f(t)), \\ f(t) &= (1 - f_w)\left(1 - \left|\sin\Big(\frac{\pi}{7}t - \frac{1}{2}\Phi_w\Big)\right|\right). \end{aligned} \tag{6.11}$$

Here, $f_w$ and $\Phi_w$ are estimated using the given data.
For the likelihood function they chose a Student's t distribution with a location parameter $\mu$, a scale parameter $\sigma$ and $\nu$ degree of freedom. We set $\nu = 4$ due to a higher stability than Gaussian, in particular if the tails are heavy and neglected noise in the dynamic process. The likelihood function can be written as

$$\mathbb{P}(C_t|\boldsymbol{\theta}) \sim \text{Student-t}(\nu, \mu = C_t(\boldsymbol{\theta}), \sigma = \eta\sqrt{C_t(\boldsymbol{\theta})}). \tag{6.12}$$

The priors were either based on previous studies or if no information was available, chosen as a wide, uninformative prior.

Xu and Tang (2021) further enhanced this model by using a SEIR model with an additional status for exposed people to incorporate the latent time from contact with an infected person to the status where the disease can be transmitted. They also added the exact number of vaccinated people and two coefficients $\alpha_1$ and $\alpha_2$, for the efficiency of the vaccination, whereby they assumed it was no longer possible

to become infected. Their system is described by the following equations:

$$\begin{aligned}
\frac{dS}{dt} &= -\beta\frac{SI}{N} - \alpha_1 V_{1,t} - \alpha_2 V_{2,t} \\
\frac{dE}{dt} &= \beta\frac{SI}{N} - \sigma E \\
\frac{dI}{dt} &= \sigma E - \gamma I \\
\frac{dR}{dt} &= \gamma I + \alpha_1 V_{1,t} + \alpha_2 V_{2,t}.
\end{aligned} \tag{6.13}$$

**SEIR with change points for long-term modeling**

Our aim was to fit a SEIR model with multiple change points over a long period of time to examine the connection between changes in the distribution of variants or the stringency of NPIs and the number of newly reported cases. The introduction of time points can help to identify points where the characteristics of the pandemic changed and might be reasoned by changes in variants or stringency of non-pharmaceutical measures. The Oxford Coronavirus Government Response Tracker (OxCGRT) project provides a Stringency Index which summarizes nine central measures and judges the strength of the interventions. Incorporated interventions are school closures, workplace closures, cancellation of public events, restrictions on public gatherings, closures of public transport, stay-at-home requirements, public information campaigns, restrictions on internal movements, and international travel controls. In Figure 6.26 the OxCGRT-score is shown during the period of the pandemic. This underlines the fact that change points can only be used approximately to determine the effectiveness of measures since interventions were frequently changed. We waived the modification of daily case numbers with a multiplicative factor by using the mean value of newly reported cases for the prior 7 days and hence did not apply a delay. The number of new infected people at day $t$ was defined as:

$$\hat{C}_t = E_{t-1} - E_t + S_{t-1} - S_t. \tag{6.14}$$

Further, we assume that the likelihood belongs to the family of Student's t distributions such that

$$\mathbb{P}(C_t|\boldsymbol{\theta}) \sim \text{Student-T}(\nu, \mu = C_t(\boldsymbol{\theta}), \sigma = \sqrt{C_t(\boldsymbol{\theta})}). \tag{6.15}$$

We fitted the SEIR model with change points with *RStan*, a package for Bayesian Analysis in R. After a burn-in of 1000 iterations, 1000 iterations were sampled from the posterior distribution. In total, 4 chains were sampled to analyze whether the chains mixed and a convergence was obtained.

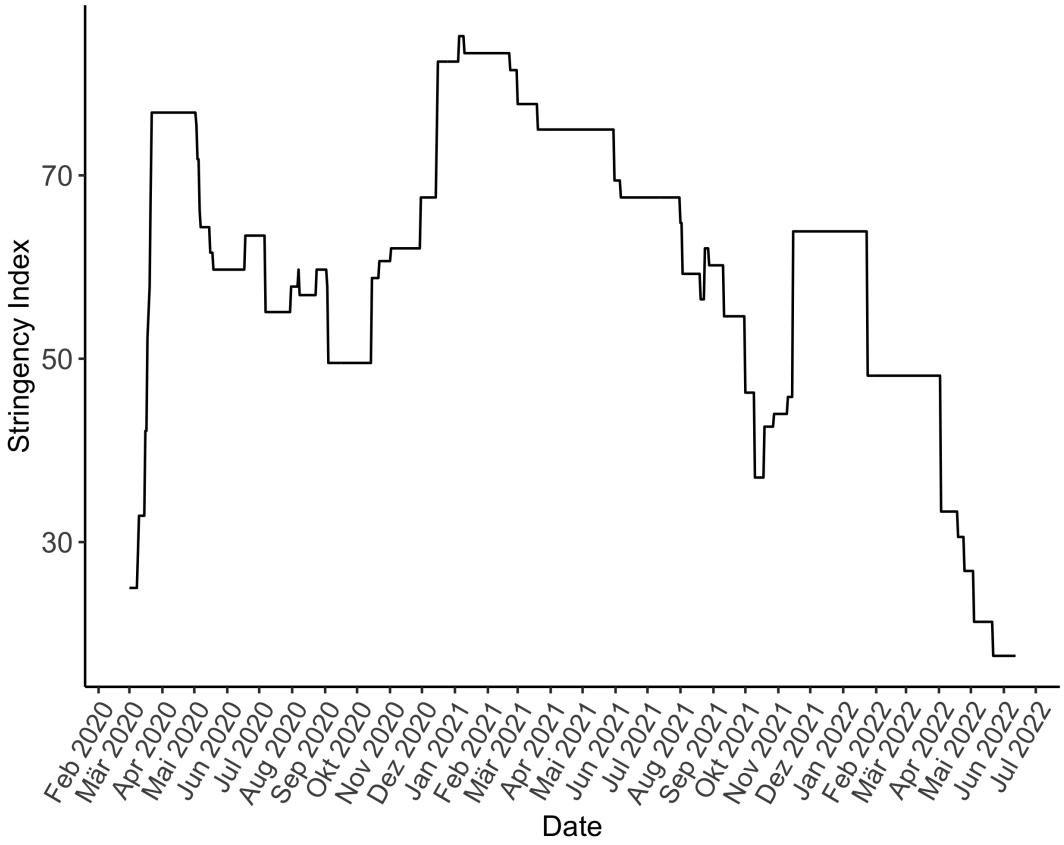


Figure 6.26: Course of the Stringency of Interventions measured with OxCGRT-Score over Time

We included 13 change points, and for the time of changes, we chose a prior normal distribution around roughly estimated values according to the course of cases, and the transmission rate for all parameters was said to be normally distributed with mean 0.4 and standard deviation 0.1. The transition around a change point was modeled by a smooth function to avoid indicator functions. All parameters were also assumed to be approximately normally distributed. The final model is described below:

$$
\begin{aligned}
C_t &\sim \text{Student-T}(\nu = 4, \mu = \hat{C}_t(\boldsymbol{\theta}), \sigma = \sqrt{\hat{C}_t(\boldsymbol{\theta}) + 1}) \\
\beta_i &\sim \mathcal{N}(0.4, 0.1) \quad i \in \{1, \ldots, 13\} \\
\gamma &\sim \mathcal{N}(0.35, 0.03) \\
\omega &\sim \mathcal{N}(0.3, 0.05) \\
i_0 &\sim \mathcal{N}(8000, 500) \\
e_0 &\sim \mathcal{N}(2000, 500) \\
t_1 &\sim \mathcal{N}(65, 3) \\
t_2 &\sim \mathcal{N}(105, 3) \\
t_3 &\sim \mathcal{N}(170, 3) \\
t_4 &\sim \mathcal{N}(210, 3) \\
t_5 &\sim \mathcal{N}(280, 3) \\
t_6 &\sim \mathcal{N}(355, 3) \\
t_7 &\sim \mathcal{N}(405, 3) \\
t_8 &\sim \mathcal{N}(450, 3) \\
t_9 &\sim \mathcal{N}(491, 3) \\
t_{10} &\sim \mathcal{N}(525, 3) \\
t_{11} &\sim \mathcal{N}(562, 3) \\
t_{12} &\sim \mathcal{N}(588, 3) \\
t_{13} &\sim \mathcal{N}(610, 3) \\
S_1 &= N - i_0 - e_0 \\
I_1 &= i_0 \\
E_1 &= e_0 \\
\beta_t &= \frac{\beta_0}{1 + 8\exp(t - t_1)} + \sum_{i=1}^{12} \beta_i \left( (1 - \frac{1}{1 + 8\exp(t - t_i)}) - (1 - \frac{1}{1 + 8\exp(t - t_{i+1})}) \right) \\
&\quad + \beta_{13}(1 - \frac{1}{1 + 8\exp(t - t_{13})}) \\
S_{t+1} &= S_t - \beta_t \frac{S_t I_t}{N} \\
E_{t+1} &= E_t + \beta_t \frac{S_t I_t}{N} - \omega E_t \\
I_{t+1} &= I_t + \omega E_t - \gamma I_t \\
\hat{C}_t &= E_{t-1} - E_t + S_{t-1} - S_t
\end{aligned}
$$

