h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbi
FACHBEREICH INFORMATIK

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmn
FACHBEREICH MATHEMATIK
UND NATURWISSENSCHAFTEN

Hochschule Darmstadt
Fachbereich Mathematik und Naturwissenschaften
&
Fachbereich Informatik

Transfer von ML-Modellen zur Vorhersage des Transplantatüberlebens aus einer US-Kohorte auf Daten des deutschen Organtransplantationsregisters: Machbarkeit, Vorhersagegüte und Rekalibrierung

---

ABSCHLUSSARBEIT ZUR ERLANGUNG DES AKADEMISCHEN
GRADES MASTER OF SCIENCE (M. SC.)
IM STUDIENGANG DATA SCIENCE

vorgelegt von
Franziska Schmidt

Referentin:	Prof. Dr. Antje Jahn
Korreferent:	Prof. Dr. Gunter Grieser
Ausgabedatum:	02.05.2022
Abgabedatum:	28.11.2022

Selbstständigkeitserklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder noch nicht veröffentlichten Quellen entnommen sind, sind als solche kenntlich gemacht. Die Zeichnungen oder Abbildungen in dieser Arbeit sind von mir selbst erstellt worden oder mit einem entsprechenden Quellennachweis versehen. Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden.

Ort, Datum	Unterschrift

Data-Disclaimer

US-amerikanische Transplantationsdaten:
The data reported here have been supplied by the United Network for Organ Sharing as the contractor for the Organ Procurement and Transplantation Network. The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the OPTN or the U.S. Government. Based on OPTN data as of June 20, 2020.

Deutsche Transplantationsdaten:
In dieser Arbeit wurden die an das nationale Transplantationsregister übermittelten transplantationsmedizinischen Daten der Jahre 2006 bis 2016 (sogenannte Altdaten) ausgewertet. Die Daten wurden in anonymisierter Form von der Transplantationsregisterstelle zu Forschungszwecken gemäß § 15g Absatz 1 TPG zur Verfügung gestellt.

Zusammenfassung

Die Kalibrierung ist neben der Diskriminierungsfähigkeit ein wichtiges Merkmal von klinischen Prognosemodellen. Verwendet man ein bereits vorhandenes Prognosemodell für eine neue Kohorte, muss stets untersucht werden, wie gut die Kalibrierung dieses Modells für die neue Kohorte ist. Schlecht kalibrierte Risikoschätzungen können zu falschen Erwartungen bei Patienten führen und sind somit aus ethischen Gründen problematisch. In bestimmten Fällen kann es dennoch von Vorteil sein, ein bereits vorhandenes Modell zu verwenden, anstatt ein neues zu trainieren. Insbesondere wenn die Stichprobenzahl in der neuen Kohorte gering ist, kann so dem Risiko einer Überanpassung entgegengewirkt werden. Dazu muss das alte Modell zunächst rekalibriert werden, um eine zufriedenstellende Kalibrierung zu erreichen. In der Überlebenszeitanalyse gibt es für die Cox-Regression und andere semi-parametrische und parametrische Modelle gut erforschte Rekalibrierungsmethoden. Bei der Entwicklung von Prognosemodellen werden zunehmend auch speziell auf zensierte Daten angepasste Methoden des maschinellen Lernens eingesetzt. Für diese Methoden gibt es jedoch bislang keine bekannten Rekalibrierungsmethoden. In dieser Arbeit wird daher eine Methode zur Rekalibrierung des Random Survival Forests entwickelt und evaluiert. Dazu sollen Prognosemodelle zu Nierentransplantationen, welche mit US-amerikanischen Daten entwickelt wurden, auf eine deutsche Kohorte angewendet werden. Hierbei werden ein Cox-Regressionsmodell und ein Random Survival Forest betrachtet. Es wird zunächst untersucht, welche Voraussetzungen erfüllt sein müssen, um die US-Modelle auf die deutsche Kohorte anwenden zu können. Anschließend wird die Kalibrierung der US-Modelle auf den deutschen Daten untersucht und eine Rekalibrierung durchgeführt. Für den Random Survival Forest wird hierfür die in dieser Arbeit entwickelte Methode verwendet. Abschließend wird die Performance der nicht rekalibrierten, rekalibrierten und auf den deutschen Daten neu trainierten Modelle verglichen.

Schlagworte: Klinische Vorhersagemodelle, Rekalibrierung, Kalibrierung, Diskriminierung, externe Validierung, Cox Proportional Hazards Model, Random Survival Forest

Abstract

Calibration, along with the ability to discriminate, is an important aspect of clinical prediction models. If one uses an existing prediction model for a new cohort, it is always necessary to investigate how well this model is calibrated for the new cohort. Poorly calibrated risk estimates can lead to false expectations among patients and are therefore problematic for ethical reasons. In certain cases, it may however be advantageous to use an existing model rather than train a new one. Especially if the sample size in the new cohort is small, this can counteract the risk of overfitting. To do this, the old model must first be re-calibrated to achieve a satisfactory calibration. In survival analysis, well-studied re-calibration methods exist for Cox regression and other semi-parametric and parametric models. Machine learning methods specifically adapted to censored data are also increasingly used in the development of clinical prediction models. However, there are no known re-calibration methods for these methods so far. In this thesis, a method for re-calibrating the random survival forest is therefore developed and evaluated. For this purpose, prediction models for kidney transplantation, which were developed with data from the USA, are to be applied to a German cohort. A Cox regression model and a random survival forest are considered. First, the requirements for applying the US-models to the German cohort are investigated. Subsequently, the calibration of the US-models on the German data is examined and a re-calibration is performed. For the random survival forest, the method developed in this thesis is used for this purpose. Finally, the performance of the non-re-calibrated, re-calibrated and newly trained model on the German data is compared.

Keywords: clinical prediction models, re-calibration, calibration, discrimination, external validation, Cox proportional hazards model, random survival forest

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Die Überlebenszeitanalyse betrachtet den Zeitraum bis zum Eintreten eines bestimmten Ereignisses von Interesse. In der medizinischen Anwendung sind das in der Regel das Überleben eines Patienten, oder auch das Überleben eines Patienten ohne Rückfall oder Verschlechterung einer Erkrankung. Herausforderungen entstehen dadurch, dass häufig eine Unvollständigkeit der Daten insofern vorliegt, dass das Ereignis von Interesse nicht für jeden Patienten beobachtet werden konnte, sondern nur bekannt ist, bis zu welchem Zeitpunkt das Ereignis noch nicht eingetreten war. Solche Beobachtungen nennt man dann *zensierte* Beobachtungen. Um dennoch Prognosemodelle erstellen zu können, bedarf es Methoden, die mit solchen zensierten Daten umgehen können. Die Cox-Regression ist das am weitesten verbreitete Modell zur Untersuchung von Überlebenszeitdaten. Bei der Entwicklung von Prognosemodellen werden jedoch zunehmend auch speziell auf zensierte Daten angepasste Methoden des maschinellen Lernens eingesetzt (vgl. [WLR17]).
Jedoch gibt es Aspekte, welche für diese vergleichsweise neuen Methoden noch kaum untersucht wurden. Ein solcher Aspekt ist die in dieser Arbeit behandelte Rekalibrierung von Modellen. Während es für die Cox-Regression und andere semi-parametrische oder parametrische Modelle gut untersuchte Methoden zur Rekalibrierung gibt, ist die Rekalibrierung für Machine Learning Modelle im Kontext der Überlebenszeitanalyse hingegen noch weitestgehend unerforscht.

1.1 Motivation

Eine Vorhersage des Ausgangs einer Nierentransplantation ist zur Organvergabe von entscheidender Bedeutung, da hierdurch medizinische Entscheidungsfindungen unterstützt werden können, wodurch die Anzahl der Menschen, die auf eine bereits überlastete Warteliste zurückkehren müssen, verkleinert werden kann. In den vergangenen Jahren haben viele Autoren Modelle für die Prognose nach einer Nierentransplantation vorgestellt. Wenn ein solches Modell auf eine neue Kohorte angewendet werden soll, muss die Validität des Modells für die neuen Daten überprüft werden. Falls dabei eine schlechte Kalibrierung festgestellt wird, kann dieses Modell nicht ohne Weiteres auf die neue Population angewendet werden. Die Kalibirierung eines Modells ist neben seiner Diskriminierungsfähigkeit ein wichtiges Merkmal von Prognosemodellen. Es wird allgemein empfohlen, diese bei der Validierung eines Modells zu analysieren. Systematische Überprüfungen von Vorhersagemodellen haben jedoch ergeben, dass über die Kalibrierung nur selten berichtet wird (vgl. [MAR$^+$15]).
Schlecht kalibrierte Risikoschätzungen können zu falschen Erwartungen bei Patienten führen und sind somit auch aus ethischen Gründen problematisch. In vielen Fällen kann es dennoch sinnvoll sein, das vorhandene Modell auch für die neue Kohorte zu verwenden, anstatt ein neues Modell mit den Daten der neuen Kohorte zu trainieren. Dazu muss das alte Modell allerdings zunächst rekalibriert werden. Dabei wird es so angepasst, dass eine zufriedenstellende Kalibrierung erreicht werden kann. Ein Vorteil der Rekalibrierung eines bestehenden Modells gegenüber des Trainierens eines neuen Modells kann sein, dass beim Neutrainieren eines Modells potentiell aussagekräftige Informationen aus dem ursprünglichen Modell ungenutzt bleiben. Gerade wenn die Stichprobenzahl in der neuen Kohorte gering ist,

riskiert man zudem eine Überanpassung. Ferner kann die Existenz einer großen Anzahl an Modellen für denselben Endpunkt verwirrend sein und die Entscheidung erschweren, welches Modell in der Praxis anzuwenden ist. Da Machine Learning Modelle einige Vorteile gegenüber klassischen statistischen Methoden aufweisen, wären Rekalibrierungsmethoden für Machine Learning Methoden, welche auf zensierte Daten spezialisiert sind, von großem Wert. Beispielsweise liegt ein Vorteil von Machine Learning Methoden darin, dass diese komplexe Interaktionen zwischen mehreren Prädiktoren selbstständig lernen können, sodass diese nicht manuell modelliert werden müssen.

1.2 Fragestellung

Ziel dieser Arbeit ist es zu untersuchen, inwiefern Modelle, welche mit Daten von Nierentransplantationen in den USA trainiert wurden, auch auf eine deutsche Kohorte angewendet werden können. Hierzu liegen zu Transplantationen in den USA Daten des United Network for Organ Sharing (UNOS) vor, während zu Transplantationen in Deutschland Daten durch das Deutsche Transplantationsregister bereitgestellt wurden. Hierbei soll sowohl ein Cox-Regressionsmodell als Vertreter von klassischen statistischen Methoden in der Überlebenszeitanalyse, als auch ein Random Survival Forest als Vertreter neuerer Machine Learning Methoden für Überlebenszeitdaten, untersucht werden.
Für die Suche nach geeigneten Modellen soll eine systematische Recherche zu publizierten Cox-Regressionsmodellen und Random Survival Forests erfolgen. Gesucht wird dabei nach Vorhersagemodellen zum Ausgang von Nierentransplantationen nach Todspende, bei denen die UNOS-Daten verwendet wurden. Damit eine Untersuchung der Kalibrierung eines US-Modells auf der deutschen Kohorte und das Umsetzen einer Rekalibrierung realisierbar sind, müssen zudem die folgenden zwei Voraussetzungen erfüllt sein, die es zu überprüfen gilt. Zum Einen muss das veröffentlichte Prognosemodell zur Verfügung stehen, also alle Informationen vorliegen, die notwendig sind, um dieses Modell auf neue Daten anwenden zu können. Zum Anderen müssen im deutschen Datensatz die für das US-Modell benötigten Daten vorliegen. Anschließend soll die Frage beantwortet werden, wie gut diese Modelle für die deutsche Kohorte kalibriert sind und ob eine Motivation zur Rekalibrierung vorliegt.
Des Weiteren soll untersucht werden, inwiefern die Kalibrierung der US-Modelle auf die deutsche Kohorte durch Rekalibrierung verbessert werden kann. Hierzu soll für den Random Survival Forest, für den es bisher keine bekannten Methoden zu Rekalibrierung gibt, zunächst ein solches Verfahren entwickelt werden. Zu diesem Zweck wird eine 2016 von Dankowski und Ziegler (vgl. [DZ16]) vorgestellte Methode für Random Forests zur Wahrscheinlichkeitsschätzung so erweitert, dass diese auch für rechtszensierte Daten anwendbar ist.
Außerdem sollen für die deutschen Daten auch neue Modelle trainiert werden, um abschließend die Performance der nicht rekalibrierten, rekalibrierten und neu trainierten Modelle zu vergleichen.

2 Theoretische Grundlagen

2.1 Grundlagen der Überlebenszeitanalyse

In der *Überlebenszeitanalyse* wird die Zeitspanne untersucht, die vergeht, bis ein bestimmtes *Ereignis (Event)* von Interesse eintritt. Hierbei werden Einflussfaktoren auf den Zeitpunkt dieses Ereignisses analysiert. Anders als der Name vermuten lässt, muss dieses Ereignis nicht der Tod sein, sondern kann im Kontext klinischer Studien bspw. auch ein Rezidiv bei der Behandlung von Krebs oder das Versagen eines Spenderorgans sein. Die Überlebenszeitanalyse kann außerdem, unter dem Namen *Ereigniszeitanalyse*, auch in anderen Disziplinen, wie beispielsweise der Betriebswirtschaftslehre zum Einsatz kommen.
Die Definitionen und Herleitungen dieses Kapitels sind angelehnt an [KM03] sowie [KK05]. Wurden weitere Quellen verwendet, wird im Text darauf verwiesen.

2.1.1 Endpunkt bei Überlebenszeitdaten und Zensierung

Die Besonderheit von Überlebenszeitdaten liegt darin, dass in manchen Fällen das Ereignis nicht beobachtet wird. Mögliche Ursachen hierfür sind beispielsweise, dass Patienten die Studie verlassen oder das Ereignis bis zum Ende der Studie nicht eingetreten ist. Dieses Phänomen wird *Zensierung* genannt. In solchen Fällen ist zwar unbekannt, wann genau das Ereignis eingetreten ist, dennoch liegen Informationen darüber vor, dass das Ereignis bis zum Zeitpunkt der Zensierung nicht eingetreten ist. Die unvollständigen Beobachtungen werden nicht ignoriert, da sonst der Zeitpunkt des Ereignisses unterschätzt werden würde, Patienten also eine zu pessimistische Prognose erhalten würden.
Es gibt im medizinischen Anwendungsbereich drei Haupttypen der Zensierung: *Linkszensierung*, *Intervallzensierung* und *Rechtszensierung*. Ein Patient wird rechtszensiert,wenn der Nachbeobachtungszeitraum endet, bevor er das Ereignis von Interesse erlebt, wobei der genaue Zeitpunkt des Ereignisses nicht bekannt ist. Eine Intervallzensierung liegt vor, wenn das Ereignis innerhalb des Zeitintervalls eintritt, der genaue Zeitpunkt des Ereignisses innerhalb dieses Intervalls jedoch nicht bekannt ist. Die Linkszensierung liegt vor, falls der genaue Zeitpunkt des Ereignisses nicht bekannt ist, das Ereignis aber vor dem Untersuchungszeitpunkt eingetreten ist. Da in den in dieser Arbeit betrachteten Daten ausschließlich Fälle der Rechtszensierung betrachtet werden, werden im Folgenden Zensierung und Rechtszensierung synonym genutzt.
Für jeden Patienten ist entweder der Zeitpunkt des Ereignisses oder der Zeitpunkt der Zensierung bekannt. Zensierte Beobachtungen können also als Tripel (X, T, δ) aufgefasst werden, bei welchem X den Kovariatenvektor, T die beobachtete Zeit und δ den Zensierungsindikator darstellt. Letzterer gibt an, ob zum Zeitpunkt T ein Ereignis stattgefunden hat ($\delta = 1$), oder ob es sich hierbei um den Zeitpunkt der Zensierung handelt ($\delta = 0$).
Der Endpunkt bei Überlebenszeitdaten besteht also aus zwei Komponenten (T, δ), für die gilt:

$$\delta = \mathbb{1}_{T^* \leq C}, \quad T = \min(T^*, C),$$

wobei C der Zeitpunkt der Zensierung und T^* der Zeitpunkt des Ereignisses ist. Falls ein Patient zensiert wurde, ist der Ereigniszeitpunkt zwar unbekannt, aber sicher später als der

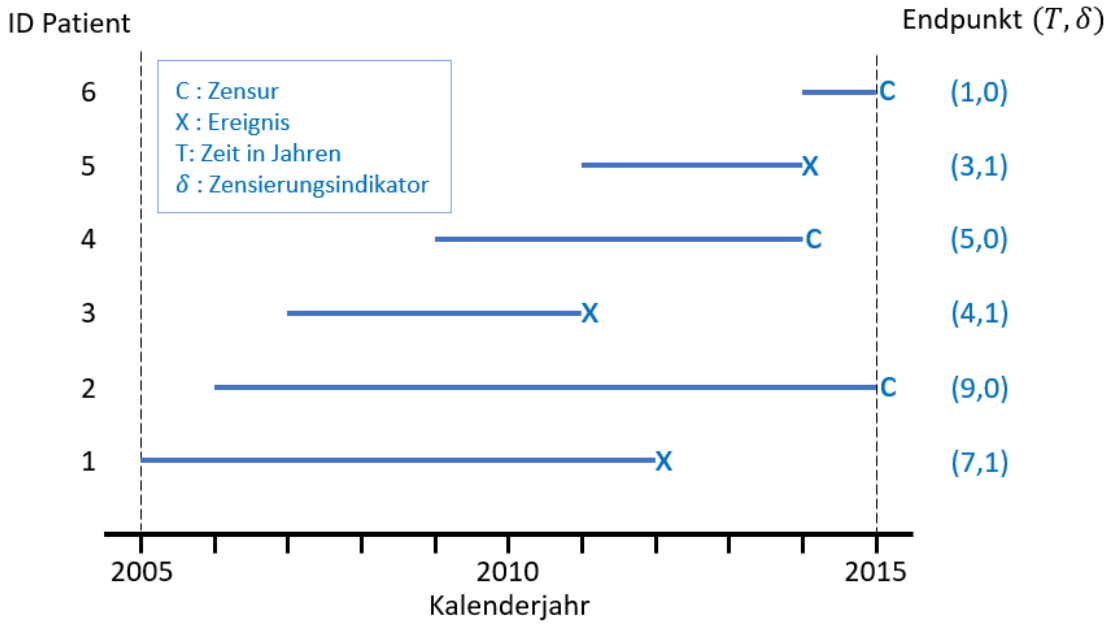

Abbildung 1: Beispiel zur Definition des Endpunktes bei Überlebenszeitdaten

Zensierungszeitpunkt. Falls das Ereignis beobachtet wurde, existiert keine Zensierung und der Zensierungszeitpunkt ist in diesem Fall als ∞ definiert.
In Abbildung 1 ist ein Beispiel zur Definition des Endpunktes gegeben. Bei einer klinischen Studie wurden Patienten zwischen 2005 und 2015 beobachtet. Dabei sind die Patienten zu unterschiedlichen Zeitpunkten in die Studien aufgenommen worden. Die Patienten mit den IDs 1, 3 und 5 haben während der Studie das Ereignis erlitten. Für sie gilt also $\delta = 1$ und der Wert T gibt die Zeit zwischen Einstieg in der Studie und Ereignis an. Bei den Patienten mit den IDs 2 und 6 ist das Ereignis bis zum Ende der Studie im Jahr 2015 nicht eingetreten und sie sind daher zensiert. Der Patient mit der ID 4 hat die Studie vor 2015 verlassen, ohne dass zuvor das Ereignis beobachtet wurde. Daher ist auch er zensiert. Bei diesen drei Beobachtungen gibt T die Zeitspanne vom Einstieg in die Studie bis zur Zensierung an. Daher gilt für diese Beobachtungen $\delta = 0$.
Um Prognosemodelle erstellen zu können, müssen klassische statistische Methoden oder Verfahren des maschinellen Lernens so erweitert werden, dass sie für diesen speziellen Datentyp anwendbar sind.

2.1.2 Wichtige Funktionen der Überlebenszeitanalyse

Sei T^* eine stetige Zufallsvariable, welche den Zeitpunkt des untersuchten Ereignisses misst und Werte im Intervall $[0; \infty)$ annehmen kann. Seien f und F die zugehörige Dichte- bzw. Verteilungsfunktion:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T^* < t + \Delta t)}{\Delta t} \tag{1}$$

$$F(t) = \int_0^t f(u)du \tag{2}$$

Durch die Verteilungsfunktion $F(t)$ wird die Wahrscheinlichkeit angegeben, dass das Ereignis bis zum Zeitpunkt t eingetreten ist. Daher wird diese Funktion auch *Sterbefunktion* genannt. Die *Überlebensfunktion* $S(t)$ kann also als Gegenstück der Verteilungsfunktion $F(t)$ angesehen werden. $S(t)$ ist definiert als die Wahrscheinlichkeit, dass das Ereignis bis zum Zeitpunkt t noch nicht eingetreten ist.

$$S(t) = P(T^* > t) = 1 - F(t) = 1 - \int_0^t f(u)du, \qquad t \in [0; \infty) \tag{3}$$

Da f eine Dichtefunktion ist, gilt $\int_0^\infty f(u)du = 1$ und die Überlebensfunktion lässt sich folglich auch darstellen als:

$$S(t) = \int_t^\infty f(u)du \tag{4}$$

Da es sich bei F um eine Verteilungsfunktion handelt, ist die Überlebensfunktion wegen $S(t) = 1 - F(t)$ eine monoton fallende Funktion, für welche $S(0) = 1$ und $\lim_{t\to\infty} S(t) = 0$ gilt. Außerdem gilt zwischen f und S folgender Zusammenhang:

$$f(t) = -\frac{\partial(S(t))}{\partial t} \tag{5}$$

Beweis:

$$\begin{aligned}
f(t) &= \lim_{\Delta t\to 0} \frac{P(t \le T^* < t + \Delta t)}{\Delta t} \\
&= \lim_{\Delta t\to 0} \frac{P(T^* < t + \Delta t) - P(T^* < t)}{\Delta t} \\
&= \lim_{\Delta t\to 0} \frac{1 - P(T^* \ge t + \Delta t) - (1 - P(T^* \ge t))}{\Delta t} \\
&= \lim_{\Delta t\to 0} \frac{1 - P(T^* > t + \Delta t) - (1 - P(T^* > t))}{\Delta t} \\
&= \lim_{\Delta t\to 0} \frac{-P(T^* > t + \Delta t) + P(T^* > t))}{\Delta t} \\
&= \lim_{\Delta t\to 0} - \frac{S(t + \Delta t) - S(t))}{\Delta t} \\
&= -\frac{\partial(S(t))}{\partial t}
\end{aligned}$$

Neben der Überlebensfunktion ist auch die *Hazardfunktion* $h(t)$ eine elementare Funktion in der Überlebenszeitanalyse. Sie gibt die Rate an, dass zum Zeitpunkt t ein Ereignis eintritt, unter

der Voraussetzung, dass bis zu diesem Zeitpunkt kein Ereignis eingetreten ist. Die Hazardfunktion wird auch Ausfallrate genannt und ist wie folgt definiert:

$$h(t) := \lim_{\Delta t \to 0} \frac{P(t \leq T^* < t + \Delta t | T^* \geq t)}{\Delta t} \tag{6}$$

Zwischen Überlebens- und Hazardfunktion gilt folgender Zusammenhang:

$$h(t) = -\frac{\partial \ln(S(t))}{\partial t} \tag{7}$$

Beweis:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T^* < t + \Delta t | T^* \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t \leq T^* < t + \Delta t)}{P(T^* \geq t) \cdot \Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(T^* < t + \Delta t) - P(T^* < t)}{P(T^* \geq t) \cdot \Delta t} \\ &= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \cdot \frac{1}{P(T^* \geq t)} \\ &= \frac{f(t)}{S(t)} = -\frac{\partial(S(t))}{\partial t} \frac{1}{S(t)} = -\frac{\partial \ln(S(t))}{\partial t} \end{aligned}$$

Die kummulative Hazardfunktion ist definiert als

$$H(t) := \int_0^t h(u) du. \tag{8}$$

Sie lässt sich als die „Ansammlung des Hazards“ im Laufe der Zeit interpretieren.
Aufgrund von 7 gilt $H(t) = -\ln(S(t))$.

Bisher wurde die Zeit T^* als stetige Zufallsvariable definiert. Betrachtet man die Zeit als diskrete Größe mit m eindeutigen Ereigniszeitpunkte t_1 bis t_m und
$t_0 = 0 < t_1 < t_2 < ... < t_m$, so ist die Wahrscheinlichkeit, bis zum Zeitpunkt t ein Ereignis zu erleiden, gegeben durch:

$$P(T^* \leq t) = \sum_{t_j \leq t} P(t_j),$$

wobei $\sum_{j=1}^{m} P(t_j) = 1$ und $P(T^* = t_0) = 0$ gilt.

Im diskreten Fall gilt dann für die Überlebensfunktion:

$$S(t) = 1 - P(T^* \leq t) = 1 - \sum_{t_j \leq t} P(t_j) = \sum_{t_j > t} P(t_j) \tag{9}$$

Die Hazardfunktion und die kummulierte Hazardfunktion sind im diskreten Fall gegeben durch

$$h(t) = P(T^* = t | T^* \geq t) = \frac{P(T^* = t)}{P(T^* \geq t)} \tag{10}$$

und

$$H(t) = \sum_{t_j \leq t} h(t_j). \tag{11}$$

Für den Zusammenhang zwischen der diskreten Überlebens- und der diskreten Hazardfunktion gilt:

$$S(t) = \prod_{t_j \leq t} (1 - h(t_j)) \tag{12}$$

Beweis:
O.B.d.A. wird die Aussage für $S(t_m)$ gezeigt.
Zunächst gilt folgender Zusammenhang:

$$\begin{aligned} h(t_i) &= \frac{P(T^* = t_j)}{P(T^* \geq t_j)} \\ &= \frac{P(T^* = t_j)}{P(T^* > t_{j-1})} \\ &= \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} \\ &= 1 - \frac{S(t_j)}{S(t_{j-1})} \end{aligned}$$

Daraus folgt:

$$\begin{aligned} \prod_{t_j \leq t_m} 1 - h(t_j) &= \prod_{t_j \leq t_m} \frac{S(t_j)}{S(t_{j-1})} \\ &= \prod_{j=1}^{m} \frac{S(t_j)}{S(t_{j-1})} \\ &= \frac{S(t_m)}{S(t_0)} \\ &\overset{S(t_0)=1}{=} S(t_m) \end{aligned}$$

2.1.3 Kaplan-Meier- und Nelson-Aalen-Schätzer

Zum nichtparametrischen Schätzen der Überlebens- und Hazardfunktion wird auf die Definitionen im diskreten Fall zurückgegriffen. Auch wenn die unbekannte aber wahre Verteilung stetig ist, kann die Zielvariable T^* in der Stichprobe nur endlich viele Werte annehmen und somit als diskret interpretiert werden. Ferner wird die Zielvariable in vielen

Fällen ohnehin als diskrete Einheit erhoben, beispielsweise wenn die Zeitspanne bis zum Ereignis in ganzen Tagen gemessen wird.

Seien hierzu weiterhin die t_j mit $j = 1, ..., m$ die sortierten Ereigniszeitpunkte. Mit n_j wird die Anzahl an Individuen bezeichnet, die unmittelbar vor dem Zeitpunkt t_j unter Risiko stehen. Das heißt, dass bei ihnen bis zu diesem Zeitpunkt weder ein Ereignis noch eine Zensierung stattgefunden hat. Zuletzt sei d_j die Anzahl an Individuen, bei denen das Ereignis genau zum Zeitpunkt t_j eintritt.
Die geschätzte Ereignisrate (Hazard) zum Zeitpunkt t_j ist dann durch den Quotienten aus d_j und n_j gegeben:

$$\hat{h}(t_j) = \frac{d_j}{n_j} \tag{13}$$

Aus 12 folgt hieraus ein Schätzer für die Überlebensfunktion:

$$\hat{S}_{KM}(t) = \prod_{j:t_j \leq t} \frac{n_j - d_j}{n_j} \tag{14}$$

Diese Methode zum nichtparametrischen Schätzen der Überlebensfunktion wurde 1958 von Kaplan und Meier [KM58] vorgestellt und wird daher Kaplan-Meier-Schätzer (oder KM-Schätzer) genannt.
Varianz und Standardabweichung können mithilfe der Greenwood-Formel geschätzt werden (vgl. [KM58] und [Gre26]).

$$\hat{Var}\left(\hat{S}(t)\right) = \left(\hat{S}(t)\right)^2 \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)} \tag{15}$$

$$\hat{SE}(\hat{S}(t)) = \sqrt{\left(\hat{S}(t)\right)^2 \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}} \tag{16}$$

Analog dazu kann durch Einsetzen der Ereignisrate in 11 der Nelson-Aalen-Schätzer (vgl. [Nel72], [Nel69] und [Aal78]) hergeleitet werden, welcher ein nichtparametrischer Schätzer für die kummulative Hazardfunktion ist:

$$\hat{H}_{0,NA}(t) = \sum_{j:t_j \leq t} \frac{d_j}{n_j} \tag{17}$$

Tabelle 1 zeigt einen Beispieldatensatz samt der für die Bestimmung des Kaplan-Meier und des Nelson-Aalen-Schätzers notwendigen Rechenschritte. Die zu dem Beispiel passende geschätzte Überlebensfunktion, die sogenannte Kaplan-Meier-Kurve, und die geschätzte kummulierte Hazardfunktion werden in Abbildung 2 gezeigt.

