

Aus dem Institut für Medizinische Biometrie,
Informatik und Epidemiologie
Direktor: Prof. Dr. Matthias Schmid

Fortgeschrittene Methoden zur Modellierung
von diskreten Ereigniszeiten

Habilitationsschrift
zur Erlangung der Venia Legendi
der Medizinischen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn
für das Lehrgebiet
“Medizinische Biometrie”

vorgelegt von

Dr. rer. nat. Moritz Maximilian Berger

aus Traunstein

2023

Die folgenden Originalarbeiten sind Grundlage der vorliegenden kumulativen Habilitationsschrift zum Thema “Fortgeschrittene Methoden zur Modellierung von diskreten Ereigniszeiten”:

1. Berger, M., and Schmid, M. (2018). Semiparametric regression for discrete time-to-event data, *Statistical Modelling* 18, 322-345.
doi: 10.1177/1471082X17748084. (2.039)¹
2. Puth, M.-T., Tutz, G., Heim, N., Schmid, M., and Berger, M. (2020). Tree-based modeling of time-varying coefficients in discrete time-to-event models, *Lifetime Data Analysis* 26, 545-572.#
doi: 10.1007/s10985-019-09489-7. (1.588)¹
Reproduced with permission from Springer Nature.
3. Berger, M., Welchowski, T., Schmitz-Valckenberg, S., and Schmid, M. (2019). A classification tree approach for the modeling of competing risks in discrete time, *Advances in Data Analysis and Classification* 13, 965-990.#
doi: 10.1007/s11634-018-0345-y. (2.134)¹
Reproduced with permission from Springer Nature.
4. Berger, M.*, Schmid, M.*, Welchowski, T., Schmitz-Valckenberg, S., and Beyersmann, J. (2020). Subdistribution hazard models for competing risks in discrete time, *Biostatistics* 21, 449-466.
doi: 10.1093/biostatistics/kxy069. (5.899)¹
*equal contributions
5. Berger, M., and Schmid, M. (2022). Assessing the calibration of subdistribution hazard models in discrete time, *Canadian Journal of Statistics* 50, 572-591.
doi: 10.1002/cjs.11633. (0.875)¹

Habitationskolloquium: 27. Oktober 2022

¹2020 Journal Impact Factor des SCI Journal Citation Reports

Inhaltsverzeichnis

1	Einleitung	4
1.1	Zensierungsmechanismen	5
1.2	Diskrete Ereigniszeiten	7
1.3	Konkurrierende Ereignisse	9
1.4	Parametrische und nicht-parametrische Regressionsmodelle	11
1.5	Anwendungsbeispiele	14
1.6	Ziele der vorliegenden Arbeit	17
1.7	Software	18
2	Ergebnisse	19
2.1	Berger et al., Statistical Modelling 18, 322-345	19
2.2	Puth et al., Lifetime Data Analysis 26, 545-572	44
2.3	Berger et al., Advances in Data Analysis and Classification 13, 965-990	73
2.4	Berger, Schmid et al., Biostatistics 21, 449-466	100
2.5	Berger et al., Canadian Journal of Statistics 50, 572-591	147
3	Diskussion	184
4	Zusammenfassung	188
	Literatur	189

1 Einleitung

In klinischen und epidemiologischen Studien unterscheidet man zwischen Querschnittserhebungen, bei denen Daten einmalig zu einem bestimmten Zeitpunkt erhoben werden, und Längsschnitterhebungen, sogenannten longitudinalen Studien, bei denen die Datenerhebung mehrfach über die Zeit erfolgt. In longitudinalen Kohortenstudien wird eine bestimmte Gruppe von Patienten/Patientinnen, die spezifische Einschlusskriterien erfüllt, über einen vorgegebenen Zeitraum hinweg beobachtet (Petrie und Sabin, 2019). Die Untersuchungen erfolgen dabei üblicherweise zu Beginn der Studie (bezeichnet als “Baseline”-Untersuchung) und in regelmäßigen Abständen zu vorab festgelegten, späteren Zeitpunkten (bezeichnet als “Follow-up”-Untersuchungen). Im Rahmen dieser Arbeit werden die Analysen dreier longitudinaler, klinischer Studien vorgestellt, darunter eine Studie unter Patienten/Patientinnen mit odontogenen Infektionen (Heim et al., 2019), die MODIAMD-Studie (Steinberg et al., 2016) über altersbedingte Makuladegeneration sowie eine Studie zum Auftreten von Lungenentzündungen unter Patienten/Patientinnen, die auf eine Intensivstation aufgenommen werden mussten (Wolkewitz et al., 2008). Neben der Bestimmung von klinischen Variablen (z.B. der Funktion von Organen) und Laborparametern (z.B. Cholesterin-Werten) wird in vielen Studien erfasst ob, und, wenn ja, wann ein bestimmtes Ereignis aufgetreten ist. Die statistische Analyse der Zeit bis zum Eintreten dieses interessierenden Ereignisses bezeichnet man als Ereigniszeit- oder Überlebenszeitanalyse (Klein und Moeschberger, 2003). Klassische Beispiele in klinischen Studien sind die Zeit bis zum Tod und das Auftreten, Fortschreiten oder der Rückfall einer Krankheit.

