Junior Management Science 7(2) (2022) 338-353

Junior Management Science

journal homepage: www.jums.academy

Survival Analysis: An Investigation of Covid-19 Patient Data

Survival Analysis: Eine Untersuchung von Covid-19 Patientendaten

Akira Karimkhani

Freie Universität Berlin

Abstract

The aim of this work is to test the feasibility of a model based on survival analysis for Covid- 19 patients. To investigate the feasibility, a Cox regression (CPH-Model) was constructed and evaluated using introduced diagnostic methods and modified using presented extensions. It is shown that disregarding the model assumptions can lead to biased estimation results. Furthermore, a sample analysis of the current literature in which CPH-Model was used revealed that the underlying model assumptions were comprehensibly tested in 40% of the articles reviewed. The novelty value of this work is based on the data analysis showing that the conventional CPH-Model is inappropriate for the Covid-19 dataset studied. In order to apply CPH-Model, the model had to be extended. It was necessary to adjust the functional form of a variable, remove outliers, include time interactions and stratify the data set. Finally, this allowed the creation of a final model that met all assumptions. However, four of the estimated coefficients appear questionable. Therefore, the adequacy of the extended model is doubtful. This implies that when CPH-Model is applied, the fulfillment of the model assumptions should be checked most carefully, and more robust estimation methods should be used in case of nonfulfillment.

Zusammenfassung

Ziel dieser Arbeit ist es die Realisierbarkeit einer Cox-Regression (CPH-Modell) für Covid-19 Patienten zu prüfen. Dafür wird das konstruierte Modell anhand von eingeführten Diagnostik-Methoden ausgewertet und mittels vorgestellter Erweiterungen modifiziert. Weiterhin wurde eine stichprobenartige Analyse der relevanten Literatur durchgeführt. Die Literaturanalyse hat aufgezeigt, dass die zugrundeliegenden Modell-Annahmen in lediglich 40% der untersuchten Artikel nachvollziehbar geprüft wurden. Der Neuigkeitswert dieser Arbeit begründet sich darin, dass gezeigt werden konnte, dass ein konventionelles CPH-Modell für den untersuchten Covid-19 Datensatz unangemessen ist. Um das CPH-Modell anwenden zu können war es notwendig die funktionale Form einer Variable anzupassen, Ausreißer zu entfernen, Zeitinteraktionsterme aufzunehmen und den Datensatz aufzuteilen. Schließlich konnte so ein finales Modell erstellt werden, welches alle Annahmen erfüllt. Allerdings erscheinen vier der geschätzten Koeffizienten fragwürdig. Daher ist die Angemessenheit des erweiterten Modells zweifelhaft. Dies impliziert, dass bei Anwendung des CPH-Modells auf Covid-19 Datensätzen die Erfüllung der Modell-Annahmen genauesten überprüft und bei Nichterfüllung robustere Schätzmethoden verwendet werden sollten.

Keywords: Covid-19; Cox-Regression; CPH-Modell; Proportional Hazards Model; Survival Analysis.

1. Einleitung

Aufgrund der derzeit herrschenden SARS-CoV- 2 (Covid-19) Pandemie kommt es global zu einer beispiellosen Belastung des Gesundheitswesens. Durch steigende Infektionszahlen und den damit einhergehenden schweren Krankheitsverläufen werden die verfügbaren Betten auf Intensivstationen zunehmend knapp. In zahlreichen Ländern der Welt übersteigt die Anzahl der versorgungsbedürftigen Patienten die vorhandenen Kapazitäten der medizinischen Einrichtungen. Damit Krankenhäuser handlungsfähig bleiben können, ist eine effiziente Allokation von verfügbaren Ressourcen unabdingbar. In Krankenhäusern stehen dabei Ärzte in der Pflicht, Patienten nach Erfolgsaussichten zu priorisieren. Um medizinisches Fachpersonal zu entlasten, könnte ein datenbasiertes,

DOI: https://doi.org/10.5282/jums/v7i2pp338-353

statistisches Modell bei der rationalen Entscheidungsfindung unterstützend wirken.

1.1. Problemumfeld

Um die Beziehung zwischen medizinischen Einflussgrößen und dem Tod eines Patienten zu untersuchen, bieten sich die statistischen Verfahren der Survival-Analyse an. Anhand einer Analyse der Überlebenszeiten von Covid-19 Patienten, können Rückschlüsse auf das Mortalitätsrisiko bestimmter Subgruppen gezogen werden. Der Kaplan-Meier Schätzer und die Cox-Regression eignen sich dabei besonders, um anhand von Patientendaten, Aussagen über die Überlebenswahrscheinlichkeiten zu treffen.

1.2. Zielsetzung

Ziel dieser wissenschaftlichen Arbeit ist es, die Realisierbarkeit eines auf der Survival-Analyse basierenden Modells für Covid-19 Patienten zu prüfen. Dabei soll anhand einer Kaplan-Meier Schätzung und Cox-Regression verschiedene Einflüsse auf die Mortalität von Patienten untersucht werden. Besonderes Gewicht soll bei der Modell-Konstruktion den zugrundeliegenden Annahmen zukommen, um eine Angemessenheit des Modells zu begutachten.

1.3. Aufbau der Arbeit

Konkret wird das Ziel der Arbeit erreicht, in dem zunächst in Kapitel 2 die theoretischen Grundlagen der Survival-Analyse erläutert werden. Unter Kapitel 2.1 werden dabei die Begriffe der Hazardrate und Survival-Funktion eingeführt. In Kapitel 2.2 wird explizit auf die Verfahren der Survival-Analyse eingegangen. Dabei wird zunächst der Kaplan-Meier Schätzer als nicht-parametrisches und die Cox-Regression als semi-parametrisches Verfahren vorgestellt. Nach der Einführung beschäftigt sich Kapitel 2.2.3 mit einigen Erweiterungen und Kapitel 2.2.4 mit Diagnostiken der Cox-Regression. Der theoretische Teil der Arbeit endet mit Kapitel 2.3, in dem der aktuelle Stand der Forschung umrissen wird. Kapitel 3 kennzeichnet den empirischen Teil der Arbeit, welcher in Kapitel 3.1 mit einem Konzept zur Datenanalyse beginnt. In Kapitel 3.2 werden die notwendigen Anpassungen des Datensatzes vor Beginn der Analyse erläutert sowie die verwendeten RStudio Pakete vorgestellt. Kapitel 3.3 bildet den Hauptgegenstand dieser Arbeit. Die zuvor eingeführten theoretischen Konzepte werden hier sukzessive angewendet. Dabei wird ein Modell anhand der im theoretischen Teil erläuterten Methoden konstruiert und evaluiert. In Kapitel 4 werden die geschätzten Koeffizienten des finalen Modells interpretiert und das Modell als Ganzes diskutiert. Abschließend wird der methodische und empirische Teil der Arbeit in Kapitel 5 zusammengefasst und weiterhin ein finaler Ausblick auf weitere Forschungsmöglichkeiten gegeben.

2. Notwendigkeit der Survival-Analyse

Die Survival-Analyse bezeichnet eine Klasse von statistischen Verfahren, welche zeitabhängige Ereignisse untersuchen. Dabei ist das Ereignis von Interesse in den meisten Anwendungen der Todeszeitpunkt eines Individuums. Die Methoden der Survival-Analyse wurden ursprünglich vor allem in klinischen Studien zur Analyse der Überlebenszeit von Patienten verwendet, jedoch finden sie auch in nicht-medizinischen Kontexten Verwendung (Sedlacek, 2018, S. 21 f.). Die Datengrundlage von Survival-Analysen (Survival-Daten) unterscheidet sich dabei von konventionellen Daten, die für statistische Verfahren genutzt werden, dahingehend, dass Survival-Daten für gewöhnlich nicht symmetrisch verteilt sind. Daher wäre es unangemessen, diesen Daten eine Normalverteilung zu unterstellen bzw. Modelle, die auf der Normverteilungs-Annahme basieren, anzuwenden (Collett, 2015, S. 1 f.). Weiterhin unterscheiden sich die Methoden der Survival-Analyse gegenüber herkömmlichen Modellierungen darin, dass die abhängige Variable häufig nur begrenzt beobachtet werden kann. So kann es sein, dass das Ereignis von Interesse bei bestimmten Individuen auch bei unendlicher Beobachtung nicht auftritt. Bei Individuen mit endlicher Beobachtungszeit kann es hingegen sein, dass das Ereignis vor oder auch nach Studienende eintritt. Dadurch kommt eine Zensur der Daten zustande. Das bedeutet, dass vollständige Informationen über die Ereigniszeiten aufgrund des Studiendesigns nicht verfügbar sind (Frees, 2009, S. 383 f.). Darüber hinaus kann es sein, dass bestimmte Individuen, deren Ereignis von Interesse nicht im beobachteten Zeitraum liegt, nicht repräsentiert werden. In diesem Fall spricht man von einer Verkürzung der Daten (Klein & Moeschberger, 2006, S. 64). Die Nicht-Berücksichtigung von Zensur und Verkürzung durch schlichtes Auslassen der entsprechenden Beobachtungen kann zu stark verzerrten Schätzern führen (Sedlacek, 2018, S. 26).

Dadurch, dass Survival-Daten Überlebenszeiten von Individuen abbilden, kommt es sowohl bei der Erhebung als auch der Verteilung zu kennzeichnenden Charakteristiken. Eine Analyse von Survival-Daten bedarf daher der Einführung von theoretischen Grundkonzepten, auf die sich die Anwendung im späteren Teil der Arbeit beziehen wird.

2.1. Grundlegende theoretische Konzepte

Es sei T die Zeitdauer bis zum Eintritt eines Ereignisses von Interesse, so ist T eine nicht-negative, stetige Zufallsvariable einer homogenen Population. Es charakterisieren drei Funktionen die Verteilung von T, nämlich die Survival-Funktion, die Hazardrate bzw. Hazardfunktion und die Wahrscheinlichkeits-Dichtefunktion (Dichtefunktion). Sobald eine der drei Funktionen bekannt ist, können die anderen eindeutig bestimmt werden. In der Praxis werden diese drei Funktionen gemeinsam mit einer weiteren, der kumulativen Hazardrate verwendet, um die verschiedenen Aspekte der Verteilung von T zu veranschaulichen (Klein & Moeschberger, 2006, S. 21 f.).

2.1.1. Survival-Funktion

Es sei T die Zeitdauer bis zum Eintritt eines Ereignisses von Interesse, so ist T eine nicht-negative, stetige Zufallsvariable mit der Dichtefunktion $f(t)$ und Verteilungsfunktion

F(t). Die Verteilungsfunktion *F(t)* ist dabei gegeben als

$$F(t) = P(T < t) = \int_0^t f(u)\mathrm{d}u \tag{1}$$

und gibt die Wahrscheinlichkeit an, dass die Überlebenszeit eines Individuums geringer als ein Wert t ist. Danach ist

$$S(t) := P(T \geq t) = 1 - F(t) \tag{2}$$

definiert als die Überlebens- bzw. Survival-Funktion. Sie gibt die Wahrscheinlichkeit an, dass die Überlebenszeit eines Individuums größer oder gleich t ist. Die Survival-Funktion kann daher verwendet werden um die Wahrscheinlichkeit anzugeben, dass ein Individuum bis zu einem Zeitpunkt, welcher nach t liegt, überlebt hat (Collett, 2015, S. 10).
Da das Ereignis von Interesse hier der Tod eines Individuums ist, gibt die Survival-Funktion an, dass ein Individuum den Zeitpunkt t „erlebt" hat. Für $S(t)$ gilt $S(0) = 1$, sowie ein monoton fallender Verlauf (Sedlacek, 2018, S. 29).