Tabelle 1: Beispieldaten für die Erläuterung des Kaplan-Meier-Schätzers und des Nelson-Aalen-Schätzers und die Berechnung zu den eindeutigen Ereigniszeitpunkten. Die zugehörigen Kurven sind in Abbildungen 2 zu sehen.

i	T_i	δ_i	n_i	d_i	$(n_i - d_i)/n_i$	d_i/n_i	$\hat{S}_{KM}(T_i)$	$\hat{H}_{0,NA}(T_i)$
1	2	1	10	1	$(10-1)/10 = 0.9$	$1/10$	$1 \cdot 0.9 = 0.9$	0.1
2	7	0	9	0				
3	16	1	8	2	$(8-2)/8 = 0.75$	$2/8$	$0.9 \cdot 0.75 = 0.675$	$0.1 + 0.25 = 0.35$
4	16	1	8	2				
5	23	0	6	0				
6	24	1	5	1	$(5-1)/5 = 0.8$	$1/5$	$0.8 \cdot 0.675 = 0.54$	$0.35 + 0.2 = 0.55$
7	28	1	4	1	$(4-1)/4 = 0.75$	$1/4$	$0.75 \cdot 0.54 = 0.405$	$0.55 + 0.25 = 0.8$
8	29	0	3	0				
9	35	1	2	1	$(2-1)/2 = 0.5$	$1/2$	$0.5 \cdot 0.405 = 0.2025$	$0.8 + 0.5 = 1.3$
10	35	0	2	1				

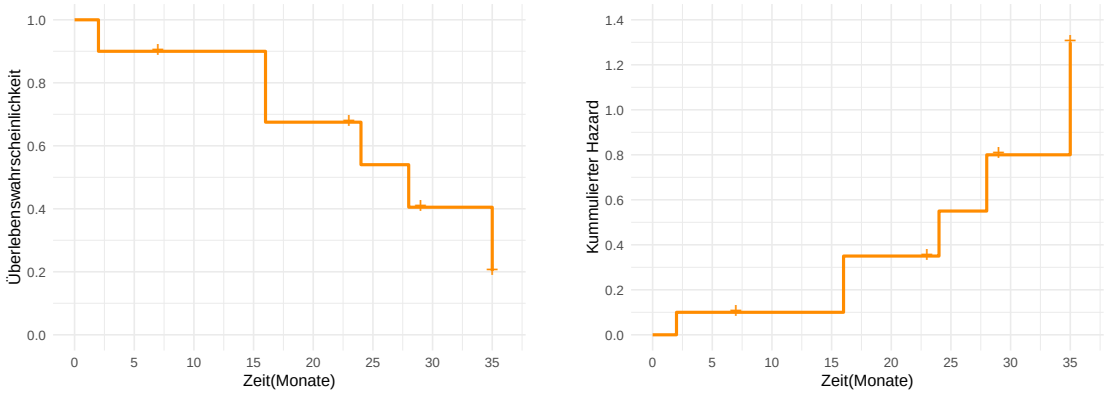

Abbildung 2: Beispiel einer einer Kaplan-Meier Kurve (links) und einer durch den Nelson-Aalen-Schätzer geschätzten kummulativen Hazardfunktion (rechts) für die Daten aus Tabelle 1

2.2 Cox-Regression

Die Cox-Regression wurde 1972 von Cox [Cox72] vorgestellt. Dieses Regressionsmodell ist ein semiparametrisches Modell, in welchem ein linearer und additiver Zusammenhang zwischen den einzelnen Kovariaten und der logarithmierten Hazardfunktion modelliert wird. An die Verteilung der Ereigniszeit werden in der Cox-Regression keine Annahmen gestellt und diese muss nicht bekannt sein. Dafür wird angenommen, dass die erklärenden Variablen zu jeder Zeit denselben Effekt auf das Modell haben. Die zu schätzenden Parameter sind also unabhängig von der Zeit.
Sei $X = (X_1, X_2, ..., X_p)$ der Kovariatenvektor. Dann ist das Modell definiert als

$$\ln\left(\frac{h(t|X)}{h_0(t)}\right) = \beta^T X = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \tag{18}$$

bzw.

$$h(t|X) = h_0(t) e^{\beta^T X}. \tag{19}$$

Dabei ist β der Parametervektor und $h_0(t)$ der sogenannte Baseline-Hazard, welcher nicht bekannt sein muss.
Die Änderung der Kovariaten bewirkt eine proportionale Änderung der Hazardfunktion, der Hazard-Ratio bleibt daher über die Zeit konstant:

$$\frac{h(t|X)}{h(t|X^*)} = \frac{h_0(t) e^{\beta^T X}}{h_0(t) e^{\beta^T X^*}} = e^{\beta^T (X - X^*)}$$

Aufgrund dieser Eigenschaft wird die Cox-Regression auch Cox-Proportional-Hazards-Modell genannt. Unter der Annahme eines proportionalen Hazards können also die Einflüsse von Kovariaten auf die Ereigniszeit geschätzt werden, ohne dass Annahmen an die Grundverteilung gemacht werden müssen. Die geschätzten Parameter sind durch das Hazard-Ratio einfach zu interpretieren. Ein Nachteil des Cox-Modells ist es hingegen, dass nichtlineare Effekte von Variablen durch Transformationen oder die Erweiterung der Design-Matrix um spezielle Basisfunktionen modelliert werden müssen. Beispielsweise muss eine schrittweise Regression verwendet werden, um festzustellen, ob nichtlineare Effekte vorliegen. Wechselwirkungen können nur über einen Brute-Force-Ansatz oder Domänenwissen identifiziert werden.

2.2.1 Parameterschätzung mit partieller Likelihood-Funktion

Seien $(T_1, X_1, \delta_1), ..., (T_n, X_n, \delta_n)$ die Datenpunkte der Stichprobe, bei denen T_i die beobachtete Zeit, X_i der Kovariatenvektor und δ_i der Zensierungsindikator des Individuums i ist. Da an die Verteilung der Überlebenszeit bzw. an den Baseline-Hazard keine Voraussetzungen gestellt werden, kann der Parametervektor β unabhängig von $h_0(t)$ mit der Maximum-Likelihood-Methode geschätzt werden. Man spricht daher von der partiellen Likelihood-Funktion:

$$L(\beta) = \prod_{i=1}^{n} \frac{e^{\beta X_i}}{\sum\limits_{j:T_j \geq T_i} e^{\beta X_j}} \tag{20}$$

Um die Parameter β zu ermitteln, wird die partielle Log-Likelihood-Funktion

$$l(\beta) = \sum_{i=1}^{n} \beta X_i - \sum_{i=1}^{n} \ln\left(\sum_{j:T_j \geq T_i} e^{\beta X_i}\right) \tag{21}$$

unter Zuhilfenahme von numerischen Algorithmen maximiert.

2.2.2 Schätzung des Baseline-Hazards

Möchte man mithilfe des Cox-Regressionsmodells Prognosen für einzelne Individuen tätigen, so muss der Baseline-Hazard ebenfalls geschätzt werden.
Als Erweiterung der Arbeit von Cox hat Breslow 1972 [Bre72] einen Maximum-Likelihood-Schätzer für den Baseline-Hazard vorgestellt, welcher in der Praxis häufig verwendet wird (vgl. [Lin07]).
Der Baseline-Hazard nach Breslow ist definiert als

$$\hat{h}_{0,BR}(t) = \frac{d_t}{\sum\limits_{i:T_i \geq t} e^{\hat{\beta}^T X_i}}, \tag{22}$$

wobei d_t die Anzahl der Individuen ist, bei denen das Ereignis genau zum Zeitpunkt t beobachtet wird.

2.3 Log-Rank-Test

Der Log-Rank-Test nach Mantel [Man72], welcher an den Chi-Quadrat-Test angelehnt ist, überprüft die Überlebensfunktion zweier Stichproben für einen festen Zeitpunkt t auf statistisch signifikante Unterschiede:

$$H_0 : S_0(t) = S_1(t) \quad \text{gegen} \quad H_1 : S_0(t) \neq S_1(t)$$

Für den Zeitpunkt t_j seien $n_{j,0}$, $n_{j,1}$ und n_j die Anzahl der Fälle unter Risiko in Gruppe 0, Gruppe 1 bzw. insgesamt. Analog dazu seien $d_{j,0}$, $d_{j,1}$ und d_j die Anzahl an Ereignissen zu diesem Zeitpunkt in den Gruppen bzw. insgesamt. Dies lässt sich in einer Kontingenztabelle wie folgt darstellen:

		Gruppe		
		1	0	Gesamt
Ereignis	ja	$d_{j,1}$	$d_{j,0}$	d_j
	nein	$n_{j,1} - d_{j,1}$	$n_{j,0} - d_{j,0}$	$n_j - d_j$
	Unter Risiko	$n_{j,1}$	$n_{j,0}$	n_j

Falls beide Gruppen zum Zeitpunkt t_j tatsächlich die gleiche Überlebensfunktion haben, also falls H_0 wahr ist, gilt für den Erwartungswert von $d_{j,1}$:

$$\mathbb{E}[d_{j,1}]_{H_0} = \frac{n_{j,1} d_j}{n_j}$$

Untersucht wird also die Differenz $D_j := (d_{j,1} - \frac{n_{j,1} d_j}{n_j})$.
Ein Schätzer für die Varianz von D_j kann über die hypergeometrische Verteilung abgeleitet werden:

$$\hat{Var}(D_j) = \frac{n_{j,0} n_{j,1} d_j (n_j - d_j)}{n_j^2 (n_j - 1)} \quad = \quad \frac{n_{j,1}}{n_j} \left(1 - \frac{n_{j,1}}{n_j}\right) \left(\frac{n_j - d_j}{n_j - 1}\right) d_j$$

Falls die Zensierungen in beiden Gruppen dem gleichen Muster folgen und es sich um eine ausreichend große Stichprobe handelt, ist $Q = \frac{(\sum_{j=1}^{m} D_j)^2}{\sum_{j=1}^{m} \hat{Var}(D_j)}$ asymptotisch χ^2-verteilt mit einem Freiheitsgrad.
Die Teststatistik für den Log-Rank-Test ist daher gegeben durch:

$$Q = \frac{\left(\sum_{j=1}^{m} \left(d_{j,1} - n_{j,1}\frac{d_j}{n_j}\right)\right)^2}{\sum_{j=1}^{m} \frac{n_{j,1}}{n_j}\left(1 - \frac{n_{j,1}}{n_j}\right)\left(\frac{n_j - d_j}{n_j - 1}\right) d_j} \tag{23}$$

2.4 Random Survival Forest

Genau wie beim Random Forest nach Breiman [Bre01] zur Klassifikation oder Regression, ist ein Random Survival Forest (RSF) ein Ensemble aus baumbasierten Lernern, welches auf den folgenden Prinzipien beruht:

1. Für jeden Baum wird ein Bootstrap-Sample gezogen, mit welchem der Baum trainiert wird.
2. An jedem Knoten im Baum wird eine Teilmenge der Merkmale zufällig gewählt, die als Kandidat für einen möglichen Split betrachtet werden soll
3. Bäume können tiefer verzweigt werden als bei Modellen, die aus einem einzigen Baum bestehen, da hier ein geringeres Risiko für eine Überanpassung besteht.
4. Um für neue Daten Prognosen zu erstellen, wird über die Ergebnisse der einzelnen Bäume aggregiert.

Der Random Survival Forest wurde von Ishwaran *et al.* 2008 [IKBL08] eingeführt, um den Random Forest für rechtszensierte Überlebenszeitdaten zu erweitern. Hierfür wurden die Endknotenstatistik und die Regeln zum Splitten von Knoten entsprechend angepasst.
Die Ausweitung des Random Forests auf rechtszensierte Überlebensdaten bietet die Möglichkeit, ein nicht-parametrisches Modell zur Überlebenszeit zu erstellen. Im Gegensatz zu parametrischen und semi-parametrischen Modellen können so auch nichtlineare Effekte und Interaktionen mit in das Modell einfließen. Durch die Parameterfreiheit ist es allerdings schwieriger, das Modell zu interpretieren und den Einfluss der Kovariaten zu quantifizieren.

2.4.1 Split-Regeln

Seien $(X_1, T_1, \delta_1), ..., (X_n, T_n, \delta_n)$ die Datenpunkte, bei denen T_i die beobachtete Zeit, X_i der Kovariatenvektor und δ_i der Zensierungsindikator des Individuums i ist.
Ziel eines Splits ist es, einen Baumknoten in einen linken und einen rechten Tochterknoten aufzuteilen, sodass sich die beiden Tochterknoten bzgl. der zeitlichen Verläufe der Ereignisse möglichst stark unterscheiden. Ishwaran *et al.* schlagen hierfür die Log-Rank-Statistik (vgl. Kapitel 2.3) als ein geeignetes Kriterium vor. Diese Methode wurde bereits von anderen

Autoren für den Split von Baumlernern im Kontext von Ereigniszeitdaten verwendet (vgl. [Seg88] und [LC93]).
Für eine Kovariate k und einen Split-Wert c erfolgt eine Aufteilung in $L = \{X_i | x_{i,k} \leq c\}$ und $R = \{X_i | x_{i,k} > c\}$.
Seien $t_1 < t_2 < ... < t_m$ die eindeutigen Zeitpunkte der Ereignisse, d_j die Anzahl der Ereignisse und n_j die Anzahl der Individuen unter Risiko zum Zeitpunkt t_j im Elternknoten. Seien außerdem $d_{j,L}, d_{j,R}$ die Anzahl der Ereignisse und $n_{j,L}, n_{j,R}$ die Anzahl der Individuen unter Risiko zum Zeitpunkt t_j in den Tochterknoten L bzw. R. Dann ist die Log-Rank-Statistik definiert als:

$$L(k,c) = \frac{\sum_{j=1}^{m} \left(d_{j,L} - n_{j,L} \frac{d_j}{n_j} \right)}{\sqrt{\sum_{j=1}^{m} \frac{n_{j,L}}{n_j} \left(1 - \frac{n_{j,L}}{n_j} \right) \left(\frac{n_j - d_j}{n_j - 1} \right) d_j}}$$

Zur Beurteilung des Splits wird die Größe $|L(k,c)|$ herangezogen. Je größer der Wert von $|L(k,c)|$ ist, desto unähnlicher sind sich die Beobachtungen in den beiden Knoten im Bezug auf die Zielgröße.
Gesucht werden folglich die Kovariate k^* und der Wert c^*, sodass gilt:
$|L(k^*, c^*)| \geq |L(k,c)| \quad \forall k, c$

2.4.2 Endknoten-Statistiken

Für einen Baum des Random Survival Forests kann dann für die einzelnen Blätter h die Überlebensfunktion mit dem Kaplan-Meier-Schätzer und die kummulative Hazardfunktion mit den Nelson-Aalen-Schätzer geschätzt werden:

$$\hat{H}_h(t) = \sum_{t_{h,j} \leq t} \frac{d_{j,h}}{n_{j,h}}, \qquad \hat{S}_h(t) = \prod_{t_{h,j} \leq t} \frac{n_{j,h} - d_{j,h}}{n_{j,h}}$$

Dabei sind $t_{1,h} < t_{2,h} < ... < t_{m(h),h}$ die sortierten eindeutigen Zeitpunkte der Ereignisse im Blatt h. Es wird also für jede Beobachtung im Blatt h die gleiche Überlebens- und Hazardfunktion geschätzt.
Für einen Patienten mit dem Kovariatenvektor X werden die Überlebens- und die kummulierte Hazardfunktion durch die Mittelwertbildung der einzelnen Bäume ermittelt:

$$\hat{H}(t|X) = \frac{1}{B} \sum_{b=1}^{B} H_b(t|X), \qquad \hat{S}(t|X) = \frac{1}{B} \sum_{b=1}^{B} S_b(t|X)$$

Hierbei sind $H_b(t|X)$ bzw. $S_b(t|X)$ die Schätzer der Blätter, in die die Beobachtung X in Baum b fällt. B bezeichnet die Anzahl der Bäume.

2.4.3 Mortalität

Zur Ermittlung des Concordance-Indexes (vgl. Kapitel 2.5.2) ist eine Kennzahl nötig, die für jeden individuellen Patienten aus dem Prognosemodell abgeleitet werden kann. Ein solcher

Risiko-Score muss unabhängig von der Zeit sein und soll für jeden paarweisen Vergleich zweier Beobachtungen aussagen, welcher Patient eine bessere bzw. schlechtere Prognose hat. Hierfür wird das einfach zu interpretierende Maß der *Mortalität (Mortality)* eingeführt. Seien $t_1 < t_2 < ... < t_m$ die sortierten eindeutigen Zeitpunkte der Ereignisse in dem Datensatz, auf welchem das Modell evaluiert werden soll. Die Mortalität eines Individuums X ist gegeben durch

$$M(X) = \sum_{j=1}^{m} \hat{H}(t_j|X).$$

Der Wert der Mortalität ist das geschätzte Risiko eines Patienten, angepasst an die Anzahl der Ereignisse. Das bedeutet, dass die Anzahl der Ereignisse im Beobachtungszeitraum dieses Individuums (sowohl für zensierte als auch für unzensierte Beobachtungen) in Erwartung genau diesem Wert entsprechen würde, wenn alle Beobachtungen die gleichen Kovariatenausprägungen hätten wie dieses Individuum. Diese Eigenschaft wird auch als *conservation-of-events principle* (etwa Prinzip der Erhaltung von Ereignissen) bezeichnet.

2.5 Evaluationsmetriken für Prognosemodelle in der Überlebenszeitanalyse

Um ein gegebenes Prognosemodell zu evaluieren oder verschiedene Modelle bzgl. ihrer Modellgüte miteinander zu vergleichen, bedarf es Evaluationsmetriken, welche die Vorhersageleistung der Modelle quantifizieren. Hierbei sind herkömmliche Metriken zur Beurteilung von Regressionsmodellen ungeeignet, da sie für den speziellen Endpunkt in der Überlebenszeitanalyse nicht anwendbar sind. Im Folgenden werden daher Metriken vorgestellt, welche für Modelle entwickelt oder erweitert wurden, die auf zensierten Daten basieren.

2.5.1 Brier Score

Der *Brier Score (prediction error)* gibt den mittleren quadratischen Abstand zwischen beobachteten und vorhergesagten Werten für einen Patienten zu einem bestimmten Zeitpunkt an. Ursprünglich wurde er von Brier 1950 [BSR$^+$50] für multikategorielle Klassifikationsaufgaben entwickelt. Für Daten ohne Zensierung lässt sich der Brier Score darstellen als

$$BS(t) = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{1}_{T_i > t} - \hat{S}(t, x_i) \right)^2. \tag{24}$$

Bei einem guten Modell sollte die geschätzte Überlebensfunktion zum Zeitpunkt t nahe 0 sein, wenn das Ereignis bereits stattgefunden hat, und im umgekehrten Fall nahe 1. Der Brier Score basiert auf genau dieser Überlegung. Da $\mathbb{1}_{T_i>t} \in \{0, 1\}$ und $\hat{S}(t, x_i) \in [0, 1]$, kann der Brier Score Werte zwischen 0 und 1 annehmen. Kleine Werte sprechen hierbei für eine gute Vorhersage, größere Werte hingegen für ein weniger gutes Modell.
Wenn für alle Patienten eine Überlebenswahrscheinlichkeit von 0.5 vorhergesagt wird, so ist das Resultat ein Brier Score von 0.25. Diesen Wert erhält man also bei einem Modell, bei welchem man die Ereignisfreiheit der Individuen bis zum analysierten Zeitpunkt zufällig

vorhersagt. Ein gutes Modell muss somit einen Brier Score haben, der unterhalb dieses Schwellenwerts liegt.
Falls die Daten einer Zensierung unterliegen, muss der Brier Score angepasst werden. Hierzu wurde 1999 von Graf *et al.* [GSSS99] eine Methode vorgestellt, welche die von Robins [Rob93] publizierte *IPCW-Methode (Inverse Probability Censoring Weighting)* verwendet. Dabei wird eine Funktion G geschätzt, welche die Wahrscheinlichkeit eines Patienten angibt, bis zu einem bestimmten Zeitpunkt noch nicht zensiert worden zu sein. Dies kann man sich vorstellen wie eine Überlebensfunktion, bei der die Zensur das Ereignis darstellt. Die Schätzung von G kann beispielsweise mit der Kaplan-Meier-Methode realisiert werden. Mit dieser Funktion wird der Brier Score entsprechend gewichtet. Dieser kann wie folgt berechnet werden:

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left(0-\hat{S}(t,x_i)\right)^2 \cdot \mathbb{1}_{T_i\leq t,\delta_i=1}}{\hat{G}(T_i)} + \frac{\left(1-\hat{S}(t,x_i)\right)^2 \cdot \mathbb{1}_{T_i>t}}{\hat{G}(t)}\right) \tag{25}$$

Möchte man einen Wert für den Vorhersagefehler ermitteln, der nicht nur einen diskreten Zeitpunkt betrachtet, sondern eine Beobachtungszeitspanne, so kann man auf den Integrierten Brier Score (IBS) zurückgreifen. Sei t_{max} der späteste beobachtete Zeitpunkt aller Patienten, dann ist der Integrierte Brier Score für den stetigen Fall definiert als

$$IBS = \frac{1}{t_{\text{max}}}\int_0^{t_{\text{max}}} BS(t)dt. \tag{26}$$

2.5.2 Concordance-Index

Der *C-Index* nach Harrell *et al.* 1982 [HCP+82] (auch *Concordance-Index*) vergleicht aus einem statistischen Modell geschätzte Risiko-Scores von einzelnen Patienten. Für die Cox-Regression wird als Risiko-Score der Hazard-Ratio verwendet, für den Random Survival Forest eignet sich die Mortalität als Maß für das Risiko (vgl. [IKBL08]). Die Idee des C-Indexes ist es, die Patienten der Studie paarweise zu vergleichen und zu überprüfen, ob bei dem Patienten mit dem höheren Risiko-Score das Ereignis tatsächlich früher eingetreten ist. Ist dies der Fall, spricht man von Konkordanz (Übereinstimmung). Man kann den C-Index also als Rangkorrelation für zensierte Daten interpretieren. Sei dazu η_i der Risiko-Score des Patienten i, dann gelten folgende Regeln für die Ermittlung des C-Indexes.

1. Wenn keines der beiden Individuen des Paares (i, j) zensiert ist, handelt es sich um Konkordanz, falls $\eta_i > \eta_j$ und $T_i < T_j$. Falls $\eta_i > \eta_j$ und $T_i > T_j$, handelt es sich um ein nicht konkordantes Paar.

2. Falls beide Beobachtungen des Paares (i, j) zensiert sind, kann nicht festgestellt werden, bei welchem Patienten das Ereignis früher eingetreten ist.

3. Zudem gibt es die Möglichkeit, dass das Individuum i zum Zeitpunkt T_i zensiert ist und das Individuum j zum Zeitpunkt T_j ein Ereignis erleidet (für den umgekehrten Fall gilt das gleiche Prinzip). In diesem Fall wird zwischen zwei Szenarien unterschieden:

(a) Für $T_i < T_j$, kann nicht ermittelt werden, welches Individuum das Ereignis zuerst erlebt hat.

(b) Wenn $T_i > T_j$ gilt, dann ist das Ereignis bei Patient j früher eingetreten als bei Patient i. Daher gilt das Paar für den Fall $\eta_i < \eta_j$ als konkordant und für $\eta_i > \eta_j$ als nicht konkordant.

In der Berechnung des C-Indexes gehen nur Paare ein, die als konkordant oder nicht konkordant klassifiziert werden können. Die Fälle 2 und 3a werden also nicht berücksichtigt. Der C-Index ist definiert als der Quotient aus der Anzahl der konkordanten Paare und der Anzahl aller Paare, die in die Berechnung eingehen.

$$C = \frac{|\text{konkordante Paare}|}{|\text{konkordante Paare}| + |\text{nicht konkordante Paare}|}$$

Der Concordance-Index lässt sich also wie folgt berechnen:

$$C = \frac{\sum_{i \neq j} \mathbb{1}_{T_j < T_i} \mathbb{1}_{\eta_j > \eta_i} \delta_j}{\sum_{i \neq j} \mathbb{1}_{T_j < T_i} \delta_j} \tag{27}$$

Es ergeben sich somit für C Werte zwischen 0 und 1, wobei ein Wert von 0.5 für eine zufällige Einordnung der Patienten steht und größere Werte für ein besseres Modell sprechen.

3 Bewertung der Kalibrierung und Methoden zur Rekalibrierung

Die Definitionen und Herleitungen dieses Kapitels sind angelehnt an [Ste09], [CMS+19] sowie [Hou00]. Wurden weitere Quellen verwendet, wird im Text darauf verwiesen.
Wenn ein bereits vorhandenes Modell auf eine neue Kohorte angewendet werden soll, muss die Validität des Modells für die neuen Daten überprüft werden. Falls dabei eine schlechte Kalibrierung festgestellt wird, kann dieses Modell nicht ohne Weiteres auf die neue Population angewendet werden. Neben der Möglichkeit des Erstellens eines neuen Modells, gibt es auch die Option das bereits vorhandene Modell für die neue Kohorte zu rekalibrieren, also so anzupassen, dass eine zufriedenstellende Kalibrierung erreicht werden kann. Hierbei kann es sich bei der neuen Kohorte beispielsweise um eine Population zu einem späteren Zeitpunkt (vgl. [BRE+20]) oder aus einer anderen geographischen Region (vgl. [LKH+18], [DBP+19]) handeln.
In diesem Kapitel wird zunächst beleuchtet, wie Kalibrierung definiert ist, was der Unterschied zwischen Kalibrierung und Diskriminierung ist, und wie ein Modell auf dessen Kalibrierung untersucht werden kann. Anschließend werden Methoden zur Rekalibrierung der logistischen Regression, der Cox-Regression und des Random Forests vorgestellt.

3.1 Diskriminierung und Kalibrierung

Die Diskriminierung stellt die Fähigkeit eines Modells dar, Beobachtungen auf individueller Ebene zu vergleichen. Der C-Index ist die wichtigste Metrik zur Quantifizierung der Diskriminierung. Diese Metrik berücksichtigt sowohl das Auftreten des Ereignisses als auch die Dauer der Nachbeobachtung und gibt ein Maß für die Prognosegüte des gesamten Modells an, anstatt verschiedene Zeitpunkte einzeln zu bewerten. Allerdings kann es vorkommen, dass ein Modell in der Lage ist, die Patienten korrekt nach ihrem Risiko zu sortieren, dabei aber das Risiko allgemein über- oder unterschätzt. Ein solches Modell erreicht trotz dieser Problematik einen hohen C-Index.
Daher ist die Bewertung der Kalibrierung, eine weitere Messgröße für die Vorhersagegenauigkeit, von wesentlicher Bedeutung. Dennoch wird diese bei Modellen für Ereigniszeitdaten in vielen Studien nicht ausreichend untersucht (vgl. [MAR+15], [CMS+19], [CAT+16] ,[Hou00]).

Die Kalibrierung bewertet, ob das tatsächlich beobachtete Ergebnis mit der Vorhersage übereinstimmt. In der Überlebenszeitanalyse kann die Kalibrierung nur für einzelne Zeitpunkte und nicht für das Modell im Allgemeinen ausgewertet werden. Hierbei vergleicht man die zu dem Auswertungszeitpunkt durchschnittlich vorhergesagte Ereigniswahrscheinlichkeit mit dem beobachteten Anteil an Patienten, die bis zu diesem Zeitpunkt ein Ereignis erlitten haben. Die Kalibrierung kann durch den Calibration-Intercept γ_0 und den Calibration-Slope γ_1 quantifiziert und mit Hilfe von Kalibrierungskurven graphisch analysiert werden.

Der Brier Score bestimmt die mittlere quadratische Abweichung zwischen beobachteten und vorhergesagten Werten zu einem bestimmten Zeitpunkt. Hierduch werden sowohl Kalibrierung als auch Diskriminierung beachtet. Er bietet die Möglichkeit durch das Mitteln des Scores für

verschiedene Zeitpunkte (Integrated Brier Score) eine Metrik für den ganzen Beobachtungszeitraum zu ermitteln. Er kann als Bewertung der „Gesamt-Performance“ eines Prognosemodells angesehen werden, ist jedoch wenig intuitiv und schwieriger zu interpretieren.
Eine Erläuterung weiterer Metriken und Analysemethoden zu Diskriminierung, Kalibrierung und Gesamt-Performance ist in [PPKP21] zu finden.