We also provided starting values to the change points which had previously been

Table 6.4: Parameter setting for two SEIR models with change points

| Parameter | model 1 | model 2 |
|---|---|---|
| $\gamma$ | 0.1 | 0.17 |
| $\beta_0$ | 0.2 | 0.34 |
| $\beta_1$ | 0.35 | 0.59 |
| $\beta_2$ | 0.2 | 0.34 |
| $t_1$ | 80 | 80 |
| $t_2$ | 120 | 120 |
| $i_0$ | 3 | 1.66 |
| $e_0$ | 1 | 1 |
| $\omega$ | 0.15 | 0.10 |

tested and shown to be in the correct range. A significant impediment in our model was that different parameter settings can produce similar results, and this made the optimization of our model more difficult. Chains of the MCMC algorithm can become stuck at different local optima, and hence the chains might not mix. We illustrate this behaviour with an example of two situations with different parameters (Table 6.4), which lead to nearly the same result even if the change points occur at the same time (Figure 6.27). If we consider model 1 as the reference, the mean error of model 2 was 0.08 and the absolute mean error of model 2 was 1.00. This problem of the difficult identification of parameters shows that the impact between an association of a single intervention and the incidence cannot be determined entirely or with certainty.

Parameters from the final fitted model are displayed in Table 6.5. The fitted values were close to the actual values (Figure 6.28) and showed a good performance.
If we plot the distribution of important variants of the virus against the time, we can see how the pandemic developed. Interestingly, five of our determined change points coincide with a peak of a variant and the following descent (Figure 6.29). In all cases, the spreading rate decreases to a lower level and after some time rises again. This might be understood from the circumstance whereby the old variant has evolved and the new one needs to develop.

### 6.2.4 Discussion

In this section, we provided a SEIR model with several change points to depict not only a short phase of the pandemic but to provide an expanded discussion of the development of COVID-19 incidence. Compared to the previous section, our model contained more underlying background, resulting from the movement of individuals

Table 6.5: Parameters of the fitted Bayesian SEIR model with change points

| Parameter | Mean | 95%-credible interval |
|---|---|---|
| $\gamma$ | 0.36 | $[0.33, 0.40]$ |
| $\beta_0$ | 0.41 | $[0.39, 0.45]$ |
| $\beta_1$ | 0.65 | $[0.60, 0.69]$ |
| $\beta_2$ | 0.37 | $[0.34, 0.41]$ |
| $\beta_3$ | 0.28 | $[0.23, 0.32]$ |
| $\beta_4$ | 0.44 | $[0.41, 0.47]$ |
| $\beta_5$ | 0.20 | $[0.14, 0.25]$ |
| $\beta_6$ | 0.57 | $[0.53, 0.60]$ |
| $\beta_7$ | 0.35 | $[0.31, 0.39]$ |
| $\beta_8$ | 0.56 | $[0.52, 0.59]$ |
| $\beta_9$ | 0.30 | $[0.24, 0.34]$ |
| $\beta_{10}$ | 0.65 | $[0.61, 0.69]$ |
| $\beta_{11}$ | 0.39 | $[0.35, 0.43]$ |
| $\beta_{12}$ | 0.55 | $[0.51, 0.59]$ |
| $\beta_{13}$ | 0.38 | $[0.30, 0.44]$ |
| $\omega$ | 0.17 | $[0.11, 0.22]$ |
| $i_0$ | 4955 | $[3937, 6220]$ |
| $e_0$ | 1065 | $[348, 2067]$ |
| $t_1$ | 72.1 | $[71.7, 72.6]$ |
| $t_2$ | 100.9 | $[100.6, 101.2]$ |
| $t_3$ | 173.8 | $[173.1, 174.6]$ |
| $t_4$ | 209.3 | $[208.9, 209.7]$ |
| $t_5$ | 281.7 | $[281.4, 282.1]$ |
| $t_6$ | 345.2 | $[344.7, 345.5]$ |
| $t_7$ | 403.6 | $[402.9, 404.7]$ |
| $t_8$ | 444.8 | $[444.4, 445.2]$ |
| $t_9$ | 493.8 | $[493.6, 494.1]$ |
| $t_{10}$ | 526.0 | $[525.7, 526.3]$ |
| $t_{11}$ | 559.3 | $[559.1, 559.6]$ |
| $t_{12}$ | 586.5 | $[586.4, 586.7]$ |
| $t_{13}$ | 610.5 | $[609.4, 613.4]$ |

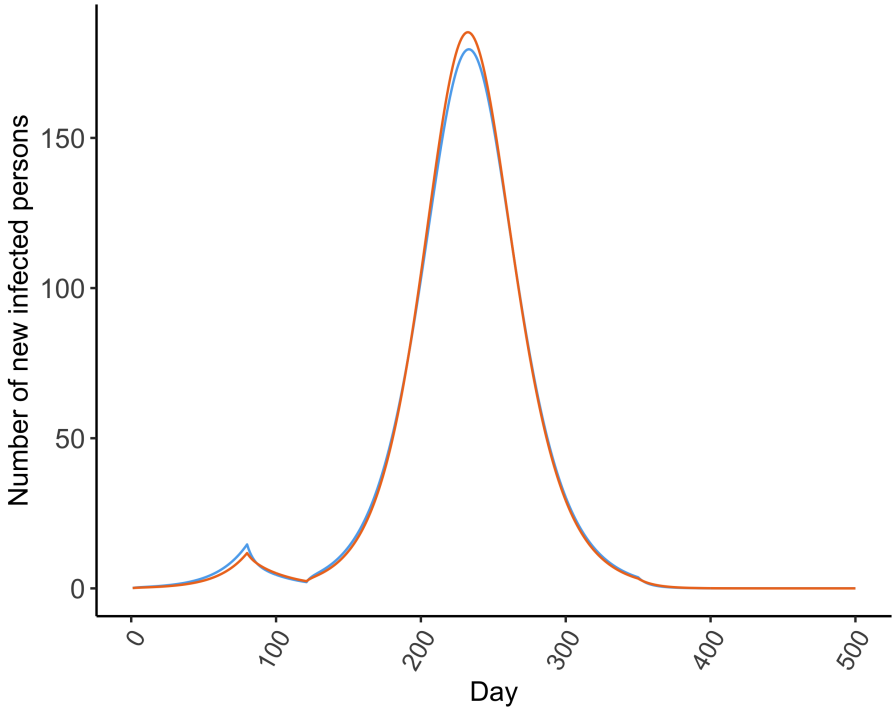


Figure 6.27: Comparison of the simulated course of COVID in two different parameter settings: model 1 (blue line) and model 2 (orange line)

between compartments. We were able to show that an approximation of the course via our model is possible and that the change points show an association with a change in the distribution of variants.

Due to the high multimodality and the combination of different interventions at every time point the interpretation of single measurements does not seem possible and we rely on associations. Besides the considered factors, many other parameters could be influencing the course of the incidence. The longer the duration of the pandemic, the more testing capacities have been available, and this has led to a more reliable overview of the number of cases. Furthermore, even with implemented NPIs it cannot be stated to what degree the interventions are being adhered to. As a result of this impact of external, non-measurable factors, not every change point can be interpreted.