2.1.2. Hazardrate

Neben der Survival-Funktion ist die Hazardrate eine zentrale Größe in der Survival-Analyse. Die Hazardrate gibt das augenblickliche Risiko zum Zeitpunkt $T = t$ an, dass das Ereignis von Interesse eintritt, unter der Bedingung, dass es bisher nicht eingetreten ist (Sedlacek, 2018, S. 29). In anderen Worten gibt die Hazardrate die Wahrscheinlichkeit an, dass ein Individuum an Zeitpunkt t verstirbt unter der Bedingung, dass es bis zu diesem Zeitpunkt überlebt hat (Collett, 2015, S. 10). Definiert wird die Hazardrate als

$$h(t) := \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{3}$$

Aus Formel (3) lassen sich dabei nützliche Beziehungen zwischen der Survival-Funktion und der Hazardrate herleiten. Unter Berücksichtigung einer der Grunderkenntnisse der Wahrscheinlichkeitstheorie nach Bayes und Price (1763) gilt

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \tag{4}$$

wobei $P(A|B)$ die Wahrscheinlichkeit des Auftretens des Ereignisses A unter der Bedingung von B , $P(A \cap B)$ die Wahrscheinlichkeit für das gemeinsame Auftreten der Ereignisse A und B und $P(B)$ die Wahrscheinlichkeit für das alleinige Auftreten von B ist. Unter der Verwendung von (4) lässt sich der Zähler in (3) umformen zu

$$P(t \leq T < t + \Delta t | T \geq t) = \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)}. \tag{5}$$

Nenner und Zähler aus (5) können unter Verwendung von (1) und (2) ausgedrückt werden als

$$\frac{P(t \leq T < t + \Delta t)}{P(T \geq t)} = \frac{F(t + \Delta t) - F(t)}{S(t)}.$$

Für $h(t)$ gilt daher

$$h(t) = \lim_{\Delta t \to 0} \Big[\frac{F(t + \Delta t) - F(t)}{\Delta t} \Big] \frac{1}{S(t)}, \tag{6}$$

wobei der erste Term dem Differentialquotient der Verteilungsfunktion $F(t)$ entspricht. Daher gilt

$$h(t) = \frac{f(t)}{S(t)}. \tag{7}$$

Aus (7) folgt unmittelbar $h(t) = -\big[\log[S(t)]\big]'$, da

$$\begin{aligned}
-\big[\log[S(t)]\big]' &= -\frac{1}{S(t)} S'(t) \\
-\big[\log[S(t)]\big]' &= -\frac{1}{S(t)} [1 - F(t)]' \\
-\big[\log[S(t)]\big]' &= -\frac{1}{S(t)} [-f(t)] \\
-\big[\log[S(t)]\big]' &= \frac{f(t)}{S(t)}.
\end{aligned}$$

Stellt man nun $h(t) = -\big[\log[S(t)]\big]'$ nach $S(t)$ um erhält man

$$S(t) = \exp[-H(t)], \tag{8}$$

wobei $H(t) = \int_0^t h(u)\, \mathrm{d}u$ ist. Die Funktion H(t) ist definiert als die kumulative Hazardrate und kann durch die Survival-Funktion erhalten werden, da $H(t) = -\log[S(t)]$ gilt (Collett, 2015, S. 12).

2.2. Statistische Verfahren der Survival-Analyse

Als erster Schritt in der Analyse von Survival-Daten wird für gewöhnlich ein grafischer und numerischer Überblick über die Überlebenszeiten von Individuen gegeben. Typischerweise umfasst der Überblick die Schätzung der Survival-Funktion und der Hazardrate mittels der Verteilung angemessener Verfahren. Die Hazardrate und die Survival-Funktion werden dabei aus den beobachteten Überlebenszeiten geschätzt (Collett, 2015, S. 13 f.). Methoden, die zur Schätzung keine Annahmen bezüglich der Form der Verteilung machen, zählen zu den nicht-parametrischen Verfahren der Survival-Analyse. Methoden, die spezifische Annahmen treffen, zählen zu den (semi-)parametrischen Verfahren.

2.2.1. Kaplan-Meier Schätzer

Die in der Anwendung am weitesten verbreitete nicht-parametrischen Verfahren bilden dabei die Sterbetafelmethode (Lifetable Schätzer) und der von Kaplan und Meier (1958) vorgestellte Produkt-Grenzwert Schätzer (Kaplan-Meier Schätzer). Dabei verwendet man die Sterbetafelmethode bei diskreter und den Kaplan-Meier Schätzer bei stetiger Zeitmessung (Windzio, 2013, S. 90). Da im empirischen Teil dieser Arbeit die Überlebenszeit als stetige Variable erfasst wird, wird im folgenden nur auf den Kaplan-Meier Schätzer als nicht-parametrisches Verfahren eingegangen.

Der Kaplan-Meier Schätzer ist ein nicht-parametrischer Schätzer der Survival-Funktion $S(t)$. Er ist für alle Werte von t, die beobachtet werden können, definiert als

$$\hat{S}(t) = \begin{cases} 1 & \text{wenn } t < t_1, \\ \prod_{t_i \leq t}[1 - \frac{d_i}{Yi}], & \text{wenn } t_1 \leq t. \end{cases}$$

Wobei Y_i die Anzahl der noch nicht verstorbenen Individuen und d_i die Anzahl der Todesfälle zum Zeitpunkt t_i ist. Der Kaplan-Meier Schätzer ist damit eine Treppenfunktion, welche Sprünge bei den beobachteten Todesfällen aufweist (Klein & Moeschberger, 2006, S. 84).

Mit dem Kaplan-Meier Schätzer können die Survival-Funktionen verschiedene Subgruppen geschätzt und anschließend verglichen werden. Für einen deskriptiven Vergleich bietet es sich an die Kurven-Verläufe unterschiedlicher Gruppen in einer Grafik abzubilden und zu vergleichen. Für einen statistischen Vergleich durch Hypothesentests bietet sich der Log-Rank Test an (Zwiener, Blettner & Hommel, 2011, S. 166). Der Log-Rank wird dabei verwendet um die Null-Hypothese zu testen, ob zwischen Populationen die gleiche Wahrscheinlichkeit für den Eintritt eines Todesfalls zu beliebigen Zeitpunkten besteht (Bland & Altman, 2004, S. 1073). Die Ablehnung der Null-Hypothese impliziert dabei, dass sich die Hazardraten der Populationen während einem oder mehreren Zeitpunkten unterscheiden (Gad & Rousseaux, 2002, S. 384). Für weiterführende Informationen bezüglich des Log-Rank Test siehe Mantel (1966).

Der Kaplan-Meier Schätzer ist ein geeignetes Verfahren zur Schätzung der Survival-Funktion bei simplen Datensätzen, in denen lediglich die Überlebenszeiten von Individuen erfasst werden. Ein häufig auftretendes Problem bei der Analyse von Survival-Daten ist die Berücksichtigung begleitender Informationen durch Anpassung der Survival-Funktion (Klein & Moeschberger, 2006, S. 45). Um Zusammenhänge zwischen der Überlebenszeit eines Individuums und erklärenden Variablen zu untersuchen, kann ein auf der Regressions-Analyse basierender Ansatz verwendet werden (Collett, 2015, S. 53). Das wohl prominenteste Modell ist hierfür das von Cox (1972) präsentierte proportionale Hazard Modell (CPH-Modell) bzw. Cox-Regression als semi-parametrisches Verfahren.

2.2.2. Cox-Regression

Neben den Überlebenszeiten werden in medizinischen Anwendungen typischerweise weitere Variablen wie etwa das Alter, Geschlecht und Gewicht erhoben. Ähnlich wie in der linearen Regression lässt sich mit der Cox-Regression der Einfluss von erklärenden Variablen auf die abhängige Variable quantitativ ermitteln. Die Cox-Regression basiert dabei auf der Annahme von proportionalen Hazards, jedoch wird den Überlebenszeiten keine bestimmte Wahrscheinlichkeitsverteilung unterstellt. Daher zählt das Modell zu den semi-parametrischen Verfahren (Collett, 2015, S. 54).

Sei der Vektor $\mathbf{x}$ das Tupel der erklärenden Variablen x, so dass $\mathbf{x} = (x_1, x_2, ..., x_p)^T$ und sei $h_0(t)$ die Hazardfunktion für ein Individuum, dessen Wert aller Kovariablen des Vektors $\mathbf{x}$ Null sind. So heißt die unspezifizierte Funktion $h_0(t)$ die Baseline Hazardfunktion. Die Hazardfunktion für ein Individuum i kann definiert werden als

$$h_i(t) = \varphi(\mathbf{x}_i)h_0(t), \tag{9}$$

wobei $\varphi(\mathbf{x}_i)$ eine Funktion der erklärenden Variablen für das Individuum i ist. Die Funktion $\varphi(\mathbf{x}_i)$ kann dabei als Hazardrate für ein Individuum zum Zeitpunkt t mit den erklärenden Variablen $\mathbf{x}_i$, relativ zu einem Individuum bei dem $\mathbf{x} = \mathbf{0}$ gilt, aufgefasst werden, da aus (9)

$$\frac{h_i(t)}{h_0(t)} = \varphi(\mathbf{x}_i) \tag{10}$$

folgt. Die relative Hazardrate $\varphi(\mathbf{x}_i)$ kann nicht negativ werden, daher ist es zweckmäßig sie als $\exp[\eta_i]$ auszudrücken. Dabei ist η_i eine Linearkombination der p erklärenden Variablen in $\mathbf{x}_i$. So dass

$$\eta_i = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}$$

gilt, bzw. in Matrix-Notation

$$\eta_i = \beta^T \mathbf{x}_i.$$

Das proportionale Hazard Modell wird somit zu

$$h_i(t) = \exp[\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}]h_0(t) \text{ bzw.} \tag{11}$$

$$\log\Big[\frac{h_i(t)}{h_0(t)}\Big] = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} \tag{12}$$

(Collett, 2015, S. 55 ff.).

Um das in (11) erhaltene Modell auf einen Datensatz anwenden zu können, müssen die unbekannten Parameter $\beta_1, \beta_2, ..., \beta_p$ und die Baseline Hazardfunktion $h_0(t)$ bestimmt werden. Dabei können beide Komponenten des Modells separat voneinander geschätzt werden. Zunächst werden die β-Koeffizienten durch eine Maximum-Likelihood Schätzung bestimmt. In einem zweiten Schritt werden die geschätzten β-Koeffizienten dazu verwendet, eine Schätzung der Baseline Hazardfunktion zu konstruieren (Collett, 2015, S. 61 f.). Da bei den meisten statistischen Programmen die Konstruktion und Maximierung der Likelihood-Funktion unproblematisch ist, wird dies im Weiteren nicht vertieft. Für weiterführende Information zur Form der Likelihood-Funktion des proportionalen Hazard Modells siehe Cox (1972). Die Maximierung der Likelihood-Funktion erfolgt für gewöhnlich durch das Newton-Raphson-Verfahren. Für weiterführende Informationen zum Newton-Raphson-Verfahren siehe Ben-Israel (1966) und Galántai (2000).

Voraussetzungen der Cox-Regression

Um die bisher vorgestellten Verfahren der Survival-Analyse

anwenden zu können, müssen bestimmte Voraussetzungen erfüllt sein. So darf sowohl beim Kaplan-Meier Schätzer als auch bei der Cox-Regression keine informative Zensur vorliegen. Das heißt, dass wenn eine Zensur der Daten vorliegt, diese zufällig erfolgen muss. Eine nicht-informative Zensur stellt sicher, dass die Überlebenszeiten zensierter Individuen so repräsentativ wie die der nicht zensierten sind (Collett, 2015, S. 4). Weiterhin müssen die beobachteten Überlebenszeiten in beiden Modellen unabhängig voneinander sein. Neben diesen zwei Anforderungen verlangt das CPH-Modell proportionale Hazards. Die Annahme von proportionalen Hazards (PH-Annahme) sagt aus, dass das Verhältnis der Hazardraten verschiedener Beobachtungen im Zeitverlauf konstant ist (Kleinbaum & Klein, 2011, S. 123). Zur Veranschaulichung sei β ein p-dimensionaler Koeffizienten-Vektor und $\mathbf{x}_i$ sowie $\mathbf{x}_j$ p-dimensionale Kovariablen-Vektoren zweier Beobachtungen. So lässt sich das Verhältnis der Hazards (Hazardratio bzw. HR) darstellen als

$$HR = \frac{h_i(t)}{h_j(t)} = \frac{h_0(t)\exp[\beta^T \mathbf{x}_i]}{h_0(t)\exp[\beta^T \mathbf{x}_j]} = \exp[\beta^T(\mathbf{x}_i - \mathbf{x}_j)] = \theta, \tag{13}$$

wobei θ eine Konstante sei (Xue & Schifano, 2017. S. 584). Darüber hinaus verlangt das CPH-Modell, so wie in der linearen Regression, einen linearen Zusammenhang zwischen den kontinuierlichen, erklärenden Variablen und der abhängigen Variable. Formel (12) veranschaulicht diese Annahme. Des Weiteren nimmt das CPH-Modell an, dass die Werte der Kovariablen im Zeitverlauf konstant sind. Ersichtlich ist diese Annahme in Formel (9), da die Funktion $\varphi(\mathbf{x}_i)$ wohl für Individuen i variiert, jedoch nicht für Zeitpunkte t (im Gegensatz zu der Baseline Hazardfunktion $h_0(t)$).