3.2 Kalibrierungskurven

Bei einer Kalibrierungskurve werden beobachtete Ereigniswahrscheinlichkeiten und vorhergesagte Ereigniswahrscheinlichkeiten graphisch gegenübergestellt. Die Beobachtungen X_i werden dazu nach ihrer vom Modell vorhergesagten Ereigniswahrscheinlichkeit

$$\hat{F}_{model}(t|X_i) = 1 - \hat{S}_{model}(t|X_i)$$

zum betrachteten Zeitpunkt t in Quantile unterteilt. Die vorhergesagte Ereigniswahrscheinlichkeit der Quantil-Gruppe q kann dann über die durchschnittlich vorhergesagte Ereigniswahrscheinlichkeit der Patienten dieser Gruppe ausgedrückt werden:

$$\hat{F}_{model,q}(t) = \frac{1}{n_q} \sum_{X_i \in q} \hat{F}_{model}(t|X_i)$$

Dabei ist n_q die Anzahl an Patienten in Quantil-Gruppe q
Für die Patienten des Quantils q wird anschließend mit dem Kaplan-Meier-Schätzer eine Überlebensfunktion $\hat{S}_{KM,q}$ geschätzt. Den Wert

$$\hat{F}_{KM,q}(t) = 1 - \hat{S}_{KM,q}(t)$$

kann als „beobachtete Ereigniswahrscheinlichkeit“ der zu diesem Quantil gehörenden Patienten zum Zeitpunkt t interpretiert werden.
Für jede Gruppe wird dann der Wert $\hat{F}_{model,q}(t)$ auf der Abszisse abgetragen, während der Wert $\hat{F}_{KM,q}(t)$ auf der Ordinate abgetragen wird. Dadurch wird jede Quantil-Gruppe durch einen Punkt im Koordinatensystem dargestellt. Zusätzlich wird häufig das 95%-Konfidenzintervall des Kaplan-Meier-Schätzers in der Kalibrierungskurve gezeigt. Beispiele für Kalibrierungskurven sind in Abbildung 3 zu sehen.
Die Ideallinie ist durch die identische Abbildung gegeben, bei einer guten Kalibrierung sollten also alle Punkte möglichst nah an dieser Linie liegen.

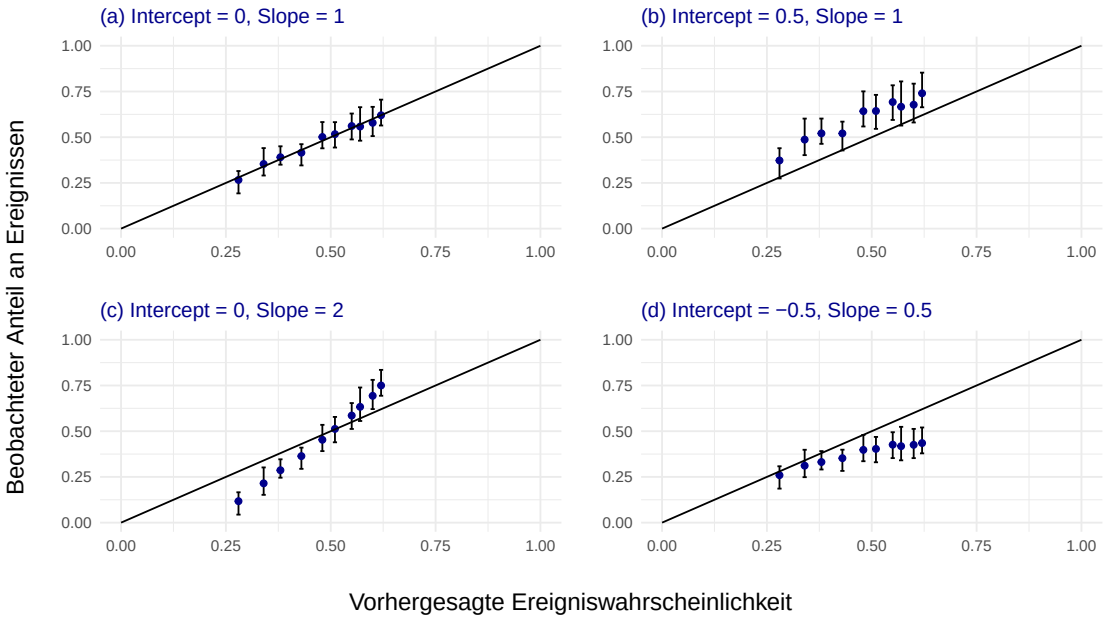

Abbildung 3: Beispiele für Kalibrierungskurven von verschieden gut kalibrierten Modellen: (a) perfekte Kalibrierung, (b) durchschnittlich zu gering vorhergesagtes Risiko, (c) zu moderate Risikoschätzung, (d) Risikoschätzung durchschnittlich zu hoch und außerdem zu extrem

Die Ergebnisse der Kalibrierungskurven lassen sich durch *Calibration-Intercept* und *Calibration-Slope* quantifizieren.
Die Idee bei der Berechnung von Calibration-Slope und Calibration-Intercept ist es, mit den Validierungsdaten eine erneute Regression anzupassen, bei welcher der linearen Prädiktor des ursprünglichen Modells die einzige Kovariate darstellt. Bei guter Kalibrierung müsste so ein Intercept von 0 und ein Slope von 1 geschätzt werden (vgl. [Roy14]). Die Cox-Regression kann dargestellt werden als

$$S(t|X) = S_0(t)^{\exp(\beta^T X)}. \tag{28}$$

Durch Anwendung der Link-Funktion $g(u) := ln(-ln(u))$ kann diese Gleichung wie folgt dargestellt werden:

$$g(S(t|X)) = g(S_0(t)) + \beta^T X \tag{29}$$

Die rechten Seite von Gleichung 29 ist gegeben durch die Summe aus einer Konstanten und der Linearkombination aus den erklärenden Variablen und den dazugehörigen durch die Cox-Regression geschätzten Parametern, also durch einen linearen Prädiktor. Folglich können Calibration-Intercept und Calibration-Slope bei der Cox-Regression bestimmt werden, indem ein verallgemeinertes lineares Modell (GLM) der $\hat{F}_{model,q}(t)$ auf die Zielvariable $\hat{F}_{KM,q}(t)$ mit Link-Funktion g angepasst wird.

Der Calibration-Intercept gibt Aufschluss darüber, ob das durchschnittlich vorhergesagte

Risiko mit dem durchschnittlich beobachteten Risiko übereinstimmt, also ob das Vorhersagemodell das Risiko im Allgemeinen über- oder unterschätzt. Man spricht in diesem Zusammenhang auch von *mean calibration* oder *calibration-in-the-large*. Bei negativen Werten des Calibration-Intercepts ist das durchschnittlich vorhergesagte Risiko zu hoch, bei negativen Werten zu niedrig.
Für die *weak calibration* wird zusätzlich der Calibration-Slope betrachtet. Diese Metrik bewertet die Verteilung der geschätzten Risiken. Eine Steigung kleiner 1 deutet darauf hin, dass die geschätzten Ereigniswahrscheinlichkeiten zu extrem sind. Das bedeutet, dass bei Patienten mit hohem Risiko die Ereigniswahrscheinlichkeit über- und bei Patienten mit niedrigem Risiko unterschätzt wird. Eine Steigung größer 1 hingegen spricht für eine zu moderate Risikoschätzung. In diesen Fällen wird bei Patienten mit hohem Risiko die Ereigniswahrscheinlichkeit unter- und bei Patienten mit niedrigem Risiko überschätzt.
In Abbildung 3 sind die Kurven für vier unterschiedlich gut kalibrierte Modelle zu sehen. Für das Modell in (a) befinden sich die Punkte nahe an der Ideallinie und der Wert von Intercept und Slope ist 0 bzw. 1. Es handelt sich also um ein sehr gut kalibriertes Modell. Für das Modell in Abschnitt (b) liegt ein Intercept von 0.5 und ein Slope von 1 vor. Das durchschnittlich vorhergesagte Risiko ist in diesem Modell also zu niedrig, während die Einschätzungen dabei aber nicht zu extrem oder zu moderat sind. Das zum Abschnitt (c) gehörende Modell weist einen Intercept von 0 und einen Slope von 2 auf. Hier liegt also eine zu moderate Risikoschätzung vor, wobei das durchschnittlich vorhergesagte Risiko allerdings nicht von dem durchschnittlich beobachteten Risiko abweicht. Zuletzt zeigt Abschnitt (d) ein Modell, bei welchem das Risiko durchschnittlich zu hoch und dabei zu extrem eingeschätzt wird. Dies zeigt sich durch einen Intercept von -0.5 und einen Slope von 0.5.

3.3 Methoden zur Rekalibrierung

Falls bei einer neuen Kohorte für ein bereits vorhandenes Modell eine schlechte Kalibrierung festgestellt wird, können mit diesem Modell nicht ohne Weiteres Prognosen für Patienten aus der neuen Kohorte getätigt werden. Bei einem neu trainierten Modell bleiben potentiell aussagekräftige Informationen aus dem ursprünglichen Modell ungenutzt und können verloren gehen. Insbesondere wenn die Stichprobenzahl in der neuen Kohorte gering ist, riskiert man zudem eine Überanpassung. Unter Umständen kann ein rekalibriertes Modell also für die neue Kohorte eine bessere Vorhersagegüte haben als ein auf diesen Daten neu trainiertes Modell. Ferner kann die Existenz einer große Anzahl an Modellen für denselben Endpunkt verwirrend sein und die Entscheidung erschweren, welches in der Praxis anzuwenden ist. Anstatt ein neues Modell mit diesen Daten anzupassen, kann es daher vorteilhaft sein, stattdessen das bereits vorhandene Modell für die neue Kohorte zu rekalibrieren.

3.3.1 Rekalibrierung der Cox-Regression

Verschiedene Ansätze zur Rekalibrierung von klinischen Vorhersagemodellen, unter anderem auch für die logistische Regression und die Cox-Regression, wurden 2000 von Van Houwelingen [Hou00] vorgeschlagen. Um ein gegebenes Cox-Modell $\hat{h}(t|X) = \hat{h}_0(t)e^{\hat{\beta}^T X}$ auf neue Daten anzupassen, kann man mithilfe der neuen Beobachtungen einen neuen

Baseline-Hazard $\hat{h}_{0,\text{new}}$ schätzen. Auf diese Weise erhält man eine rekalibrierte Schätzung der Hazardfunktion:

$$\hat{h}_{\text{recalibrated}}(t|X) = \hat{h}_{0,\text{new}}(t) \cdot e^{\hat{\beta}^T X} \tag{30}$$

Zusätzlich besteht die Möglichkeit, einen Parameter $\hat{\beta}_{\text{recalibration}}$ zum Anpassen des Slopes zu bestimmen. Dazu wird ein neues Cox-Modell trainiert, welches eine einzige Kovariate hat, nämlich den linearen Prädiktor $\hat{\beta}^T X$ des Ursprungsmodells. Die so rekalibrierte Hazardfunktion ist gegeben durch

$$\hat{h}_{\text{recalibrated}}(t|X) = \hat{h}_{0,\text{new}}(t) \cdot e^{\hat{\beta}_{\text{recalibration}}\hat{\beta}^T X}. \tag{31}$$

Während die Dimension von $\hat{\beta}$ der Anzahl der Kovariaten entspricht, die in das Modell eingehen, ist $\hat{\beta}_{\text{recalibration}}$ eindimensional. Daher werden alle Parameter des Modells mit dem gleichen Skalierungsfaktor angepasst.

3.3.2 Rekalibrierung der logistischen Regression

Möchte man ein gegebenes logistisches Regressionsmodell $\hat{P}(Y = 1|X) = \frac{1}{1+e^{-(\hat{\alpha}+\hat{\beta}^T X)}}$ für einen neuen Datensatz anpassen, gibt es wie bei der Cox-Regression die Möglichkeit den Intercept $\hat{\alpha}$ oder sowohl Intercept $\hat{\alpha}$ also auch Slope $\hat{\beta}$ zu rekalibrieren.
Soll nur eine Neuberechnung des Intercepts erfolgen, wird eine neue logistische Regression mit $\hat{\alpha}_{\text{new}}$ als einzigem freien Parameter angepasst. Der lineare Prediktor wird hierbei als sogenannte *offset variable* behandelt. Dabei wird der Wert $\hat{\beta}^T X$ als einzige Kovariate betrachtet, deren Parameter allerdings auf den Wert 1 fixiert ist. Die Wahrscheinlichkeitsschätzung des so rekalibrierten Modells ist

$$\hat{P}_{\text{recalibrated}}(Y = 1|X) = \frac{1}{1 + e^{-(\hat{\alpha}_{\text{new}}+\hat{\beta}^T X)}}. \tag{32}$$

Für eine zusätzliche Anpassung des Slopes wird ebensfalls der Wert $\hat{\beta}^T X$ als einzige Kovariate betrachtet, deren Parameter $\hat{\beta}_{\text{recalibration}}$ ist in diesem Fall jedoch ein freier Parameter. Somit erhält man ein rekalibriertes Modell, welches gegeben ist durch

$$\hat{P}_{\text{recalibrated}}(Y = 1|X) = \frac{1}{1 + e^{-(\hat{\alpha}_{\text{new}}+\hat{\beta}_{\text{recalibration}}\hat{\beta}^T X)}}. \tag{33}$$

Genau wie beim Cox-Modell ist $\hat{\beta}_{\text{recalibration}}$ ein eindimensionaler Skalierungsfaktor, der also auf alle Steigungsparameter des ursprünglichen Modells den gleichen multiplikativen Einfluss hat.

3.3.3 Rekalibrierung des Random Forests

Eine Methode zur Rekalibrierung von Random Forests zur Wahrscheinlichkeitsschätzung wurde 2016 von Dankowski und Ziegler [DZ16] vorgestellt. Hierbei werden die einzelnen Bäume des Random Forests in logistische Regressionsmodelle umgewandelt. Diese Modelle

können dann mit den bekannten Verfahren zur Rekalibrierung der logistischen Regression (vgl. Kapitel 3.3.2) angepasst werden. Dabei schlagen die Autoren vor, sich bei der Rekalibrierung auf den Intercept zu beschränken. Möchte man allerdings neben der *calibration-in-the-large* auch die *mean calibration* untersuchen, kann auch eine Anpassung des Slopes erfolgen. Abbildung 4 stellt die einzelnen Schritte der Rekalibrierungsmethode dar.
Um einen Entscheidungsbaum in ein logistisches Regressionsmodell umzuwandeln, wird der Baum im ersten Schritt mit den Daten, mit welchen er trainiert wurde, erneut durchlaufen. Dabei wird ermittelt, in welche Blätter die einzelnen Beobachtungen fallen. Im zweiten Schritt wird eine neue Darstellung der Daten erzeugt, indem für jede Beobachtung über Dummy-Variablen angegeben wird, in welches Blatt sie fällt. Dieses Verfahren wird in Abbildung 5 veranschaulicht. Eines der Blätter stellt dabei, wie bei Dummy-Variablen üblich, die Referenz-Kategorie dar. So erhält man eine veränderte Darstellung dieser Daten, bei welcher über eine Dummy-Variable die Zugehörigkeit zu den einzelnen Blättern kodiert ist. In Schritt 3 wird eine logistische Regression auf der neuen Datendarstellung angepasst. Dabei sind die Dummy-Variablen der Blattzugehörigkeiten die neuen Kovariaten. Ein solches Modell schätzt die Wahrscheinlichkeit also in Abhängigkeit der Blätter, in welche die jeweiligen Beobachtungen fallen. Anschließend wird im vierten Schritt auch der Datensatz mit den Beobachtungen aus der neuen Kohorte durch den Baum geschickt und es werden in Schritt 5 erneut Dummy-Variablen für die Blattzugehörigkeiten erstellt. Im abschließenden sechsten Schritt wird das logistische Regressionsmodell aus Schritt 3 mit dem Datensatz aus Schritt 5 rekalibriert. Dieses Vorgehen wird für jeden einzelnen Baum des Random Forests angewendet. Die vorhergesagten Wahrscheinlichkeiten des so rekalibrierten Random Forests werden durch den Mittelwert der Ergebnisse der rekalibrierten Regressionsmodelle bestimmt.

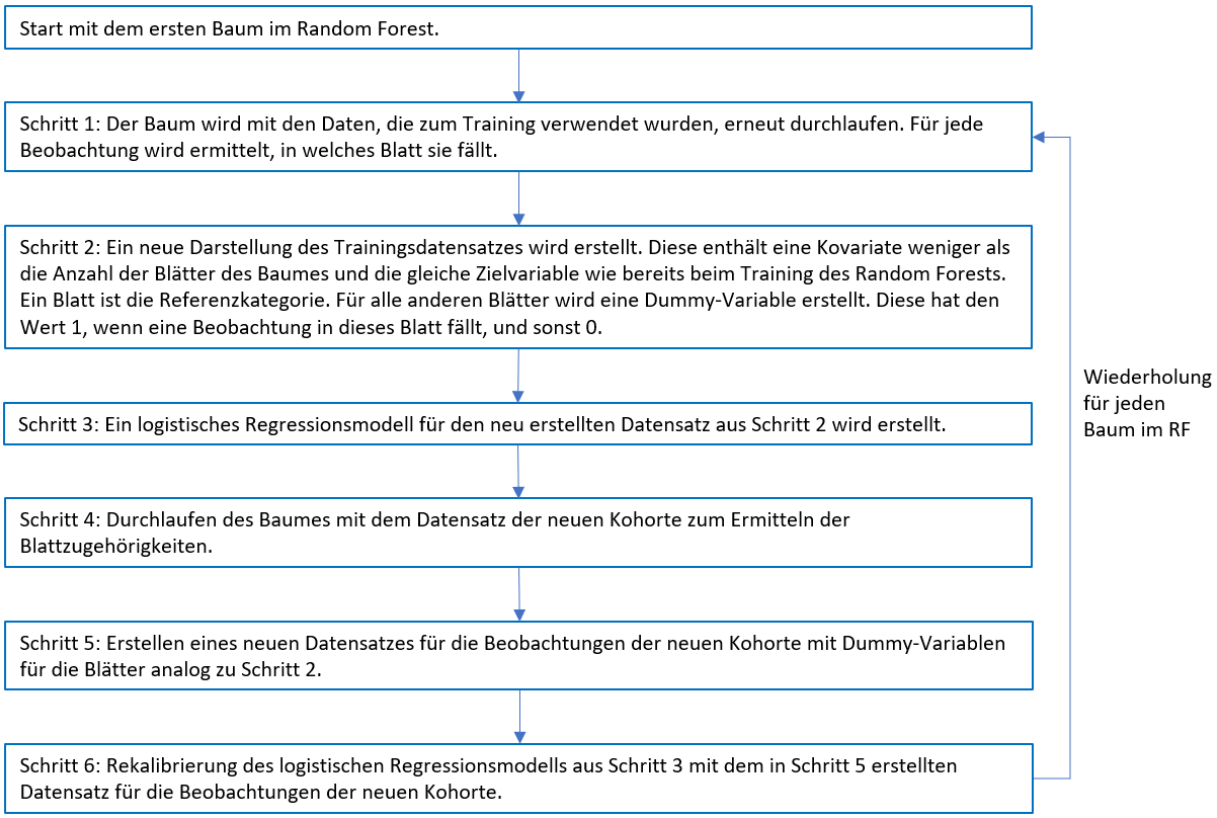

Abbildung 4: Rekalibrierung eines Random Forests nach der Methode von Dankowski und Ziegler

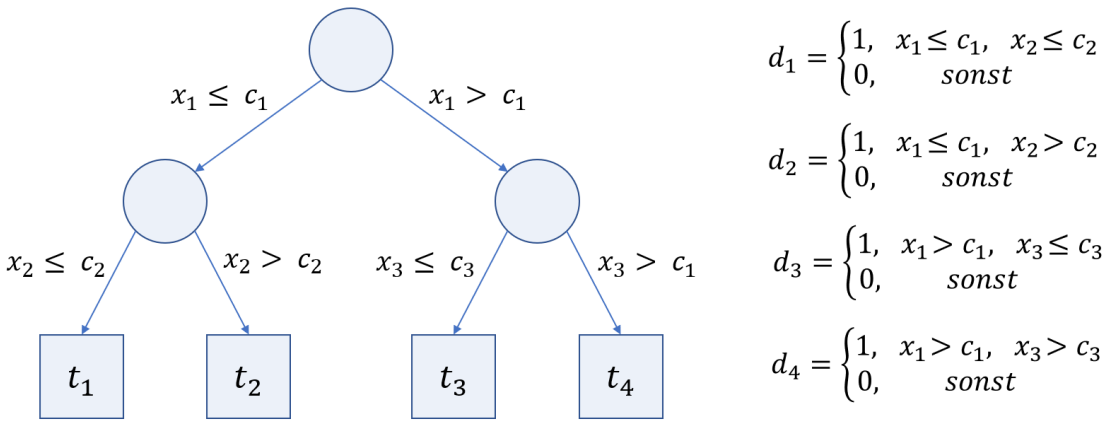

Abbildung 5: Beispiel zur Erstellung der Dummy-Variablen für die Umwandlung eines Entscheidungsbaumes in eine logistische Regression nach der von Dankowski und Ziegler vorgeschlagenen Methode

4 Datengrundlage

4.1 Herkunft der Daten

Die in dieser Arbeit verwendeten deutschen Daten wurden vom nationalen Transplantationsregister zur Verfügung gestellt. Das Transplantationsregister ist eine Institution, die Daten über Organspenden, Transplantationen, Spender und Empfänger in Deutschland an einer zentralen Stelle speichert. Die Erschaffung des Registers wurde 2016 durch das Gesetz zur Errichtung eines Transplantationsregisters (TxRegG) angestoßen. Für Transplantationen aus den Jahren 2006 bis 2016 wurden dazu Daten der DSO, IQTIG und Eurotransplant zusammengetragen. Die in dieser Arbeit untersuchten Daten sind die aus diesen Jahren stammenden sogenannten Altdaten.

Die *Deutsche Stiftung Organtransplantation (DSO)* ist die bundesweite Koordinierungsstelle für die postmortale Organspende. Sie ist mit der Organisation sämtlicher Schritte des Organspendeablaufs von der Mitteilung eines potentiellen Spenders im Krankenhaus bis zur Übergabe des Organs an das Transplantationszentrum betraut.
Das *Institut für Qualitätssicherung und Transparenz im Gesundheitswesen (IQTIG)* ist das zentrale Institut für die gesetzlich verankerte Qualitätssicherung im deutschen Gesundheitswesen. Sie agiert im Auftrag des Gemeinsamen Bundesausschusses, welcher das höchste Beschlussgremium der gemeinsamen Selbstverwaltung im deutschen Gesundheitswesen ist. Das IQTIG entwickelt Qualitätssicherungsverfahren und ist auch an der Durchführung dieser Verfahren beteiligt.
Die *Stiftung Eurotransplant (ET)* ist die Vermittlungsstelle für Organspenden in acht europäischen Ländern. Als internationale Non-Profit-Organisation vermittelt sie zwischen Spenderkrankenhäusern und Transplantationszentren in allen Mitgliedsstaaten. Diese sind die Benelux-Länder, Deutschland, Österreich, Slowenien, Kroatien und Ungarn. Im Einzugsgebiet von ET leben circa 137 Millionen Menschen.

Die in dieser Arbeit verwendeten US-amerikanischen Daten wurden vom *United Network for Organ Sharing (UNOS)* zur Verfügung gestellt. UNOS ist eine Non-Profit-Organisation in den USA, welche die Organvergabe des ganzen Landes organisiert und sich mit dem Management der zugehörigen Wartelisten befasst. Eine weitere Aufgabe von UNOS ist die Pflege der Datenbank, die alle Daten zu Organtransplantationen in den USA enthält. UNOS arbeitet im Auftrag des *Organ Procurement and Transplantation Network (OPTN)*, welches die Transplantationsrichtlinien in den USA erstellt und pflegt.

4.2 Datenstruktur

Die Altdaten des deutschen Transplantationsregisters stammen von ET, DSO und IQTIG, also aus drei verschiedenen Quellen. Dabei wurden die erhobenen Werte in gemeinsamen Tabellen zusammengetragen, bei welchen jedoch jede erhobene Variable durch ihre Benennung eindeutig der Quelle zugeordnet werden kann. Hierdurch gibt es in den einzelnen Tabellen häufig mehrere Felder, welche die gleichen Bedeutungen haben. So gibt es beispielsweise in

der Tabelle, welche Eigenschaften der Spender beinhaltet, drei verschiedene Felder, welche dessen Größe angeben. Häufig finden sich jedoch auch Felder mit derselben Bedeutung je nach Quelle in verschiedenen Tabellen, oder es müssen benötigte Informationen je nach Quelle über verschiedene Algorithmen hergeleitet werden. Diese Eigenschaften führen beim Data-Mining-Prozess für den deutschen Datensatz zu diversen Herausforderungen.

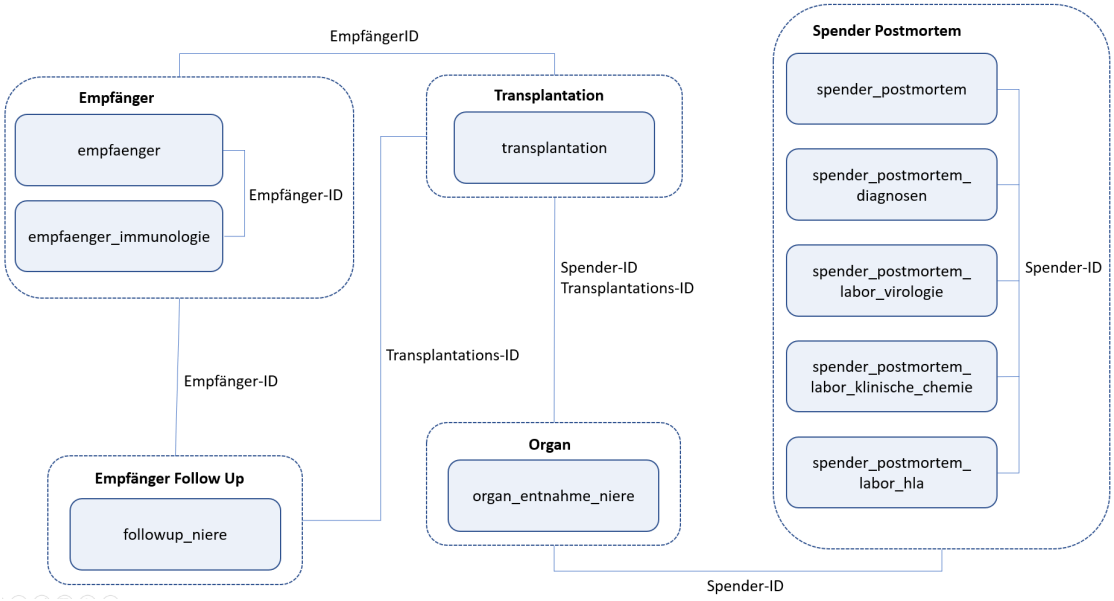

Abbildung 6: Modell der Daten des deutschen Transplantationsregisters mit allen Tabellen, die zur Kovariatenerhebung, Bestimmung des Endpunktes und Selektion der Analysepopulation benötigt wurden. Die Namen der IDs für Empfänger, Spender und Transplantation variieren zwischen den Tabellen. Eine Zuordnung ist in Tabelle 17 zu finden.

In Abbildung 6 ist eine Darstellung der Datenstruktur der Daten des deutschen Transplantationsregisters zu sehen. Diese enthält alle Tabellen, welche zur Erhebung der Kovariaten und zur Erstellung der Analysepopulation (vgl. Kapitel 4.3) für die in dieser Arbeit verwendeten Modelle benötigt werden. Für die Verknüpfung der Tabellen gibt es IDs für Empfänger, Spender und Transplantation. Diese sind zwischen den drei Quellen einheitlich, jedoch variiert ihre Benennung je nach Tabelle und Quelle. Eine Zuordnung der IDs ist im Anhang in Tabelle 17 zu finden.

In den UNOS-Daten gibt es mit der Tabelle *kidpan_data* ein zentrales Element im Bezug auf Nieren- und Pankreastransplantationen. Die Zusammenfassung dieser beiden Organe in eine Tabelle hat den Hintergrund, dass Niere und Pankreas häufig simultan in einer Operation transplantiert werden. Die Tabelle kidpan_data enthält Informationen zu Empfänger, Spender, Transplantation und Organ. Alle Informationen, die zum Erstellen der hier betrachteten Modelle benötigt werden, können bereits in dieser einzigen Tabelle gefunden werden. Die weiteren Tabellen der US-amerikanischen Transplantationen enthalten zusätzliche

Informationen zu Spendern, Follow-Up-Untersuchungen, Warteliste oder Histokompatibilität, werden aber für die Zwecke dieser Arbeit nicht verwendet.

4.3 Auswahl der Kovariaten sowie Definition von Analysepopulation und Endpunkt

Für die in dieser Arbeit verwendeten Vorhersagemodelle wurden die Definition der Analysepopulation, die Definition des Endpunktes sowie die betrachteten Kovariaten angelehnt an die Publikation von Rao *et al.* ([RS09]).
Die **Analysepopulation** wird eingeschränkt auf Nierentransplantationen, bei welchen

- es sich um eine Transplantation nach einer Todspende handelt.
- der Empfänger mindestens 18 Jahre alt ist.
- die Blutgruppen von Spender und Empfänger kompatibel sind.
- es sich um die erste Transplantation des Empfängers handelt, auch keine anderen Organe außer der Niere wurden zuvor transplantiert.
- nicht in der gleichen Operation auch andere Organe transplantiert wurden
- die Transplantation zwischen dem 1.1.2006 und dem 31.12.2016 stattgefunden hat
- das Spenderorgan von einem hirntoten Spender stammt

In der Publikation von Rao *et al.* werden beide Spendertypen, hirntote und herztote Spender, betrachtet. In Deutschland sind nur Organspenden von hirntoten Spendern erlaubt. Daher wurde die Analysepopulation (anders als in Rao *et al.*) auf hirntote Spender beschränkt. In der US-Kohorte wurden aus diesem Grund Spenden von herztoten Spendern exkludiert.
Auch der betrachtete Zeitraum der Transplantationen ist anders als im Modell von Rao *et al.* Die Daten des Transplantationsregisters enthalten Transplantationen, die zwischen dem 1.1.2006 und dem 31.12.2016 stattgefunden haben. Daher wurde dies als Untersuchungszeitraum festgelegt.