Many different approaches have been suggested for the estimation of the effects of NPIs. Flaxman et al. (2020) and Brauner et al. (2021) introduced renewal equations to model the process of infections which is also based on Bayesian hierarchical models and comparable to a SIR model but not as restricted. They created their model on the basis of a range of countries and the number of new cases and deaths. Furthermore, they transformed the basic reproduction rate into daily growth and

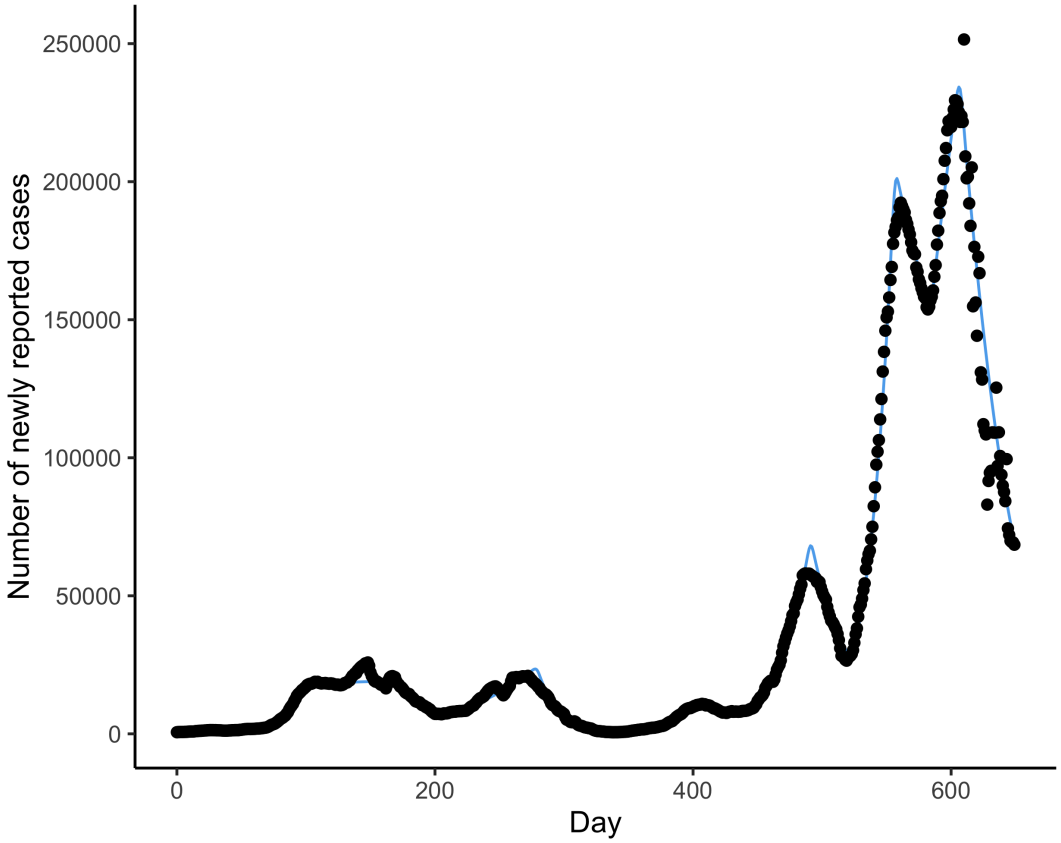


Figure 6.28: Fitted values of the Bayesian SEIR model with change points (blue line) compared with the actual reported numbers (black)

obtained the expected number of daily confirmed cases via a discrete convolution with a relevant delay distribution. So far, there have been only analyses of the first phase of the virus but no further investigation about later waves. Flaxman et al. (2020) could show an effect of gatherings limited to 10, 100, 1000, schools closed, business closed, most businesses closed, universities closed, and stay-at-home orders. However, as mentioned above, it remains questionable if all important facts have been incorporated and how well individual interventions can be identified.

SIR models have been expanded by several additional components to included different situations such as a reinfection, vaccination, and other compartments. Y. Li et al. (2021) provided a time-dependent SEIR model with incubation period, immunity, reinfection, and vaccination and developed a new model named SEVIS. With this model, the trajectories of time-changing parameters (transmission rate, recovery rate, basic reproduction number) could be analysed. Poonia et al. (2022) enhanced a SEIR model with vaccination and introduced several adaptions of these models (with quarantined, hospitalized, vaccinated,...). It is questionable whether models with more compartments are better suited than the classic SEIR model because

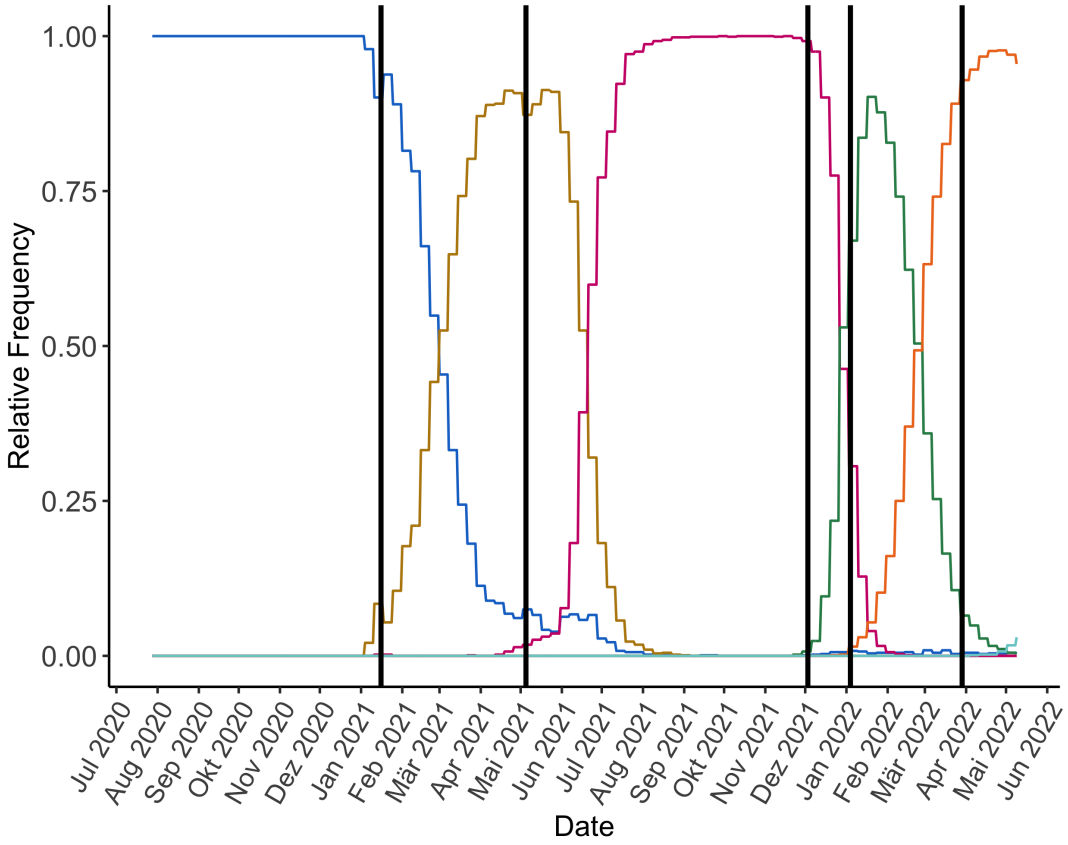


Figure 6.29: Distribution of Variants over time with selected change points from the SEIR model: Wildtype (blue), Alpha variant (brown), Delta variant (purple), Omicron BA.1 variant (green), Omicron BA.2 variant (orange), Omicron BA.5 variant (cyan)

these models need a higher number of parameters and hence more identification problems might arise. Further studies are needed to examine complex compartment models and compare their capability to model the course of disease incidence with measures like AIC.

Finally, we compared the models introduced in this chapter. All introduced models have their specific advantages and a comparison of their performance is difficult because of their diverse structure. External variables can be easier included into a log-linear autoregressive Poisson model than in change point models. By expanding the model equation with terms for external variables, a detection of associations between variables and the course of disease incidence is possible. Another advantage of a log-linear autoregressive Poisson model can be the missing structure that is given in a SEIR model. In particular, when several external factors influence the disease incidence, a high number of change points is needed to provide an approximation of the disease incidence course.

However, change point models are more flexible for external factors through the introduction of change points and can incorporate factors that are unknown so far or are not measurable such as the acceptance of interventions. The location of change points can then be used to search for relevant factors.

Further, a SEIR model with change points facilitates a long-term prognosis. We can predict the further course of disease incidence in case of more restrictive or less restrictive interventions directly through a changed transmission rate. This enables the simulation of different scenarios and an assessment of how strict interventions might be necessary to avoid an exceeding of the clinical capacity limit.

In summary, we have presented several statistical models based on different concepts which are helpful for the modeling of newly reported case numbers of COVID-19. This can help to predict the further course of the disease.