2.2.3. Erweiterungen der Cox-Regression

Sind die im vorherigen Abschnitt genannten Voraussetzungen der Cox-Regression nicht erfüllt, so ist es nicht zwingend notwendig, das Modell zu verwerfen. Stattdessen bietet es sich an, das Grundmodell mittels Modifikationen zu erweitern. Im Folgenden werden einige Erweiterungen des CPH-Modell kurz erläutert, welche im empirischen Teil Anwendung finden.

Nicht-Linearität der kontinuierlichen Kovariablen
Liegt bei einer kontinuierlichen Kovariable ein nicht-linearer Zusammenhang vor, kann durch Transformation der funktionalen Form der Kovariable der Zusammenhang linear ausgedrückt werden. Dabei ist es wie in der linearen Regression erforderlich, dass der Koeffizient der jeweiligen Kovariable linear verbleibt (Casson & Farmer, 2014, S. 592). Für weiterführende Informationen bezüglich der Annahmen der linearen Regression und der Überprüfung dieser siehe Casson und Farmer (2014).

Die stratifizierte Cox-Regression
In manchen Fällen kann es sein, dass die PH-Annahme nicht für die gesamten Daten erfüllt ist. Es kann beispielsweise sein, dass sich die Hazardrate von Patienten, die in einer Studie teilnehmen, in der Placebos eingesetzt werden, unverhältnismäßig zu der Hazardrate von Patienten, die einen Wirkstoff erhalten, verhält. In einer solchen Situation kann es sinnvoll sein anzunehmen, dass die Baseline Hazardfunktion $h_0(t)$ zwischen Subgruppen variiert, während die erklärenden Variablen die PH-Annahme befriedigen. In so einem Fall leistet eine Aufteilung der Beobachtungen in Schichten (Strata) Abhilfe. Es sei $h_{0j}(t)$, die Baseline Hazardfunktion für Schicht (Stratum) j, mit $j = 1, 2, ..., g$ wobei g die Anzahl der verschiedenen Strata ist. So kann das CPH-Modell als $h_{ij}(t)$ für das i-te Individuum in dem j-Stratum dargestellt werden, wobei $i = 1, 2, ..., n_j$ die Anzahl der Beobachtungen in dem j-Stratum ist. So ergibt sich das geschichtete (stratifizierte) CPH-Modell als Erweiterung von Formel (11) als

$$h_{ij}(t) = \exp[\beta^T \mathbf{x}_{ij}]h_{0j}(t), \tag{14}$$

wobei $\mathbf{x}_{ij}$ ein p-Dimensionaler Vektor der erklärenden Variablen $x_1, x_2, ..., x_p$ der Individueen i des j-Stratums ist (Collett, 2015, S. 269 f.).

Cox-Regression mit zeitabhängigen Kovariablen
Wie in Abschnitt 2.2.2 erläutert, nimmt das CPH-Modell an, dass die erklärenden Variablen im Zeitverlauf konstant sind. Ist diese Annahme nicht erfüllt, kann die Cox-Regression um zeitabhängige Kovariablen (TDC) erweitert werden. Es sei $\mathbf{x}_i(t)$ der Kovariablen-Vektor zeitabhängiger erklärenden Variablen $x_i(t)$ zum Zeitpunkt t der i Individuen. So kann das Modell aus Formel (11) für zeitabhängige Kovariablen definiert werden als

$$h_i(t) = \exp[\beta^T \mathbf{x}_i(t)]h_0(t). \tag{15}$$

Dadurch, dass die $x_i(t)$ zeitabhängig sind, ist das Verhältnis der Hazards $\frac{h_i(t)}{h_0(t)}$ nicht mehr konstant (Collett, 2015, S.). Um zeitabhängige Kovariablen in einem CPH-Modell zu berücksichtigen, ist in den meisten Fällen eine Anpassung der Datengrundlage notwendig. Dies rührt daher, dass die TDC ihre Werte schrittweise über die Zeit ändern. Zu jedem Zeitpunkt, indem eine TDC ihren Wert ändert, ist daher ein Intervall erforderlich, in dem die TDC konstant bleibt. Anhand einer Zusammenfassung der Daten in Form eines Zählprozesses kann dem gerecht werden (Box-Steffensmeier, Box-Steffensmeier & Jones, 2004, S. 97 f.).
Da die technische Implementierung eines Zählprozesses unproblematisch ist, wird eine detaillierte Ausführung ausgelassen. Für weiterführende Informationen zur Formulierung von Zählprozessen für die Cox-Regression siehe Andersen und Gill (1982).

2.2.4. Modell Diagnostik

Nach der Konstruktion eines CPH-Modells ist es notwendig, die Modellgüte zu beurteilen. Zum einen darf keine der zuvor genannten Modellannahmen verletzt werden, zum anderen sollte die Anpassung des Modells stimmig sein. Für eine Beurteilung der Modellgüte gibt es unterschiedliche An-

sätze. Eine weit verbreitete Methode ist die graphische Analyse von Residuen. In dem CPH-Modell sind die Cox-Snell- (Cox & Snell, 1968), Deviance- (Therneau, Grambsch & Fleming, 1990), Martingale- (Lagakos, 1981), Schoenfeld- (Schoenfeld, 1982) und Score-Residuen (Cain & Lange, 1984) bzw. Df-Betas besonders relevant. Da im empirischen Teil der Arbeit jedes der genannten Residuen Anwendung findet, werden sie im Folgenden kurz aufgegriffen.

Cox-Snell-Residuen
Die Cox-Snell-Residuen sollten besonders hervorgehoben werden, da durch simple Transformation dieser die Deviance- und Martingale-Residuen hergeleitet werden können. Die Cox-Snell-Residuen sind, für ein Individuum i mit $i = 1, 2, ..., n$ definiert als

$$r_{Ci} = \exp[\hat{\beta}^T \mathbf{x}_i]\, \hat{H}_0(t_i) = \hat{H}_i(t_i), \tag{16}$$

wobei $\hat{H}_0(t_i)$ und $\hat{H}_i(t_i)$ die geschätzten kummulativen Hazard- und Baseline Hazardfunktionen zum Zeitpunkt t für ein Individuum i sind. Aus Formel (8) geht für $\hat{H}_i(t_i)$ hervor, dass

$$r_{Ci} = \hat{H}_i(t_i) = -\log[\hat{S}_i(t_i)] \tag{17}$$

ist, wobei $\hat{S}_i(t_i)$ die geschätzte Survival-Funktion des i-ten Individuums zum Zeitpunkt t ist (Collett, 2015, S. 150 f.).

Die Cox-Snell-Residuen eignen sich zur Überprüfung der allgemeinen Anpassung des Modells. Bei einer korrekten Anpassung des Modells folgen die Cox-Snell-Residuen approximativ einer Exponential-Verteilung im Einheitsintervall. Um zu überprüfen, ob die Cox-Snell-Residuen dieser Verteilung folgen, kann ein Residuenplot der geschätzten kumulativen Hazardfunktion, basierend auf den Cox-Snell Residuen $\hat{H}_r(r_{Ci})$ gegen die Cox-Snell-Residuen r_{Ci}, erstellt werden. Bei einer korrekten Anpassung ergibt sich im Graphen eine Gerade durch den Ursprung mit einer Steigung von eins (Box-Steffensmeier et al., 2004, S. 120). Für eine Herleitung der approximativen Verteilung der Cox-Snell-Residuen siehe Crowley und Hu (1977).

Martingale-Residuen
Durch Subtraktion der Cox-Snell-Residuen von einer Indikatorvariable δ_i, welche 0 bei Zensierung und 1 bei dem Tod eines Individuums annimmt, erhält man die Martingale-Residuen. Die Martingale-Residuen werden folglich definiert als

$$r_{Mi} = \delta_i - r_{Ci}. \tag{18}$$

Die r_{Mi} nehmen dabei Werte zwischen $(-\infty, 1]$ an und haben bei großem n einen Erwartungswert von Null. Allerdings sind sie nicht symmetrisch um Null verteilt (Collett, 2015, S. 153).

Die Martingale-Residuen sind hilfreich, um die Modell-Annahme des linearen Zusammenhangs zwischen dem logarithmierten Hazard und den kontinuierlichen Kovariablen zu untersuchen (Wollschläger, 2010, S. 369). Da die Martingale-Residuen bei großem n einen Erwartungswert von Null haben, können systematische Abweichungen von Null auf eine falsche funktionale Form hinweisen (Box-Steffensmeier et al., 2004, S. 126). Für weiterführende Informationen und einer allgemeineren Herleitung der Martingale-Residuen siehe Lagakos (1981) und Barlow und Prentice (1988).

Deviance-Residuen
Die Deviance-Residuen sind skalierte Martingale-Residuen, welche im Gegensatz zu den Martingale-Residuen approximativ symmetrisch um Null verteilt sind. Die Deviance-Residuen können dabei als eine Verallgemeinerung der kleinsten-Quadrat-Residuen aus der linearen Regression betrachtet werden. Besonders große Deviance-Residuen entsprechen dabei Beobachtungen, welche durch das Modell nicht gut angepasst sind (Collett, 2015, S. 153). Dabei sagt ein negatives Deviance-Residuum aus, dass die vorhergesagte Überlebenszeit geringer ist als die tatsächliche. Das bedeutet, dass für diese Beobachtungen die Hazardrate überschätzt wird. In anderen Worten überlebt die Beobachtung länger, als sie nach dem CPH-Modell sollte. Bei Beobachtungen mit positivem Deviance-Residuum ergibt sich die umgekehrte Logik (Box-Steffensmeier et al., 2004, S. 130 f.).
Die Herleitung der Deviance-Residuen geht über den Rahmen dieser Arbeit hinaus und wird aus diesem Grund ausgelassen. Für eine formale Definition und Herleitung der Deviance-Residuen siehe Therneau et al. (1990).

2.2.5. Score-Residuen/Df-Betas

Um ein CPH-Modell auf einflussreiche Beobachtungen zu untersuchen, können ähnlich wie in der linearen Regression Df-Betas verwendet werden (Wollschläger, 2010, S. 369). Die Df-Betas liefern für jede Beobachtung ein standardisiertes Maß, wie stark sich die geschätzten Parameter ändern, wenn die jeweilige Beobachtung aus den Daten ausgeschlossen wird (Wollschläger, 2010, S. 215). Man erhält die Df-Betas für ein CPH-Modell, indem man eine Matrix von Score-Residuen erstellt und durch Multiplikation der Varianz-Kovarianz-Matrix skaliert. Durch die Skalierung geben die Df-Betas die verursachte Änderung der einflussreichen Beobachtungen in den geschätzten Parametern als Änderung der Standardabweichung an (Box-Steffensmeier et al., 2004, S. 123 ff.). Die Herleitung der Score-Residuen geht über den Rahmen dieser Arbeit hinaus und wird daher ausgelassen. Für eine formale Definition und Herleitung der Score-Residuen siehe Cain und Lange (1984).

Schoenfeld-Residuen
Die Schoenfeld-Residuen können im Wesentlichen als die erwarteten, subtrahiert von den beobachteten Werten der Kovariablen des CPH-Modells betrachtet werden. Anhand einer grafischen Darstellung der i-ten Schoenfeld-Residuen gegen die t_i-ten Überlebenszeiten lassen sich Verletzungen der PH-Annahme aufzeigen. Dabei deuten Veränderungen der Residuen im Zeitverlauf auf eine Zeitabhängigkeit hin (Box-Steffensmeier et al., 2004, S. 121). Darüber hinaus ist

der in Abschnitt 2.2.1 angesprochene Log-Rank Test geeignet um die PH-Annahme anhand der Schoenfeld-Residuen statistisch zu untersuchen. Anhand der Korrelation zwischen Überlebenszeiten und den Schoenfeld-Residuen testet der Log-Rank Test mit einem Chi-Quadrat-Test (Chi-Sq) die Null-Hypothese, ob die PH-Annahme erfüllt ist (Wollschläger, 2010, S. 369). Daher weist ein signifikantes Ergebnis die PH-Annahme zurück, während ein nicht-signifikantes die PH-Annahme bestärkt. Die Herleitung der Schoenfeld-Residuen geht über den Rahmen dieser Arbeit hinaus und wird daher im Folgenden ausgelassen. Für eine formale Definition und Herleitung der Schoenfeld-Residuen siehe Schoenfeld (1982).