Die betrachteten **Kovariaten** sind:

- Alter des Spenders in Jahren
- Finaler Kreatininwert in mg/dl
- Größe des Spenders in cm
- Gewicht des Spenders in kg
- CVA (Cerebrovascular Accident): War die Todesursache ein Zerebrovaskulärer Unfall?
- Hypertonie: Litt der Spender an einer Form des Bluthochdrucks?

- Hepatitis C: Waren beim Spender zum Zeitpunkt der Organentnahme HCV-Antikörper vorhanden?
- Typ: Handelt es sich um die Transplantation einer Niere, zweier isolierter Nieren oder zweier Nieren, die als eine Einheit transplantiert werden (EnBloc)?
- Diabetes: Litt der Spender an Diabetes?
- Kalte Ischämiezeit: Zeit in Stunden zwischen der Organentnahme und der -transplantation, in der das Transplantat gekühlt gelagert und transportiert wird.
- HLA-B Mismatches: Anzahl der Mismatches der Humanen Leukozytenantigene B zwischen Spender und Empfänger
- HLA-DR Mismatches: Anzahl der Mismatches der Humanen Leukozytenantigene DR zwischen Spender und Empfänger

In der Publikation von Rao *et al.* werden zwei weitere Kovariaten verwendet. Zum einen wird der Spendertyp betrachtet, also unterschieden, ob es sich um einen herztoten oder um einen hirntoten Spender handelt. Da es in der hier betrachteten Analysepopulation jedoch nur hirntote Spender gibt, wird diese Kovariate nicht verwendet. Des Weiteren kann die Kovariate Ethnizität (ethnische Zugehörigkeit des Spenders) nicht verwendet werden, da diese in den Daten des deutschen Transplantationsregisters nicht erhoben wird.

Als **Endpunkt** sind Organversagen oder Tod definiert. Dabei spielt die Funktionalität des Organes im Falle des Todes keine Rolle.
Es wurde eine administrative Zensierung für sieben Jahre nach der Transplantation durchgeführt. Das bedeutet, dass Patienten, die das Ereignis nach sieben Jahren noch nicht erlitten haben, zensiert wurden. Für sie wurde die beobachtete Zeit also auf sieben Jahre gesetzt und der Endpunkt dieser Patienten wird mit $(T = 2566 \text{ Tage }, \delta = 0)$ angegeben. Die Erhebung der deutschen Daten endete am 31.12.2016, diese beinhalten also bereits eine administrative Zensierung zu diesem Datum. Daher wurde dies für die US-amerikanischen Daten ebenfalls durchgeführt, um die beiden Analysepopulationen anzugleichen.

4.4 Datenvorverarbeitung

In diesem Kapitel wird erläutert, wie der Data-Mining-Prozess auf den deutschen und den US-amerikanischen Daten durchgeführt wurde, um die Analysepopulationen zu erhalten. Dabei wird darauf eingegangen, mit welchen Feldern der Datensätze und mit welchen Algorithmen die betrachteten Kovariaten ermittelt wurden. Außerdem wird berichtet, wie die Analysepopulationen selektiert und wie der Endpunkt ermittelt wurde.

4.4.1 Daten des deutschen Transplantationsregisters

Zunächst soll erläutert werden, wie die **numerischen Kovariaten** für die Daten des deutschen Transplantationsregisters ermittelt wurden. Tabelle 2 zeigt eine Übersicht der hierzu verwendeten Variablen und Tabellen.

Dabei wurde für alle Kovariaten zunächst der von Eurotransplant übermittelte Wert betrachtet. Falls dieser Wert fehlte, wurde der von der DSO bereitgestellte Wert verwendet. Falls auch dieser Wert fehlte, wurde, falls vorhanden, der von IQTIG bereitgestellte Wert verwendet. Für die Ermittlung der Kovariate *Kreatinin* mussten zudem die angegebenen Werte bei Bedarf in die Einheit mg/dL umgewandelt werden. Diese Kovariate beschreibt den letzten gemessenen Kreatininwert im Serum des Spenders. Für die Angaben von ET und DSO sind jedoch mehrere gemessene Kreatininwerte pro Spender gegeben und es ist nicht zu ermitteln, bei welchem es sich um den spätesten Messwert handelt. Daher wurde, falls die von ET oder DSO stammenden Werte genutzt wurden, der Mittelwert aller für den Spender gemeldeten Kreatininwerte verwendet. Bei der Angabe des Kreatininwertes von IQTIG handelt es sich bereits um den gesuchten letzten gemessenen Wert. Allerdings sind hier bzgl. der verwendeten Einheit einige nicht plausible Angaben zu finden. Deshalb wurde dennoch vorrangig auf die Werte von ET oder DSO zurückgegriffen, falls diese vorhanden waren. In allen drei Fällen wurde außerdem das Minimum aus dem gegebenen Wert und 8 mg/dL verwendet.

Tabelle 2: Beschreibung der Variablen, welche zur Bestimmung der numerischen Kovariaten für die deutschen Daten verwendet wurden

Kovariate	Variable	Tabelle	Bemerkung
Kalte Ischämie-zeit	TIschaemiezeitKaltWertET	transplantation	
	ONQualityFormEvaluierung-IschaemiezeitKaltWertDSO	organ_entnahme_niere	
Alter	SPostmBasisGeburtsdatumET	spender_postmortem	Differenz zu SPostmBasisZeitpunktHirntodET geteilt durch 365.25
	SPostmBasisAlterDSO	spender_postmortem	
	SPostmBasisAlterIQTIG	spender_postmortem	
Gewicht	SPostmBasisGewichtWertET	spender_postmortem	
	SPostmBasisGewichtWertDSO	spender_postmortem	
	SPostmBasisGewichtWertIQTIG	spender_postmortem	
Größe	SPostmBasisGroesseWertET	spender_postmortem	
	SPostmBasisGroesseWertDSO	spender_postmortem	
	SPostmBasisGroesseWertIQTIG	spender_postmortem	
Kreatinin	SPostmLaborKCKreatinin-WertET	spender_postmortem_ labor_klinische_chemie	In SPostLaborKCKreatininEinheitET, SPostmLaborKCKreatininEinheitDSO
	SPostmLaborKCKreatinin-WertDSO	spender_postmortem_ labor_klinische_chemie	SPostmLaborKCKreatininEinheitDSO geg. Einheiten ggf. in mg/dL
	SPostmLaborKCKreatinin-LetzterWertIQTIG	spender_postmortem_ labor_klinische_chemie	umwandeln, Mittelwertbildung, Minimum aus diesem Wert und 8mg/dL

Die für die Ermittlung der **kategorischen Kovariaten** verwendeten Variablen und Tabellen

sind in Tabelle 3 aufgelistet. Dabei ist vermerkt, welche Ausprägungen dieser Variablen zu welcher Ausprägung der zu ermittelnden Kovariate abgeleitet wurden. Für die Kovariaten *Diabetes*, *Hypertonie* und *Hepatitis* wurde davon ausgegangen, dass der Spender unter der jeweiligen Krankheit gelitten hatte, wenn eines der hierfür betrachteten Felder dies angab. Für die Kovariate *CVA* wurden zunächst die von ET gegebenen Variablen betrachtet. Falls diese leer waren, wurde der von DSO gegebene Wert verwendet.

Aus den von Spendern und Empfängern gegebenen *HLA-Antigenen* wurden die HLA-B- und HLA-DR-Antigene extrahiert. Die Anzahlen der Mismatches wurden anschließend nach dem Zuordnungsverfahren von Eurotransplant bestimmt. Dieses kann im Eurotransplant Manual ([ET22]) in Kapitel 10 nachgelesen werden. Die entsprechenden Tabellen 23 und 24 sind im Anhang zu finden. Die Kovariate *Transplantationstyp* wurde hergeleitet, indem für jede Transplantation überprüft wurde, ob bei der Operation ein oder zwei Organe transplantiert wurden. Im Fall von einer doppelten Transplantation waren in der Tabelle *transplantation* für eine Transplantations-ID zwei Einträge mit einer transplantierten Niere zu finden. Ob es sich bei den Transplantationen um EnBloc-Transplantationen handelt, ist mit den Daten des Transplantationsregisters nicht festzustellen. Die Gesetzeslage in Deutschland sieht vor, dass bei Organen von Spendern bis zum Alter von zwei Jahren nur beide Nieren als eine Einheit transplantiert werden dürfen. Für Spender im Alter von bis zu fünf Jahren wird dieses Vorgehen zudem dringend empfohlen. Daher wurde für die Kovariate *Transplantationstyp* bei Spendern bis zu einem Alter von zwei Jahren immer von einer EnBloc-Transplantation ausgegangen. Bei Spendern bis zum Alter von fünf Jahren wurde im Fall einer Doppeltransplantation ebenfalls davon ausgegangen, dass beide Organe als eine Einheit transplantiert wurden.

Für den **Endpunkt** wurden acht verschiedene im Transplantationsregister erhobene Variablen betrachtet, welche in Tabelle 4 aufgeführt werden. Dabei wurde der Endpunkt durch ein Ereignis definiert, wenn mittels dieser Informationen ein Zeitpunkt für den Tod des Patienten, ein Organversagen oder eine Retransplantation festgestellt werden konnte. Dabei wurde das jeweils frühste hierbei ermittelte Datum als Zeit festgelegt und der Zensierungsindikator wurde auf nicht *zensiert* gesetzt. Falls kein Zeitpunkt für ein Ereignis festgestellt werden konnte, wurde der Patient als zensiert betrachtet. Der Zeitpunkt der Zensierung wurde dabei als das späteste der durch die drei zugehörigen Felder ermittelten Werte definiert. Abschließend wurde eine administrative Zensierung für den Tag sieben Jahre nach der jeweiligen Transplantation durchgeführt.

Tabelle 3: Beschreibung der Variablen, welche zur Bestimmung der kategorialen Kovariaten für die deutschen Daten verwendet wurden

Kovariate	Variable	Tabelle	Bemerkung
Diabetes	SPostmAnamnVorerkrankungen-DiabetesMellitusET	spender_postmortem	Ja, falls 'Yes', 'Type 1', 'Type 2'; Nein, falls 'No'
	SPostmAnamnVorerkrankungen-DiabetesMellitusDSO	spender_postmortem	Ja, falls 'Ja'; Nein, falls 'Nein'
	SPostmDiagnosenDiagnose-BeschreibungDSO	spender_postmortem_ diagnosen	es gibt eine Diagnose zu diesem Spender, welche den Begriff 'Diabetes' beinhaltet
Hypertonie	SPostmAnamnVorerkrankungen-HypertonieET	spender_postmortem	Ja, falls 'Yes' ; Nein, falls 'No'
	SPostmAnamnVorerkrankungen-HypertonieDSO	spender_postmortem	Ja, falls 'Ja'; Nein, falls 'Nein'
	SPostmDiagnosenDiagnose-BeschreibungDSO	spender_postmortem_ diagnosen	es gibt eine Diagnose zu diesem Spender, welche die Begriffe 'Hypertonie' oder 'Hypertensiv' beinhaltet
Hepatitis	SPostmAnamnVorerkrankungen-HepatitisDSO	spender_postmortem	Ja, falls 'Ja'; Nein, falls 'Nein'
	SPostmDiagnosen-DiagnoseICD10DSO	spender_postmortem_ diagnosen	ICD10 Codes 'B18.2' oder 'B17.1'
	SPostmLaborVIRHCVAbET	spender_postmortem_ labor_virologie	Ja, falls 'Positive' ; Nein, falls 'Negative'
	SPostmLaborVIRHCVRNADSO	spender_postmortem_ labor_virologie	Ja, falls 'Pos' ; Nein, falls 'Neg'
	SPostmLaborVIRHCVAKRDSO	spender_postmortem_ labor_virologie	Ja, falls 'Pos' ; Nein, falls 'Neg'
	SPostmLaborVIRHCVAKDSO	spender_postmortem_ labor_virologie	Ja, falls 'Pos' ; Nein, falls 'Neg'
CVA	SPostmBasisTodesursache-ICD10ET	spender_postmortem_ diagnosen	Ja, falls Code mit 'I61', 'I63', 'I64' oder 'I67' beginnt
	SPostmBasisTodesursacheET	spender_postmortem_ diagnosen	Ja, falls Text mit 'CVA' beginnt
HLA-Antigene	EImmHLAPhaenotypisierungET	empfaenger_immunologie	zur Berechnung
	SPostmLaborHLAAntigeneET	spender_postmortem_ labor_hla	der Mismatches

Tabelle 4: Beschreibung der Variablen, welche bei den deutschen Daten zur Ermittlung des Endpunktes verwendet wurden

Bedeutung	Name der Variable	Tabelle	Erläuterung
Todesdatum	FNTodesdatumIQTIG	followup_niere	falls Endpunkt durch Tod definiert
Todesdatum	EBasisTodesdatumET	empfaenger	falls Endpunkt durch Tod definiert
Datum Organversagen	TPostOPOrganversagenDateET	transplantation	falls Endpunkt durch Organversagen definiert
Datum Organversagen	FNGraftversagenDateIQTIG	followup_niere	falls Endpunkt durch Organversagen definiert
Datum Transplantation	TTxDateET	transplantation	falls Endpunkt durch Retransplantation definiert
Follow-Up Termine	FNDateET	followup_niere	falls Endpunkt durch Zensierung definiert, nur spätester Eintrag pro Empfänger wird betrachtet
Follow-Up Termine	FNDateIQTIG	followup_niere	falls Endpunkt durch Zensierung definiert, nur spätester Eintrag pro Empfänger wird betrachtet
letzter Follow-Up-Termin	TFollowUpLetztesDateET	transplantation	falls Endpunkt durch Zensierung definiert

Für die Selektion der **Analysepopulation** wurden die in Tabelle 5 aufgeführten Variablen verwendet. Dabei wurde zunächst für jeden Empfänger mithilfe dessen Empfänger-ID und den Datumswerten der Transplantationen nur die erste Transplantation selektiert. Anschließend wurden nur die Beobachtungen ausgewählt, bei denen Nieren nicht in Kombination mit anderen Organen transplantiert wurden. Empfänger unter 18 Jahren und Spender-Empfänger-Paare mit nicht kompatiblen Blutgruppen (gemäß Tabelle 25, welche im Anhang zu finden ist) wurden anschließend aussortiert. Durch den Merge mit der Tabelle *spender_postmortem* wurden Beobachtungen aussortiert, bei welchen es sich um eine Lebendspende handelt. Zuletzt wurden Beobachtungen mit fehlenden Werten sowie Patienten, bei welchen das Ereignis bereits am Tag der Transplantation beobachtet wurde oder die bereits an diesem Tag zensiert wurden, exkludiert. Ein Flussdiagramm, welches den Verlauf der Entwicklung der Analysepopulation darstellt, wird in Abbildung 7 gezeigt. Hierbei ist die Anzahl der initialen Beobachtungen gegeben durch die Anzahl der Beobachtungen in der Tabelle *transplantation*. Das Diagramm zeigt, wie sich die Anzahl der Beobachtungen durch die einzelnen Selektionskriterien verringert.

Tabelle 5: Beschreibung der Variablen, welche bei den deutschen Daten zur Selektion der Analysepopulation verwendet wurden

Bedeutung	Variable	Tabelle	Erläuterung
Alter Empfänger	EBasisGeburtsdatumET	empfaenger	Differenz zu TTxDateET geteilt durch 365.25
Alter Empfänger	EBasisGeburtsdatumIQTIG	empfaenger	Differenz zu TTxDateET geteilt durch 365.25
Blutgruppe Empfänger	EBasisBlutgrET	empfaenger	Ermittlung der Kompatibilität zw. Spender und Empfänger
Blutgruppe Spender	SPostmBasisBlutgrET	spender_postmortem	Ermittlung der Kompatibilität zw. Spender und Empfänger
transplantiertes Organ	TOrganET	transplantation	nach Nieren selektieren, simultane mit anderen Organen aussortieren
Datum der Transplantation	TTxDateET	transplantation	nur Ersttransplantationen Bestimmung des Empfängeralters

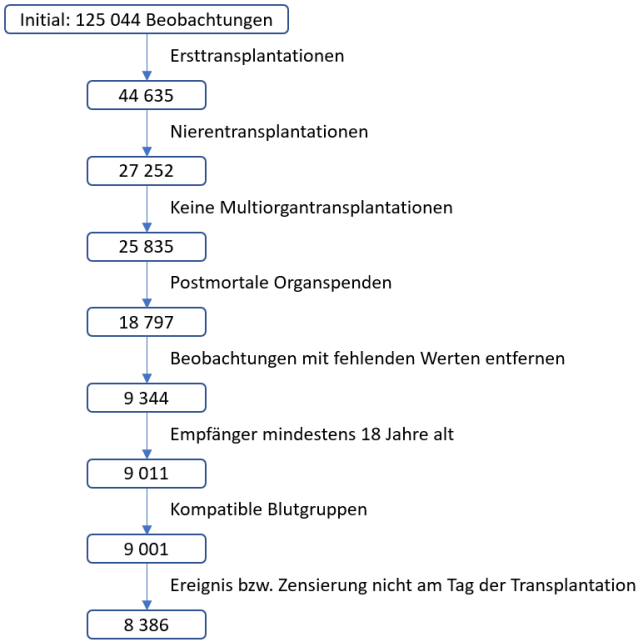

Abbildung 7: Das Flussdiagramm zeigt, wie aus den 125 044 Beobachtungen der Tabelle *transplantation* durch die Anwendung verschiedener Selektionskriterien die deutsche Analysepopulation entsteht. Diese besteht aus 8 386 Beobachtungen.

4.4.2 US-amerikanische UNOS-Daten

Alle Kovariaten konnten bei den UNOS-Daten direkt aus einem einzelnen Feld der Tabelle *kidpan_data* abgelesen werden. Tabelle 6 zeigt, um welche Felder es sich dabei handelt und wie dabei die Einträge der kategorischen Kovariaten kodiert sind. Um eine einheitliche Definition zu den für die deutschen Daten extrahierten Kovariaten zu gewährleisten, wurden für den Kreatininwert ebenfalls das Minimum des gegebenen Wertes und 8 mg/dL verwendet.

Tabelle 6: Beschreibung der Variablen, welche zur Bestimmung der Kovariaten für die US-amerikanischen Daten verwendet wurden

Bedeutung	Variable	Bemerkung
Todesursache	COD_CAD_DON	CVA: 2 ; kein CVA: sonst
Transplantationstyp	TX_PROCEDUR_TY_KI	Einzeln: 101/102; EnBloc: 103; Doppelt: 104
Kreatininwert	CREAT_DON	auf 8 beschränken
Hypertonie	HIST_HYPERTENS_DON	Ja: "P"; Nein: "N"
Diabetes	DIABETES_DON	Ja: "P"; Nein: "N"
Hepatitis C	HEP_C_ANTI_DON	Ja: "P"; Nein: "N"
Kalte Ischämiezeit	COLD_ISCH_KI	
HLA-DR Mismatches	DRMIS	
HLA-B Mismatches	BMIS	
Alter	AGE_DON	
Größe	HGT_CM_DON_CALC	
Gewicht	WGT_KG_DON_CALC	

Die beiden Komponenten des **Endpunktes** konnten ebenfalls direkt aus der Tabelle *kidpan_data* abgelesen werden. Tabelle 7 zeigt die beiden hierfür benötigten Variablen. Es erfolgte zusätzlich eine administrative Zensierung auf den 31.12.2016 und auf den Tag 7 Jahre nach der Transplantation.

Tabelle 7: Beschreibung der Variablen, welche bei den US-amerikanischen Daten zur Ermittlung des Endpunktes verwendet wurden

Bedeutung	Variable
Zeit bis zum Ereignis oder bis zur Zensierung (Tage)	GTIME_KI
Wurde Ereignis beobachtet (Zensierungsindikator)	GSTATUS_KI

Für die Selektion der **Analysepopulation** wurden die in Tabelle 8 aufgeführten Variablen verwendet und nach den dort angegebenen Kriterien gefiltert. Zusätzlich wurde für jede Empfänger-ID nur der Eintrag mit dem frühesten Transplantationstermin verwendet.
Da in Deutschland nur Organspenden von hirntoten Spendern erlaubt sind, Spenden von herztoten Spendern hingegen verboten sind, enthält die deutsche Analysepopulation nur Spenden von hirntoten Spendern. Daher wurde die US-amerikanische Analysepopulation ebenfalls auf diesen Spendertypen eingeschränkt.

Ein Flussdiagramm, welches den Verlauf der Entwicklung der Analysepopulation darstellt, wird in Abbildung 8 gezeigt.

Tabelle 8: Beschreibung der Variablen, welche bei den US-amerikanischen Daten zur Selektion der Analysepopulation verwendet wurden

Bedeutung	Name der Variable	Bemerkung
Anzahl vorheriger Transplantationen	NUM_PREV_TX	Anzahl vorheriger Transplantationen muss 0 sein, nur ersten Eintrag pro Empfänger betrachten
transplantiertes Organ	ORGAN	nach Nierentransplantationen selektieren: "KI"
Multiorgantransplantation	MULTIORG	Beobachtungen mit "Y" aussortieren
Spendertyp	DON_TY	nach postmortalen Spenden selektieren: "C"
Spende nach Herztod	NON_HRT_DON	keine Spenden nach Herztod: "N"
Transplantationsdatum	TX_DATE	zwischen "2006-01-01" und "2016-12-31"
Blutgruppenkompatibilität	ABO_MAT	selektieren nach 1(identical) oder 2(compatibel)
Alter des Empfängers	AGE	Muss größer oder gleich 18 Jahre sein

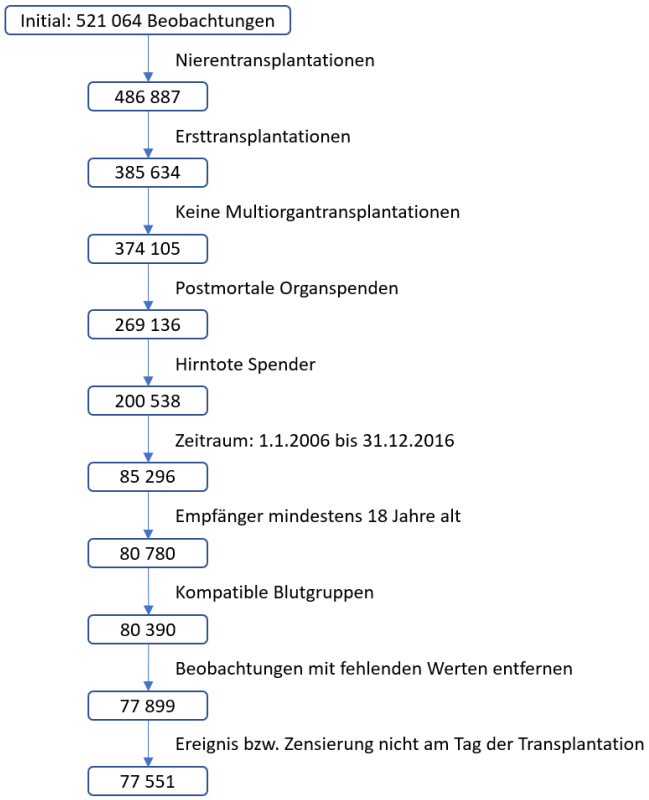

Abbildung 8: Das Flussdiagramm zeigt, wie aus den 521 064 in der Tabelle *kidpan_data* gelisteten Transplantationen durch die Anwendung verschiedener Selektionskriterien die US-amerikanische Analysepopulation entsteht. Diese besteht aus 77 551 Beobachtungen.

4.5 Explorative Datenanalyse: Vergleich der beiden Kohorten

In diesem Abschnitt werden die beiden Kohorten, welche nach den im letzten Kapitel erläuterten Regeln hergeleitet wurden, miteinander verglichen. Dabei werden die beobachteten Ereigniszeiten und die Kovariatenausprägungen miteinander verglichen.
In Abbildung 9 werden die mit der Kaplan-Meier-Methode geschätzten Überlebensfunktionen des Transplantatüberlebens der beiden Kohorten gezeigt. Am unteren Ende der Grafik ist die Anzahl der Individuen unter Risiko in Einjahresschritten für die beiden Kohorten gezeigt. Dabei fällt auf, dass es in der deutschen Kohorte zu jedem Zeitpunkt eine geringere Transplantatüberlebenswahrscheinlichkeit gibt als in der US-amerikanischen Kohorte. Dabei fällt die Kurve der deutschen Kohorte zu Beginn wesentlich stärker als die zu der US-Kohorte gehörende Kurve. Im späteren Verlauf hebt sich dieser Unterschied auf und die Kurven laufen streckenweise annähernd parallel.

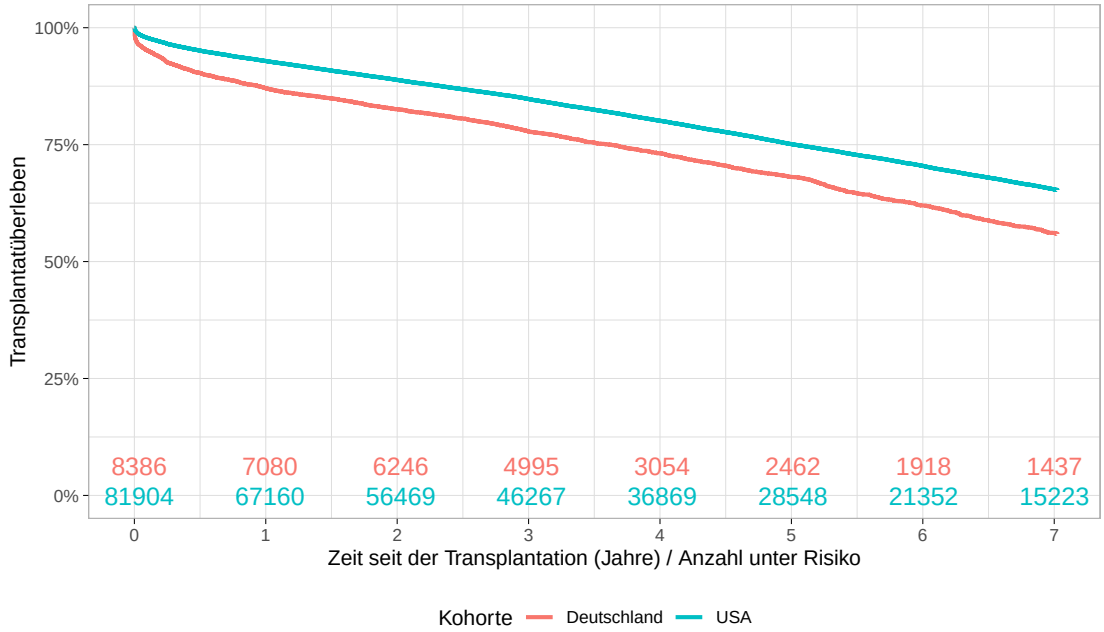

Abbildung 9: Kaplan-Meier-Kurve zum Vergleich des Transplantatüberlebens zwischen den beiden Kohorten im Verlauf über die Zeit

Beim Vergleich der kontinuierlichen Kovariaten, welche in Abbildung 10 gezeigt werden, ist zunächst ein auffälliger Unterschied der Verteilung des Spenderalters zwischen den beiden Kohorten zu erkennen. Während das mediane Alter der deutschen Spender bei 57 Jahren liegt, ist ein Spender in der US-Kohorte im Median 41 Jahre alt. Für den Kreatininwert ist ein Unterschied des Interquartilsabstandes festzustellen. Die kalte Ischämiezeit weist in den USA im Median und im oberen sowie im unteren Quartil höhere Werte auf. Zudem liegen häufigere und stärkere Abweichungen nach oben vor.

Bei den kategorischen Kovariaten gibt es für die Hypertonie und die beiden HLA-Mismatches die stärksten Abweichungen. Dabei sind die HLA-Mismatches zwischen den Ländern nur schwer zu vergleichen, da die Vorgaben der Zuordnung von passenden und nicht passenden Antigenen nicht einheitlich definiert sind und zwischen den beiden Ländern variieren. Innerhalb der US-Kohorte, bei welcher die Anzahl der Mismatches nicht selbst abgeleitet, sondern direkt aus entsprechenden Feldern im Datensatz entnommen wurde, gibt es möglicherweise sogar variierende Definitionen je nach Transplantationsjahr. Dies ist darin begründet, dass Zuordnung von passenden und nicht passenden Antigenen auf einer jährlichen Basis überarbeitet wird. Die Zahl der HLA-Mismatches wird durch die Zuordnung der Spenderorgane zu den Empfängern bestimmt. Dabei wird die Verteilung der Mismatch-Anzahl in einer Population nicht zuletzt durch deren Rolle beim Allokationsverfahren bestimmt. Die Allokationsverfahren der beiden Länder werden im nächsten Kapitel erläutert.