# Chapter 7

# Further Studies

This chapter presents further studies which are based on a collaboration with the University Hospital Augsburg. Because similar methods are already presented in this thesis, only an overview of biological background and of used methods is provided.

**Keywords:** biomarker, survival analysis, generalized linear models, sensitivity and specificity

## 7.1 Autopsies of COVID-19 patients

COVID-19 has been a central part of multiple projects for data analysis. At the initial phase hardly any characteristics of COVID-19 were known and research groups in all fields collected information about the virus. In this project, the analysis was focused on the consequences of an infection for the organs. This work resulted in a publication: Hirschbühl et al. (2021). The Institute of Pathology of University Hospital Augsburg performed autopsies for 19 deceased patients and investigate in which organs SARS-CoV-2 could be detected. These autopsies were able to demonstrate that the lungs in particular showed relevant histological changes. These changes were scored according to their severity. We developed a linear regression model to identify variables which are associated with a higher severity of lung damage.

## 7.2 New biomarker for gastric and colon cancer

Biomarkers have an important role for detection of cancer and the prognosis of its further course. A prognosis has a significant value for decisions concerning a patient's therapy. Biomarkers should be easily accessible and inexpensive, so that they can be used in a daily routine. Together with the Institute of Pathology, we worked

on a new biomarker for gastric and colon cancer (Grosser et al. (2022) and Martin et al. (2021)). A histological sample is labeled positive if a cluster of tumor glands/cells comprising at least five tumor cells and inconspicuous surrounding adipose tissue are noted at the invasion front. This biomarker is named SARIFA, i.e. Stroma AReactive Invasion Front Areas. SARIFA was investigated in two samples: one for gastric cancer and one for colon cancer.

In a sample of 480 adenocarcinomas of the stomach and the gastroesophageal junction, 20% tested positive for SARIFA with a high level of agreement between different pathologists, who had each independently judged each tissue sample (Kappa values: 0.74 and 0.78). A survival analysis showed that patients classified as SARIFA-positive had a shorter overall survival compared to patients that were SARIFA-negative. A Cox proportional hazards regression confirmed SARIFA as an important and independent prognostic factor for the prediction of overall survival (HR= $1.64; 95\%$-CI $1.15 - 2.33$ ;p= 0.006). Furthermore, a transcriptome analysis was performed to identify associated genes.

We found similar results in another analysis for colon cancer ($n = 449$). This time, 25% of all histological samples were SARIFA-positive, and the interobserver variability was low (Kappa: 0.77 and 0.87). We performed a survival analysis to compare patients for their SARIFA status. Patients that were SARIFA-positive had a shorter disease-specific survival, a shorter absence of metastasis and a shorter overall survival. Likewise, we conducted a Cox proportional hazard regression which showed SARIFA as an independent prognostic parameter for colon-cancer-specific survival.

Overall, we were able to conclude that SARIFA is a promising biomarker for gastric cancer and colon cancer. Further studies are required to examine the performance of SARIFA in other types of tumors.

## 7.3 Lymphocyte subsets in patients with colorectal carcinoma

We have already presented an analysis of lymphocyte subsets in a previous chapter. We applied these methods not only for COVID-19, but also for the immune response for colon cancer. Our collaboration with the Second Medical Clinic of University Hospital Augsburg led to another publication: Waidhauser et al. (2021). The research focused on the manner in which a carcinoma influences the immune system

of a patient with cancer compared to that of a healthy individual. The sample consisted of 47 patients and 50 healthy individuals, in whom different lymphocyte subsets were measured by flow cytometry. Characteristics such as age, gender, tumor stage, sidedness of the tumor and microsatellite instability status (MSI) were collected.

We performed unadjusted and adjusted linear regression models on the log-transformed lymphocyte measurements to obtain a closer approximation of the residuals to the normal distribution. Moreover, this transformation led to a more realistic interpretation of the results because the baseline value was multiplied by factor which enabled us to maintain the positivity of the response variable. The adjusted model contained age and gender for the basic analysis, as well as other tumor characteristics for further analysis. We were able to show that B cells, helper T cells and NK cells were lowered for individuals with cancer.

## 7.4 Comparison of surgery techniques for parotidectomy

Generalized linear models are an important tool in medical research. We applied these in a project that compared different surgery techniques according to their complication rate (cf. Thölken et al. (2021)). This prospective study was a collaboration with the Department of Otorhinolaryngology – Head and Neck Surgery and included 300 patients who had been treated for benign neoplasms. A part of the sample received an extracapsular dissection (ECD), whereas others underwent a surgery with a standard surgery technique. Primary endpoints were the incidence rates of transient and permanent (18 months after surgery) facial palsy. Simple and multiple logistic regressions were performed for both endpoints, with the parameters age, number of lesions, size of lesions, duration of surgery, and type of surgery. All parameters except age were dichotomized in the model. We were able to show that ECD had lower incidence rates for complications compared to other surgery techniques, even after adjustment for other variables.

## 7.5 Accuracy of ultrasound-guided core needle biopsy

Diagnostic tools which can be used preoperatively to diagnose patients are of significant importance for the medical treatment. In a collaboration with the Department of Otorhinolaryngology – Head and Neck Surgery, we investigated the accuracy of

an ultrasound-guided core needle biopsy of a parotid lesion in a retrospective study. Our work led to a publication: Jering et al. (2021). A core needle biopsy was conducted in patients, and the results of the biopsy were verified either in a necessary surgery or during the follow-up. Sensitivity as well as specificity were computed for the group and showed high values, indicating the suitability of a core-needle biopsy for diagnosis. In addition, the accuracy of the histological classification was high and only a small proportion of all patients suffered from post-procedural complications.

## 7.6 Survival analysis for parotid gland

Studies are not only needed to evaluate diagnostic tools but also for analysis of overall survival and cancer-specific survival of patients in order to support the identification of relevant risk factors. In a retrospective study with the Department of Otorhinolaryngology – Head and Neck Surgery, the survival of patients with primary malignancies and metastatic cutaneous squamous cell carcinoma of the parotid glands during follow-up was investigated (cf.Jering et al. (2022)). 94 patients with a follow-up of at least two years were included. The mean follow-up according to the inverse Kaplan-Meier method was 50 months.
We tested for differences between patients with a primary malignancy and those with a metastatic malignancy using chi-squared tests or Fisher's exact tests, as well as t-tests or Wilcoxon-Mann-Whitney tests. Overall and disease-free survival was compared with Kaplan-Meier curves and results were presented for two years and five years after diagnosis. Univariable Cox proportional hazard regressions were performed to determine factors associated with disease-free survival in both subgroups of tumor types separately. Schoenfeld residuals were examined to ensure proportional hazards. Patients with metastatic malignancies had a low survival rate, and such patients might benefit from an earlier diagnosis of the metastases.

## 7.7 VR-based relaxation for enhancement of perioperative well-being

Hospital stays, and surgeries, in particular, cause a lot of stress for a patient and can also have an impact on their well-being. Patients would benefit from approaches that can reduce stress and increase quality of life. Listening to classical music might be a possible measure, but there are also new tools such as an intervention based on virtual reality (VR). In collaboration with the Department of General, Visceral and Transplantation Surgery, we evaluated the impact of a VR intervention on patients with colorectal cancer and compared the performance to a music intervention and

a control group without any intervention. This work resulted in a publication: Schrempf et al. (2022). Patients were assigned randomly to a study group. Due to the varying number of interventions the measurements of each patient were averaged. We used non-parametrical tests to compare patients before and after an intervention. After the intervention, patients showed a reduction in heart rate and respiratory rate. Furthermore, their overall mood improved. Quality of life was similar across all groups. Patients with a VR intervention experienced a greater improvement in mood and vital signs than those in the music group.

# Chapter 8

# Summary

In this thesis, several methods from machine learning and advanced statistical methods have been presented for the development of biomarkers and for models of the course of disease incidence. Theoretical concepts were applied in projects with University Hospital Augsburg.

Machine learning has acquired importance in research because of its algorithms that are able to approximate complex structures and hidden features. In this thesis, a biomarker for patients with colon cancer was developed based on a machine learning algorithm, and this was shown to be well suited for classifying the risk of the occurrence of metastases. We used histologically stained images of tumor tissue, binarized the images, and trained a CNN to predict the probability of the occurrence of metastases. Patients in the high-risk group had a shorter metastases-free survival and our risk factor was an independent prognostic factor. As the number of digital histological images and the computing performance further increase, machine learning algorithms will be possible for many different applications, and these will help physicians not only with patient diagnoses but also with their prognosis.