2.2.6. Stand der Forschung

Nach der Einführung von Cox (1972) wurde das CPH-Modell stetig um Modifikationen erweitert. Andersen und Gill (1982) haben die Verwendung der Zählprozess- Formulierung von Daten zur Verwendung im CPH-Modell vorgestellt, um die Aufnahme von zeitabhängigen Kovariablen zu realisieren. Die Integration von zeitabhängigen Kovariablen im CPH-Modell wurde dabei erstmals von Zucker und Karr (1990) sowie Gamerman (1991) vorgestellt. Die Aufnahme von Kovariablen, welche mit der Überlebenszeit interagieren, ergibt meist dann Sinn, wenn die PH-Annahme im regulären Modell verletzt ist und eine Zeitabhängigkeit plausibel erscheint.

Um die PH-Annahme grafisch zu überprüfen, hat Schoenfeld (1982) eine verallgemeinerte Form von Residuen vorgestellt, die zeitunabhängig sind. Ergänzend zu den grafischen Verfahren haben Grambsch und Therneau (1994) einen statistischen Test, welcher auf skalierten Schoenfeld-Residuen beruht, zur Überprüfung der PH-Annahme vorgestellt. Liegt eine Verletzung der PH-Annahme vor, ist zu klären, ob und in welcher Form Zeitabhängigkeit auftritt. Fisher und Lin (1999) diskutieren dafür das CPH-Modell mit zeitabhängigen Kovariablen. Dabei bemerken sie, dass die technische Implementierung problematisch sein kann und Potential für Verzerrungen bietet. Insbesondere unterstreichen sie die Bedeutsamkeit, bei der Aufnahme von zeitabhängigen Kovariablen die funktionale Form der Zeitabhängigkeit korrekt zu bestimmen. Fisher und Lin (1999) fassen zusammen, dass die Form der Zeitabhängigkeit nicht unbedingt offensichtlich ist, jedoch durch inhaltliches Verständnis erfasst werden kann. Box-Steffensmeier et al. (2004) teilen die Auffassung von Fisher und Lin (1999) hinsichtlich dem Verzerrungs-Potenzial in der funktionalen Form von Zeitabhängigkeiten. In diesem Zusammenhang nennen Box-Steffensmeier et al. (2004) $\log(t)$, als die in der Anwendung favorisierte funktionale Form, um den Interaktionseffekt zwischen einer Kovariable und der Überlebenszeit zu modellieren.

Keele (2010) demonstriert, dass Nicht-Linearität von metrischen Kovariablen ebenfalls zu der Verletzung der PH-Annahme führen kann. Daher wird die Korrektur der funktionalen Form von Nicht-Linearität als ein wichtiger Aspekt zur Verbesserung der Modell-Spezifikation betrachtet. Bereits Therneau et al. (1990) haben dafür ein auf den Martingale-Residuen basierendes grafisches Verfahren zur Überprüfung der funktionalen Form von kontinuierlichen Kovariablen vorgestellt. Farcomeni und Viviani (2011) weisen ergänzend darauf hin, dass Ausreißer in den Daten ebenfalls zur Verletzung der PH-Annahme führen können.

Bei der Konstruktion eines CPH-Modell sollten die angesprochenen Risiken berücksichtigt werden, um eine Verletzung der PH-Annahme zu vermeiden. Es sei anzunehmen, dass in praktischen Anwendungen, in denen das CPH-Modell Verwendung findet, zumindest die Bekräftigung der PH-Annahme ausgesprochen wird. Nach einer Untersuchung von Altman, De Stavola, Love und Stepniewska (1995) wurden jedoch bei lediglich 5% der Veröffentlichungen in medizinischen Fachzeitschriften die zugrunde liegenden Annahmen des CPH-Modells geprüft. Altman et al. (1995) schlagen daher Richtlinien zur Veröffentlichung von Survival-Analysen vor. Rulli et al. (2018) haben eine ähnliche Untersuchung angestellt. Dabei führen sie eine systematische Analyse klinischer Studien mit dem Ziel durch, die Angemessenheit der dort verwendeten Cox-Regressionen basierend auf der PH-Annahme zu beurteilen. Rulli et al. (2018) stellen darin fest, dass in lediglich 4 von 115 Artikeln, in denen die Cox-Regression verwendet wurde, die Bestätigung der PH-Annahme erwähnt wurde.

Aktuell wird die Cox-Regression in einer Vielzahl von Studien verwendet, um Forschungsfragen bezüglich der Covid-19 Pandemie nachzugehen. Stand Dezember 2020 gibt es auf der medizinischen Datenbank PubMed (*https://pubmed.ncbi.nlm.nih.gov/*) auf die Suchanfrage „cox regression proportional hazard covid 19“ 125 veröffentlichte Artikel. Zu den 10 relevantesten Artikel zählen Veröffentlichungen von Bhandari et al. (2020), Chen et al. (2020), Cheng et al. (2020), Cummings et al. (2020), Grasselli et al. (2020), Ji et al. (2020), Li et al. (2020), Rosenberg et al. (2020), Shi et al. (2020) und Zha et al. (2020). Dabei wird die Relevanz anhand der „Best Match“-Suchoption gemessen. Diese berücksichtigt die frühere Verwendung, das Publikationsdatum, die Relevanz-Bewertung und die Art des Artikels. In jedem der veröffentlichten Artikel wird ein (erweitertes) CPH-Modell verwendet, was nahelegen sollte, dass zumindest die Überprüfung der PH-Annahme eine Erwähnung findet. Bei lediglich 40% (Rosenberg et al., 2020 , Cummings et al. (2020), Cheng et al. (2020) und Grasselli et al. (2020)) werden die Modell-Annahmen jedoch nachvollziehbar im Sinne der Richtlinien von Altman et al. (1995) geprüft. Bezüglich der Annahmen des CPH-Modells heißt das, dass die Autoren zumindest eine Bekräftigung der PH-Annahme aussprechen bzw. deren Überprüfung erläutern. Tabelle 7 im Appendix fasst die zehn Studien hinsichtlich der Überprüfung der PH-Annahme und deren Nachvollziehbarkeit zusammen. Dabei sind neben den Autoren zusätzlich relevante Zitate, welche die Nachvollziehbarkeit der Untersuchung unterstützen, abgebildet. Da 60% der Veröffentlichungen aus Tabelle 7 im Appendix die PH-Annahme nicht nachvollziehbar prüfen, ist es fraglich, ob die Anwendung des CPH-Modells für Covid-19 Patienten in diesen Fällen

angemessen ist. Im Folgenden soll daher ein (erweitertes) CPH-Modell für einen beispielhaften Covid-19 Patienten-Datensatz konstruiert werden. Insbesondere sollen dabei die Modell-Annahmen im Einzelnen geprüft werden, um die Angemessenheit des Modells einschätzen zu können.

3. Survival-Analyse an Covid-19 Patientendaten

Im Folgenden werden die zuvor eingeführten Methoden der Survival-Analyse an einem beispielhaften Datensatz angewendet. Bei dem betrachteten Datensatz handelt es sich um offizielle, mit Covid-19 assoziierte Patientendaten, die von der mexikanischen Regierung erhoben worden sind. Dabei stammt der Datensatz nicht unmittelbar von der Website des mexikanischen Gesundheitsamtes (*www.gob.mx/salud/*). Stattdessen wurden die Daten von der Online-Plattform Kaggle (*www.kaggle.com*) bezogen. Dort wurde eine ins englische übersetzte Version des Datensatzes zur Verfügung gestellt, auf den sich die folgende Analyse bezieht. Der Datensatz enthält insgesamt 566.602 Beobachtungen bzw. Patienten, welche im Zeitraum vom 01.01.2020 - 29.06.2020 in mexikanische Krankenhäuser eingeliefert worden sind. Zu jedem Patient wurden 23 verschiedene Kovariablen erfasst, auf die in Kapitel 3.2 genauer eingegangen wird. Voranzustellen ist allerdings, dass für jeden Patienten ein Einlieferungs- und falls eingetreten Sterbedatum, vorliegend ist. Aus technischer Sicht eignet sich der Datensatz daher für eine Survival-Analyse.

3.1. Konzept

Im Weiteren werden zunächst einige Anpassungen des Datensatzes aus Gründen der Praktikabilität vorgenommen. Anschließend wird der in Abschnitt 2.2.1 eingeführte Kaplan-Meier Schätzer konstruiert und ausgewertet. Für eine tiefgreifendere Analyse wird ein CPH-Modell im Sinne des Abschnitt 2.2.2 erstellt. Um die Anwendbarkeit des CPH-Modells zu gewährleisten, müssen die zugrunde liegenden Annahmen der Cox-Regression (siehe Abschnitt 2.2.2) und die Angemessenheit des Modells (siehe Abschnitt 2.2.4) bestätigt werden. Daher wird nach Aufnahme der Kovariablen die allgemeine Angemessenheit des Modells mittels Cox-Snell-Residuen überprüft. Daraufhin wird die funktionale Form der kontinuierlichen Kovariablen mittels Martingale-Residuen untersucht. Einflussreiche Beobachtungen werden anhand der Df-Betas und Ausreißer mittels der Deviance-Residuen identifiziert. Schließlich wird die für das CPH-Modell entscheidende PH-Annahme anhand der Schoenfeld-Residuen überprüft. Bei Unzulänglichkeiten soll das Modell anhand der in Abschnitt 2.2.3 eingeführten Erweiterungen modifiziert werden. Abschließend werden die Ergebnisse des finalen Modells kritisch gewürdigt und interpretiert.

3.2. Vorbereitungen

Um das in Abschnitt 3.1 beschriebene Konzept durchführen zu können, müssen einige Anpassungen am Datensatz vorgenommen werden. Neben der Datenaufbereitung werden im Folgenden die verwendeten RStudio-Pakete aufgegriffen. Verweise auf den beigefügten RStudio-Code werden durch Fettdruck kenntlich gemacht.

3.2.1. R-Pakete

In dem empirischen Teil der Arbeit wird RStudio Version 1.2.5001 verwendet. In der Analyse finden dabei die folgenden Pakete Anwendung. Das Paket **survival** (Therneau, 2020) liefert den Großteil der benötigten Funktionen für die Survival-Analyse. Das Paket **ggplot2** (Wickham, 2016) bzw. ergänzend **survminer** (Kassambara & Kosinski, 2020) dient der Erstellung von Visualisierungen. Das Paket **MASS** (Venables & Ripley, 2002) wird für eine schrittweise Variablen-Selektion für das optimale Modell nach dem Akaike-Informationskriterium (AIC) benötigt. Schließlich wird das Paket **Greg** (Gordon & Seifert, 2020) benötigt, um die Überlebenszeiten von den Beobachtungen in Teilintervalle zu zerlegen, falls Zeitabhängigkeit vorliegt. Bevor die genannten RStudio-Pakete auf den Datensatz angewendet werden können, müssen die zugrunde liegenden Daten für die Survival-Analyse aufbereitet werden. Im Folgenden wird das genaue Vorgehen als unterstützende Erläuterung des Codes beschrieben.