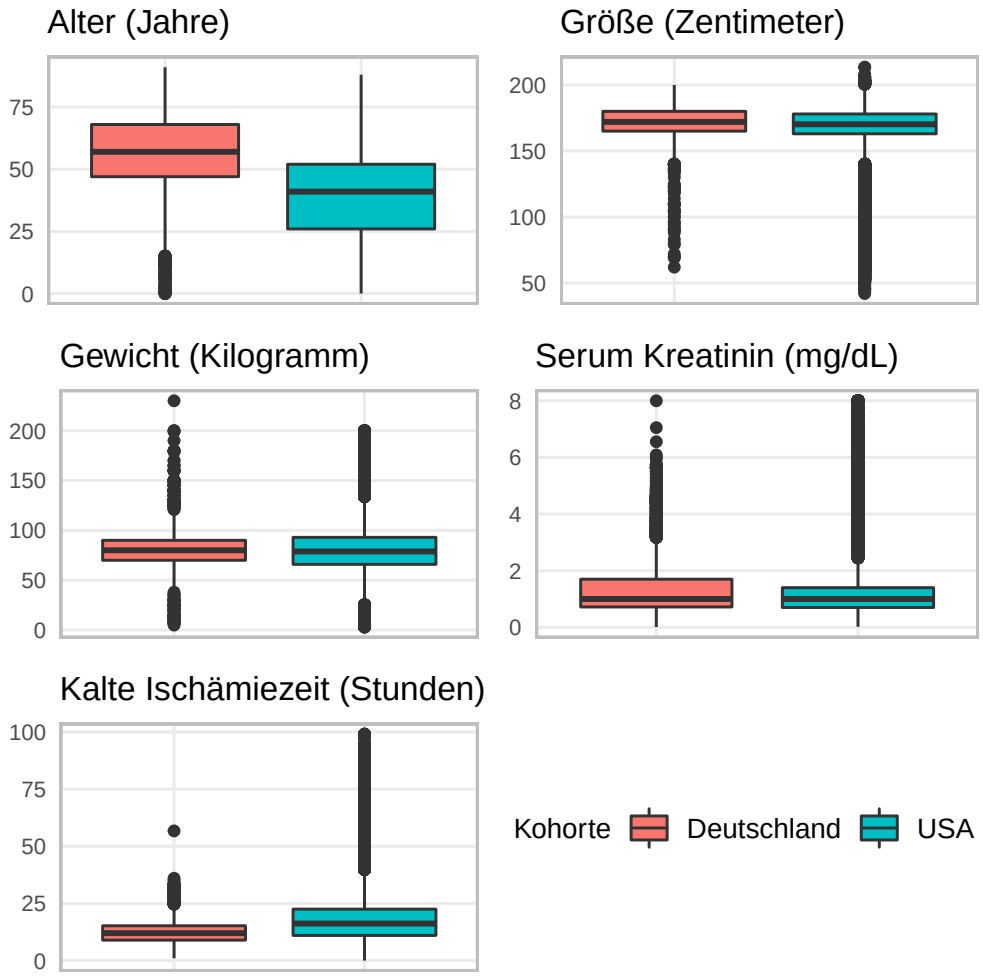

Abbildung 10: Vergleich der stetigen Kovariaten zwischen den beiden Kohorten

Abbildung 11: Vergleich der kategorialen Kovariaten zwischen den beiden Kohorten

4.6 Deutscher und US-amerikanischer Allokationsprozess im Vergleich

In den folgenden beiden Abschnitten werden die Allokationsverfahren der beiden Länder vorgestellt. Diese können auf die Verteilung der Kovariaten einen entscheidenden Einfluss haben. Die Frage, welcher Spender welchem Empfänger zugeordnet wird, beeinflusst die Kovariaten, welche die Kompatibilität zwischen Empfänger und Spender ausdrücken und durch die resultierenden Transportwege wird die kalte Ischämiezeit beeinflusst. Ferner kann das Allokationsverfahren einen Einfluss darauf haben, ob ein Empfänger ein ihm angebotenes Spenderorgan annimmt oder nicht. Auch das hat indirekt einen Einfluss auf die Verteilung der Kovariaten.

Die Informationen zur Nierenallokation in Deutschland basieren auf Kapitel 4 des Eurotransplant Manuals ([ET22]). Die Informationen zum US-amerikanischen Allokationssystems basieren auf den Quellen [SSA+17] und [PP20].

4.6.1 Allokationsprozess in Deutschland über Eurotransplant

Die Allokation der postmortalen Spenderorgane erfolgt für Organe von unter 65-jährigen Spendern über das *Eurotransplant Kidney Allocation System (ETKAS)* und für Organe von mindestens 65-jährigen Spendern über das *Eurotransplant Senior Program (ESP)*. Zudem gibt es das *Acceptable Mismatch Program (AM)*, welches ein spezielles Allokationsverfahren für hoch sensibilisierte Patienten ist.
Das ETKAS-Programm wurde im Jahr 1996 von Eurotransplant eingeführt und das ESP-Programm kam 1999 hinzu. Das AM-Programm gibt es bereits seit 1985. Bei allen in dieser Arbeit untersuchten deutschen Transplantationen kam also der in diesem Kapitel vorgestellte Allokationsprozess zur Anwendung

Im Eurotransplant Kidney Allocation System (ETKAS) erfolgt die Allokation der Spenderorgane über ein Punktesystem, bei welchem für jeden Transplantationskandidaten ein Punktewert berechnet wird. In diesen fließen die folgenden Merkmale ein:

- Grad der Übereinstimmung der Antigene HLA-A, HLA-B und HLA-DR
- Mismatch-Wahrscheinlichkeit: Wahrscheinlichkeit des Kandidaten ein Organ zu erhalten, welches weitgehend mit seinen eigenen HLA-Antigenen übereinstimmt
- Penal Reactive Antibodies (PRA): Anteil der Bevölkerung, gegen den bei dem Kandidaten vorgebildete Antikörper vorliegen
- Medizinische Dringlichkeit unterteilt in normale und hohe Dringlichkeit
- Distanz zwischen Organspender und Kandidaten
- Bonuspunkte für pädiatrische Kandidaten (Kinder)
- Bonuspunkte für Kandidaten, die in der Vergangenheit Lebendspender waren
- Organaustauschbalance zwischen den ET-Mitgliedsstaaten

Der Empfänger mit dem höchsten Punktewert für das betrachtete Spenderorgan bekommt dieses als erstes angeboten.

Im Gegensatz zu ETKAS werden die HLA-Merkmale beim Eurotransplant Senior Program (ESP) nicht berücksichtigt. Stattdessen liegt die Priorität bei diesem Allokationssystem für über 65-jährige Spender darauf, die kalte Ischämiezeit durch möglichst kurze Transportwege gering zu halten. Gibt es in der Spenderumgebung mehrere potentielle Empfänger, entscheidet die Wartezeit darüber, wer das Organ als erstes angeboten bekommt. Wird im Rahmen des ESP kein passender Empfänger für das Organ gefunden, wird es an ETKAS weitergegeben. In

Deutschland können Patienten mit einem Alter von 65 Jahren oder höher auswählen, ob sie über ESP oder über ETKAS bei der Organallokation berücksichtigt werden möchten.
Über das Acceptable Mismatch Program (AM) werden Nierentransplantate an Patienten mit einem besonders hohen Sensibilisierungsgrad vergeben. Dabei handelt es sich um Patienten, welche einen PRA-Wert von 85% oder größer haben. Diese Patienten erhalten Spenderorgane bevorzugt vor anderen Empfängern, um ihre geringere Chance auf ein geeignetes Spenderorgan auszugleichen.

4.6.2 Allokationsprozess in den USA

Das *Kidney Allocation System (KAS)* ist das aktuelle Allokationsverfahren in den USA, welches seit dem Jahr 2014 eingesetzt wird. Der vorherige Allokationsprozess beruhte primär auf der Wartezeit. Der hier beschriebene Allokationsprozess betrifft also nur einen Teil der in dieser Arbeit betrachteten Transplantationen in der US-Kohorte.
Zwei zentrale Elemente des KAS sind der *Kidney Donor Profile Index (KDPI)* und der *Estimated Post-Transplant Survival Score (EPTS)*.

Der KDPI gibt an, wie lange die Spenderniere eines Todspenders im Vergleich zu allen Nieren, die im vergangenen Jahr in den USA entnommen wurden, voraussichtlich funktionieren wird. Niedrigere KDPI-Werte sind dabei mit einer längeren geschätzten Funktion, höhere KDPI-Werte mit einer kürzeren geschätzten Funktion verbunden. Bei einer Niere mit einem KDPI von 20% wird beispielsweise eine kürzere Transplantatsüberlebenszeit erwartet als bei 20% aller gespendeten Nieren (bzw. eine längere Transplantatsüberlebenszeit als bei 80% der gespendeten Nieren). Der KDPI basiert auf dem von Rao *et al.* vorgestellten Risiko-Skore. Jedoch werden nur die Informationen zum Spender genutzt, die Informationen zu HLA-Kompatibilität und Transplantat gehen hingegen nicht in die Berechnung ein. Somit setzt sich der KDPI aus zehn Werten zusammen, nämlich Alter, Größe, Gewicht, Diabetesdiagnose, Hepatitis-C-Status, Hypertoniediagnose, Ethnizität, Kreatininwert, Spendertyp (Herztod oder Hirntod), Todesursache (CVA oder andere Todesursache).

Der EPTS schätzt die Überlebenszeit des Empfängers. Es gehen dabei Dialysezeit, Alter, Diabetesdiagnose und die Information, ob bei diesem Patienten in der Vergangenheit bereits eine Transplantation stattgefunden hat, ein. Der errechnete EPTS-Wert ist analog zum KDPI zu interpretieren: Für einen Patienten mit einem ETPS von 20% wird eine kürzere Überlebenszeit geschätzt als bei 20% aller Empfänger des vergangenen Jahres (bzw. eine längere Überlebenszeit als bei 80% der Empfänger).
Für die Organallokation wird je nach Bereich, in welchen der KDPI des Organs fällt, ein anderes Punktesystem verwendet. Dabei werden die besten 20% Spendernieren (KDPI ≤ 20) an die 20% der Patienten vergeben, die voraussichtlich am meisten davon profitieren werden (EPTS ≤ 20).
Außerdem gehen in das Punktesystem die Blutgruppen von Spender und Empfänger, die HLA-DR Mismatches, die geographische Distanz zwischen Spender und Empfänger, geschätzter PRA, die Wartezeit und die Dringlichkeit ein. Zudem gibt es Bonuspunkte, wenn der Kandidat ein Kind ist oder in der Vergangenheit Lebendspender war.

5 Methodik

In diesem Kapitel sollen das Vorgehen bei der Erstellung sowie der Rekalibrierung der Modelle erläutert werden. Dabei wird insbesondere die neu entwickelte Rekalibrierungsmethode für Random Survival Forests vorgestellt.

5.1 Publizierte Vorhersagemodelle

Ziel dieser Arbeit ist es, auf UNOS-Daten beruhende Prognosemodelle zur Ereigniszeitanalyse nach Nierentransplantationen auf deutschen Daten extern zu validieren. Dabei soll zudem untersucht werden, inwiefern eine Rekalibrierung zu einem besser auf die neue Kohorte angepassten Modell führen kann. Zudem soll ein Vergleich zu einem mit den deutschen Daten neu trainierten Modell erfolgen. Hierbei liegt der Fokus auf zwei verschiedenen Modelltypen, der Cox-Regression und dem Random Survival Forest.
Dazu erfolgte eine Suche nach entsprechenden publizierten Prognosemodellen, also Modellen, welche auf der Cox-Regression oder dem Random Survival Forest beruhen und welche mit den von UNOS bereitgestellten US-amerikanischen Daten erstellt worden sind.
Dies wurde umgesetzt, indem aktuelle Publikationen mit systematischen Reviews solcher Prognosemodelle untersucht wurden und die in diesen Publikationen vorgestellten Modelle nach den oben genannten Kriterien selektiert wurden. Dabei wurden für die Suche nach Random Survival Forest Modellen die Publikationen von Díez-Sanmartín und Cabezuelo (2020) [DSC20] und Senanayake *et al.* [SWG+19] herangezogen. Für die Suche nach auf der Cox-Regression beruhenden Modellen wurde die Publikation von Ramspek *et al.* (2021) [RMW+21] betrachtet.
Tabelle 9 zeigt einen Vergleich der gefundenen Publikationen, welche dieses Kriterium erfüllen. Betrachtet man die Modelle im Detail, kann man eine Heterogenität in den Analysepopulationen, den Kovariaten und der Definition des Endpunktes feststellen. Bei der Analysepopulation gibt es neben den betrachteten Transplantationsjahren auch Unterschiede in der Selektion der Patienten. Beispielsweise werden unterschiedliche Kriterien für das Alter der Empfänger und die Art der Spende angelegt. In der Spalte zu den Kovariaten ist vermerkt, aus welchen Bereichen die verwendeten Kovariaten stammen. Bei den Endpunkten wird in manchen Fällen nur der Tod als Ereignis betrachtet, in anderen Fällen ist das Ereignis als Tod oder Organversagen definiert. Hierbei ist jedoch die Definition des Organversagens zwischen den einzelnen Publikationen nicht einheitlich.
Das Validieren, Rekalibrieren und Aktualisieren von Modellen ist nur möglich, wenn das veröffentlichte Prognosemodell zur Verfügung gestellt wird. In der Praxis wird beispielsweise der Baseline-Hazard für ein Cox-Modell nur selten veröffentlicht, oder es werden bei Random Survival Forests nur Ergebnisse zu bestimmten vordefinierten Zeitpunkten zugänglich gemacht. Außerdem können bereits vorhandene Modelle dann nicht auf neuen Daten extern validiert oder rekalibriert werden, wenn in diesem Modell verwendete Kovariaten in der neuen Kohorte nicht erhoben werden oder nötige Informationen zum Endpunkt oder zur Selektion der Analysepopulation nicht vorhanden sind.
Bei den hier in Betracht gezogenen Publikationen stellte sich heraus, dass keines der publizierten Modelle auf die deutschen Daten angewendet werden kann. Der Grund hierfür ist

Tabelle 9: Vergleich verschiedener Publikationen, bei welchen auf Basis der US-amerikanischen UNOS-Daten zu Nierentransplantationen nach Todspende Cox-Regressionmodelle oder Random Survival Forest Modelle erstellt wurden

Publikation	Methoden	Analysepopulation	Kovariaten	Endpunkt	Evaluation
Rao *et al.* 2009 [RS09]	Cox-Regression	erste Transplantation, erwachsene Empfänger, Todspende, 1995 - 2005	Spender, Organ, Kompatibilität	Organversagen oder Tod	C-Index
Kasiske *et al.* 2010 [KIS$^+$]	Cox-Regression	einzelne Transplantation, erwachsene Empfänger, Todspende, 2000-2016	Spender, Empfänger, Kompatibilität	Organversagen oder Tod	C-Index
Molnar *et al.* 2017 [MNC$^+$17]	Cox-Regression	erste Transplantation, erwachsene Empfänger an der Dialyse, Todspende, 2001-2006	Spender, Empfänger, Kompatibilität	Organversagen oder Tod	C-Index, Kalibrierungs-kurven
Vinson *et al.* 2019 [VKDT19]	Cox-Regression	einzelne Transplantation, erwachsene Empfänger, Todspende, 2000-2014	Spender, Empfänger, Organ, Kompatibilität	Organversagen oder Tod	C-Index, Kalibrierungs-kurven
Bae *et al.* 2019 [BMT$^+$19]	Random Survival Forest und Weibull-Regression	Todspende, 2005-2016	Spender, Empfänger, Organ	Tod	C-Index
Mark *et al.* 2019 [MGG$^+$19]	Kombination aus Random Survival Forest und Cox-Regression	erwachsene und minderjährige Empfänger, Todspende und Lebendspenden, 1987-2014	Spender, Empfänger, Organ Kompatibilität	Tod	C-Index, IBS

zum einen, dass alle Modelle die ethnische Zugehörigkeit von Spender oder Empfänger als Kovariate verwenden, diese jedoch in den deutschen Daten nicht erhoben wird. Außerdem werden in [RS09] und [VKDT19] die Baseline-Hazards nicht zur Verfügung gestellt. In [BMT+19] wird die geschätzte Überlebensfunktion nur für zwei Zeitpunkte (5 Jahre und 10 Jahre) veröffentlicht.
Deshalb wurden neue Modelle auf den UNOS-Daten erstellt, um diese anschließend auf den deutschen Daten zu validieren und zu rekalibrieren.
Das Neuerstellen der Modelle auf den UNOS-Daten bietet zudem den Vorteil, dass für die beiden untersuchten Modelltypen Modelle mit gleich definierter Analysepopulation, gleich definierten Endpunkten, sowie den gleichen eingehenden Kovariaten erstellt werden konnten.

Ferner erfolgte eine Suche nach Veröffentlichungen, bei denen auf den US-Daten basierende Modelle mit deutschen Transplantationsdaten extern validiert wurden. Verschiedene Autoren haben in den vergangenen Jahren eine externe Validierung des KDPI (vgl. Allokationsprozess Kapitel 4.6) für europäische Kohorten vorgenommen. Dabei betrachteten Lehner *et al.* 2018 [LKH+18] sowie Dahmen [DBP+19] *et al.* 2019 Kohorten aus Deutschland.
Beide Arbeiten zeigen, dass der KDPI zur Unterstützung klinischer Entscheidungsfindungen auch für europäische Kohorten geeignet ist. Jedoch zeigt sich, dass der mediane KDPI und das mediane Spenderalter in der deutschen Kohorte von den US-Transplantationen abweicht, was auf das andersgeartete Allokationsverfahren zurückzuführen sein könnte. Die Autoren empfehlen daher weitergehende Untersuchungen auf größeren Datensätzen.

Reviews mit weiteren publizierten Prognosemodellen für den Ausgang nach Nierentransplantationen, auch für Kohorten weiterer Länder und mit anderen statistischen Methoden, können in den Arbeiten [RZT+22] und [KHH+17] gefunden werden.

5.2 Trainieren der Vorhersagemodelle

Für das Trainieren der auf den UNOS-Daten basierenden Modelle wurde für den US-amerikanischen Datensatz ein Train-Test-Split in einem Verhältnis von 2:1 vorgenommen. Zwei Drittel der Beobachtungen wurden also zum Trainieren der Modelle verwendet, während ein Drittel der Daten zum Evaluieren der erstellten Modelle diente.
Für beide Modelltypen wurden beim Training Zeitschritte von 3 Monaten verwendet. Die geschätzten Survivalfunktionen sind also durch eine geschätzte Wahrscheinlichkeit der Ereignisfreiheit alle drei Monate gegeben.

Für den Random Survival Forest wurde ferner für die Wahl der Hyperparameter eine 10-Fold-Crossvalidation durchgeführt. Dabei wurden verschiedene Kombinationen an Anzahl der Bäume und Mindestanzahl der Beobachtungen pro Blatt getestet. Die Kombination mit dem besten integrierten Brier Score wurde ausgewählt. Für die Anzahl der Beobachtungen pro Blatt kamen erst Werte ab 1000 in Frage, da die verwendeten Rekalibrierungsverfahren ansonsten nicht anwendbar gewesen wären (vgl. Kapitel 5.3.2).

Auch für die deutschen Transplantationsdaten wurden eine Cox-Regression und ein Random

Survival Forest angepasst. Die Aufteilung in Trainingsdaten und Testdaten erfolgte auch hier in einem Verhältnis von zwei Dritteln Trainingsdaten zu einem Drittel Testdaten. Genau wie bei dem Random Survival Forest für die US-amerikanischen Daten erfolgte die Auswahl der Hyperparameter für den deutschen Random Survival Forest mittels einer 10-Fold-Crossvalidation. Allerdings gab es in diesem Fall keine vorherigen Einschränkungen für die Mindestanzahl der Beobachtungen in den Blättern.

Mithilfe der Testdaten der deutschen Kohorte wurden die US-Modelle anschließend extern evaluiert. Es sollte hierbei die Frage beantwortet werden, inwiefern eine Motivation für eine Rekalibrierung besteht. Es wurde also untersucht, inwiefern das auf den US-Daten angepasste Modell für die deutschen Daten angewendet werden kann. Dabei wurden Kalibrierung, C-Index und Brier Score untersucht und die Performance zwischen dem nicht rekalibrierten US-Modell und dem auf den deutschen Daten neu trainierten Modell verglichen. Zur Bewertung der Kalibrierung wurden Kalibrierungskurven erstellt. Zudem wurden Calibration-Slope und Calibration-Intercept bestimmt.
Dabei erfolgte die Berechnung der beiden Werte für beide Modelltypen nach der in Kapitel 3.2 beschriebenen Methode. Es wurde also auch für den Random Survival Forest die mit der Cox-Regression assoziierte Link-Funktion verwendet. Die Kalibrierungskurven wurden für die Zeitpunkte zu jedem vollen Jahr nach der Transplantation betrachtet.
Zu den ermittelten Werten von Calibration-Slope γ_0 und Calibration-Intercept γ_1 wurden zusätzlich Wald-Tests durchgeführt. Für den Calibration-Intercept wurde die Hypothese $\gamma_0 = 0$ und für den Calibration Slope die Hypothese $\gamma_1 = 1$ geprüft. Außerdem erfolgte ein multivariater Test mit der Hypothese $(\gamma_0, \gamma_1) = (0, 1)$.

Die Vorverarbeitung der Datensätze sowie die Auswahl der in die Modelle eingehenden Kovariaten ist in Kapitel 4.3 beschrieben.

5.3 Rekalibrierung der US-Modelle auf die deutsche Kohorte

Anschließend wurden die auf die US-Daten angepassten Modelle für die deutschen Daten rekalibriert. Dazu wurden die Trainingsdaten für die Rekalibrierung verwendet und die Evaluation der rekalibrierten Modelle erfolgte mit den Testdaten. Dabei wurden erneut C-Index, Brier Score, Kalibrierungskurven, Calibration-Slope und Calibration-Intercept betrachtet.
Zum Rekalibrieren des Cox-Regressionsmodells wurde der Ansatz von Van Houwelingen ([Hou00]) verwendet.
Für den Random Survival Forest wurden zwei Methoden angewendet. Zum einen wurde die Methode zum Rekalibrieren des Random Forests zur Wahrscheinlichkeitsschätzung von Dankowski und Ziegler ([DZ16]) aus Kapitel 3.3.3 verwendet. Dafür wurde der für rechtszensierte Daten ausgelegte Random Survival Forest durch eine naive Binarisierung auf ein Klassifikationsverfahren reduziert.
Zum Anderen wurde eine neue Methode angewendet, welche als eine Erweiterung der vorherigen Methode auf rechtszensierte Daten angesehen werden kann.

5.3.1 Rekalibrierung des Cox-Modells

Zum Rekalibrieren des Cox-Regressionsmodells wurde der Ansatz von Van Houwelingen ([Hou00]) verwendet, welcher in Kapitel 3.3.1 vorgestellt wurde. Es wurden dabei wie in Gleichung 31 sowohl der Baseline-Hazard neu geschätzt, als auch ein Skalierungsfaktor für die Modellparameter berechnet.

5.3.2 Rekalibrierung des Random Survival Forests

Die von Dankowski und Ziegler vorgeschlagene Rekalibrierungsmethode, welche in Kapitel 3.3.3 vorgestellt wurde, ist für die Anpassung von Random Forests zur Wahrscheinlichkeitsschätzung geeignet. Es handelt sich also um die Rekalibrierung von Modellen zur Klassifikation mit einer binären Zielgröße.
Zum Rekalibrieren eines Random Survival Forests kann diese Methode für verschiedene Zeitpunkte einzeln vorgenommen werden, sodass zu diesen Zeitpunkten die vorhergesagte Wahrscheinlichkeit dafür, dass das Ereignis noch nicht eingetreten ist, angepasst wird. Für die Rekalibrierung der aus den Entscheidungsbäumen abgeleiteten logistischen Regressionsmodelle können dazu wie in Gleichung 33 Intercept und Slope angepasst werden. Um ein Modell, das eine Überlebensfunktion schätzt, zu rekalibrieren, muss diese Methode also für jeden Zeitpunkt, an dem die Überlebensfunktion ausgewertet werden soll, angewendet werden. Um den Überlebenszeit-Endpunkt (T, δ) der Beobachtungen für einen bestimmten Zeitpunkt t als binäre Zielgröße zu analysieren, muss dieser entsprechend transformiert werden. Dabei soll der neue Endpunkt Y_t aussagen, ob das Ereignis bis zum Zeitpunkt t bereits eingetreten ist. Wenn ein Patient bis zum Zeitpunkt t das Ereignis noch nicht erlitten hat gilt $Y_t = 1$, andernfalls gilt $Y_t = 0$. Falls es sich um eine zensierte Beobachtung handelt, bei welcher die Zensierung vor dem Zeitpunkt t stattgefunden hat, kann der Wert von Y_t nicht ermittelt werden. Die entsprechende Beobachtung kann für eine Klassifikation der Ereignisfreiheit zum Zeitpunkt t nicht verwendet werden.

$$Y_t = \begin{cases} 0, & \text{falls } T < t,\ \delta = 1 \\ 1, & \text{falls } T \geq t \\ \text{NA}, & \text{sonst} \end{cases} \tag{34}$$

Durch die Binarisierung wird die Anzahl der Beobachtungen, die zum Rekalibrieren des Modells verwendet werden können, mit zunehmender Zeit t weniger. Diese systematische Selektion der Beobachtungen kann mit sinkender Zahl an eingehenden Beobachtungen zu einer Verzerrung der Verteilung des Datensatzes führen.
Aus diesem Grund wird eine Ausweitung dieser Methode auf Daten mit einem Überlebenszeit-Endpunkt (T, δ) entwickelt, welcher nun vorgestellt und erläutert wird.

In einem neu entwickelten Ansatz zur Rekalibrierung des Random Survival Forests soll die Anpassung des Modells auf eine neue Kohorte auch direkt für rechtszensierte Daten möglich sein. Dies bietet neben einem weniger komplexen Verfahren den Vorteil, dass keine Binarisierung notwendig ist und damit eine durch Selektion der Beobachtungen ausgelöste

systematische Verzerrung verhindert werden kann. Dazu werden die einzelnen Bäume nicht in ein logistisches Regressionsmodell, sondern in eine Cox-Regression umgewandelt. Die einzelnen Schritte dieses Verfahrens werden in Abbildung 12 gezeigt. In Schritt 1 und Schritt 2 durchläuft dazu der Datensatz, welcher zum Training des Random Survival Forests verwendet wurde, erneut den Baum. Dabei wird ermittelt, in welche Blätter die einzelnen Beobachtungen fallen. Genau wie in der Methode von Dankowski und Ziegler werden Dummy-Variablen für die Blattzugehörigkeit als neue Zielvariable erstellt (vgl. Abbildung 5). Eines der Blätter stellt dabei die Referenz-Kategorie dar.
In Schritt 3 wird eine Cox-Regression mit den Dummy-Variablen der Blattzugehörigkeiten als neuen Kovariaten erstellt.
In Schritt 4 und Schritt 5 wird auch für die Beobachtungen der neuen Kohorte bestimmt, zu welchem Blatt des Baumes sie gehören. Entsprechend werden auch für diese Daten Dummy-Variablen für die einzelnen Blätter erzeugt. Im abschließenden sechsten Schritt kann das Cox-Modell aus Schritt 3 mit der Dummy-Darstellung der neuen Kohorte aus Schritt 5 rekalibriert werden. Dazu werden wie in Gleichung 31 Intercept und Slope angepasst. Das Verfahren wird anschließend für jeden weiteren Baum des Random Survival Forest wiederholt. Für einen bestimmten Zeitpunkt ist der Wert der vom rekalibrierten Random Survival Forest vorhergesagten Überlebensfunktion dann der Mittelwert der von den einzelnen Cox-Regressionen vorhergesagten Werte zu diesem Zeitpunkt.
Für sehr tief verzweigte Bäume kann es dazu kommen, dass diese nicht in ein Cox-Regressionsmodell umgewandelt werden können. Das liegt daran, dass der Maximum-Likelihood-Schätzer bei einer hohen Anzahl an Dummy-Kovariaten mit nur wenigen zugehörigen Beobachtungen pro Dummy-Variable nicht konvergiert. Um dies zu verhindern, muss beim Trainieren des Random Survival Forests darauf geachtet werden, eine ausreichende Mindestanzahl an Beobachtungen pro Blatt als Hyperparameter anzugeben. Wenn es sich nur um einen sehr geringen Anteil der Bäume des Random Survival Forests handelt, für welche keine Cox-Regression angepasst werden kann, so können diese Bäume bei dem rekalibrierten Modell auch verworfen werden. Dementsprechend verringert sich die Gesamtzahl der Lerner des Ensembles.

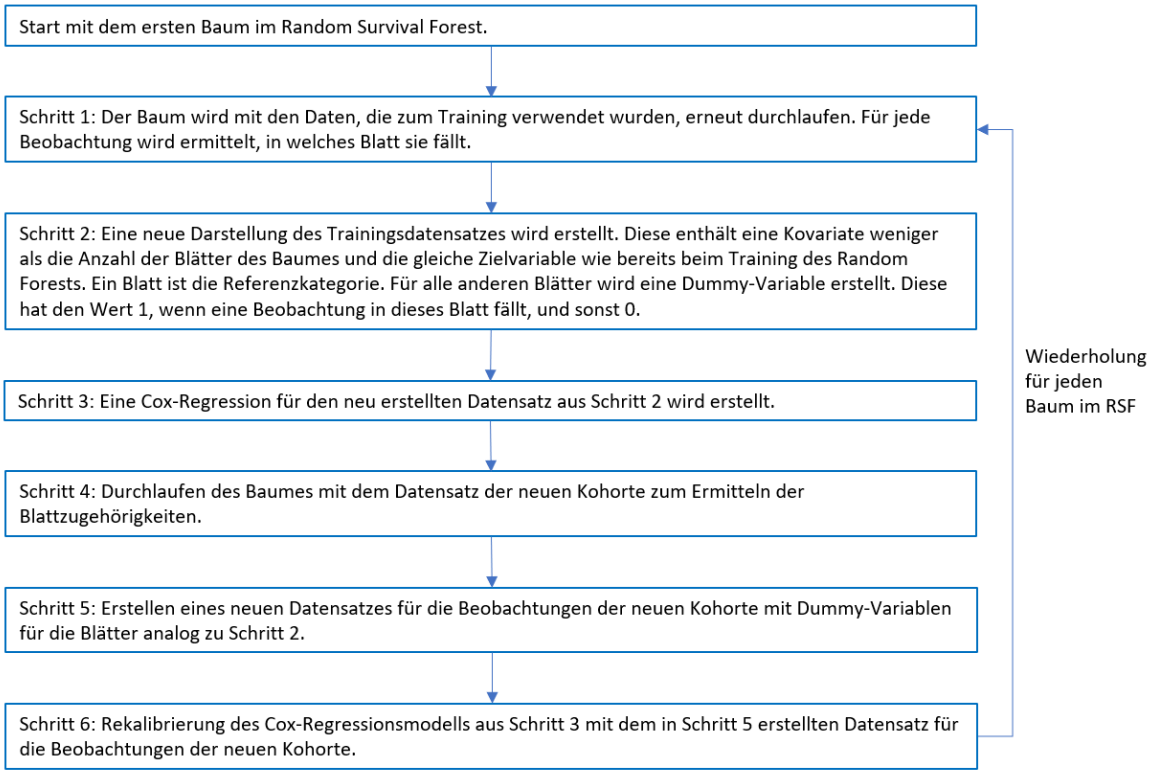

Abbildung 12: Das Diagramm zeigt analog zu Abbildung 4 den Ablauf der in dieser Arbeit entwickelten Methode zur Rekalibrierung eines Random Survival Forests mittels Umwandlung der Bäume in einzelne Cox-Regressionsmodelle

6 Ergebnisse

In diesem Kapitel werden die Ergebnisse der im vorherigen Kapitel erläuterten Methoden vorgestellt. Dabei wird zunächst auf die Machbarkeit einer Rekalibrierung der US-Modelle auf die deutsche Kohorte eingegangen. Anschließend werden die erstellten Modelle vorgestellt. Zuletzt werden die Ergebnisse der Rekalibrierung vorgestellt.