Scores are essential in clinical routine since they are easily accessible and able to distinguish patients. This thesis introduced two theoretical concepts for scores. The first was based on GLMs and was applied to the risk of a permanent shunt implantation. Parameters of the patient captured during admission or measurements in the brain were shown to be relevant for a prognostic score.
The second project used approaches from survival analysis to stratify patients with oligometastatic colon cancer according to their overall survival after the surgical removal of metastases. Risk factors could be identified that enable the identification of a subset of patients who are likely to benefit from a surgery. Applications such as this underline the value of scores for patient treatment and how they can support the formulation of a prognosis.

The COVID-19 pandemic has caused many infections and deaths. In this thesis, we examined how lymphocyte counts are affected by a COVID-19 infection. Nearly all subsets were reduced in patients with an infection compared to a healthy control group. We observed that the lymphocyte counts needed to be transformed with a logarithmic function to ensure a normal distribution of the residuals of the regression model.
Furthermore, we compared different strategies to model the daily reported number of new infections. Non-seasonal and seasonal ARIMA models as well as a log-linear autoregressive Poisson model showed an adequate approximation of the actual case numbers. We included the distribution of variants at each time point as an external covariate.
All these models performed well but encountered problems with the simulation of different settings of non-pharmaceutical interventions. We relied on a Bayesian SEIR model, which has a mechanistic structure and incorporated several change points to account for changes in the virus variants and the stringency index. The obtained fit was close to the actual case numbers and changes in the variants of interest could be found as change points in our model. This method enabled the simulation of effects of interventions according to influences on the transmission rate.

In summary, all these theoretical concepts underline the importance of advanced statistical methods in medicine for the development of biomarkers and models of the evolution of disease incidence.

# List of abbreviations

| | |
|---|---|
| ACA | anterior cerebral artery |
| ACF | autocorrelation function |
| AIC | akaike information criterion |
| ARDS | acute respiratory distress syndrome |
| AUC | area under the receiver operator curve |
| aSAH | aneurysmatic subarachnoid hemorrhag |
| BIg-CoMet | Binary ImaGe Colon Metastasis classifier |
| CI | confidence interval |
| CT | computer tomography |
| CNN | convolutional neural network |
| COVID-19 | Corona Virus Disease 2019 |
| Cox PH model | Cox proportional hazards model |
| DFS | disease-free survival |
| GCS | Glasgow-Coma-Scale |
| GLM | generalized linear model |
| H&E | haematoxylin and eosin stain |
| HR | hazard ratio |
| ICA | internal carotid artery |
| IRT | inflammatory response to the tumor |
| KM | Kaplan-Meier curve |
| MCA | middle cerebral artery |

| | |
|---|---|
| MERS-CoV | Middle east respiratory syndrome coronavirus |
| MLE | maximum likelihood estimator |
| NN | neural network |
| OR | odds ratio |
| OS | overall survival |
| OxCGRT | Oxford COVID-19 Government Response Tracker |
| PACF | partial autocorrelation function |
| RKI | Robert Koch Institute |
| ROC | receiver operator curve |
| SARS-CoV-2 | Severe acute respiratory syndrome coronavirus type 2 |
| SIR | Susceptible-Infected-Recovered Model |
| SEIR | Susceptible-Exposed-Infected-Recovered Model |
| TNM | Tumor-Node-Metastasis classification |
| TS | training sample |
| TSR | tumor-stroma ratio |
| UICC | Union for International Cancer Control |
| VS | validation sample |

# Bibliography


Abràmoff, M. D., P. J. Magalhães, and S. J. Ram (2004). "Image processing with ImageJ". In: *Biophotonics international* 11.7, pp. 36–42.

Agosto, A. and P. Giudici (2020). "A Poisson autoregressive model to understand COVID-19 contagion dynamics". In: *Risks* 8.3, p. 77.

Alaimo Di Loro, P., F. Divino, A. Farcomeni, G. Jona Lasinio, G. Lovison, A. Maruotti, and M. Mingione (2021). "Nowcasting COVID-19 incidence indicators during the Italian first outbreak". In: *Statistics in Medicine* 40.16, pp. 3843–3864.

Barrıa-Sandoval, C., G. Ferreira, K. Benz-Parra, and P. López-Flores (2021). "Prediction of confirmed cases of and deaths caused by COVID-19 in Chile through time series techniques: A comparative study". In: *Plos one* 16.4, e0245414.

Bederson, J. B., E. S. Connolly Jr, H. H. Batjer, R. G. Dacey, J. E. Dion, M. N. Diringer, J. E. Duldner Jr, R. E. Harbaugh, A. B. Patel, and R. H. Rosenwasser (2009). "Guidelines for the management of aneurysmal subarachnoid hemorrhage: a statement for healthcare professionals from a special writing group of the Stroke Council, American Heart Association". In: *Stroke* 40.3, pp. 994–1025.

Benvenuto, D., M. Giovanetti, L. Vassallo, S. Angeletti, and M. Ciccozzi (2020). "Application of the ARIMA model on the COVID-2019 epidemic dataset". In: *Data in brief* 29, p. 105340.

Bland, J. M. and D. G. Altman (1996). "Transformations, means, and confidence intervals." In: *BMJ: British Medical Journal* 312.7038, p. 1079.

Brauner, J. M., S. Mindermann, M. Sharma, D. Johnston, J. Salvatier, T. Gavenčiak, A. B. Stephenson, G. Leech, G. Altman, V. Mikulik, et al. (2021). "Inferring the effectiveness of government interventions against COVID-19". In: *Science* 371.6531, eabd9338.

Brierley, J. D., M. K. Gospodarowicz, and C. Wittekind (2017). *TNM classification of malignant tumours*. John Wiley & Sons.

Brockwell, P. J. and R. A. Davis (2002). *Introduction to time series and forecasting*. Springer.

Bychkov, D., N. Linder, R. Turkki, S. Nordling, P. E. Kovanen, C. Verrill, M. Walliander, M. Lundin, C. Haglund, and J. Lundin (2018). "Deep learning based

tissue analysis predicts outcome in colorectal cancer". In: *Scientific reports* 8.1, pp. 1–11.

Castelvecchi, D. (2016). "Can we open the black box of AI?" In: *Nature News* 538.7623, p. 20.

Changyong, F., W. Hongyue, L. Naiji, C. Tian, H. Hua, L. Ying, et al. (2014). "Log-transformation and its implications for data analysis". In: *Shanghai archives of psychiatry* 26.2, p. 105.

Chen, G., D. Wu, W. Guo, Y. Cao, D. Huang, H. Wang, T. Wang, X. Zhang, H. Chen, H. Yu, et al. (2020). "Clinical and immunological features of severe and moderate coronavirus disease 2019". In: *The Journal of clinical investigation* 130.5, pp. 2620–2629.

Czado, C., T. Gneiting, and L. Held (2009). "Predictive model assessment for count data". In: *Biometrics* 65.4, pp. 1254–1261.

Dehning, J., J. Zierenberg, F. P. Spitzner, M. Wibral, J. P. Neto, M. Wilczek, and V. Priesemann (2020). "Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions". In: *Science* 369.6500, eabb9789.

Desson, Z., L. Kauer, T. Otten, J. W. Peters, and F. Paolucci (2022). "Finding the way forward: COVID-19 vaccination progress in Germany, Austria and Switzerland". In: *Health Policy and Technology* 11.2, p. 100584.

Diao, B., C. Wang, Y. Tan, X. Chen, Y. Liu, L. Ning, L. Chen, M. Li, Y. Liu, G. Wang, et al. (2020). "Reduction and functional exhaustion of T cells in patients with coronavirus disease 2019 (COVID-19)". In: *Frontiers in immunology* 11, p. 827.

Dunn, P. K. and G. K. Smyth (2018). *Generalized linear models with examples in R*. Vol. 53. Springer.

Felsenstein, S., J. A. Herbert, P. S. McNamara, and C. M. Hedrich (2020). "COVID-19: Immunology and treatment options". In: *Clinical immunology* 215, p. 108448.

Feng, C., H. Wang, N. Lu, and X. M. Tu (2013). "Log transformation: application and interpretation in biomedical research". In: *Statistics in medicine* 32.2, pp. 230–239.

Fischer, A. H., K. A. Jacobson, J. Rose, and R. Zeller (2008). "Hematoxylin and eosin staining of tissue and cell sections". In: *Cold spring harbor protocols* 2008.5, pdb–prot4986.