3.2.2. Datenaufbereitung

In dem Datensatz **covid.csv** sind ursprünglich 566.602 Beobachtungen enthalten. Um zu gewährleisten, dass in der Analyse ausschließlich Covid-19 Erkrankte berücksichtigt werden, wird der Datensatz auf Beobachtungen mit positivem Covid-19 Test reduziert (**covid_res** $= 1$). Daraufhin verkürzt sich der Datensatz auf 220.657 Beobachtungen. Um für die nachfolgende Analyse Überlebenszeiten zu konstruieren werden die Variablen Einlieferungs- (**entry_date**) und Sterbedatum (**date_died**) als Datumstyp formatiert. Im Datensatz wäre alternativ zu dem Einlieferungsdatum ein Symptom-Eintrittsdatum vorhanden. In vielen Fällen entspricht dieses jedoch dem Einlieferungsdatum, was darauf hindeutet, dass für jede Beobachtung, für die kein Symptom-Eintrittsdatum festgestellt werden konnte, schlicht das Einlieferungsdatum gewählt worden ist. Aus diesem Grund fällt die Entscheidung bei der Konstruktion der Überlebenszeiten auf das Einlieferungsdatum. Um die Überlebenszeiten zu konstruieren, wird die Variable **dfentrydeath** erstellt, welche die Differenz zwischen dem Einlieferungs- und dem Sterbedatum in Tagen abbildet. Da nicht für alle Beobachtungen ein Tod im Erhebungszeitraum festgestellt worden ist, enthält die Variable **dfentrydeath** für alle zensierten Einträge den Wert „NA“. Da Berechnungen mit dem Wert „NA“ nicht möglich sind, werden die Variablen **time** und **death** erstellt. Die Variable **time** nimmt bei zensierten Überlebenszeiten den Wert 0 und für die beobachteten Todeszeitpunkte den Wert von **dfentrydeath** an. Die Variable **death** stellt eine Indikatorvariable dar, welche den Wert 0 bei Zensur und 1 für einen Todesfall annimmt. Anschließend werden 60 Beobachtungen, welche negative Überlebenszeiten aufweisen (**time** < 0) aus dem Datensatz entfernt, da diese unplausibel

erscheinen und auf Fehler in der Erhebung hindeuten. Weiterhin werden 2032 Beobachtungen, welche bei Einlieferung einem sofortigen Tod erliegen (Überlebenszeit $= 0$) aus der Analyse ausgeschlossen, da für sie $S(0) = 1$ nach Formel (1) nicht gilt. Anschließend werden 454 Beobachtungen mit **age** kleiner 1 und größer 100 entfernt, um unplausible Altersangaben auszuschließen. Weiterhin werden die Ausprägungen von den nominal-skalierten Variablen aufgrund geringen Aussagegehalts von *„1 = Ja , 2 = Nein, 97 = Nicht Anwendbar, 98 = Ignoriert, 99 = Unbestimmt“* zu *„1 = Ja und 0 = Nein (inklusive 97, 98 und 99)“* zusammengefasst. Abschließend werden die nominal-skalierten Variablen faktorisiert und die nicht benötigten Variablen vom Datensatz entfernt. Zu den entfernten Variablen zählen **outpatient**, **other_disease** und **other_contact_covid**. Außerdem wurden die zuvor erwähnten, zur Konstruktion der Überlebenszeiten verwendeten Variablen entfernt. Die Variable **outpatient** gibt dabei an, ob es sich um ambulante oder stationär behandelte Patienten handelt. Die Variable **other_disease** gibt an, ob bei dem Patienten weitere Krankheiten vorliegen, welche nicht durch die restlichen Variablen erfasst worden sind. Die Variable **other_contact_covid** gibt an, ob ein Kontakt zwischen dem Patienten und einer Covid-19 infizierten Person bestand. Da sich die Survival-Analyse auf spezifische, medizinische Einflussfaktoren beschränken soll, sind die genannten Variablen für die weitere Untersuchung irrelevant. Für die anschließende Analyse enthält der Datensatz schließlich noch 218.111 Beobachtungen mit den in Tabelle 1 aufgezählten 17 Variablen.

3.3. Survival-Analyse von Covid-19 Patientendaten

Die angepassten Daten erfüllen nach der Aufbereitung in Abschnitt 3.2.2 aus technischer Sicht die Voraussetzung für die anschließende Survival-Analyse.

3.3.1. Kaplan-Meier Schätzung

Um eine Kaplan-Meier Schätzung zu konstruieren, wird zunächst eine **Surv()**-Funktion erstellt, welche die Indikatorvariable **death** und die Überlebenszeit-Variable **time** in einem Survival-Objekt zusammenfasst. Nach der Erstellung der **Surv()**-Funktion wird mittels der Funktion **survfit()** die Survival-Funktion per Kaplan-Meier Methode geschätzt. Beide Funktionen sind in dem **survival**-Paket enthalten. Tabelle 2 sind die geschätzten Survival-Funktion $\hat{S}(t)$ für $t = 0 - 5, 10, 15, 20$ und 25 Tagen, sowie die jeweiligen 0.95-Konfidenzintervalle (0.95-KI) zu entnehmen. Die Notation in der Kopfzeile erfolgt dabei nach Abschnitt 2.2.1. Neben dem numerischen Überblick in Tabelle 2 wird in Abbildung 1 im Appendix eine Visualisierung von $\hat{S}(t)$ für die Überlebenszeit t dargestellt. Aus Abbildung 1 im Appendix wird ersichtlich, dass die mediane Überlebenszeit der Beobachtungen bei 6 Tagen liegt. Die mediane Überlebenszeit sagt hier aus, dass 50% der Covid-19 Patienten nach 6 Tagen Krankenhausaufenthalt verstorben sind. Die Kaplan-Meier Schätzung gibt einen ersten deskriptiven Einblick über das Mortalitätsrisiko von Covid-19 Patienten. Um die in Tabelle 1 erhobenen Variablen als Einflussfaktoren auf die Überlebenszeiten zu betrachten, wird im Folgenden eine Cox-Regression konstruiert.

3.3.2. Konstruktion der Cox-Regression

Die Konstruktion des CPH-Modells erfolgt in der Praxis nach einer ähnlichen Vorgehensweise, wie in der konventionellen Regressions-Analyse üblich. Die erklärenden Variablen werden auf die abhängige Variable regressiert, wobei die abhängige Variable hier das zuvor erwähnte Survival-Objekt ist. Anhand einer beidseitigen Variablenselektion nach dem AIC findet dabei die Modellauswahl für ein Step-Modell (**CoxStep**) statt. In Tabelle 1 im Appendix sind die Kovariablen des **CoxStep**-Modells, deren Koeffizienten im Exponenten der Exponentialfunktion, das 0.95-KI der Koeffizienten und der probabilitas-Wert (p-Wert) dieser zu entnehmen. Aus Tabelle 1 im Appendix geht hervor, dass alle Koeffizienten abgesehen von **copd** auf einem Alpha-Niveau von 5% einen signifikanten Einfluss auf die Hazardraten haben. Des Weiteren ist das Gesamt-Modell nach Likelihood-Quotienten- (LQ-Test), Wald- (W-Test) und Log-Rank-Test (LR-Test) signifikant und weist einen AIC von 456261.5 auf. Allerdings sind die Koeffizienten an dieser Stelle noch nicht interpretierbar, da zunächst die Modellannahmen geprüft werden müssen.

3.3.3. Überprüfung der Annahmen der Cox-Regression

Um valide Schätzergebnisse des CPH-Modells zu gewährleisten, müssen die Modellannahmen und eine allgemeine Angemessenheit der Cox-Regression bestätigt werden. Im Folgenden werden die in Abschnitt 2.2.4 eingeführten Residuen zur Modell-Diagnostik herangezogen.

Allgemeine Anpassung des Modells

Um in einem ersten Schritt die allgemeine Anpassung des Modells zu untersuchen können die Cox-Snell-Residuen verwendet werden. In RStudio existiert derzeit keine simple Funktion zur Ausgabe der Cox-Snell-Residuen. Diese lassen sich allerdings unproblematisch durch Umstellung der Martingale-Residuen in Ausdruck (18) herleiten. Dafür werden zunächst die Martingale-Residuen für jede Beobachtung mit der Funktion **residuals(coxstep, type = “martingale“)** berechnet und anschließend von der Indikatorvariable **death** subtrahiert. Nach Erstellung der Cox-Snell-Residuen wird eine Survival-Funktion vom Kaplan-Meier Typ geschätzt (siehe Abschnitt 3.3.1), welche auf den zuvor erstellten Cox-Snell-Residuen basiert. Anschließend wird der negative Logarithmus der Survival-Funktion berechnet, um wie nach Formel (8) die kumulative Hazardrate zu erhalten. Schließlich wird die kumulative Hazardrate gegen die Cox-Snell-Residuen in einem Graph dargestellt. Dies veranschaulicht Abbildung 2 im Appendix. Im Graphen zeigt sich eine Gerade mit Ursprung in Null und einer konstanten Steigung von circa Eins. Nach Abschnitt 2.2.4 deutet dies auf eine allgemeine Angemessenheit der Modell-Anpassung hin. Da die grafische Darstellung der Cox-Snell-Residuen nur ein ungefähres Bild über die Eignung des Modells gibt, sind weitere Untersuchungen der Modellannahmen notwendig.

Tabelle 1: Übersicht der in der Analyse verwendeten Variablen (eigene Darstellung)

Variable	Bedeutung	Skalenniveau	Ausprägung
age	Alter	metrisch	Alter in Jahren
asthma	Asthma-Erkrankung	nominal	ja = 1, nein = 0
cardiovascular	Herz-Kreislauf-Erkrankung	nominal	ja = 1, nein = 0
copd	Lungenerkrankung (COPD)	nominal	ja = 1, nein = 0
death	Todesfall	nominal	ja = 1, nein = 0
diabetes	Diabetes-Erkrankung	nominal	ja = 1, nein = 0
hypertension	Bluthochdruck	nominal	ja = 1, nein = 0
icu	Aufnahme in die Intensivstation	nominal	ja = 1, nein = 0
inmsupr	Autoimmunerkrankung	nominal	ja = 1, nein = 0
intubed	künstliche Beatmung	nominal	ja = 1, nein = 0
obesity	Übergewicht	nominal	ja = 1, nein = 0
pneumonia	Lungenentzündung	nominal	ja = 1, nein = 0
pregnancy	Schwangerschaft	nominal	ja = 1, nein = 0
renal_chronic	chronisches Nierenversagen	nominal	ja = 1, nein = 0
time	Überlebenszeit	metrisch	überlebte Tage
tobacco	Tabakkonsum	nominal	ja = 1, nein = 0
sex	Geschlecht	nominal	w = 1, m = 2

Tabelle 2: Kaplan-Meier Schätzung von $\hat{S}(t)$ (eigene Darstellung)

t	Y_i	d_i	$\hat{S}(t)$	0.95-KI
0	218111	0	1	(1.0; 1.0)
1	25010	2745	0.8902	(0.8864; 0.8941)
2	22265	2525	0.7893	(0.7842; 0.7944)
3	19740	2326	0.6963	(0.6906; 0.7020)
4	17414	2082	0.6130	(0.6070; 0.6191)
5	15332	8817	0.5377	(0.5316; 0.5439)
10	7567	6942	0.2601	(0.2548; 0.2656)
15	3485	3524	0.1192	(0.1153, 0.1233)
20	1591	1641	0.0536	(0.0509; 0.0565)
25	718	725	0.0246	(0.0228; 0.0266)

Linearität der kontinuierlichen Kovariablen

Wie in Abschnitt 2.2.6, angesprochen haben Therneau et al. (1990) vorgeschlagen, die Linearität von kontinuierlichen Kovariablen anhand der Martingale-Residuen zu überprüfen. Demnach wird zunächst eine Schätzung des **Cox-Step**-Modells ohne die Kovariable **age** in einem Test-Modell (**testFitAge**) ermittelt. Anschließend werden die Martingale-Residuen von **testFitAge** berechnet. Schließlich werden die Martingale-Residuen von **testFitAge** gegen die ausgelassene Kovariable **age** dargestellt. Abbildung 3 im Appendix visualisiert dieses Vorgehen. Anhand der Grafik sind systematische Abweichungen von Null erkennbar, was einen Hinweis dafür liefert, dass für die Kovariable **age** Nicht-Linearität vorliegt. Um sich der funktionalen Form von **age** anzunähern ist die Funktion **ggcoxfunctional()** aus dem Paket **survminer** geeignet. Die Funktion **ggcoxfunctional()** erstellt dabei Graphen für verschiedene Transformationen der kontinuierlichen Kovariablen, die gegen die Martingale-Residuen dargestellt werden sollen. Denkbar sind hier Transformationen von **age**, wie etwa $\mathbf{age}^2$, $\sqrt{\mathbf{age}}$ oder log[**age**]. Da die Martingale-Residuen einen Erwartungswert von Null haben, werden Abweichungen von Null mit den vorgeschlagenen Transformationen verglichen. Abbildung 4 im Appendix veranschaulicht dieses Vorgehen. Der Grafik ist zu entnehmen, dass die Transformation von **age** in Form von $\sqrt{\mathbf{age}}$ und log[**age**] am ehesten die funktionale Form der Martingale-Residuen abbildet. Folglich werden beide Transformationen für die weitere Modell-Konstruktion berücksichtigt.