6.1 Machbarkeit

Das Rekalibrieren eines publizierten Modells ist nur möglich, wenn das gesamte Modell zur Verfügung gestellt wird. Außerdem kann eine externe Validierung und eine Rekalibrierung nur dann erfolgen, wenn die in diesem Modell verwendeten Kovariaten in der neuen Kohorte ebenfalls erhoben wurden und außerdem auch nötige Informationen zum Endpunkt oder zur Selektion der Analysepopulation gegeben sind. Da in der Literatur kein publiziertes Modell gefunden werden konnte, welches alle diese Kriterien erfüllt, war eine Rekalibrierung eines bereits vorhandenen auf den UNOS-Daten beruhenden Modells auf die deutsche Kohorte nicht machbar.
Stattdessen mussten für die UNOS-Daten zunächst eigene Modelle trainiert werden, um diese anschließend für die deutschen Daten zu rekalibrieren.
Große Hürden bei der Umsetzung der Rekalibrierung lagen in der Datenvorverarbeitung und dem Feature Engineering Prozess bei den Daten des deutschen Transplantationsregisters. So mussten beispielsweise für die Ermittlung der Kovariate *Transplantationstyp* Annahmen getroffen werden und für die Kovariaten, welche die Histokompatibilität messen, bleibt unklar, wie stark die Definitionen der Mismatches für die beiden Kohorten voneinander abweichen. Grundsätzlich konnte eine Rekalibrierung vorgenommen werden, jedoch mit der Einschränkung, dass die Validität der deutschen Daten und die Einheitlichkeit der Definition der Kovariaten aufgrund der genannten Schwierigkeiten nicht vollständig sichergestellt ist.

6.2 Trainierte Modelle

In diesem Abschnitt werden die auf den US-Daten und auf den deutschen Daten trainierten Modelle vorgestellt. Dabei wird deren Kalibrierung auf den eigenen Testdaten untersucht, die Modelle werden miteinander verglichen und die *Variable Importance* der Modelle wird betrachtet. Letztere wird quantifiziert, indem für die einzelnen Kovariaten analysiert wird, wie stark sich der Brier Score durch Permutation dieser Kovariate ändert. So kann ausgedrückt werden, wie relevant eine Kovariate für das Modell ist.

6.2.1 Cox-Regressionsmodelle

Die Abbildungen 10 und 11 zeigen die Modellzusammenfassungen der Cox-Regressionsmodelle für die US-amerikanische bzw. die deutsche Kohorte. Dabei werden die geschätzten Modellparameter sowie p-Werte und Signifikanzniveaus berichtet. Beim US-Modell zeigt sich für den Parameter jeder Kovariate ein signifikanter p-Wert. Beim

deutschen Modell werden hingegen für die Parameter der Kovariaten Gewicht, Hepatitis, Diabetes und CVA keine p-Werte unterhalb des Signifikanzniveaus 0.1 festgestellt.

Tabelle 10: Zusammenfassung des auf den deutschen Daten angepassten Cox-Regressionsmodells

Kovariate	geschätzter Parameter	p-Wert
Groesse	-0.013020	2.5e-05 ***
Gewicht	0.002518	0.161056
Alter	0.023388	<2e-16 ***
Kalte Ischämiezeit	0.018591	0.000158 ***
Transplantationstyp = Einzeln	0.543568	0.019639 *
Transplantationstyp = EnBloc	0.367441	0.643697
Hepatitis = Nein	0.079868	0.628663
Hypertonie = Nein	-0.095968	0.074029 .
Diabetes = Nein	-0.060282	0.353980
Kreatinin	0.047733	0.054013 .
CVA = Nein	-0.037209	0.486768
HLA-B-Mismatches = 1	0.201970	0.019896 *
HLA-B-Mismatches = 2	0.329516	0.000311 ***
HLA-DR-Mismatches = 1	0.060966	0.386434
HLA-DR-Mismatches = 2	0.190182	0.013522 *

Signifikanzniveau: 0.001(***), 0.01(**), 0.05(*), 0.1(.)

Das deutsche Cox-Modell weist einen C-Index von 0.6313 und einen Brier Score von 0.1641 auf den eigenen Testdaten auf. Bei dem US-Cox-Modell wird ein C-Index von 0.5989 und ein Brier Score von 0.1319 auf den Testdaten der US-Kohorte erreicht. Das US-Modell hat somit den geringeren Vorhersagefehler, während das deutsche Modell besser diskriminiert.
Für beide Modelle wurden zur Untersuchung der Kalibrierung Calibration-Slope und Calibration-Intercept berechnet. Zusätzlich wurde mit dem Wald-Test untersucht, ob sich die berechneten Werte signifikant von den idealen Werten unterscheiden. Diese wären null für den Intercept und eins für den Slope. Auch ein multivariater Test wurde durchgeführt, bei dem überprüft wird, ob sowohl der Intercept Eins als auch der Slope Null ist. Die Ergebnisse dazu sind im Anhang in Tabelle 18 zu finden. Das deutsche Modell ist auf den deutschen Testdaten gut kalibriert. Das US-Modell weist hingegen für die Zeitpunkte ein Jahr und zwei Jahre nach der Transplantation eine weniger gute Kalibrierung auf.

Die *Variable Importance* der beiden Cox-Regressionsmodelle wird in Abbildung 13 gezeigt. Die angegebenen Werte beschreiben, um wie viel Prozent sich der Brier Score für die Testdaten verschlechtert, wenn die entsprechende Variable permutiert wird. Es fällt auf, dass die Kovariate *Alter* bei beiden Kohorten mit Abstand den größen Effekt hat. Hierbei ist der Effekt für die deutschen Daten noch extremer als bei den US-amerikanischen. Bei anderen Kovariaten, wie beispielsweise der *Hepatitis*, zeigt sich ein deutlicher Unterschied bei den

Tabelle 11: Zusammenfassung des auf den US-amerikanischen Daten angepassten Cox-Regressionsmodells

Kovariate	geschätzter Parameter	p-Wert
Groesse	-0.0042660	9.98e-08 ***
Gewicht	-0.0019181	0.000316 ***
Alter	0.0134972	<2e-16 ***
Kalte Ischämiezeit	0.0051795	5.438 5.39e-08 ***
Transplantationstyp = Einzeln	0.1733836	0.023711 *
Transplantationstyp = EnBloc	0.0258768	0.823140
Hepatitis = Nein	-0.5195675	<2e-16 ***
Hypertonie = Nein	-0.1597156	3.54e-12 ***
Diabetes = Nein	-0.2772662	<2e-16 ***
Kreatinin	0.0300936	0.011807 *
CVA = Nein	-0.050041	0.022938 *
HLA-B-Mismatches = 1	0.0920078	0.025072 *
HLA-B-Mismatches = 2	0.0604673	0.120067
HLA-DR-Mismatches = 1	0.1215850	0.000168 ***
HLA-DR-Mismatches = 2	0.1805894	5.65e-08 ***

Signifikanzniveau: 0.001(***), 0.01(**), 0.05(*), 0.1(.)

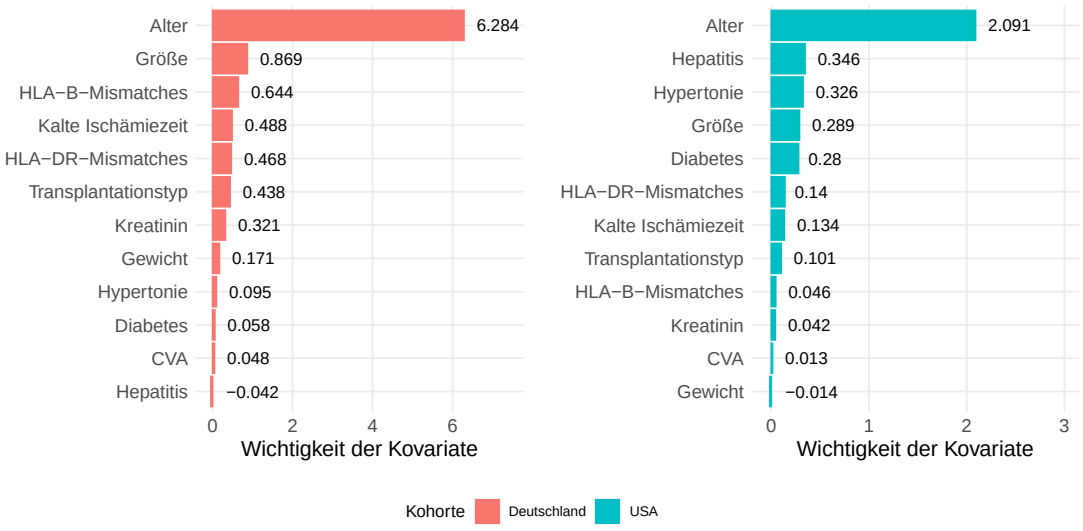

Abbildung 13: Die Abbildung zeigt die Variable Importance für die Cox-Regressionsmodelle. Die angegebenen Werte zeigen die prozentuale Verschlechterung des Brier Scores an, die man erhält, wenn man die jeweilige Variable permutiert

einzelnen Modellen. Während diese bei dem US-Modell den zweitgrößten Einfluss hat, verbessert sich das deutsche Modell durch deren Permutation sogar leicht.

6.2.2 Random Survival Forest Modelle

Zum Trainieren der Random Survival Forests wurde sowohl bei den US-amerikanischen als auch bei den deutschen Daten zum Bestimmen der Hyperparameter eine 10-Fold-Crossvalidation durchgeführt. Dabei wurden verschiedene Kombinationen für die Anzahl der Bäume und die Mindestanzahl der Beobachtungen pro Blatt getestet. So wurden für beide Datensätze die Kombinationen mit dem niedrigsten integrierten Brier Score ausgewählt. Für die deutschen Daten wurde so eine Mindestanzahl an Beobachtungen pro Blatt von 50 und eine Baumanzahl von 250 ausgewählt. Für die US-Kohorte mussten die Hyperparameter so gewählt werden, dass die in dieser Arbeit entwickelte Rekalibrierungsmethode für das Modell anwendbar ist. Hierfür musste ein Random Survival Forest mit einer Mindestblattgröße von 1000 gewählt werden. Unter dieser Nebenbedingung war ein Modell mit einer Blattgröße von 1000 und einer Baumanzahl von 500 bzgl. des Brier Scores die beste Wahl. Daher wurde dieses Modell für den US-amerikanischen Random Survival Forest gewählt.
Das auf den deutschen Daten basierende Modell weist einen integrieten Brier Score von 0.1639 und einen C-Index von 0.6366 auf. Für das auf den US-Daten beruhende Modell beträgt der Brier Score 0.1320 und der C-Index 0.5963. Genau wie beim Cox-Modell ist auch beim Random Survival Forest das deutsche Modell besser diskriminierend, während das US-Modell besser kalibriert ist.
Auch für die Random Survival Forests wurden Calibration-Slope und Calibration-Intercept bestimmt und ein Wald-Test durchgeführt. Abbildung 19 zeigt die zugehörigen Ergebnisse. Der deutsche Random Survival Forest ist gut kalibriert. Bei US-amerikanischen Random Survival Forest wird hingegen Für die Jahre 1 und 4 nach der Transplantation eine weniger gute Kalibrierung festgestellt.
Die Feature Importance der beiden Random Survival Forest Modelle wird in Abbildung 14 gezeigt. Es fällt auf, dass wie auch bei der Cox-Regression die Kovariate *Alter* bei beiden Kohorten mit Abstand den größen Effekt hat. Der Effekt für die deutschen Daten ist dabei erneut noch extremer als bei den US-amerikanischen Daten.

6.3 Kalibrierung der US-Modelle auf der deutschen Kohorte

In diesem Anschnitt soll untersucht werden, ob die auf den US-Daten erstellten Vorhersagemodelle auch für die deutsche Kohorte anwendbar sind. Es soll also untersucht werden, wie gut die US-Modelle für die deutschen Daten kalibriert sind und ob es eine Motivation zur Rekalibrierung gibt. Dazu werden diese Modelle auf die deutschen Testdaten angewendet und die Ergebnisse werden anschließend evaluiert.
Zur Untersuchung der Kalibrierung wurden Calibration-Intercept und Calibration-Slope bestimmt. Die Abweichungen dieser Werte von den Werten eines perfekt kalibrierten Modells wurden mithilfe des Wald-Tests analysiert. Tabelle 12 zeigt die Ergebnisse dieser Untersuchungen.
Sowohl auf den US-Daten trainierter Random Survival Forest als auch auf den US-Daten trainierte Cox-Regression weisen auf den deutschen Testdaten keine gute Kalibrierung auf. Dabei sind bei beiden Modelle besonders die Jahren eins bis vier problematisch. Es besteht somit Bedarf zur Rekalibrierung.

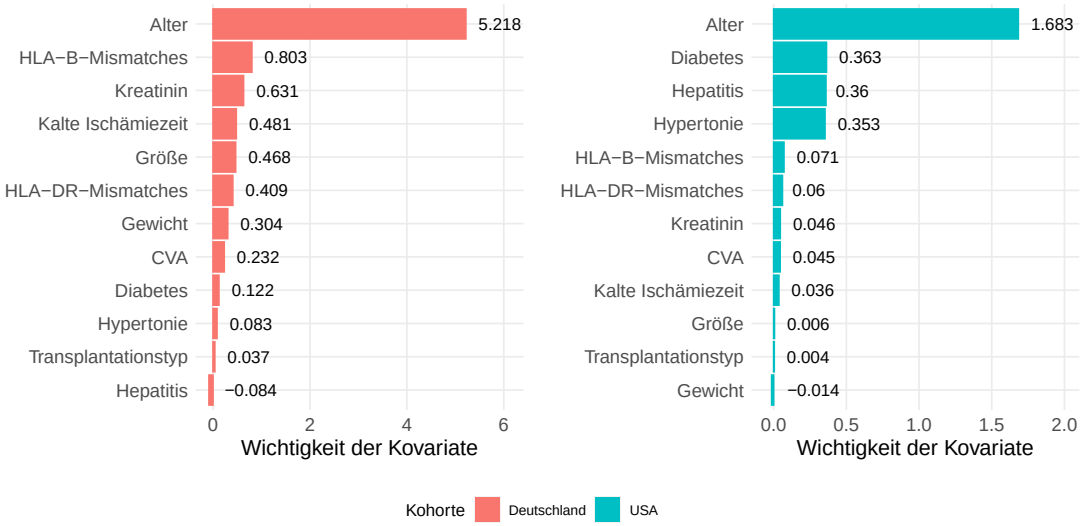

Abbildung 14: Die Abbildung zeigt die Variable Importance für die Random Survival Forests. Die angegebenen Werte zeigen die prozentuale Verschlechterung des Brier Scores an, die man erhält, wenn man die jeweilige Variable permutiert

6.4 Ergebnisse der Rekalibrierungen

Für die Anpassung der US-Modelle auf die deutsche Kohorte wurden diese Modelle mittels der deutschen Trainingsdaten rekalibriert und die rekalibrierten Modelle wurden anschließend mittels der deutschen Testdaten validiert. Beim Random Survival Forest konnten für das rekalibrierte Modell nur 415 der ursprünglichen 500 Bäume verwendet werden, da für die restlichen Bäume das Verfahren zur Umwandlung in Cox-Regressionsmodelle nicht konvergierte.
Zunächst zeigen die Tabellen 13 und 14 die Vergleiche des unveränderten US-Modells, des rekalibrierten US-Modells und des auf den deutschen Daten neu trainierten Modells in Bezug auf den integrierten Brier Score und den C-Index. Im Bezug auf den integrierten Brier Score konnte durch die Rekalibrierung sowohl für das Cox-Modell als auch für den Random Survival Forest eine Verbesserung erzielt werden. Jedoch wurden in beiden Fällen die besten Ergebnisse durch das Neuerstellen eines Modells auf den deutschen Trainingsdaten erzielt.
Auch im Bezug auf den C-Index schnitten die neu trainierten Modelle am besten ab. Für den Random Survival Forest führt die Rekalibrierung zu einer leichten Verschlechterung des C-Indexes. Der C-Index des Cox-Modells ändert sich für das rekalibrierte Modell nicht, da dieser eine Rangkorrelation ist. Er bewertet lediglich die Diskriminierung, für welche nur die Risiko-Scores betrachtet werden. Die individuellen Risiko-Scores ändern sich zwar durch die Rekalibrierung, allerdings bei der hier verwendeten Methode nur so, dass sich die Rangfolge der Patienten dadurch nicht ändert.
Die Abbildungen 26 und 27 zeigen den Brier Score im Verlauf über die Zeit für die Cox-Regression bzw. den Random Survival Forest. Dabei sind jeweils die Werte für das

Tabelle 12: In der Tabelle werden Calibration-Slope, Calibration-Intercept und die Ergebnisse des zugehörigen Wald-Testsgezeigt. Es wird untersucht, wie gut die beiden auf den US-Daten trainierten Modelle für die deutschen Kohorte kalibriert sind.

		Cox-Regression		Random Survival Forest	
Jahr	Parameter	Wert	p-Wert	Wert	p-Wert
1	Intercept	1.4805	0.0033 **	1.0485	0.0014 **
	Slope	1.4286	0.0317 *	1.2578	0.0486 *
	Multivariater Test		3.79e-04 ***		1.14e-06 ***
2	Intercept	0.9217	0.0403 *	0.7037	0.1081
	Slope	1.3422	0.1221	1.2369	0.2755
	Multivariater Test		0.0015 **		0.0025 **
3	Intercept	0.7577	0.0172 *	0.9182	0.0072 **
	Slope	1.3423	0.0698	1.4377	0.0313 *
	Multivariater Test		0.0016 **		0,0006 ***
4	Intercept	0.5886	0.0060 **	0.8307	0.0061 **
	Slope	1.3534	0.0219 *	1.5177	0.0177 *
	Multivariater Test		0.0064 **		0.007 **
5	Intercept	0.2643	0.0889	0.5345	0.0356 *
	Slope	1.1688	0.2164	1.3962	0.0757
	Multivariater Test		0.0929		0.0591
6	Intercept	0.1987	0.2354	0.4575	0.1178
	Slope	1.125	0.4828	1.4089	0.1917
	Multivariater Test		0.2473		0.2334
7	Intercept	0.2325	0.0909	0.4306	0.0472 *
	Slope	1.2067	0.2583	1.4738	0.1053
	Zusammen		0.1598		0.1111

Signifikanzniveau: 0.001(***), 0.01(**), 0.05(*)

Tabelle 13: Werte des integrierten Brier Scores für Cox-Regression und Random Survival Forest, jeweils für ursprüngliches, rekalibriertes und neues Modell

	Cox-Regression	Random Survival Forest
Nicht rekalibriertes Modell	0.16729	0.16806
Rekalibriertes Modell	0.16588	0.16641
Neu trainiertes Modell	0.16412	0.16385

unveränderte US-Modell, das rekalibrierte Modell und das für die deutsche Kohorte neu trainierte Modell zu sehen. Es ist zu erkennen, dass die Unterschiede zwischen diesen drei Modellen für verschiedene Zeitpunkte unterschiedlich sind.
Für die rekalibrierten Modelle wurden anschließend Calibration-Slope und Calibration-Intercept geschätzt. In den Tabelle 15 und 16 sind diese zusammen mit den zugehörigen Ergebnissen des Wald-Tests zu sehen. Außerdem werden sie in Abbildung 15

Tabelle 14: Werte des Concordance Indexes für Cox-Regression und Random Survival Forest, jeweils für ursprüngliches, rekalibriertes und neues Modell

	Cox-Regression	Random Survival Forest
Nicht rekalibriertes Modell	0.62299	0.62336
Rekalibriertes Modell	0.62299	0.62139
Neu trainiertes Modell	0.63126	0.63658

graphisch dargestellt.

Tabelle 15: Calibration-Intercept, Calibration-Slope und Ergebnisse der zugehörigen Wald-Tests für die nicht rekalibrierten, rekalibrierten und neu trainierten Cox-Modelle

		unverändertes Modell		Rekalibriertes Modell		Neu trainiertes Modell	
Jahr	Parameter	Wert	p-Wert	Wert	p-Wert	Wert	p-Wert
1	Intercept	1.4805	0.0033 **	0.2848	0.4002	0.0303	0.9227
	Slope	1.4286	0.0317 *	1.1592	0.3254	1.0283	0.8462
	Multivariater Test		3.79e-04 ***		0.5028		0.9069
2	Intercept	0.9217	0.0403 *	0.1036	0.7436	-0.1688	0.4205
	Slope	1.3422	0.1221	1.0893	0.6188	0.9156	0.465
	Multivariater Test		0.0015 **		0.6984		0.7133
3	Intercept	0.7577	0.0172 *	0.1226	0.5956	-0.1472	0.3764
	Slope	1.3423	0.0698	1.0896	0.5584	0.8877	0.2926
	Multivariater Test		0.0016 **		0.8362		0.5387
4	Intercept	0.5886	0.0060 **	0.1202	0.4606	-0.0775	0.6351
	Slope	1.3534	0.0219 *	1.0987	0.4306	0.9317	0.5707
	Multivariater Test		0.0064 **		0.7325		0.8423
5	Intercept	0.2643	0.0889	-0.0235	0.8489	-0.152	0.3854
	Slope	1.1688	0.2164	0.9491	0.6459	0.8251	0.2415
	Multivariater Test		0.0929		0.7259		0.4523
6	Intercept	0.1987	0.2354	-0.0468	0.7222	-0.1467	0.3991
	Slope	1.125	0.4828	0.9139	0.5511	0.7890	0.2363
	Multivariater Test		0.2473		0.7717		0.4657
7	Intercept	0.2325	0.0909	0.0324	0.7702	-0.0501	0.6979
	Slope	1.2067	0.2583	0.9805	0.8953	0.8433	0.3197
	Multivariater Test		0.1598		0.7500		0.4845

Signifikanzniveau: 0.001(***), 0.01(**), 0.05(*)

Bei der Cox-Regression haben sich alle Parameter verbessert. Es sind also sowohl Calibration-Intercept als auch Calibration-Slope zu jedem Zeitpunkt näher an den idealen Werten als beim nicht rekalibrierten Modell. Zudem ist das rekalibrierte Modell zu vielen Zeitpunkten etwa gleich gut oder sogar besser kalibriert als das neu trainierte Modell. Dies gilt besonders für die späteren Zeitpunkte.
Beim Random Survival Forest konnte die Rekalibrierung für den Intercept zu allen Zeitpunkten

Tabelle 16: Calibration-Intercept, Calibration-Slope und Ergebnisse der zugehörigen Wald-Tests für die nicht rekalibrierten, rekalibrierten und neu trainierten Random Survival Forest Modelle

		unverändertes Modell		Rekalibriertes Modell		Neu trainiertes Modell	
Jahr	Parameter	Wert	p-Wert	Wert	p-Wert	Wert	p-Wert
1	Intercept	1.0485	0.0014 **	0.6639	0.2149	0.0815	0.6948
	Slope	1.2578	0.0486 *	1.3629	0.1667	1.0574	0.5563
	Multivariater Test		1.14e-06 ***		0.2738		0.6254
2	Intercept	0.7037	0.1081	0.406	0.2507	-0.0358	0.8702
	Slope	1.2369	0.2755	1.2811	0.1732	1.0049	0.9675
	Multivariater Test		0.0025 **		0.2292		0.7603
3	Intercept	0.9182	0.0072 **	0.4596	0.1122	-0.0331	0.7995
	Slope	1.4377	0.0313 *	1.3462	0.0814	0.9764	0.7777
	Multivariater Test		0,0006 ***		0.1905		0.9602
4	Intercept	0.8307	0.0061 **	0.3998	0.0653	0.0219	0.804
	Slope	1.5177	0.0177 *	1.3546	0.0417 *	1.0220	0.7384
	Multivariater Test		0.007 **		0.1102		0.9311
5	Intercept	0.5345	0.0356 *	0.1996	0.3726	0.0111	0.9284
	Slope	1.3962	0.0757	1.2020	0.3410	1.0007	0.9948
	Multivariater Test		0.0591		0.6343		0.9784
6	Intercept	0.4575	0.1178	0.1099	0.6368	0.0339	0.7284
	Slope	1.4089	0.1917	1.1533	0.5764	1.0351	0.7426
	Multivariater Test		0.2334		0.8484		0.9406
7	Intercept	0.4306	0.0472 *	0.1558	0.3591	0.0578	0.4988
	Slope	1.4738	0.1053	1.2317	0.3539	1.0560	0.6188
	Multivariater Test		0.1111		0.6375		0.7911

Signifikanzniveau: 0.001(***), 0.01(**), 0.05(*)

eine Verbesserung erzielen, für den Slope hingegen nur für die Zeitpunkte ab 3 Jahren nach der Transplantation. Jedoch war das neu trainierte Modell stets besser kalibriert als das neu trainierte Modell. Insgesammt ist der auf den deutschen Daten neu trainierte Random Survival Forest im Bezug auf alle untersuchten Kriterien, C-Index, integrierter Brier Score und Kalibrierung, das Vorhersagemodell mit der besten Performance für die deutsche Kohorte.

Abbildung 15: Vergleich von Calibration-Slope und Calibration-Intercept für Random Survival Forest und Cox-Regression im Verlauf über die Zeit. Es werden die Werte für die nicht rekalibrierten, rekalibrierten und neu trainierten Modelle gezeigt. Die schwarze Linie zeigt den idealen Wert für Slope bzw. Intercept

Ferner wurden für verschiedene Zeitpunkte, jeweils für jedes volle Jahr nach der Transplantation, Kalibrierungskurven angefertigt. In Abbildung 16 sind die Kurven für ein Jahr nach der Transplantation zu sehen. Die oberen beiden Kurven zeigen die Kalibrierung für die Modelle, wenn die US-Modelle ohne Kalibrierung auf den deutschen Daten angewendet werden. Zum einen kann man erkennen, dass die Steigung der eingezeichneten Punkte steiler ist als die Winkelhalbierende und zudem verlaufen die Punkte überhalb der Winkelhalbierenden. Die vorhergesagten Ereigniswahrscheinlichkeiten sind also durchschnittlich zu hoch und dabei zu moderat. Dies gilt sowohl für das Cox-Modell als auch für den Random Survival Forest. Die beiden Kurven in der zweiten Zeile zeigen die Kalibrierung der beiden rekalibrierten Modelle. Bei der Cox-Regression ist eine deutliche Verbesserung von *mean calibration* und *calibration-in-the-large* zu sehen. Beim Random Survival Forest hat sich die *calibration-in-the-large* ebenfalls sichtbar verbessert. Es ist allerdings weiterhin eine zu moderate Risikoschätzung zu sehen. In der unteren Zeile sind die Kurven der neu trainierten Modelle abgebildet, welche eine gute Kalibrierung beider Modelle zeigen.
Die Kalibrierungskurven für das Jahr 4 nach der Transplantation werden in Abbildung 17 gezeigt. Genau wie bei den Kurven vom Jahr 1 zeigt sich für die nicht rekalibrierten Modelle sowohl für das Cox-Modell, als auch für den Random Survival Forest eine durchschnittlich zu hohe und zu moderate Risikoschätzung, allerdings in einer weniger starken Ausprägung. Die Kurven der rekalibrierten Modelle zeigen für beide Modelle eine Verbesserung in Bezug auf

mean calibration und *calibration-in-the-large*. Das am Besten kalibrierte Modell ist der auf den deutschen Daten neu trainierte Random Survival Forest.
Die Kalibrierungskurven anderer Zeitpunkte werden im Anhang in den Abbildungen 18 bis 22 gezeigt.