Fisher, C., J. Kistler, and J. Davis (1980). "Relation of cerebral vasospasm to subarachnoid hemorrhage visualized by computerized tomographic scanning". In: *Neurosurgery* 6.1, pp. 1–9.

Flaxman, S., S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, C. Whittaker, H. Zhu, T. Berah, J. W. Eaton, et al. (2020). "Estimating the

effects of non-pharmaceutical interventions on COVID-19 in Europe". In: *Nature* 584.7820, pp. 257–261.

Fokianos, K. and D. Tjøstheim (2011). "Log-linear Poisson autoregression". In: *Journal of Multivariate Analysis* 102.3, pp. 563–578.

Fong, Y., J. Fortner, R. L. Sun, M. F. Brennan, and L. H. Blumgart (1999). "Clinical score for predicting recurrence after hepatic resection for metastatic colorectal cancer: analysis of 1001 consecutive cases". In: *Annals of surgery* 230.3, p. 309.

Geessink, O. G., A. Baidoshvili, J. M. Klaase, B. E. Bejnordi, G. J. Litjens, G. W. van Pelt, W. E. Mesker, I. D. Nagtegaal, F. Ciompi, and J. A. van der Laak (2019). "Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer". In: *Cellular Oncology* 42.3, pp. 331–341.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. MIT Press.

Greene, F. L., C. C. Compton, A. G. Fritz, J. P. Shah, D. P. Winchester, et al. (2006). *AJCC cancer staging atlas*. Vol. 293. Springer.

Grosser, B., M.-I. Glückstein, C. Dhillon, S. Schiele, S. Dintner, A. VanSchoiack, D. Kroeppler, B. Martin, A. Probst, D. Vlasenko, et al. (2022). "Stroma AReactive Invasion Front Areas (SARIFA)–a new prognostic biomarker in gastric cancer related to tumor-promoting adipocytes". In: *The Journal of Pathology* 256.1, pp. 71–82.

Günther, F., A. Bender, K. Katz, H. Küchenhoff, and M. Höhle (2021). "Nowcasting the COVID-19 pandemic in Bavaria". In: *Biometrical Journal* 63.3, pp. 490–502.

Harrell Jr, F. E., K. L. Lee, and D. B. Mark (1996). "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors". In: *Statistics in medicine* 15.4, pp. 361–387.

Hasan, D., M. Vermeulen, E. Wijdicks, A. Hijdra, and J. van Gijn (1989). "Management problems in acute hydrocephalus after subarachnoid hemorrhage." In: *Stroke* 20.6, pp. 747–753.

Hirschbühl, K., S. Dintner, M. Beer, C. Wylezich, J. Schlegel, C. Delbridge, L. Borcherding, J. Lippert, S. Schiele, G. Müller, et al. (2021). "Viral mapping in COVID-19 deceased in the Augsburg autopsy series of the first wave: a multiorgan and multimethodological approach". In: *PLoS One* 16.7, e0254872.

Hoffman, M. D., A. Gelman, et al. (2014). "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." In: *J. Mach. Learn. Res.* 15.1, pp. 1593–1623.

Hu, B., H. Guo, P. Zhou, and Z.-L. Shi (2021). "Characteristics of SARS-CoV-2 and COVID-19". In: *Nature Reviews Microbiology* 19.3, pp. 141–154.

Huang, W., J. Berube, M. McNamara, S. Saksena, M. Hartman, T. Arshad, S. J. Bornheimer, and M. O'Gorman (2020). "Lymphocyte subset counts in COVID-19 patients: a meta-analysis". In: *Cytometry Part A* 97.8, pp. 772–776.

Huijbers, A., R. Tollenaar, G. v Pelt, E. Zeestraten, S. Dutton, C. McConkey, E. Domingo, V. Smit, R. Midgley, B. Warren, et al. (2013). "The proportion of tumor-stroma as a strong prognosticator for stage II and III colon cancer patients: validation in the VICTOR trial". In: *Annals of Oncology* 24.1, pp. 179–185.

Hunt, W. E. and R. M. Hess (1968). "Surgical risk as related to time of intervention in the repair of intracranial aneurysms". In: *Journal of neurosurgery* 28.1, pp. 14–20.

Jering, M., M. Mayer, R. Thölken, S. Schiele, A. Maccagno, and J. Zenk (2021). "Diagnostic accuracy and post-procedural complications associated with ultrasound-guided core needle biopsy in the preoperative evaluation of Parotid tumors". In: *Head and Neck Pathology* 16.3, pp. 651–656.

Jering, M., M. Mayer, R. Thölken, S. Schiele, G. Müller, and J. Zenk (2022). "Cancer-specific and overall survival of patients with primary and metastatic malignancies of the parotid gland-A retrospective study". In: *Journal of Cranio-Maxillofacial Surgery* 50.5.

Jiang, D., J. Liao, H. Duan, Q. Wu, G. Owen, C. Shu, L. Chen, Y. He, Z. Wu, D. He, et al. (2020). "A machine learning-based prognostic predictor for stage III colon cancer". In: *Scientific reports* 10.1, pp. 1–9.

Kalbfleisch, J. D. and R. L. Prentice (2011). *The statistical analysis of failure time data*. John Wiley & Sons.

Kather, J. N., J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber, et al. (2019). "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study". In: *PLoS medicine* 16.1.

Keene, O. N. (1995). "The log transformation is special". In: *Statistics in medicine* 14.8, pp. 811–819.

Kleinbaum, D. G. and M. Klein (2012). *Survival analysis: a self-learning text*. Vol. 3. Springer.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*, pp. 1097–1105.

Kverneland, A. H., M. Streitz, E. Geissler, J. Hutchinson, K. Vogt, D. Boës, N. Niemann, A. E. Pedersen, S. Schlickeiser, and B. Sawitzki (2016). "Age and gender leucocytes variances and references values generated using the standardized ONE-Study protocol". In: *Cytometry part A* 89.6, pp. 543–564.

Li, Y., L. Ge, Y. Zhou, X. Cao, and J. Zheng (2021). "Toward the impact of non-pharmaceutical interventions and vaccination on the covid-19 pandemic with time-dependent seir model". In: *Frontiers in Artificial Intelligence* 4, p. 648579.

Liboschik, T., K. Fokianos, and R. Fried (2017). "tscount: An R package for analysis of count time series following generalized linear models". In: *Journal of Statistical Software* 82, pp. 1–51.

Löhr, P., S. Schiele, T. T. Arndt, S. Grützner, R. Claus, C. Römmele, G. Müller, C. Schmid, K. M. Dennehy, and A. Rank (2021). "Impact of age and gender on lymphocyte subset counts in patients with COVID-19". In: *Cytometry Part A*.

Loukadakis, M., J. Cano, and M. O'Boyle (2018). "Accelerating deep neural networks on low power heterogeneous architectures". In: *Eleventh International Workshop on Programmability and Architectures for Heterogeneous Multicores (MULTIPROG-2018)*.

Lugli, A., R. Kirsch, Y. Ajioka, F. Bosman, G. Cathomas, H. Dawson, H. El Zimaity, J.-F. Fléjou, T. P. Hansen, A. Hartmann, et al. (2017). "Recommendations for reporting tumor budding in colorectal cancer based on the International Tumor Budding Consensus Conference (ITBCC) 2016". In: *Modern pathology* 30.9, pp. 1299–1311.

Malik, H. Z., K. R. Prasad, K. J. Halazun, A. Aldoori, A. Al-Mukhtar, D. Gomez, J. P. A. Lodge, and G. J. Toogood (2007). "Preoperative prognostic score for predicting survival after hepatic resection for colorectal liver metastases". In: *Annals of surgery* 246.5, pp. 806–814.

Martin, B., B. M. Banner, E.-M. Schäfer, P. Mayr, M. Anthuber, G. Schenkirsch, and B. Märkl (2020). "Tumor proportion in colon cancer: results from a semiautomatic image analysis approach". In: *Virchows Archiv* 477, pp. 185–193.

Martin, B., B. Grosser, L. Kempkens, S. Miller, S. Bauer, C. Dhillon, B. M. Banner, E.-M. Brendel, É. Sipos, D. Vlasenko, et al. (2021). "Stroma AReactive Invasion Front Areas (SARIFA)—a new easily to determine biomarker in Colon cancer—results of a retrospective study". In: *Cancers* 13.19, p. 4880.