Einflussreiche Beobachtungen

Wie in Abschnitt 2.2.5 erläutert, eignen sich die Df-Betas zur Identifizierung von einflussreichen Beobachtungen. Dabei sollten im Datensatz keine zu einflussreichen Beobachtungen enthalten sein, um zu gewährleisten, dass das Modell keine verzerrten Ergebnisse liefert. Die Konstruktion der Df-Betas erfolgt ähnlich wie die der Martingale-Residuen. Anstelle der Martingale-Residuen wird in der **residuals()**-

Funktion als Typ „dfbetas“ eingetragen. Nach der Berechnung werden die Df-Betas für jede Kovariable visualisiert. In Abbildung 5 und 5 im Appendix wird ersichtlich, dass es einige Beobachtungen gibt, welche im Vergleich zu den restlichen Beobachtungen einen starken Ausschlag in den Df-Betas haben. Dabei messen die auf der Y-Achse aufgetragenen Df-Betas, nach Abschnitt 2.2.5, die Veränderung der Koeffizienten durch Auslassung der einzelnen Beobachtungen in Höhe der Standardabweichung. Da in Abbildung 5 und 6 im Appendix keine Beobachtung einen Ausschlag von annähernd größer als Eins besitzt, deutet dies daraufhin, dass keine unverhältnismäßig einflussreichen Beobachtungen im Datensatz vorliegen.

Proportionales Hazard

Zur Überprüfung der für die Cox-Regression entscheidende PH-Annahme werden die Schoenfeld-Residuen verwendet. Dabei werden, wie in Abschnitt 2.2.5 erläutert, die Schoenfeld-Residuen gegen den Zeitverlauf abgebildet, um systematische Abweichungen von Null aufzuzeigen. Systematische Abweichungen von Null im Graphen sind dabei ein Hinweis auf eine Zeitabhängigkeit der Kovariablen. In Abbildung 7 und 8 im Appendix werden die Schoenfeld-Residuen für jede Kovariable gegen die Überlebenszeiten abgebildet. Aus den Graphen geht hervor, dass für alle Kovariablen bis auf **icu** eine Zeitabhängigkeit denkbar ist. Insbesondere scheinen hohe Überlebenszeiten einen systematischen Trend auszulösen.

Neben der grafischen Überprüfung der PH-Annahme ist ergänzend ein Log-Rank Test der Schoenfeld-Residuen auf Zeitabhängigkeit sinnvoll. Hierfür ist die **cox.zph()**-Funktion als Test auf Zeitabhängigkeit geeignet, da sie einen Log-Rank Test für jede Kovariable und das gesamte Modell vornimmt. In Tabelle 3 ist das Ergebnis eines Log-Rank Test für das **StepCox**-Modell zu entnehmen. Die Interpretation des p-Wert des Log-Rank Test auf Zeitabhängigkeit folgt dabei der in Abschnitt 2.2.5 erläuterten Logik. Von **asthma**, **copd** und **icu** abgesehen kann für alle Kovariablen eine Verletzung der PH-Annahme nicht verworfen werden. Des Weiteren wird ein Log-Rank Test für die transformierten Kovariablen $\sqrt{\textbf{age}}$ und log[**age**] durchgeführt, um die Erfüllung der PH-Annahme zu überprüfen. Dafür wird ein Test-Modell (**FitAgeTransform**) erstellt, welches lediglich die Kovariablen $\sqrt{\textbf{age}}$ und log[**age**] enthält. Aus Tabelle 2 im Appendix wird ersichtlich, dass die PH-Annahme auf einem Alpha-Niveau von 5% für log[**age**] erfüllt und für $\sqrt{\textbf{age}}$ verletzt ist. Die Modell-Konstruktion wird daher mit log[**age**] fortgeführt.

Ausreißer

Ausreißer im Datensatz werden mit den Deviance-Residuen identifiziert. Dabei wird erneut die **residuals()**-Funktion verwendet, jedoch wird als Typ „deviance“ eingetragen. Wie in Abschnitt 2.2.4 erläutert entsprechen Deviance-Residuen mit hohen absoluten Werten Beobachtungen, welche nicht gut durch das Modell angepasst sind. In Abbildung 9 im Appendix sind die Deviance-Residuen des **CoxStep**-Modells gegen die Überlebenszeiten abgebildet. Aus der Grafik geht hervor, dass das Modell Beobachtungen mit besonders hohen Überlebenszeiten schlecht anpasst. Dieses Ergebnis ist insoweit nachvollziehbar, da es intuitiv erscheint, dass einige wenige Patienten mit sehr hohen oder niedrigen Überlebenszeiten die Schätzergebnisse verzerren. Aus Abbildung 9 geht weiterhin hervor, dass der Großteil der Beobachtungen im Intervall [-2;2] liegt und mit zunehmender Überlebenszeit die Anpassung des Modells nachlässt. Daher wird im Datensatz eine neue Variable **Ausreißer** erstellt. Die Variable **Ausreißer** nimmt dabei für jede Beobachtung 1 an, wenn das jeweilige Deviance-Residuum Werte kleiner -2 und 0, falls es Werte größer -2 annimmt. Anschließend werden alle Beobachtungen mit **Ausreißer** $= 1$ aus dem Datensatz entfernt. Insgesamt werden dadurch 285 Ausreißer entfernt, was 0,13% aller Beobachtungen oder 1,14% der Beobachtungen mit Todesfall ausmacht.

Nachdem die Ausreißer entfernt wurden, wird ein neues Step-Modell (**CoxOptimal**) mit Transformation der kontinuierlichen Variable zu log[**age**] konstruiert. In Tabelle 3 im Appendix wird ersichtlich, dass für das **CoxOptimal**-Modell die gleichen Variablen für das Modell ausgewählt werden wie in der vorherigen Variablenselektion. Außerdem weist das **CoxOptimal**-Modell gegenüber dem **CoxStep**-Modell einen niedrigeren AIC (450342.7) und ausschließlich signifikante Koeffizienten bei einem Alpha-Niveau von 5% auf. Nach der Konstruktion erfolgt ein erneuter Log-Rank Test um die PH-Annahme des **CoxOptimal**-Modells zu überprüfen. Nach Tabelle 4 erfüllt das **CoxOptimal**-Modell global die PH-Annahme nicht, allerdings verletzen nur noch die Kovariablen **icu**, **pneumonia** und **renal_chronic** die PH-Annahme individuell. Da die Aufnahme in eine Intensivstation und die Entwicklung einer Lungenentzündung bzw. chronischen Nierenversagens im Krankheitsverlauf eines Covid-19 Patienten auftreten können, ist bei den genannten Kovariablen eine Zeitabhängigkeit denkbar. Folglich sollte das **CoxOptimal**-Modell um die in Abschnitt 2.2.3 angesprochenen Modifikationen erweitert werden.

3.3.4. Erweiterung des Cox-Regression

Im vorherigen Abschnitt wurde dargelegt, dass trotz der Entfernung von Ausreißern und einer Korrektur der funktionalen Form von **age** die PH-Annahme für die Kovariablen **icu**, **pneumonia** und **renal_chronic** verletzt ist. Da bei den drei genannten Kovariablen eine Zeitabhängigkeit denkbar ist, wird eine Interaktion der Kovariablen mit den Überlebenszeiten im Sinne von Abschnitt 2.2.3 implementiert. Dafür müssen im Datensatz zunächst die Beobachtungsintervalle in Teilintervalle aufgeteilt werden. Die Aufteilung der Beobachtungsintervalle erfolgt dabei mit der **timeSplitter()**-Funktion des **Greg**-Pakets. Im Wesentlichen wird durch die **timeSplitter()**-Funktion die Variable **time** in 0,5-Intervalle unterteilt, wobei die neuen Variablen **Start_time** und **Stop_time** die Intervall-Grenzen erfassen. Hervorzuheben ist hierbei, dass obwohl ein neues Survival-Objekt (**Surv_intervall**) erstellt wird, sich das darauf basierende Intervall-Modell nicht von dem vorherigen **CoxOptimal**-Modell unterscheidet. Nach der Aufteilung der Beobachtungsintervalle sind die technischen

Tabelle 3: Log-Rank Test der Kovariablen des **CoxStep**-Modells auf PH-Annahme (eigene Darstellung)

Kovariable	Chi-Sq	Freiheitsgrade	p-Wert
age	10.907	1	0.00096
asthma	3.647	1	0.05617
copd	3.369	1	0.06645
diabetes	8.981	1	0.00273
icu	0.155	1	0.69351
intubed	5.919	1	0.01498
obesity	4.933	1	0.02634
pneumonia	7.231	1	0.00716
renal_chronic	18.459	1	< 0.0001
GLOBAL	55.666	9	< 0.0001

Tabelle 4: Log-Rank Test der Kovariablen des **CoxOptimal**-Modells auf PH-Annahme (eigene Darstellung)

Kovariable	Chi-Sq	Freiheitsgrade	p-Wert
$\log[age]$	0.73473	1	0.39136
asthma	2.45537	1	0.11712
copd	0.36076	1	0.54808
diabetes	0.00217	1	0.96283
icu	14.10517	1	**0.00017**
intubed	0.34892	1	0.55472
obesity	0.20561	1	0.65023
pneumonia	3.41762	1	**0.06450**
renal_chronic	4.95804	1	**0.02597**
GLOBAL	32.75795	9	0.00015

Voraussetzungen für eine Zeitinteraktion der Kovariablen gegeben. Zunächst wird das zuvor konstruierte Modell um eine Zeitinteraktion mit **Start_time** bei den Kovariablen **icu**, **pneumonia** und **renal_chronic** erweitert (**CoxTDC**-Modell). Anschließend wird ein Log-Rank Test durchgeführt. Tabelle 4 im Appendix ist der Log-Rank Test des **CoxTDC**-Modell zu entnehmen. Aus diesem geht hervor, dass die PH-Annahme für die Kovariablen **intubed** und **renal_chronic** sowie das gesamte Modell verletzt ist. Wie in Abschnitt 2.2.6 erwähnt ist der Logarithmus in der Literatur eine präferierte Transformation für die Zeitinteraktion. Daher wird die Variable **logtime** erstellt, welche den Wert $\log[\textbf{Start\_Time} + \textbf{1}]$ für jede Beobachtung annimmt. Anschließend wird ein CPH-Modell, welches eine Interaktion der Kovariablen mit **logtime** besitzt, konstruiert (**CoxTDC_logtime**). Im Anschluss wird ein ein Log-Rank Test des **CoxTDC_logtime**-Modells durchgeführt. Aus Tabelle 5 im Appendix geht hervor, dass das **CoxTDC_logtime**-Modell die PH-Annahme global verletzt, allerdings ist **renal_chronic** die einzige Kovariable, bei der eine individuelle Verletzung vorliegt.

Da im **CoxTDC_logtime**-Modell bei der Kovariable **renal_chronic** durch die Aufnahme einer Zeitinteraktion keine Befriedigung der PH-Annahme zustande kommt, ist es denkbar, dass eine Stratifizierung im Sinne von Abschnitt 2.2.3 Abhilfe verschaffen kann. Durch die Aufnahme des Ausdrucks **strata(renal_chronic)** in der Formel des **CoxTDC_logtime**-Modells wird das Modell in zwei Schichten (**renal_chronic** $= 1$ und **renal_chronic** $= 0$) aufgeteilt. Dabei sollte hervorgehoben werden, dass durch die Stratifizierung für die Kovariable **renal_chronic** kein β-Koeffizient geschätzt werden kann, wodurch ein gewisser Informationsverlust entsteht. Tabelle 6 im Appendix veranschaulicht den Log-Rank Test des stratifizierten CPH-Modells mit zeitabhängigen Kovariablen (**CoxTDC_Strat**), aus dem hervorgeht, dass das **CoxTDC_Strat**-Modell für alle Kovariablen die PH-Annahme erfüllt.

4. Ergebnisse

Da das **CoxTDC_Strat**-Modell schließlich alle Modellannahmen erfüllt, ist eine Interpretation der Koeffizienten möglich. Die Koeffizienten des finalen Modells sind Tabelle 5 zu entnehmen.

4.1. Interpretation der Koeffizienten

Laut Tabelle 5 leisten alle im Modell enthaltenen Koeffizienten nach einem Wald-Test auf einem Alpha-Niveau von 5% einen signifikanten Einfluss auf die Hazardratios von Covid-19 Patienten. Die folgende Interpretation des Einflusses der einzelnen Kovariablen erfolgt dabei anhand der $\exp[\hat{\beta}]$, da diese nach Formel (14) die Hazardratios ausdrücken.