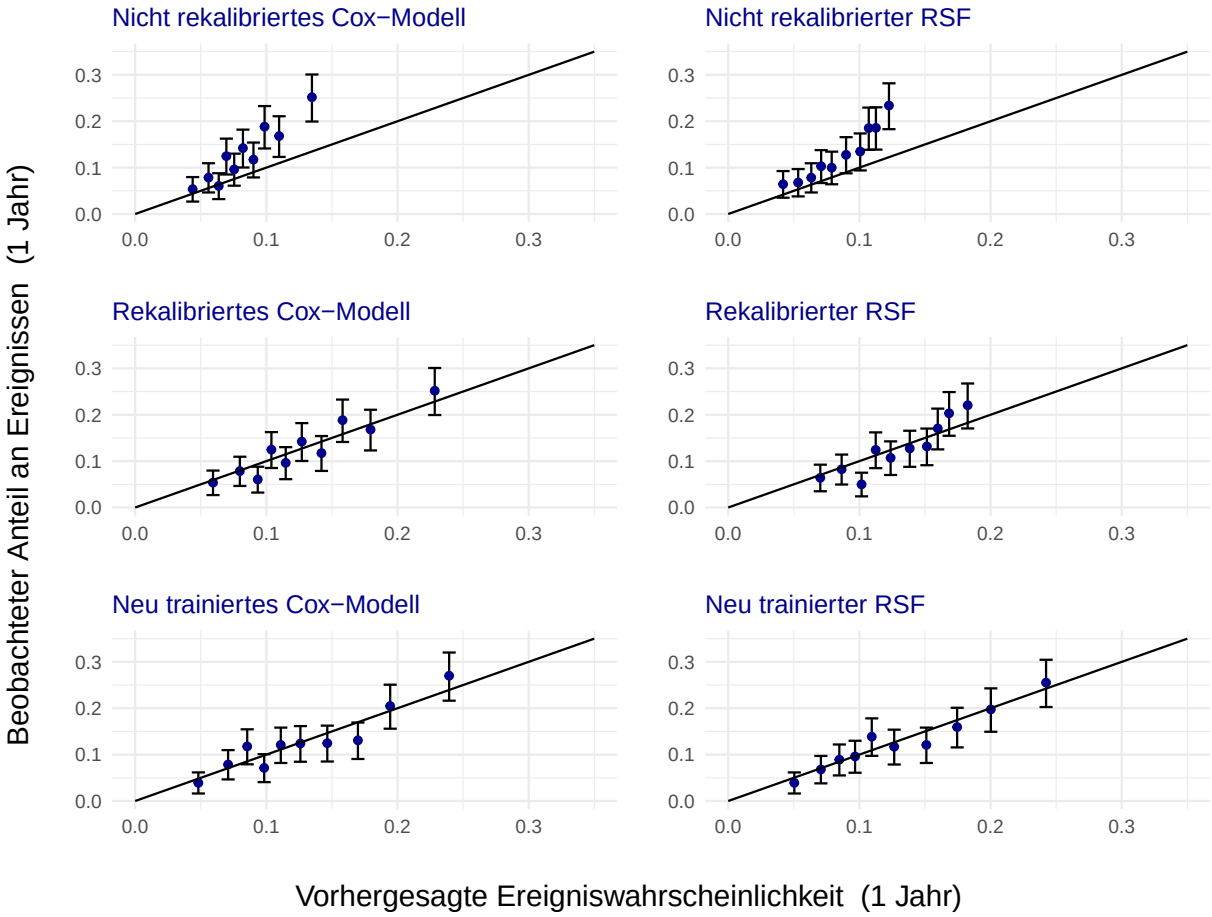

Abbildung 16: Die Abbildung zeigt die Kalibrierung der sechs Vorhersagemodelle zum Zeitpunkt 1 Jahr nach der Transplantation. Auf der linken Seite sind die Cox-Regressionsmodelle zu sehen, während die rechte Seite die Random Survival Forest Modelle zeigt.

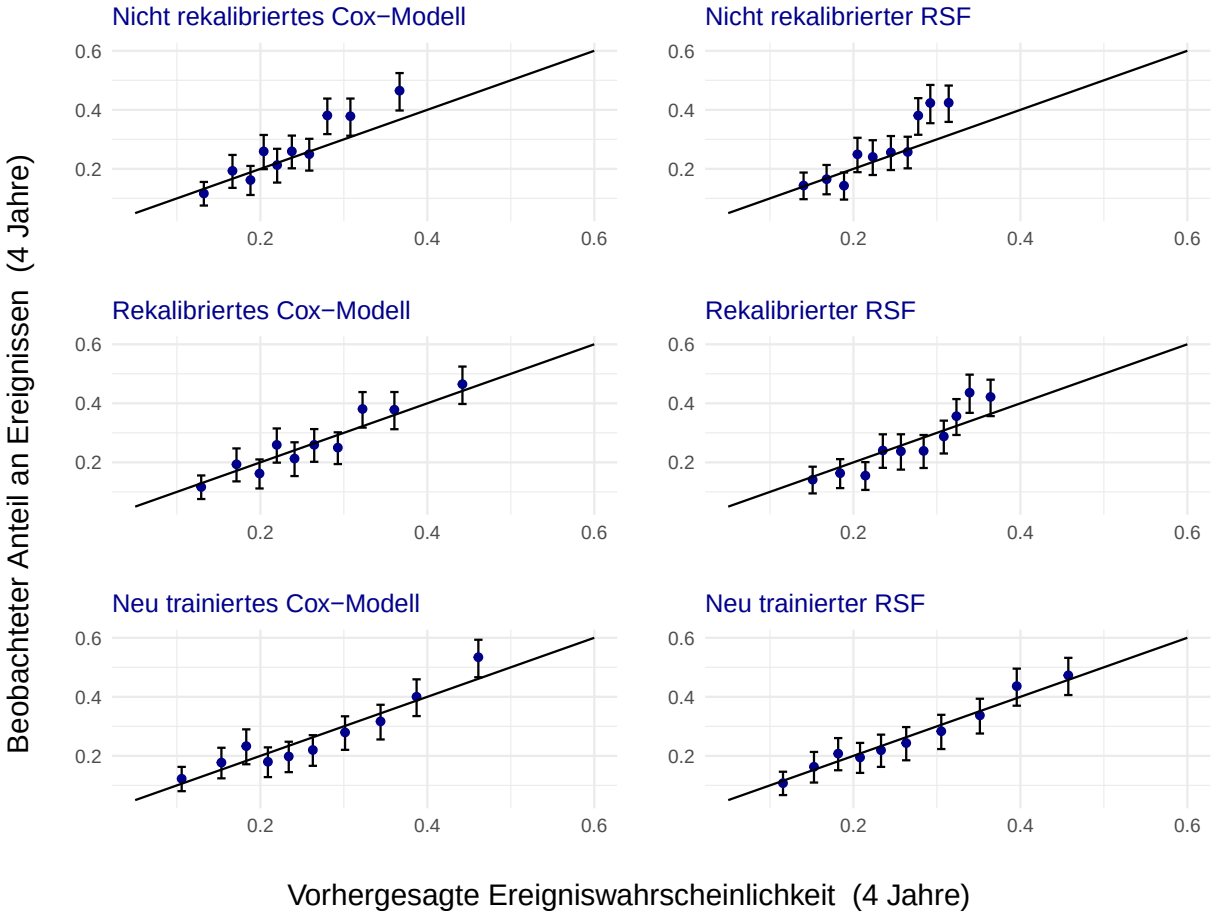

Abbildung 17: Die Abbildung zeigt die Kalibrierung der sechs Vorhersagemodelle zum Zeitpunkt 4 Jahre nach der Transplantation. Auf der linken Seite sind die Cox-Regressionsmodelle zu sehen, während die rechte Seite die Random Survival Forest Modelle zeigt.

7 Fazit und Ausblick

7.1 Fazit

In dieser Arbeit wurde untersucht, inwiefern Modelle, welche mit Daten von Nierentransplantationen in den USA trainiert wurden, auch auf eine deutsche Kohorte angewendet werden können. Hierbei lag der Fokus auf der Cox-Regression und dem Random Survival Forest. Dabei wurde untersucht, welche Voraussetzungen dafür erfüllt sein müssen, wie gut die Kalibrierung der US-Modelle für die deutschen Daten ist und wie eine Rekalibrierung realisiert werden kann.
Da es für die Rekalibrierung des Random Survial Forest keine bekannte Rekalibrierungsmethode gibt, wurde in dieser Arbeit hierfür eine eigene Methode entwickelt. Diese ist eine Erweiterung einer 2016 publizierten Methode (vgl. [DZ16]) zur Rekalibrierung von Random Forests zur Wahrscheinlichkeitsschätzung. Anschließend wurden für die deutschen Daten neue Modelle trainiert. Es folgte ein Vergleich der nicht rekalibrierten, kalibrierten und neu trainierten Modelle in Bezug auf Vorhersagegüte und Kalibrierung. Die Kalibrierung ist ein Aspekt von Vorhersagemodellen, welche trotz der dringenden Empfehlung hierzu bei der Publikation von Vorhersagemodellen häufig vernachlässigt wird.
Für das Erstellen und Rekalibrieren von Modellen lagen für die US-amerikanischen Transplantationen Daten des United Network for Organ Sharing vor und für die deutschen Transplantationen Daten des deutschen Transplantationsregisters. Die beiden Datensätze unterscheiden sich zum Teil stark in ihrer Struktur und den erhobenen Kovariaten. Um für den deutschen Datensatz Kovariaten zu ermitteln, die in den UNOS-Daten erhoben werden und häufig in Prognosemodellen verwendet werden, mussten diese durch Feature Engineering Prozesse hergeleitet werden. Zum Teil mussten hierzu Annahmen getätigt werden, welche durch durch das Aneignen von Domänenwissen hergeleitet wurden.
Es erfolgte eine Suche nach publizierten Vorhersagemodellen, welche auf den US-amerikanischen UNOS-Daten beruhen. Dazu wurden systematische Reviews verschiedener Autoren zu klinischen Vorhersagemodellen nach Nierentransplantationen betrachtet. Aus verschiedenen Gründen konnte keines der gefundenen Modelle auf die deutschen Daten angewendet werden. Zum Einen stand das veröffentlichte Prognosemodell in einigen Fällen nicht in Gänze zur Verfügung. Es fehlten also Informationen, die notwendig sind, um die Modelle auf neue Daten anwenden zu können. Zum Anderen gab es kein Modell, für welches im deutschen Datensatz alle nötigen Informationen zu den Kovariaten vorhanden gewesen wären. Daher wurden auch auf den US-Daten neue Modelle erstellt, die es anschließend zu rekalibrieren galt.
Sowohl für das auf den US-Daten trainierte Cox-Modell, als auch für den auf den US-Daten trainierten Random Survival Forest zeigte sich im Bezug auf die deutschen Transplantationsdaten eine schlechte Kalibrierung. Es bestand also der Bedarf einer Rekalibrierung. Besonders problematisch war die Kalibrierung in den ersten 4 Jahren nach der Transplantation.
Durch die Rekalibrierung der Cox-Regression konnte die Kalibrierung der US-Modelle auf der deutschen Kohorte für alle betrachteten Zeitpunkte verbessert werden. Für die Jahre 5 bis 7 nach der Transplantation wies das rekalibrierte Modell sogar eine bessere Kalibrierung auf als

das neu trainierte Modell.
Für den Random Survival Forest wurde in dieser Arbeit eine Rekalibrierungsmethode entwickelt. Mithilfe dieser konnte auch für den Random Survival Forest eine verbesserte Kalibrierung des US-Modells auf die deutsche Kohorte erzielt werden. Allerdings konnte hier für die Jahre 1 und 2 nach der Transplantation nur der Calibration-Intercept verbessert werden, während sich der Calibration-Slope für diese Zeitpunkte leicht verschlechterte. Der auf den deutschen Daten neu trainierte Random Survival Forest war für die deutsche Kohorte für jeden Zeitpunkt besser kalibriert als das nicht rekalibrierte und das rekalibrierte US-Modell.
Für alle Modelle bleibt die Frage offen, in welchem Maß die Unterschiede in den Einflüssen der Kovariaten auf die Ereigniszeit auch durch die sich unterscheidenden Allokationsverfahren begründet sind.
Die in dieser Arbeit entwickelte Rekalibrierungsmethode für Random Survival Forests hat also das Potential schlechte Kalibrierung von Modellen zu verbessern. Besonders für kleine Datensätze kann sie eine Alternative zum Neutrainieren eines Modells darstellen, um eine Überanpassung an die Trainingsdaten zu vermeiden. Hierbei ist die Einschränkung zu beachten, dass die Methode nicht für alle Random Survival Modelle anwendbar ist, da die Berechnung der Survivalfunktion für die einzelnen Bäume nicht konvergiert, wenn diese zu tief verzweigt sind.

7.2 Ausblick

In dieser Arbeit wurde die Rekalibrierung der Cox-Regression durch das Schätzen eines neuen Baseline-Hazards in Kombination mit einem Skalierungsfaktor für alle geschätzten Parameter durchgeführt. Für die Rekalibrierung des Random Survival Forests wurden die einzelnen Bäume in Cox-Regressionsmodelle umgewandelt und auf diesen Modellen wurde ebenfalls die oben beschriebene Rekalibrierungsmethode angewandt. Daher lässt sich die in dieser Arbeit entwickelte Rekalibrierungsmethode leicht erweitern, da auch jede andere für die Cox-Regression entwickelte Rekalibrierungsmethode verwendet werden kann. Hier gibt es beispielsweise die Möglichkeit, die Parameter einzeln anstatt mit einem gemeinsamen Skalierungsfaktor anzupassen. (vgl. [MAR+15])
Um die vorgestellte Rekalibrierungsmethode auch auf Random Survival Forests mit beliebig tief verzweigten Bäumen anwenden zu können, ist es denkbar, einen zusätzlichen Schritt in den Algorithmus einzubauen, welcher ein Pruning der Bäume bis zu einer hinreichend flachen Verzweigung vornimmt.
Ein anderer möglicher Ansatz zur Rekalibrierung eines Random Survival Forests könnte es sein, Methoden zur Rekalibrierung von Random Forests zur Wahrscheinlichkeitsschätzung in Kombination mit Binarisierung zu verwenden, dabei allerdings mittels IPCW-Gewichtung einer systematischen Verzerrung entgegenzuwirken.

Literatur

[Aal78] Odd Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6:701–726, 1978.

[BMT+19] Sunjae Bae, Allan B. Massie, Alvin G. Thomas, Gahyun Bahn, Xun Luo, Kyle R. Jackson, Shane E. Ottmann, Daniel C. Brennan, Niraj M. Desai, Josef Coresh, Dorry L. Segev, and Jacqueline M. Garonzik Wang. Who can tolerate a marginal kidney? predicting survival after deceased donor kidney transplant by donor–recipient combination. *American Journal of Transplantation*, 19:425–433, 2 2019.

[Bre72] Norman Breslow. Discussion on professor cox s paper. *Journal of the Royal Statistical Society*, 34:202–220, 1972.

[Bre01] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[BRE+20] Sarah Booth, Richard D. Riley, Joie Ensor, Paul C. Lambert, and Mark J. Rutherford. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. *International Journal of Epidemiology*, 49:1316–1325, 8 2020.

[BSR+50] Glenn W Brier, Charles Sawyer, F W Reichelderfer, James E Editor, and J R Caskey. Verification of forecast expressed in terms of probability. *Journal of the American Meteorological Society*, 78, 1950.

[CAT+16] Cynthia S. Crowson, Elizabeth J. Atkinson, Terry M. Therneau, Andrew B. Lawson, Duncan Lee, and Ying MacNab. Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research*, 25:1692–1706, 8 2016.

[CMS+19] Ben Van Calster, David J. McLernon, Maarten Van Smeden, Laure Wynants, Ewout W. Steyerberg, Patrick Bossuyt, Gary S. Collins, Petra MacAskill, Karel G.M. Moons, and Andrew J. Vickers. Calibration: The achilles heel of predictive analytics. *BMC Medicine*, 17, 12 2019.

[Cox72] D R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society*, 34:187–220, 1972.

[DBP+19] Maximilian Dahmen, Felix Becker, Hermann Pavenstädt, Barbara Suwelack, Katharina Schütte-Nütgen, and Stefan Reuter. Validation of the kidney donor profile index (kdpi) to assess a deceased donor's kidneys' outcome in a european cohort. *Scientific reports*, 9:11234, 8 2019.

[DSC20] Covadonga Díez-Sanmartín and Antonio Sarasa Cabezuelo. Application of artificial intelligence techniques to predict survival in kidney transplantation: A review. *Journal of Clinical Medicine*, 9, 2 2020.

[DZ16] Theresa Dankowski and Andreas Ziegler. Calibrating random forests for probability estimation. *Statistics in Medicine*, 35:3949–3960, 9 2016.

[ET22] ET. Eurotransplant manual, August 2022.

[Gre26] M. Greenwood. *A Report on the Natural Duration of Cancer*. Reports on public health and medical subjects. H.M. Stationery Office, 1926.

[GSSS99] Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. volume 18, pages 2529–2545. John Wiley and Sons Ltd, 9 1999.

[HCP+82] F.E. Harrell, R.M. Califf, D.B Pryor, K.L. Lee, and R.A. Rosati. Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247:2543–2546, 1982.

[Hou00] Hans C Van Houwelingen. Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine*, 19:3401–3415, 2000.

[IKBL08] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *Annals of Applied Statistics*, 2:841–860, 9 2008.

[KHH+17] Rémi Kaboré, Maria C. Haller, Jérôme Harambat, Georg Heinze, and Karen LeffondrÉ. Risk prediction models for graft failure in kidney transplantation: A systematic review. *Nephrology Dialysis Transplantation*, 32:ii68–ii76, 4 2017.

[KIS+] Bertram L. Kasiske, Ajay K. Israni, Jon J. Snyder, Melissa A. Skeans, Yi Peng, and Eric D. Weinhandl. *American Journal of Kidney Diseases*, pages 947–60, 11.

[KK05] D.G. Kleinbaum and M. Klein. *Survival Analysis: A Self-Learning Text*. Statistics for biology and health. Springer, 2005.

[KM58] E L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[KM03] John P. Klein and Melvin L. Moeschberger. *Survival Analysis Techniques for Censored and Truncated Data*. Second edition, 2003.

[LC93] Michael Leblanc and John Crowley. Survival trees by goodness of split. *Journal of the American Statistical Association*, 88:457–467, 1993.

[Lin07] D. Y. Lin. On the breslow estimator. *Lifetime Data Anal*, 13:471–480, 2007.

[LKH+18] Lukas Johannes Lehner, Anna Kleinsteuber, Fabian Halleck, Dmytro Khadzhynov, Eva Schrezenmeier, Michael Duerr, Kai Uwe Eckardt, Klemens Budde, and Oliver Staeck. Assessment of the kidney donor profile index in a european cohort. *Nephrology Dialysis Transplantation*, 33:1465–1472, 8 2018.

[Man72] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in itsconsideration. *Cancer Chemother. Reports*, 135:185–206, 1972.

[MAR+15] Karel G.M. Moons, Douglas G. Altman, Johannes B. Reitsma, John P.A. Ioannidis, Petra Macaskill, Ewout W. Steyerberg, Andrew J. Vickers, David F. Ransohoff, and Gary S. Collins. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): Explanation and elaboration. *Annals of Internal Medicine*, 162:W1–W73, 1 2015.

[MGG+19] Ethan Mark, David Goldsman, Brian Gurbaxani, Pinar Keskinocak, and Joel Sokol. Using machine learning and an ensemble of methods to predict kidney transplant survival. *PLoS ONE*, 14, 1 2019.

[MNC+17] Miklos Z. Molnar, Danh V. Nguyen, Yanjun Chen, Vanessa Ravel, Elani Streja, Mahesh Krishnan, Csaba P. Kovesdy, Rajnish Mehrotra, and Kamyar Kalantar-Zadeh. Predictive score for posttransplantation outcomes. *Transplantation*, 101:1353–1364, 2017.

[Nel69] Wayne Nelson. Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1(1):27–52, 1969.

[Nel72] Wayne Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14:945–966, 1972.

[PP20] Organ Procurement and Transplantation Network (OPTN) Policies. Policy 8: Allocation of kidneys, June 2020.

[PPKP21] Seo Young Park, Ji Eun Park, Hyungjin Kim, and Seong Ho Park. Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches). *Korean Journal of Radiology*, 22, 2021.

[RMW+21] Chava L. Ramspek, Mostafa El Moumni, Eelaha Wali, Martin B.A. Heemskerk, Robert A. Pol, Meindert J. Crop, Nichon E. Jansen, Andries Hoitsma, Friedo W. Dekker, M. van Diepen, and Cyril Moers. Development and external validation study combining existing models and recent data into an up-to-date prediction model for evaluating kidneys from older deceased donors for transplantation. *Kidney International*, 99:1459–1469, 6 2021.

[Rob93] JM Robins. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. *Proceedings of the Biopharmaceutical Section, American Statistical Association*, 24:24–33, 1993.

[Roy14] Patrick Royston. Tools for checking calibration of a cox model in external validation: Approach based on individual event probabilities, 2014.

[RS09] Panduranga Rao and Douglas Schaubel. A comprehensive risk quantification score for deceased donor kidneys: The kidney donor risk index. *Transplantation*, 88:231–236, 2009.

[RZT+22] Stephanie Riley, Qing Zhang, Wai-Yee Tse, Andrew Connor, and Yinghui Wei. Using information available at the time of donor offer to predict kidney transplant survival outcomes: A systematic review of prediction models. *Transplant International*, 35, 6 2022.

[Seg88] Mark Robert Segal. Regression trees for censored data. *Biometrics*, 44:35–47, 1988.

[SSA+17] Mark D. Stegall, Peter G. Stock, Kenneth Andreoni, John J. Friedewald, and Alan B. Leichtman. Why do we have the kidney allocation system we have today? a history of the 2014 kidney allocation system. *Human Immunology*, 78:4–8, 1 2017.

[Ste09] Ewout Steyerberg. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, volume 19. 01 2009.

[SWG+19] Sameera Senanayake, Nicole White, Nicholas Graves, Helen Healy, Keshwar Baboolal, and Sanjeewa Kularatna. Machine learning in predicting graft failure following kidney transplantation: A systematic review of published predictive models. *International Journal of Medical Informatics*, 130, 10 2019.

[VKDT19] Amanda Jean Vinson, Bryce A. Kiberd, Roger B. Davis, and Karthik K. Tennankore. Nonimmunologic donor-recipient pairing, hla matching, and graft loss in deceased donor kidney transplantation. *Transplantation Direct*, 5, 1 2019.

[WLR17] Ping Wang, Yan Li, and Chandan K. Reddy. Machine learning for survival analysis: A survey, 2017.

Anhang

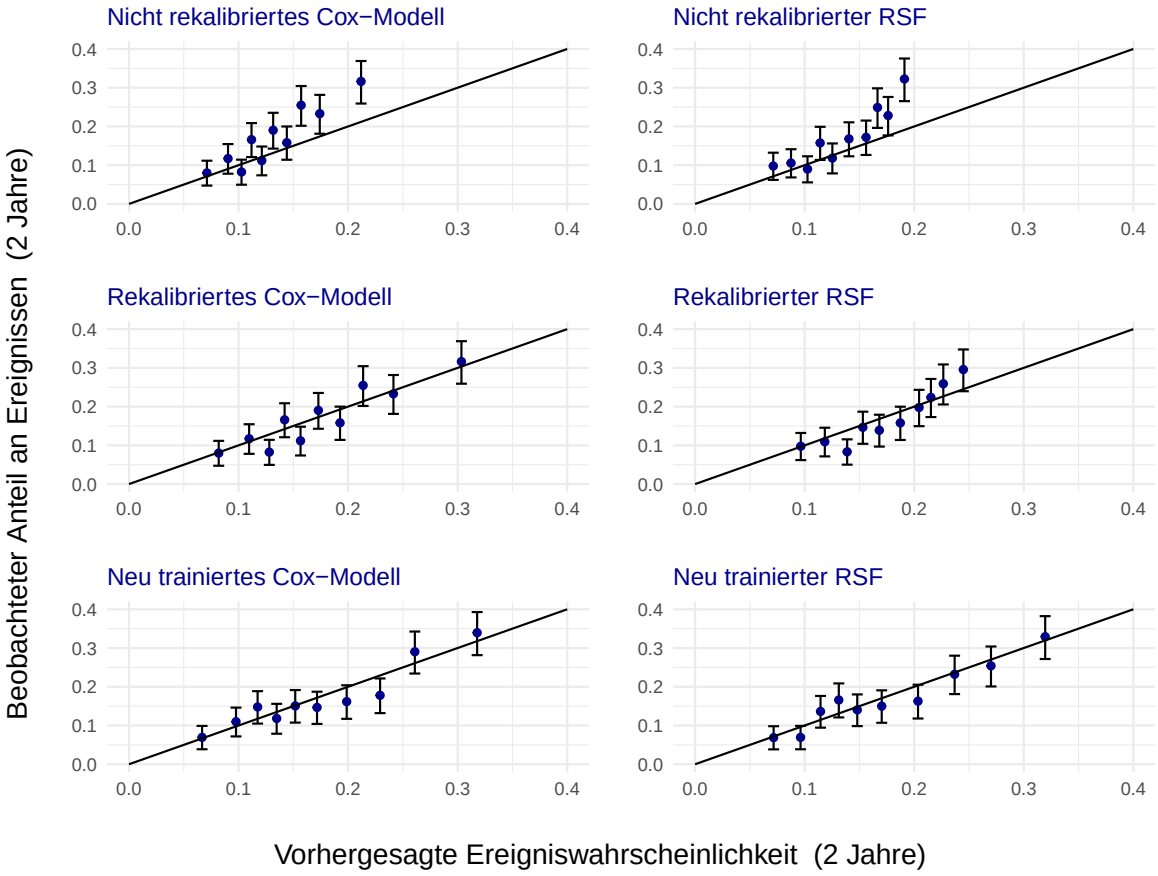

Abbildung 18: Die Abbildung zeigt die Kalibrierung der sechs Vorhersagemodelle zum Zeitpunkt 2 Jahre nach der Transplantation. Auf der linken Seite sind die Cox-Regressionsmodelle zu sehen, während die rechte Seite die Random Survival Forest Modelle zeigt.

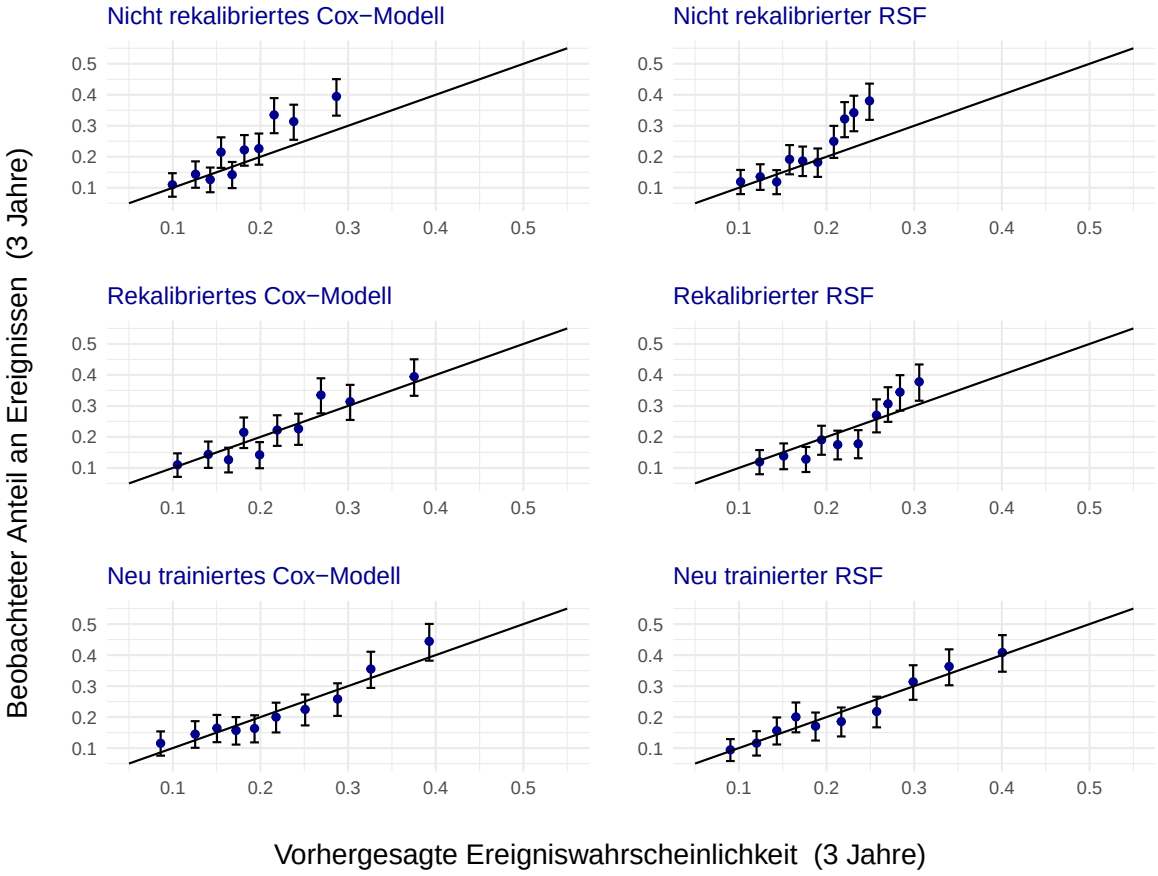

Abbildung 19: Die Abbildung zeigt die Kalibrierung der sechs Vorhersagemodelle zum Zeitpunkt 3 Jahre nach der Transplantation. Auf der linken Seite sind die Cox-Regressionsmodelle zu sehen, während die rechte Seite die Random Survival Forest Modelle zeigt.