Mesker, W. E., J. Junggeburt, K. Szuhai, P. de Heer, H. Morreau, H. J. Tanke, and R. A. Tollenaar (2007). "The carcinoma–stromal ratio of colon carcinoma is an independent factor for survival compared to lymph node status and tumor stage". In: *Analytical Cellular Pathology* 29.5, pp. 387–398.

Murphy, K. and C. Weaver (2016). *Janeway's immunobiology*. Garland science.

Neal, R. M. et al. (2011). "MCMC using Hamiltonian dynamics". In: *Handbook of markov chain monte carlo* 2.11, p. 2.

Pacal, I., D. Karaboga, A. Basturk, B. Akay, and U. Nalbantoglu (2020). "A comprehensive review of deep learning in colon cancer". In: *Computers in Biology and Medicine*, p. 104003.

Poonia, R. C., A. K. J. Saudagar, A. Altameem, M. Alkhathami, M. B. Khan, and M. H. A. Hasanat (2022). "An Enhanced SEIR Model for Prediction of COVID-19 with Vaccination Effect". In: *Life* 12.5, p. 647.

Rank, A., P. Löhr, R. Hoffmann, A. Ebigbo, S. Grützner, C. Schmid, and R. Claus (2021). "Sustained cellular immunity in adults recovered from mild COVID-19". In: *Cytometry Part A* 99.5, pp. 429–434.

Rankin, J. (1957). "Cerebral vascular accidents in patients over the age of 60: II. Prognosis". In: *Scottish medical journal* 2.5, pp. 200–215.

Rasband, W. S. et al. (1997). *ImageJ*.

Ritchie, H., E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, D. Beltekian, and M. Roser (2020). "Coronavirus Pandemic (COVID-19)". In: *Our World in Data*. https://ourworldindata.org/coronavirus.

Salzberger, B., F. Buder, B. Lampl, B. Ehrenstein, F. Hitzenbichler, T. Holzmann, B. Schmidt, and F. Hanses (2021). "Epidemiology of SARS-CoV-2". In: *Infection* 49.2, pp. 233–239.

Schaller, T., K. Hirschbühl, K. Burkhardt, G. Braun, M. Trepel, B. Märkl, and R. Claus (2020). "Postmortem examination of patients with COVID-19". In: *Jama* 323.24, pp. 2518–2520.

Schiele, S., T. T. Arndt, B. Martin, S. Miller, S. Bauer, B. M. Banner, E.-M. Brendel, G. Schenkirsch, M. Anthuber, R. Huss, et al. (2021). "Deep learning prediction of metastasis in locally advanced colon cancer using binary histologic tumor images". In: *Cancers* 13.9, p. 2074.

Schreckenbach, T., P. Malkomes, W. O. Bechstein, G. Woeste, A. A. Schnitzbauer, and F. Ulrich (2015). "The clinical relevance of the Fong and the Nordlinger scores in the era of effective neoadjuvant chemotherapy for colorectal liver metastasis". In: *Surgery today* 45.12, pp. 1527–1534.

Schrempf, M. C., J. Petzold, M. A. Petersen, T. T. Arndt, S. Schiele, H. Vachon, D. Vlasenko, S. Wolf, M. Anthuber, G. Müller, et al. (2022). "A randomised pilot trial of virtual reality-based relaxation for enhancement of perioperative well-being, mood and quality of life". In: *Scientific Reports* 12.1, pp. 1–12.

Shao, J. (2003). *Mathematical statistics*. Springer Science & Business Media.

Shorten, C. and T. M. Khoshgoftaar (2019). "A survey on image data augmentation for deep learning". In: *Journal of Big Data* 6.1, pp. 1–48.

Shumway, R. H., D. S. Stoffer, and D. S. Stoffer (2000). *Time series analysis and its applications*. Vol. 3. Springer.

Simonyan, K. and A. Zisserman (2014). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556*.

Skrede, O.-J., S. De Raedt, A. Kleppe, T. S. Hveem, K. Liestøl, J. Maddison, H. A. Askautrud, M. Pradhan, J. A. Nesheim, F. Albregtsen, et al. (2020). "Deep learning for prediction of colorectal cancer outcome: a discovery and validation study". In: *The Lancet* 395.10221, pp. 350–360.

Spendel, M. (2008). "Die aneurysmatische Subarachnoidalblutung: Epidemiologie, Ätiologie, Klinik und Komplikationen". In: *Journal für Neurologie, Neurochirurgie und Psychiatrie* 9.2, pp. 20–30.

Stemmer, B. F. L. (2019). "Klinische Kriterien zur Abschätzung einer dauerhaften Shuntpflicht nach einer aneurysmatischen Subarachnoidalblutung". PhD thesis. Martin-Luther-Universität Halle-Wittenberg.

Szegedy, C., S. Ioffe, V. Vanhoucke, and A. A. Alemi (2017). "Inception-v4, inception-resnet and the impact of residual connections on learning". In: *Thirty-first AAAI conference on artificial intelligence*.

Teasdale, G. and B. Jennett (1974). "Assessment of coma and impaired consciousness: a practical scale". In: *The Lancet* 304.7872, pp. 81–84.

Thölken, R., M. Jering, M. Mayer, S. Schiele, G. Müller, and J. Zenk (2021). "Prospective study on complications using different techniques for parotidectomy for benign tumors". In: *Laryngoscope Investigative Otolaryngology* 6.6, pp. 1367–1375.

Tomlinson, J. S., W. R. Jarnagin, R. P. DeMatteo, Y. Fong, P. Kornprat, M. Gonen, N. Kemeny, M. F. Brennan, L. H. Blumgart, and M. D'Angelica (2007). "Actual 10-year survival after resection of colorectal liver metastases defines cure". In: *Journal of Clinical Oncology* 25.29, pp. 4575–4580.

Velázquez, G. F., S. Schiele, M. Gerken, S. Neumaier, C. Hackl, P. Mayr, M. Klinkhammer-Schalke, G. Illerhaus, H. Schlitt, M. Anthuber, et al. (2022). "Predictive preoperative clinical score for patients with liver-only oligometastatic colorectal cancer". In: *ESMO open* 7.3, p. 100470.

Vittinghoff, E., D. V. Glidden, S. C. Shiboski, and C. E. McCulloch (2006). *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*. Springer.

Waidhauser, J., P. Nerlinger, T. T. Arndt, S. Schiele, F. Sommer, S. Wolf, P. Löhr, S. Eser, G. Müller, R. Claus, et al. (2021). "Alterations of circulating lymphocyte subsets in patients with colorectal carcinoma". In: *Cancer Immunology, Immunotherapy* 71.8, pp. 1937–1947.

Wang, F., R. Kaushal, and D. Khullar (2020). "Should health care demand interpretable artificial intelligence or accept "black box" medicine?" In: *Annals of internal medicine* 172.1, pp. 59–60.

Weichselbaum, R. R. and S. Hellman (2011). "Oligometastases revisited". In: *Nature reviews Clinical oncology* 8.6, pp. 378–382.

Weis, C.-A., J. N. Kather, S. Melchers, H. Al-Ahmdi, M. J. Pollheimer, C. Langner, and T. Gaiser (2018). “Automatic evaluation of tumor budding in immunohistochemically stained colorectal carcinomas and correlation to clinical outcome”. In: *Diagnostic pathology* 13.1, pp. 1–12.

Wilson, P. W., R. B. D’Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel (1998). “Prediction of coronary heart disease using risk factor categories”. In: *Circulation* 97.18, pp. 1837–1847.

Xiao, X., E. P. White, M. B. Hooten, and S. L. Durham (2011). “On the use of log-transformation vs. nonlinear regression for analyzing biological power laws”. In: *Ecology* 92.10, pp. 1887–1894.

Xu, J. and Y. Tang (2021). “Bayesian framework for multi-wave COVID-19 epidemic analysis using empirical vaccination data”. In: *Mathematics* 10.1, p. 21.

Zdilla, M. J., S. A. Hatfield, K. A. McLean, L. M. Cyrus, J. M. Laslo, and H. W. Lambert (2016). “Circularity, solidity, axes of a best fit ellipse, aspect ratio, and roundness of the foramen ovale: a morphometric analysis with neurosurgical considerations”. In: *The Journal of craniofacial surgery* 27.1, p. 222.

Zhao, K., Z. Li, S. Yao, Y. Wang, X. Wu, Z. Xu, L. Wu, Y. Huang, C. Liang, and Z. Liu (2020). “Artificial intelligence quantified tumour-stroma ratio is an independent predictor for overall survival in resectable colorectal cancer”. In: *EBioMedicine* 61, p. 103054.