Tabelle 5: Übersicht der $\hat{\beta}$-Koeffizienten des **CoxTDC_Strat**-Modells (eigene Darstellung)

Kovariable	exp[**Koeffizient**]	**0.95-KI**	**p-Wert**
$\log[age]$	1.0779	[1.0268; 1.1316]	0.00249
asthma	0.8807	[0.8061; 0.9623]	0.00495
copd	1.0771	[1.0159; 1.1420]	0.01289
diabetes	1.1233	[1.0939; 1.1535]	$<$ 0.00001
icu $*logtime$	0.8849	[0.8649; 0.9055]	$<$ 0.00001
intubed	0.9019	[0.8663; 0.9391]	$<$ 0.00001
obesity	1.0886	[1.0573; 1.1208]	$<$ 0.00001
pneumonia $*logtime$	1.0559	[1.0397; 1.0724]	$<$ 0.00001
$n = 372518$	$d_i = 24725$	AIC = 438285.2	
LQ-Test: p $<$ 0.0001	W-Test:p $<$ 0.0001	LR-Test:p $<$ 0.0001	

Die Interpretation der $\hat{\beta}$-Koeffizienten lautet wie folgt.
log[**age**]: Während alle anderen Kovariablen konstant gehalten werden (c.p.), erhöht jedes zusätzliche logarithmierte Lebensjahr die Hazardrate, also das Risiko eines Patienten, zu einem gegebenen Zeitpunkt zu sterben, durchschnittlich um 7.79% (um den Faktor 1.0779).
asthma: Zu einem gegebenen Zeitpunkt ist es für einen Patienten, bei dem Asthma diagnostiziert wurde, durchschnittlich 0.8807 mal so wahrscheinlich (11,93% weniger wahrscheinlich) zu sterben, als ein Patient bei dem kein Asthma diagnostiziert wurde (c.p.).
copd: Zu einem gegebenen Zeitpunkt ist es für einen Patienten, bei dem eine chronische Lungenerkrankung diagnostiziert wurde, durchschnittlich 1.0771 mal so wahrscheinlich (7,71% wahrscheinlicher) zu sterben, als ein Patient bei dem keine chronische Lungenerkrankung diagnostiziert wurde (c.p.).
diabetes: Zu einem gegebenen Zeitpunkt ist es für einen Patienten, bei dem Diabetes diagnostiziert wurde, durchschnittlich 1.1233 mal so wahrscheinlich (12,33% wahrscheinlicher) zu sterben, als ein Patient bei dem kein Diabetes diagnostiziert wurde (c.p.).
intubed: Zu einem gegebenen Zeitpunkt ist es für einen Patienten, der künstlich beatmet werden muss, durchschnittlich 0.9019 mal so wahrscheinlich (9.81% weniger wahrscheinlich) zu sterben, als ein Patient welcher nicht künstlich beatmet werden muss (c.p.).
obesity: Zu einem gegebenen Zeitpunkt ist es für einen Patienten, bei dem Übergewicht diagnostiziert wurde, durchschnittlich 1.0886 mal so wahrscheinlich (8,86% wahrscheinlicher) zu sterben, als ein Patient bei dem kein Übergewicht diagnostiziert wurde (c.p.).

Die Ergebnisse für die Kovariablen log[**age**], **copd**, **diabetes** und **obesity** erscheinen plausibel. Es ist nachvollziehbar, dass sich bei einer Diagnose mit einer der genannten Krankheiten die Wahrscheinlichkeit zu sterben erhöht. Gleiches gilt für die Zunahme des Alters eines Patienten. Die Wirkungsrichtung des Koeffizienten von **intubed** ist dagegen diskussionswürdig. Zum einen ist anzunehmen, dass bei der künstlichen Beatmung eines Patienten ein schwerer Krankheitsverlauf vorliegt, welcher ein höheres Sterberisiko impliziert. Zum anderen ist jedoch auch anzunehmen, dass sofern ein schwerer Krankheitsverlauf vorliegt, eine künstliche Beatmung die Überlebenschancen des Patienten positiv beeinflusst. Der Koeffizient von **intubed** erscheint nach dieser Überlegung plausibel. Widersprüchlich ist hingegen der Koeffizient von **asthma**. Aus medizinischer Sicht erscheint es unplausibel, dass eine Asthma-Erkrankung die Sterbewahrscheinlichkeit vermindert, da Asthma eine Atemwegserkrankung ist. Weitere Untersuchungen oder inhaltliche Vergleiche mit anderen Studien könnten hier neue Erkenntnisse liefern.

Die Interpretation der $\hat{\beta}$-Koeffizienten von **icu** und **pneumonia** ist durch die Aufnahme der Zeitinteraktion mit **logtime** komplexer als die der zuvor beschriebenen Kovariablen. Da die $x_i(t)$ im Zeitverlauf variieren, ist die Hazardratio nach Formel (15) abhängig von den Überlebenszeiten und muss bei einer Interpretation der Koeffizienten berücksichtigt werden.

icu $*logtime$: Die Hazardrate eines Patienten, welcher auf die Intensivstation aufgenommen wird, ist durchschnittlich um den Faktor $\mathbf{0.8849} * \log[\mathbf{Start\_Time} + \mathbf{1}]$ höher/niedriger als die eines Patienten, welcher nicht in die Intensivstation aufgenommen wird. Für Überlebenszeiten von 1, 2, 3 und 4 Tagen ergeben sich Hazardratios in Höhe von 0.6134, 0.9722, 1.2267 und 1.4242. Inhaltlich sagen die Hazardratios aus, dass ein Patient, welcher seit 1,2,3 und 4 Tagen künstlich beatmet wird, durchschnittlich um den Faktor 0.6134, 0.9722, 1.2267 und 1.4242 so wahrscheinlich verstirbt wie ein Patient, welcher nicht künstlich beatmet wird (c.p.). Insgesamt scheint im Zeitverlauf das Sterberisiko eines Patienten, welcher in eine Intensivstation eingewiesen wird, gegenüber einem, der nicht eingewiesen wird, stark anzusteigen. Inhaltlich kann dieses Ergebnis wie folgt interpretiert werden. Wird ein Patient in die Intensivstation eingeliefert, ist dessen Sterbewahrscheinlichkeit gegenüber einem Patienten, welcher nicht eingeliefert wird, zunächst geringer. Mit zunehmender Zeit auf der Intensivstation steigt die Wahrscheinlichkeit zu sterben gegenüber einem Patienten in gewöhnlicher Behandlung an. Es ist anzunehmen, dass mit einer Verlegung auf die Intensivstation ein schwerwiegenderer Krankheitsverlauf einhergeht, was eine geringere Überlebenswahrscheinlichkeit impliziert. Allerdings ist auch

anzunehmen, dass auf einer Intensivstation eine bessere medizinische Versorgung zugänglich ist, was eine Verminderung der Sterbewahrscheinlichkeit in den ersten Tagen erklärt. Bei Patienten mit voranschreitender Aufenthaltszeit auf der Intensivstation ist anzunehmen, dass sich keine Verbesserung des Gesundheitszustands einstellt, was die ansteigende Hazardratio aufzeigen könnte. Der Koeffizient von **icu** $*logtime$ erscheint nach genauerer Betrachtung plausibel.

pneumonia $*logtime$: Für den $\hat{\beta}$-Koeffizienten von **pneumonia** ergibt sich eine ähnliche Systematik wie für den von **icu**. Die Hazardrate eines Patienten mit einer Lungenentzündung und Überlebenszeiten von 1,2,3 oder 4 Tagen ist durchschnittlich um den Faktor 0.7319, 1.160, 1.4638 und 1.6994 mal so hoch wie bei einem Patienten ohne Lungenentzündung (c.p.). Während der Koeffizient für eine Überlebenszeit von einem Tag unplausibel erscheint, lässt sich die Zunahme der Hazardratio im weiteren Zeitverlauf nachvollziehen. Es ist anzunehmen, dass Patienten, bei denen eine Lungenentzündung diagnostiziert wird, unter besonderer Beobachtung stehen. Bei einem medizinischem Notfall könnten diese vermutlich schneller verpflegt werden als Patienten, bei denen keine Lungenentzündung diagnostiziert wurde. Stellt sich kurzfristig keine Verbesserung des Gesundheitszustandes ein, steigt die Hazardratio. Anhand dieser Überlegung erscheint das zunächst verminderte Sterberisiko eines Patienten mit Lungenentzündung plausibel.

4.2. Diskussion des Modells

Um valide Schätzergebnisse eines CPH-Modells zu erhalten, ist es notwendig, die Erfüllung der Modell-Annahmen genauestens zu überprüfen. In Abschnitt 3.3.3 konnte gezeigt werden, dass eine Beurteilung der allgemeinen Anpassung anhand der Cox-Snell-Residuen nicht ausreicht. Dies wird daran deutlich, dass eine Visualisierung der Cox-Snell-Residuen in Abbildung 2 im Appendix zunächst eine gute Modell-Anpassung vermuten lässt. Im weiteren Verlauf der Analyse stellt sich jedoch heraus, dass mehrere Annahmen verletzt sind. Insbesondere ist, wie in Abschnitt 3.3.3 dargelegt, die Annahme eines linearen Zusammenhangs zwischen der abhängigen und der kontinuierlichen Kovariable (**age**) verletzt. Eine Transformation der Kovariable in Form von log[**age**] gewährleistet den benötigten linearen Zusammenhang. Die funktionale Form ist hier fragwürdig, da die Kovariable für Säuglinge (**age** $= 0$) nicht definiert ist. Dies ist hier unproblematisch, da der bereinigte Datensatz keine Beobachtungen mit **age** < 1 enthält. Die in Abschnitt 3.3.3 angesprochene Entfernung von Ausreißern ist ebenfalls diskussionswürdig. Im Allgemeinen ist es nicht abwegig unplausible Beobachtungen zu entfernen, wenn diese etwa auf einen Erhebungsfehler hindeuten. Allerdings geht mit der Reduzierung des Datensatzes ein gewisser Informationsverlust einher. Nachdem die funktionale Form von **age** angepasst und die identifizierten Ausreißer aus dem Datensatz entfernt wurden, konnte die PH-Annahme für die Kovariablen **icu**, **pneumonia** und **renal_chronic** nicht bestätigt werden. Das CPH-Modell wurde daher in Abschnitt 3.3.4 erweitert. Während die Aufnahme eines Zeitinteraktions-Terms für die Kovariablen **icu** und **pneumonia** Abhilfe verschaffte, konnte für **renal_chronic** auf diese Weise keine Problemlösung erzielt werden. Daher wird für **renal_chronic** stratifiziert, was zur Folge hat, dass kein $\hat{\beta}$-Koeffizient für **renal_chronic** geschätzt werden kann. Die Interpretation der $\hat{\beta}$-Koeffizienten für **icu** und **pneumonia** wird nach Aufnahme der Zeitinteraktion komplexer als die der restlichen Kovariablen des Modells. Aussagen über die Hazardratios lassen sich für **icu** und **pneumonia** nur in Abhängigkeit von spezifischen Überlebenszeiten treffen, was eine allgemeine Interpretation erschwert. Darüber hinaus ist die Wirkungsrichtung der $\hat{\beta}$-Koeffizienten von **asthma**, **intubed**, **pneumonia** und **icu** fragwürdig. Im finalen Modell **CoxTDC_Strat** erfüllen alle Kovariablen die PH-Annahme. Allerdings sollte hervorgehoben werden, dass der LR-Test die Null-Hypothese von zeitunabhängigen Kovariablen bei **asthma** und **intubed** nur auf einem Alpha-Niveau von 10% nicht verwerfen kann. Inhaltlich deutet dies an, dass das Modell formal die PH-Annahme erfüllt, jedoch nicht mit hoher Wahrscheinlichkeit. Es ist daher also fraglich, ob die Aufnahme von zeitabhängigen Kovariablen das Vorhandensein von nicht proportionalen Hazards optimal handhabt.