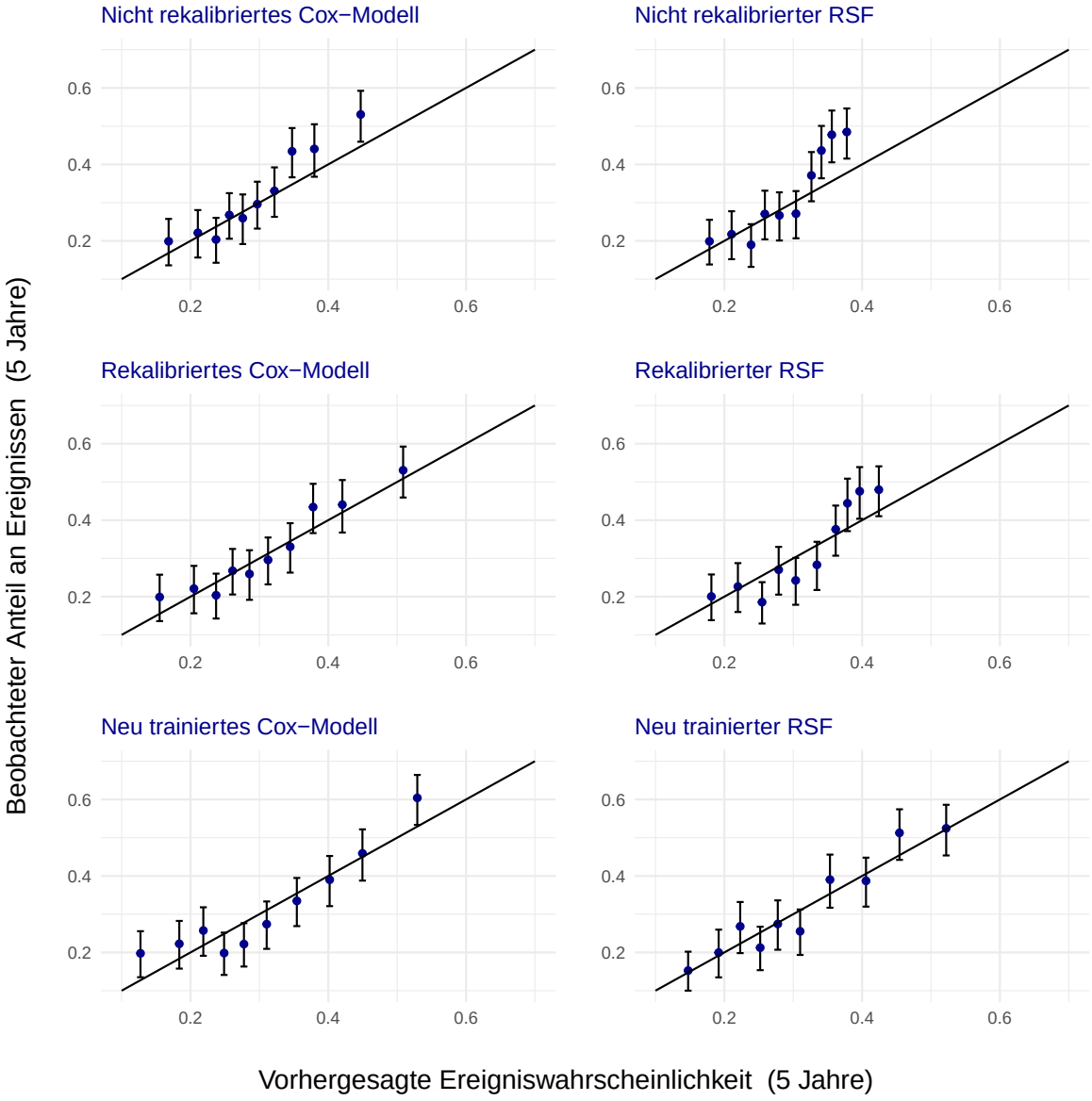

Abbildung 20: Die Abbildung zeigt die Kalibrierung der sechs Vorhersagemodelle zum Zeitpunkt 5 Jahre nach der Transplantation. Auf der linken Seite sind die Cox-Regressionsmodelle zu sehen, während die rechte Seite die Random Survival Forest Modelle zeigt.

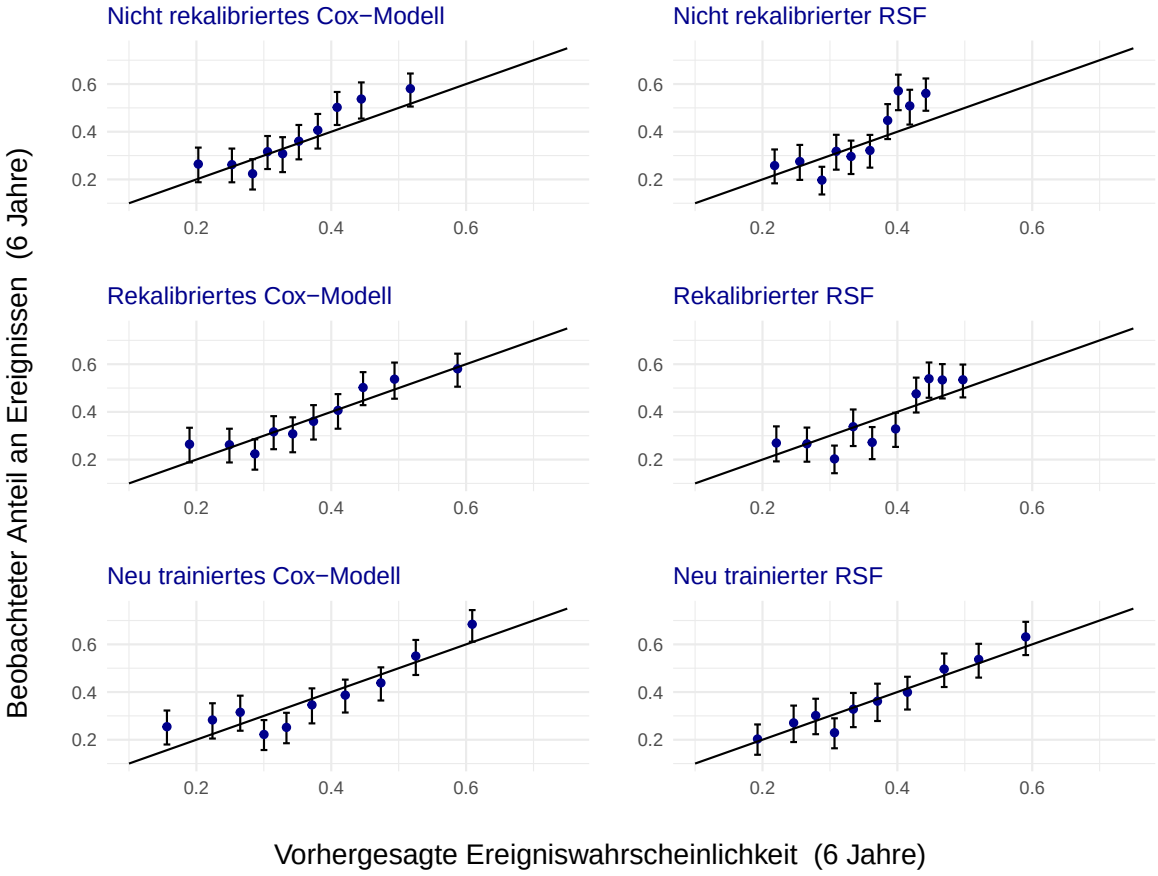

Abbildung 21: Die Abbildung zeigt die Kalibrierung der sechs Vorhersagemodelle zum Zeitpunkt 6 Jahre nach der Transplantation. Auf der linken Seite sind die Cox-Regressionsmodelle zu sehen, während die rechte Seite die Random Survival Forest Modelle zeigt.

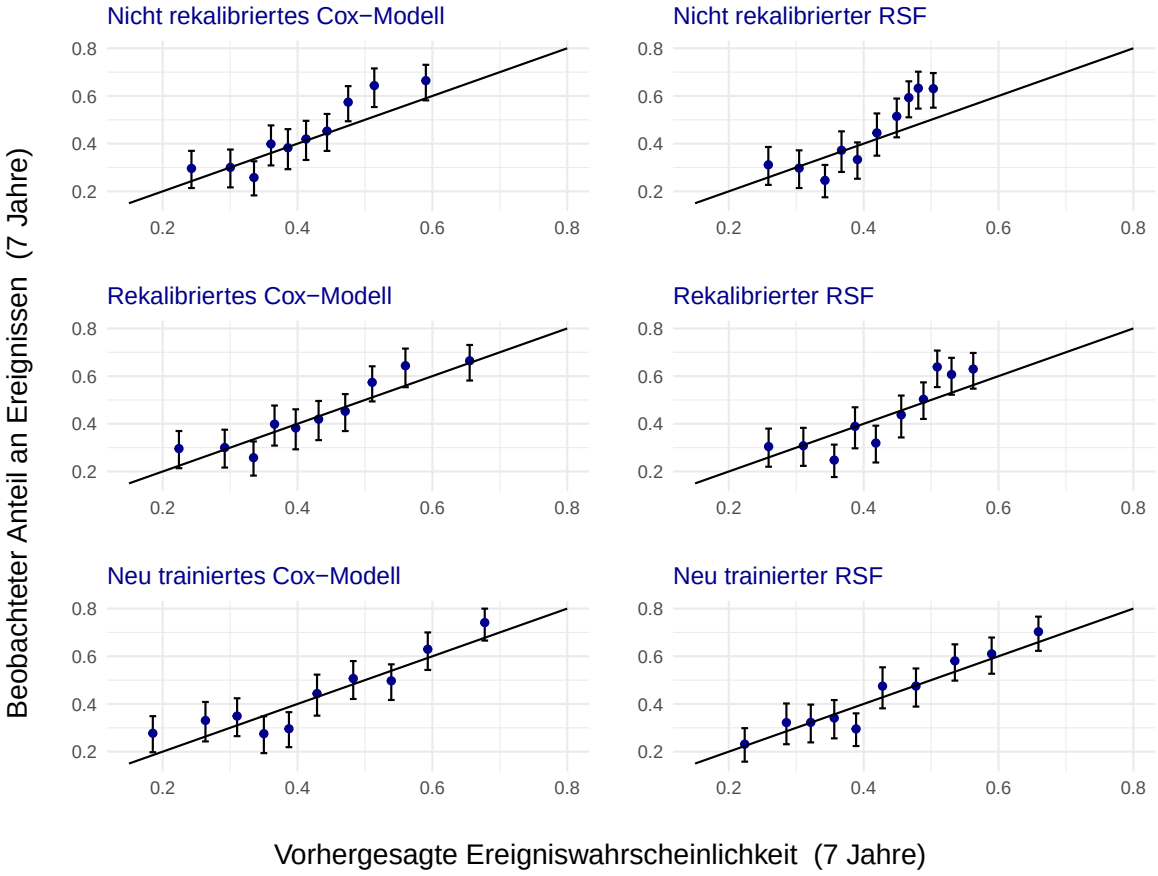

Abbildung 22: Die Abbildung zeigt die Kalibrierung der sechs Vorhersagemodelle zum Zeitpunkt 7 Jahre nach der Transplantation. Auf der linken Seite sind die Cox-Regressionsmodelle zu sehen, während die rechte Seite die Random Survival Forest Modelle zeigt.

ANTIGEN	SPLIT	BROAD	PUBLIC
B5			BW4
B51		B5	
B*51:01	B51	B5	
B*51:02	B51	B5	
B*51:07	B51	B5	
B*51:08	B51	B5	
B*51:XX	B51	B5	
B52		B5	
B*52:01	B52	B5	
B*52:XX	B52	B5	
B7			BW6
B*07:02		B7	
B*07:03		B7	
B*07:04		B7	
B*07:05		B7	
B*07:09		B7	
B*07:XX		B7	
B8			BW6
B*08:01		B8	
B*08:XX		B8	
B12			
B44		B12	BW4
B*44:02	B44	B12	
B*44:03	B44	B12	
B*44:04	B44	B12	
B*44:05	B44	B12	
B*44:10	B44	B12	
B*44:XX	B44	B12	
B45		B12	BW6
B*45:01	B45	B12	
B*45:XX	B45	B12	
B*50:02	B45	B12	
B13			BW4
B*13:01		B13	
B*13:02		B13	
B*13:XX		B13	
B14			BW6
B*14:XX		B14	
B64		B14	
B*14:01	B64	B14	
B65		B14	
B*14:02	B65	B14	
B15			
B*15:XX		B15	
B62		B15	
B*15:01	B62	B15	BW6
B*15:05	B62	B15	BW6
B*15:07	B62	B15	BW6
B*15:24	B62	B15	BW4
B*15:25	B62	B15	BW6
B*15:27	B62	B15	BW6
B*15:30	B62	B15	BW6
B63		B15	BW4
B*15:16	B63	B15	
B*15:17	B63	B15	
B75		B15	BW6
B*15:02	B75	B15	
B76		B15	BW6
B*15:12	B76	B15	
B77		B15	BW4
B*15:13	B77	B15	

ANTIGEN	SPLIT	BROAD	PUBLIC
B70			BW6
B71		B70	
B*15:10	B71	B70	
B*15:18	B71	B70	
B72		B70	
B*15:03	B72	B70	
B16			
B38		B16	BW4
B*38:01	B38	B16	
B*38:02	B38	B16	
B*38:XX	B38	B16	
B39		B16	BW6
B*39:01	B39	B16	
B*39:02	B39	B16	
B*39:05	B39	B16	
B*39:06	B39	B16	
B*39:10	B39	B16	
B*39:XX	B39	B16	
B17			BW4
B57		B17	
B*57:01	B57	B17	
B*57:02	B57	B17	
B*57:03	B57	B17	
B*57:XX	B57	B17	
B58		B17	
B*58:01	B58	B17	
B*58:02	B58	B17	
B*58:XX	B58	B17	
B18			BW6
B*18:01		B18	
B*18:XX		B18	
B21			
B49		B21	BW4
B*49:01	B49	B21	
B*49:XX	B49	B21	
B50		B21	BW6
B*40:05	B50	B21	
B*50:01	B50	B21	
B*50:XX	B50	B21	
B22			BW6
B54		B22	
B*54:01	B54	B22	
B*54:XX	B54	B22	
B55		B22	
B*55:01	B55	B22	
B*55:02	B55	B22	
B*55:XX	B55	B22	
B56		B22	
B*56:01	B56	B22	
B*56:XX	B56	B22	
B27			
B*27:02		B27	BW4
B*27:03		B27	BW4
B*27:05		B27	BW4
B*27:06		B27	BW4
B*27:07		B27	BW4
B*27:08		B27	BW6
B*27:XX		B27	BW4

ANTIGEN	SPLIT	BROAD	PUBLIC
B35			BW6
B*35:01		B35	
B*35:02		B35	
B*35:03		B35	
B*35:05		B35	
B*35:08		B35	
B*35:12		B35	
B*35:17		B35	
B*35:43		B35	
B*35:XX		B35	
B37			BW4
B*37:01		B37	
B*37:XX		B37	
B40			BW6
B*40:XX		B40	
B60		B40	
B*40:01	B60	B40	
B61		B40	
B*40:02	B61	B40	
B*40:06	B61	B40	
B41			BW6
B*41:01		B41	
B*41:02		B41	
B*41:XX		B41	
B42			BW6
B*42:01		B42	
B*42:XX		B42	
B46			BW6
B*46:01		B46	
B*46:XX		B46	
B47			BW4
B*47:01		B47	
B*47:XX		B47	
B48			BW6
B*48:01		B48	
B*48:XX		B48	
B53			BW4
B*53:01		B53	
B*53:XX		B53	
B59			BW4
B*59:01		B59	
B*59:XX		B59	
B67			BW6
B*67:01		B67	
B*67:XX		B67	
B73			BW6
B*73:01		B73	
B*73:XX		B73	
B78			BW6
B*78:01		B78	
B*78:XX		B78	
B81			BW6
B*81:01		B81	
B*81:XX		B81	
B82			BW6
B*82:01		B82	
B*82:XX		B82	
B83			BW6
B*83:01		B83	
B*83:XX		B83	
BW4			
BW6			

Abbildung 23: Bestimmung der HLA-B-Mismatches aus dem Eurotransplant Manual [ET22]

ANTIGEN	SPLIT	BROAD	DR51/52/53
DR1			
DRB1*01:01		DR1	
DRB1*01:02		DR1	
DRB1*01:03		DR1	
DRB1*01:XX		DR1	
DR2			DR51
DR15		DR2	
DRB1*15:01	DR15	DR2	
DRB1*15:02	DR15	DR2	
DRB1*15:03	DR15	DR2	
DRB1*15:XX	DR15	DR2	
DR16		DR2	
DRB1*16:01	DR16	DR2	
DRB1*16:02	DR16	DR2	
DRB1*16:XX	DR16	DR2	
DR3			DR52
DR17		DR3	
DRB1*03:01	DR17	DR3	
DRB1*03:04	DR17	DR3	
DR18		DR3	
DRB1*03:02	DR18	DR3	
DRB1*03:03	DR18	DR3	
DRB1*03:XX		DR3	
DR4			DR53
DRB1*04:01		DR4	
DRB1*04:02		DR4	
DRB1*04:03		DR4	
DRB1*04:04		DR4	
DRB1*04:05		DR4	
DRB1*04:06		DR4	
DRB1*04:07		DR4	
DRB1*04:08		DR4	
DRB1*04:09		DR4	
DRB1*04:10		DR4	
DRB1*04:11		DR4	
DRB1*04:12		DR4	
DRB1*04:XX		DR4	

ANTIGEN	SPLIT	BROAD	DR51/52/53
DR5			DR52
DR11		DR5	
DRB1*11:01	DR11	DR5	
DRB1*11:02	DR11	DR5	
DRB1*11:03	DR11	DR5	
DRB1*11:04	DR11	DR5	
DRB1*11:05	DR11	DR5	
DRB1*11:XX	DR11	DR5	
DR12		DR5	
DRB1*12:01	DR12	DR5	
DRB1*12:02	DR12	DR5	
DRB1*12:XX	DR12	DR5	
DR6			DR52
DR13		DR6	
DRB1*13:01	DR13	DR6	
DRB1*13:02	DR13	DR6	
DRB1*13:03	DR13	DR6	
DRB1*13:04	DR13	DR6	
DRB1*13:05	DR13	DR6	
DRB1*13:06	DR13	DR6	
DRB1*13:XX	DR13	DR6	
DR14		DR6	
DRB1*14:01/54	DR14	DR6	
DRB1*14:02	DR14	DR6	
DRB1*14:03	DR14	DR6	
DRB1*14:04	DR14	DR6	
DRB1*14:05	DR14	DR6	
DRB1*14:06	DR14	DR6	
DRB1*14:07	DR14	DR6	
DRB1*14:08	DR14	DR6	
DRB1*14:09	DR14	DR6	
DRB1*14:10	DR14	DR6	
DRB1*14:XX	DR14	DR6	

ANTIGEN	SPLIT	BROAD	DR51/52/53
DR7			DR53
DRB1*07:01		DR7	
DRB1*07:XX		DR7	
DR8			
DRB1*08:01		DR8	
DRB1*08:02		DR8	
DRB1*08:03		DR8	
DRB1*08:04		DR8	
DRB1*08:05		DR8	
DRB1*08:XX		DR8	
DR9			DR53
DRB1*09:01		DR9	
DRB1*09:XX		DR9	
DR10			
DRB1*10:01		DR10	
DRB1*10:XX		DR10	
DR51			
DRB5*01:01			DR51
DRB5*01:02			DR51
DRB5*02:01			DR51
DRB5*02:02			DR51
DRB5*XX			DR51
DR52			
DRB3*01:01			DR52
DRB3*02:			DR52
DRB3*03:01			DR52
DRB3*XX			DR52
DR53			
DRB4*01			DR53
DRB4*XX			DR53

Abbildung 24: Bestimmung der HLA-DR-Mismatches aus dem Eurotransplant Manual [ET22]

		Spender			
		0	A	B	AB
Empfänfer	0	✓	✗	✗	✗
	A	✓	✓	✗	✗
	B	✓	✗	✓	✗
	AB	✓	✓	✓	✓

Abbildung 25: Blutgruppenverträglichkeit zwischen Spender und Empfänger

Tabelle 17: IDs von Spender, Empfänger und Transplantation zum Verknüpfen der einzelnen Tabellen des deutschen Transplantationsregisters nach Abbildung 6

Tabelle	Variablenname	ID
spender_postmortem	SPostmIdSpenderNrETET	Spender-ID
	SPostmIdSpenderNrETIQTIG	Spender-ID
	SPostmIdSpenderNrETDSO	Spender-ID
spender_postmortem_diagnosen	SPostmDiagnosenIdSpenderNrETET	Spender-ID
	SPostmDiagnosenIdSpenderNrETDSO	Spender-ID
spender_postmortem_labor_virologie	SPostmLaborVIRIdSpenderNrETET	Spender-ID
	SPostmLaborVIRIdSpenderNrETDSO	Spender-ID
spender_postmortem_labor_ klinische_chemie	SPostmLaborKCIdSpenderNrETET	Spender-ID
	SPostmLaborKCIdSpenderNrETIQTIG	Spender-ID
	SPostmLaborKCIdSpenderNrETDSO	Spender-ID
spender_postmortem_labor_hla	SPostmLaborHLAIdSpenderNrETET	Spender-ID
	SPostmLaborHLAIdSpenderNrETDSO	Spender-ID
organ_entnahme_niere	ONIdSpenderNrETIQTIG	Spender-ID
	ONIdSpenderNrETET	Spender-ID
	ONIdSpenderNrETDSO	Spender-ID
	ONIdTransplantationsnummerETET	Transplantations-ID
transplantation	TIdSpenderNrETET	Spender-ID
	TIdTransplantationsnummerETET	Transplantations-ID
	TIdEmpfaengerNrETET	Empfänger-ID
empfänger	EIdEmpfaengerNrETET	Empfänger-ID
	EIdEmpfaengerNrETIQTIG	Empfänger-ID
empfaenger_immunologie	EImmIdEmpfaengerNrETET	Empfänger-ID
folow_up_niere	FNIdEmpfaengerNrETIQTIG	Empfänger-ID
	FNIdEmpfaengerNrETET	Empfänger-ID

		Deutschland		USA	
Jahr	Parameter	Wert	p-value	Wert	p-value
1	Intercept	0.0303	0.9227	0.8128	1.35e-04 ***
	Slope	1.0283	0.8462	1.296	1.45e-04 ***
	Multivariater Test		0.9069		6.83e-04 ***
2	Intercept	-0.1688	0.4205	0.3143	0.0199 *
	Slope	0.9156	0.465	1.1379	0.0223 *
	Multivariater Test		0.7133		0.0659
3	Intercept	-0.1472	0.3764	0.175	0.1394
	Slope	0.8877	0.2926	1.0912	0.145
	Multivariater Test		0.5387		0.3355
4	Intercept	-0.0775	0.6351	0.1120	0.2154
	Slope	0.9317	0.5707	1.0669	0.2385
	Multivariater Test		0.8423		0.4573
5	Intercept	-0.152	0.3854	-0.0075	0.9409
	Slope	0.8251	0.2415	0.9818	0.8088
	Multivariater Test		0.4523		0.7733
6	Intercept	-0.1467	0.3991	-0.0626	0.3941
	Slope	0.7890	0.2363	0.9290	0.2716
	Multivariater Test		0.4657		0.4202
7	Intercept	-0.0501	0.6979	-0.0457	0.4706
	Slope	0.8433	0.3197	0.9367	0.3447
	Multivariater Test		0.4845		0.5656

Tabelle 18: Kalibrierung der Cox-Regressionsmodelle auf eigenen Testdaten

Tabelle 19: Kalibrierung des Random Survival Forest auf den eigenen Testdaten

		USA		Deutschland	
Jahr	Parameter	Wert	p-Wert	Wert	p-Wert
1	Intercept	0.4777	0.0086 **	0.0815	0.6948
	Slope	1.1685	0.0111 *	1.0574	0.5563
	Multivariater Test		0.0277 *		0.6254
2	Intercept	0.2122	0.1654	-0.0358	0.8702
	Slope	1.0903	0.1866	1.0049	0.9675
	Multivariater Test		0.3572		0.7603
3	Intercept	0.1972	0.1491	-0.0331	0.7995
	Slope	1.1040	0.1516	0.9764	0.7777
	Multivariater Test		0.3526		0.9602
4	Intercept	0.2499	0.0459 *	0.0219	0.8040
	Slope	1.1605	0.0430 *	1.0220	0.7384
	Multivariater Test		0.1290		0.9311
5	Intercept	0.1403	0.2839	0.0111	0.9284
	Slope	11027	0.3007	1.0007	0.9948
	Multivariater Test		0.5621		0.9784
6	Intercept	0.0909	0.2637	0.0339	0.7284
	Slope	1.0780	0.2840	1.0351	0.7426
	Multivariater Test		0.5348		0.9406
7	Intercept	0.0967	0.1720	0.0578	0.4988
	Slope	1.1074	0.1626	1.0560	0.6188
	Multivariater Test		0.3751		0.7911

Signifikanzniveau: 0.001(***), 0.01(**), 0.05(*)

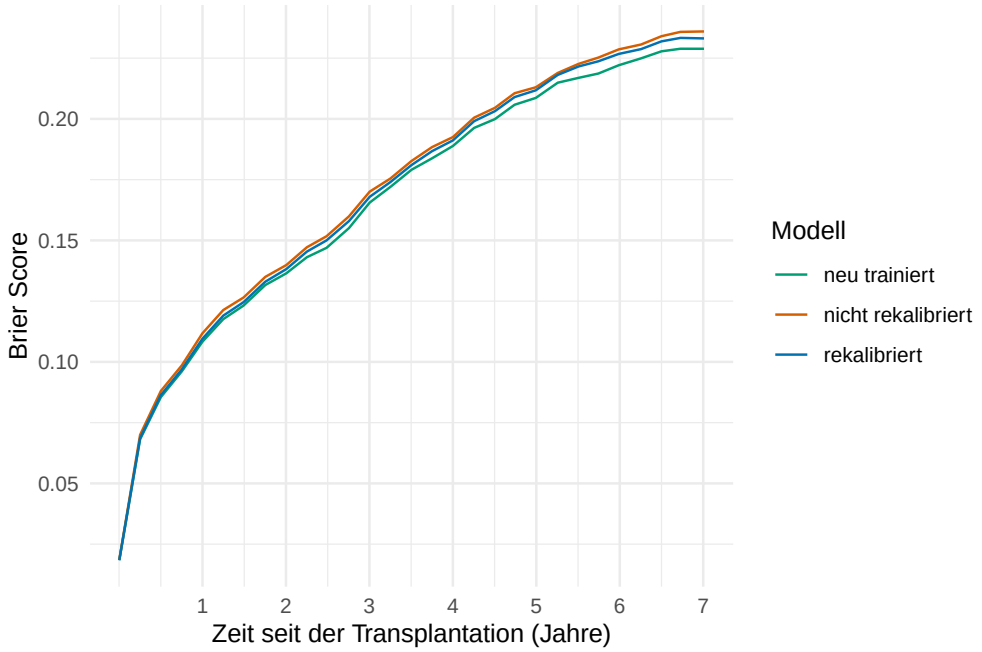

Abbildung 26: Vergleich des Brier Scores im Verlauf über die Zeit für den neu trainierten, nicht rekalibrierten und rekalibrierten Random Survival Forest

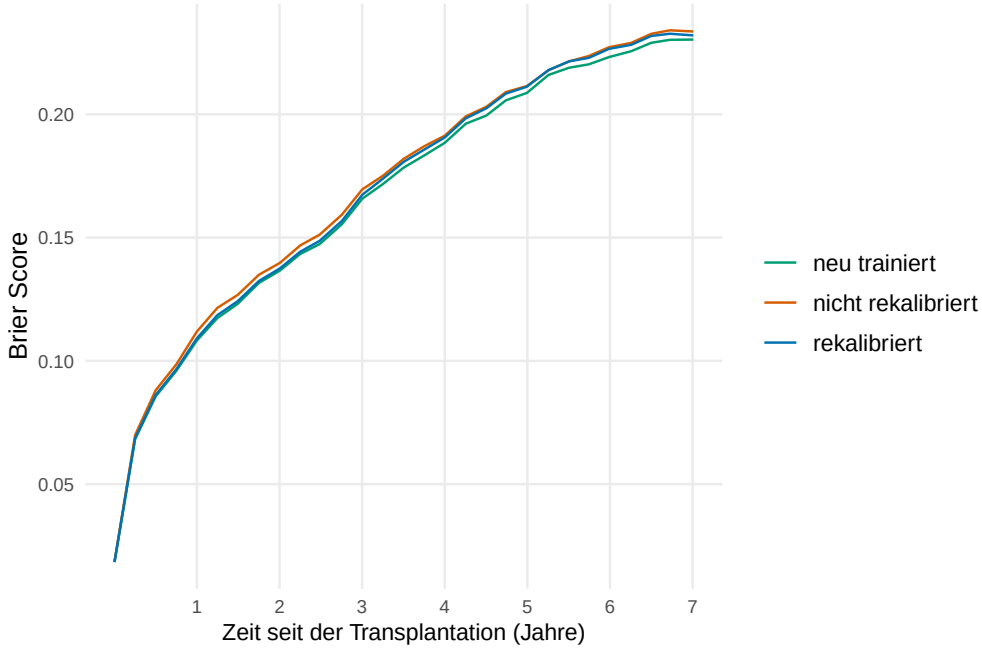

Abbildung 27: Vergleich des Brier Scores im Verlauf über die Zeit für das neu trainierte, nicht rekalibrierte und rekalibrierte Cox-Regressionsmodell