Zhong, Z., L. Zheng, G. Kang, S. Li, and Y. Yang (2020). “Random erasing data augmentation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07, pp. 13001–13008.

Zhu, N., D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, et al. (2020). “A novel coronavirus from patients with pneumonia in China, 2019”. In: *New England journal of medicine* 382.8, pp. 727–733.

# Appendices

# Appendix A

# Software Code

## A.1 Software code used for chapter 3 - training of CNN for images of tumor sections

For the implementation of our CNN we used keras which is a software package for programming in python. The CNN consists of a pretrained VGG19-net model to extract features from the images. These features are then provided to fully connected layers and a final softmax layer to predict the probabilities of survival/occurrence of metastasis. To avoid overfitting, the Image Generator implemented in keras was used to generated augmented images for training of the model.

```
import tensorflow
from keras.preprocessing.image import ImageDataGenerator
from keras.models import Sequential
from keras.layers import Conv2D, MaxPooling2D
from keras.layers import Activation, Dropout, Flatten
from keras.layers import Dense, Softmax
from keras import backend as K
from tensorflow import keras

model = keras.applications.vgg19.VGG19(include_top=False,
            weights='imagenet',
            input_tensor=None,
            input_shape=(860,648,3),
            pooling=None,
            classes=2)

model.trainable = False
for layer in model.layers:
```

```
    layer.trainable = False

gl_av_layer = tensorflow.keras.layers.GlobalAveragePooling2D()
pred_layer_1 = keras.layers.Dense(16,activation = 'relu')
pred_layer_2 = keras.layers.Dense(16,activation = 'relu')
pred_layer_3 = keras.layers.Dense(2, activation = 'softmax')

model_full = tensorflow.keras.Sequential([
            model,
            gl_av_layer,
            pred_layer_1,
            pred_layer_2,
            pred_layer_3
            ])

model_full.compile(loss='sparse_categorical_crossentropy',
    optimizer=tensorflow.keras.optimizers.RMSprop(lr=0.001)
    metrics=['accuracy'])

train_datagen = ImageDataGenerator(rotation_range=50,
                                   width_shift_range=0.3,
                                   height_shift_range=0.3,
                                   shear_range=0.01,
                                   horizontal_flip=True,
                                   vertical_flip=False,
                                   fill_mode='reflect',
                                   data_format='channels_last')

train_generator = train6_datagen.flow_from_directory(
            'directory_to_images_for_training',
            target_size=(860,648),
            batch_size=16,
            class_mode='binary')

val_datagen = ImageDataGenerator()

validation_generator = val_datagen.flow_from_directory(
            'directory_to_images_for_training',
            target_size=(860,648),
```

```
        batch_size=14,
        class_mode='binary')

checkpointer = keras.callbacks.ModelCheckpoint(
        filepath='filepath to save checkpoint',
        verbose=1, save_best_only=True)

model_full.fit_generator(
        train_generator,steps_per_epoch=10,
        validation_data=validation_generator,
        validation_steps=1,
        epochs=30,callbacks=[checkpointer])

model_full.save("file to save final model")
```

## A.2 Software code used in chapter 6 - Bayesian SEIR model with change points

The following programming code was used to define the Bayesian SEIR model with change points in RStan. The code consists of multiple blocks. The *data* block contains all variables which are provided by the dataset. All parameters and their data type are defined in the *parameter* block. Newly calculated parameters and variables such as the components of the SEIR model are defined in the section *transformed parameters*. The *model* section contains the priors and the sampling distribution of the reported number of new cases. Finally, the predicted number of cases can be obtained from the *generated quantities* section.

```
data {
    int n_days;
    int N;
    int cases[n_days];
}

parameters {
    real<lower=0, upper=1> gamma;
    real<lower=0, upper=2> beta_0;
    real beta_1;
    real beta_2;
    real beta_3;
```

```
    real beta_4;
    real beta_5;
    real beta_6;
    real beta_7;
    real beta_8;
    real beta_9;
    real beta_10;
    real beta_11;
    real beta_12;
    real beta_13;
    real<lower=0, upper=1> a;
    real<lower=0> i0; //
    real<lower=0> e0;
    real t1;
    real t2;
    real t3;
    real t4;
    real t5;
    real t6;
    real t7;
    real t8;
    real t9;
    real t10;
    real t11;
    real t12;
    real t13;
}

transformed parameters{
    real S[n_days];
    real E[n_days];
    real I[n_days];
    real incidence[n_days - 1];
    real beta[n_days-1];

    S[1] = (N - i0 - e0) ;
    I[1] = i0 ;
    E[1] = e0 ;
```

```
    for (i in 1:(n_days-1)) {
    beta[i] = beta_0*(1/(1+8*exp((i-t1)))) +
    beta_1 * ((1-1/(1+8*exp((i-t1))))-
    (1-1/(1+8*exp((i-t2)))))+
    beta_2 * ((1-1/(1+8*exp((i-t2))))-
    (1-1/(1+8*exp((i-t3)))))+
    beta_3 * ((1-1/(1+8*exp((i-t3))))-
    (1-1/(1+8*exp((i-t4)))))+
    beta_4 * ((1-1/(1+8*exp((i-t4))))-
    (1-1/(1+8*exp((i-t5)))))+
    beta_5 * ((1-1/(1+8*exp((i-t5))))-
    (1-1/(1+8*exp((i-t6)))))+
    beta_6 * ((1-1/(1+8*exp((i-t6))))-
    (1-1/(1+8*exp((i-t7)))))+
    beta_7 * ((1-1/(1+8*exp((i-t7))))-
    (1-1/(1+8*exp((i-t8)))))+
    beta_8 * ((1-1/(1+8*exp((i-t8))))-
    (1-1/(1+8*exp((i-t9)))))+
    beta_9 * ((1-1/(1+8*exp((i-t9))))-
    (1-1/(1+8*exp((i-t10)))))+
    beta_10 * ((1-1/(1+8*exp((i-t10))))-
    (1-1/(1+8*exp((i-t11)))))+
    beta_11 * ((1-1/(1+8*exp((i-t11))))-
    (1-1/(1+8*exp((i-t12)))))+
    beta_12 * ((1-1/(1+8*exp((i-t12))))-
    (1-1/(1+8*exp((i-t13)))))+
    beta_13 * (1-1/(1+8*exp((i-t13)))));
    S[i+1] = S[i] - beta[i]  * S[i]/N * I[i];
    E[i+1] = E[i] + beta[i]  * S[i]/N * I[i] - a * E[i];
    I[i+1] = I[i] + a * E[i] - gamma * I[i];
    incidence[i] = (-E[i+1] + E[i] - S[i+1] + S[i]);
    }
}

model {
  //priors
    beta_0 ~ normal(0.4, 0.1);
    beta_1 ~ normal(0.4, 0.1);
    beta_2 ~ normal(0.4, 0.1);
```

```
    beta_3 ~ normal(0.4, 0.1);
    beta_4 ~ normal(0.4, 0.1);
    beta_5 ~ normal(0.4, 0.1);
    beta_6 ~ normal(0.4, 0.1);
    beta_7 ~ normal(0.4, 0.1);
    beta_8 ~ normal(0.4, 0.1);
    beta_9 ~ normal(0.4, 0.1);
    beta_10 ~ normal(0.4, 0.1);
    beta_11 ~ normal(0.4, 0.1);
    beta_12 ~ normal(0.4, 0.1);
    beta_13 ~ normal(0.4, 0.1);
    gamma ~ normal(0.35, 0.03);
    a ~ normal(0.3, 0.05);
    i0 ~ normal(8000, 500);
    e0 ~ normal(2000, 500);
    t1 ~ normal(65,3);
    t2 ~ normal(105,3);
    t3 ~ normal(170,3);
    t4 ~ normal(210,3);
    t5 ~ normal(280,3);
    t6 ~ normal(355,3);
    t7 ~ normal(405,3);
    t8 ~ normal(450,3);
    t9 ~ normal(491,3);
    t10 ~ normal(525,3);
    t11 ~ normal(562,3);
    t12 ~ normal(588,3);
    t13 ~ normal(610,3);
    //sampling distribution
    for (i in 1:(n_days-1)) {
    cases[i] ~ student_t(4,incidence[i],
                    sqrt(incidence[i]+1));
    }
}

generated quantities {
    //real R0[n_days];
    real pred_cases[n_days-1];
    for (k in 1:(n_days-1)) {
```

```
    pred_cases[k] = student_t_rng(4,incidence[k],
                    sqrt(incidence[k]+1));
    }
}
```