5. Fazit und Ausblick

Ziel dieser wissenschaftlichen Arbeit war es, die Realisierbarkeit eines auf der Survival-Analyse basierenden Modells für Covid-19 Patienten zu prüfen. Dabei wurden anhand eines erweiterten CPH-Modells verschiedene Einflüsse auf die Mortalität untersucht. Es konnte gezeigt werden, dass die Überprüfung der Linearität der kontinuierlichen Kovariablen, die Untersuchung der PH-Annahme und die Identifikation von Ausreißern von größter Relevanz für die Modell-Konstruktion sind. Werden die zugrunde liegenden Voraussetzungen beachtet, steht die Realisierbarkeit des Modells dennoch offen. In dem untersuchten Datensatz war die entscheidende Annahme von proportionalen Hazards an mehreren Stellen verletzt. Anhand einer Visualisierung der Martingale-Residuen konnte gezeigt werden, dass zwischen der abhängigen und der kontinuierlichen Kovariable (**age**) ein nicht linearer Zusammenhang besteht. Durch Transformation zu log[**age**] konnte ein linearer Zusammenhang geschaffen werden. Mittels Deviance-Residuen konnten Ausreißer im Datensatz identifiziert werden. Dabei wurden Beobachtungen mit besonders hohen Überlebenszeiten entfernt, da sie durch das Modell schlecht angepasst werden. Um die konstruierten Modelle auf die PH-Annahme zu überprüfen wurden die Schoenfeld-Residuen und der auf ihnen basierende Log-Rank-Test herangezogen. Nachdem die funktionale Form von **age** korrigiert und die Ausreißer entfernt wurden, erfüllten alle Kovariablen abgesehen von **icu**, **pneumonia** und **renal_chronic** die PH-Annahme. Durch die Aufnahme eines Interaktionstermes mit der logarithmierten Überlebenszeit konnte die PH-Annahme für die Kovariablen **icu** und **pneumonia** befriedigt werden. Durch Stratifizierung des Datensatzes anhand der Kovariable **renal_chronic** wurde schließlich ein finales Modell konstruiert, welches die

PH-Annahme für alle Kovariablen erfüllt. Die $\hat{\beta}$-Koeffizienten des finalen Modells von log[**age**], **copd**, **diabetes** und **obesity** lassen sich plausibel interpretieren. Die $\hat{\beta}$-Koeffizienten von **intubed**, **pneumonia** und **icu** sind dagegen diskussionswürdig. Der $\hat{\beta}$-Koeffizient **asthma** erscheint dagegen unplausibel. Darüber hinaus kann das finale Modell die Null-Hypothese der PH-Annahme für die Kovariablen **asthma** und **intubed** nur auf einem Alpha-Niveau von 10% nicht verwerfen. Es ist daher also fraglich, ob die zugrunde liegende Annahme von proportionalen Hazards sinnvoll und damit das Modell angemessen ist.

Zusammenfassend lässt sich sagen, dass sich die untersuchten Covid-19 Patientendaten aus technischer Sicht für eine Survival-Analyse mittels Cox-Regression eignen. Um die Cox-Regression auf die untersuchten Covid-19 Patientendaten anwenden zu können, musste das CPH-Modell jedoch erweitert werden. Die Modifikation des Modells bringt nicht nur technischen Aufwand mit sich, sondern führt auch zu einer erschwerten Interpretation der $\hat{\beta}$-Koeffizienten. Wie in der Datenanalyse gezeigt wurde, ist die Interpretation bei zeitabhängigen Kovariablen komplexer, da nur Aussagen zu spezifischen Überlebenszeiten getroffen werden können. Es lässt sich summieren, dass das CPH-Modell durch die Aufnahme von stratifizierten und zeitabhängigen Kovariablen sowie Transformationen der kontinuierlichen Kovariablen und der Entfernung von Ausreißern an die Grenzen seiner Praktikabilität kommt. Es sollte schließlich anhand von den vorgestellten Diagnostik-Methoden abgewogen werden, ob notwendige Anpassungen in einem Verhältnis mit dem einhergehenden Informationsverlust stehen oder ob für die Analyse ein besser geeignetes Modell herangezogen werden sollte.

Zukünftige Forschung könnte mit einer Untersuchung von alternativen Schätzmethoden des CPH-Modells anknüpfen. Wie aus der Datenanalyse hervorging, traten vor allem Probleme bei der Einbindung von Ausreißern und der Verletzung der PH-Annahme auf. Als alternativen Ansatz zur Handhabung von Ausreißern und Verletzungen der PH-Annahme ist eine robuste, gewichtete Schätzung des CPH-Modells denkbar. Bei einem gewichteten CPH-Modell wird die zu maximierende Likelihood-Funktion beispielsweise so gewichtet, dass früheren Todesfällen höheres und späteren Todesfällen geringeres relatives Gewicht in der Schätzung zukommt. In der Anwendung sind bei Covid-19 Patientendaten bisher vor allem konventionelle CPH-Modelle verwendet worden. Weiterführende Forschung könnten daher die Anwendung von robusten, gewichteten CPH-Modellen für Covid-19 Patientendaten untersuchen.

Literatur

Altman, D. G., De Stavola, B. L., Love, S. B. & Stepniewska, K. A. (1995). Review of survival analyses published in cancer journals. *British Journal of Cancer*, *72* (2), 511–518.

Andersen, P. K. & Gill, R. D. (1982, 12). Cox's regression model for counting processes: A large sample study. *Ann. Statist.*, *10* (4), 1100–1120.

Barlow, W. E. & Prentice, R. L. (1988, 03). Residuals for relative risk regression. *Biometrika*, *75* (1), 65-74.

Bayes, M. & Price, M. (1763). An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, a. m. f. r. s. *Philosophical Transactions (1683-1775)*, *53*, 370–418.

Ben-Israel, A. (1966). A newton-raphson method for the solution of systems of equations. *Journal of Mathematical Analysis and Applications*, *15* (2), 243 - 252.

Bhandari, S., Tak, A., Singhal, S., Shukla, J., Shaktawat, A. S., Gupta, J., ... others (2020). Patient flow dynamics in hospital systems during times of covid-19: Cox proportional hazard regression analysis. *Frontiers in public health*, *8*.

Bland, J. M. & Altman, D. G. (2004). The logrank test. *Bmj*, *328* (7447), 1073.

Box-Steffensmeier, J. M., Box-Steffensmeier, J. M. & Jones, B. S. (2004). *Event history modeling: A guide for social scientists*. Cambridge University Press.

Cain, K. C. & Lange, N. T. (1984). Approximate case influence for the proportional hazards regression model with censored data. *Biometrics*, 493–499.

Casson, R. J. & Farmer, L. D. (2014). Understanding and checking the assumptions of linear regression: a primer for medical researchers. *Clinical & experimental ophthalmology*, *42* (6), 590–596.

Chen, R., Liang, W., Jiang, M., Guan, W., Zhan, C., Wang, T., ... others (2020). Risk factors of fatal outcome in hospitalized subjects with coronavirus disease 2019 from a nationwide analysis in china. *Chest*, *158* (1), 97–105.

Cheng, Y., Luo, R., Wang, K., Zhang, M., Wang, Z., Dong, L., ... Xu, G. (2020). Kidney disease is associated with in-hospital death of patients with covid-19. *Kidney international*, *97* (5), 829–838.

Collett, D. (2015). *Modelling survival data in medical research*. CRC press.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34* (2), 187–220.

Cox, D. R. & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, *30* (2), 248–275.

Crowley, J. & Hu, M. (1977). Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, *72* (357), 27–36.

Cummings, M. J., Baldwin, M. R., Abrams, D., Jacobson, S. D., Meyer, B. J., Balough, E. M., ... others (2020). Epidemiology, clinical course, and outcomes of critically ill adults with covid-19 in new york city: a prospective cohort study. *The Lancet*, *395* (10239), 1763–1770.

Farcomeni, A. & Viviani, S. (2011). Robust estimation for the cox regression model based on trimming. *Biometrical Journal*, *53* (6), 956–973.

Fisher, L. D. & Lin, D. Y. (1999). Time-dependent covariates in the cox proportional-hazards regression model. *Annual review of public health*, *20* (1), 145–157.

Frees, E. W. (2009). *Regression modeling with actuarial and financial applications*. Cambridge University Press.

Gad, S. C. & Rousseaux, C. G. (2002). 15 - use and misuse of statistics in the design and interpretation of studies. In W. M. HASCHEK, C. G. ROUSSEAUX & M. A. WALLIG (Hrsg.), *Handbook of toxicologic pathology (second edition)* (Second Edition Aufl., S. 327 - 418). San Diego: Academic Press.

Galántai, A. (2000). The theory of newton's method. *Journal of Computational and Applied Mathematics*, *124* (1), 25 - 44. (Numerical Analysis 2000. Vol. IV: Optimization and Nonlinear Equations)

Gamerman, D. (1991). Dynamic bayesian models for survival data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *40* (1), 63–79.

Gordon, M. & Seifert, R. (2020). Greg: Regression helper functions [Software-Handbuch]. Zugriff auf https://cran.r-project.org/web/packages/Greg/index.html

Grambsch, P. M. & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, *81* (3), 515–526.

Grasselli, G., Greco, M., Zanella, A., Albano, G., Antonelli, M., Bellani, G., ... others (2020). Risk factors associated with mortality among patients with covid-19 in intensive care units in lombardy, italy. *JAMA internal medicine*, *180* (10), 1345–1355.

Ji, D., Zhang, D., Xu, J., Chen, Z., Yang, T., Zhao, P., ... others (2020). Prediction for progression risk in patients with covid-19 pneumonia: the call score. *Clinical Infectious Diseases*, *71* (6), 1393–1399.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53* (282), 457–481.

Kassambara, A. & Kosinski, M. (2020). survminer: Drawing survival curves using "ggplot2" [Software-Handbuch]. Zugriff auf https://cran.r-project.org/web/packages/survminer/

Keele, L. (2010). Proportionally difficult: Testing for nonproportional hazards in cox models. *Political Analysis*, *18* (2), 189–205.

Klein, J. P. & Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Kleinbaum, D. G. & Klein, M. (2011). The cox proportional hazards model and its characteristics. In *Survival analysis* (S. 97–159). New York, NY: Springer New York.

Lagakos, S. W. (1981). The graphical evaluation of explanatory variables in proportional hazard regression models. *Biometrika*, *68* (1), 93–98.

Li, X., Xu, S., Yu, M., Wang, K., Tao, Y., Zhou, Y., ... others (2020). Risk factors for severity and mortality in adult covid-19 inpatients in wuhan. *Journal of Allergy and Clinical Immunology*, *146* (1), 110–118.

Mantel, N. (1966). Models for complex contingency tables and polychotomous dosage response curves. *Biometrics*, *22* (1), 83–95.

Rosenberg, E. S., Dufort, E. M., Udo, T., Wilberschied, L. A., Kumar, J., Tesoriero, J., ... others (2020). Association of treatment with hydroxychloroquine or azithromycin with in-hospital mortality in patients with covid-19 in new york state. *Jama*, *323* (24), 2493–2502.

Rulli, E., Ghilotti, F., Biagioli, E., Porcu, L., Marabese, M., D'Incalci, M., ... Torri, V. (2018). Assessment of proportional hazard assumption in aggregate data: a systematic review on statistical methodology in clinical trials using time-to-event endpoint. *British journal of cancer*, *119* (12), 1456–1463.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, *69* (1), 239–241.

Sedlacek, G. (2018). *Analyse der studiendauer und des studienabbruchrisikos: unter verwendung der statistischen methoden der ereignisanalyse*. Peter Lang International Academic Publishers.

Shi, Q., Zhang, X., Jiang, F., Zhang, X., Hu, N., Bimu, C., ... others (2020). Clinical characteristics and risk factors for mortality of covid-19 patients with diabetes in wuhan, china: a two-center, retrospective study. *Diabetes care*, *43* (7), 1382–1391.

Therneau, T. M. (2020). A package for survival analysis in r [Software-Handbuch]. Zugriff auf https://CRAN.R-project.org/package=survival

Therneau, T. M., Grambsch, P. M. & Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, *77* (1), 147–160.

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth Aufl.). New York: Springer. (ISBN 0-387-95457-0)

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Zugriff auf https://ggplot2.tidyverse.org

Windzio, M. (2013). *Regressionsmodelle für zustände und ereignisse: Eine einführung* (2013. Aufl.). Wiesbaden: Springer Fachmedien.

Wollschläger, D. (2010). *Grundlagen der datenanalyse mit r*. Springer.

Xue, Y. & Schifano, E. D. (2017). Diagnostics for the cox model. *Communications for statistical Applications and Methods*, *24* (6), 583–604.

Zha, L., Li, S., Pan, L., Tefsen, B., Li, Y., French, N., ... Villanueva, E. V. (2020). Corticosteroid treatment of patients with coronavirus disease 2019 (covid-19). *Medical Journal of Australia*, *212* (9), 416–420.

Zucker, D. M. & Karr, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach. *The Annals of Statistics*, *18* (1), 329–353.

Zwiener, I., Blettner, M. & Hommel, G. (2011). Überlebenszeitanalyse. *Dtsch Arztebl International*, *108* (10), 163-169.