In dieser Arbeit werden neuartige Methoden der Ereigniszeitanalyse vorgestellt, die auf den Fall zugeschnitten sind, dass die Ereigniszeiten auf einer diskreten Skala gemessen wurden (siehe Kapitel 1.2). Ziel ist es dabei immer, ein Regressionsmodell aufzustellen, das die Beziehung zwischen der Ereigniszeit T und einer Menge erklärender Variablen X beschreibt (siehe Kapitel 1.4), und damit zusätzlich relevante Risikofaktoren für das Auftreten des interessierenden Ereignisses zu identifizieren. Ausgangspunkt aller Entwicklungen sind die grundlegenden Methoden der Arbeit von Tutz und Schmid (2016). Die charakterisierenden Eigenschaften von Ereigniszeitdaten, die im Folgenden näher erläutert werden, sind (i) die sequentielle, longitudinale Struktur, (ii) das Vorhandensein von zensierten Beobachtungen, und (iii) das eventuelle Auftreten mehrerer, möglicher Ereignisse.

1.1 Zensierungsmechanismen

Eine wichtige Besonderheit von Ereigniszeitdaten ist die sogenannte Zensierung. Zensierte Daten liegen vor, wenn der exakte Ereigniszeitpunkt für einen Teil der Patienten/Patientinnen nicht bekannt ist (Klein und Moeschberger, 2003).

Man spricht von einer rechtszensierten Beobachtung, wenn der Zeitpunkt des Eintritts in die Studie bekannt ist, der Zeitpunkt, zu dem das Ereignis aufgetreten ist, jedoch nicht. Dies kann der Fall sein, weil ein/e Patient/in vorzeitig aus der Studie ausscheidet (z.B. aufgrund von Zeitkonflikten oder Überbelastung) oder, weil das interessierende Ereignis erst nach Beendigung der Studie realisiert wird. Andererseits spricht man von einer linkszensierten Beobachtung, wenn das interessierende Ereignis bereits vor Beginn der Studie eingetreten ist und man daher den genauen Zeitpunkt nicht zurückverfolgen kann. Dies ist z.B. der Fall, wenn bei einer Untersuchung eine Infektionskrankheit diagnostiziert wird, jedoch nicht nachvollzogen werden kann, wann der/die Patient/in infiziert worden ist.

Bei allen in dieser Arbeit vorgestellten Anwendungen handelt es sich um rechtszensierte Studiendaten. Bezeichne T die Zeit bis zum Eintreten des interessierenden Ereignisses und C die Zensierungszeit, so ergibt sich die beobachtete Ereigniszeit als $\tilde{T} = \min(T, C)$. Das heißt, falls T kleiner oder gleich C ist, beobachtet man die wahre Ereigniszeit, anderenfalls die Zensierungszeit (Klein und Moeschberger, 2003).

Ein zweites Phänomen, dem man in Ereigniszeitdaten begegnen kann, ist Trunkierung. Trunkierte Daten liegen vor, wenn bestimmte Patienten/Patientinnen aufgrund ihrer Ereigniszeit systematisch aus der Studie ausgeschlossen werden (Klein und Moeschberger, 2003). Dabei spricht man von Rechtstrunkierung, falls nur Patienten/Patientinnen berücksichtigt werden, deren Ereigniszeit kleiner ist als ein bestimmter Schwellenwert. Ein Beispiel hierfür stellt die Schätzung der Inkubationszeit von AIDS nach einer HIV-Infektion dar. In eine Stichprobe zur Untersuchung dieser Fragestellung können nur Patienten/Patientinnen eingeschlossen werden, bei denen die AIDS-Erkrankung vor einem fest definierten Zeitpunkt ausgebrochen ist (vgl. Kalbfleisch und Lawless, 1991). Andererseits spricht man von Linkstrunkierung, falls nur Beobachtungen berücksichtigt werden, deren Ereigniszeit größer ist als ein bestimmter Schwellenwert. Ein Beispiel dazu stellen Studien unter älteren Patienten/Patientinnen dar. Aus einer solchen Stichprobe werden alle Patienten/Patientinnen ausgeschlossen, die vor Erreichen eines

bestimmten Lebensalters verstarben. Methoden für trunkierte Daten stellen ein eigenes Forschungsgebiet dar und werden in dieser Arbeit nicht behandelt.

Für den Zusammenhang zwischen der Ereigniszeit T und der Zensierungszeit C unterscheidet man im Allgemeinen drei mögliche Mechanismen, siehe Kalbfleisch und Prentice (2002).

Zufällige Zensierung (engl. “random censoring”)

Die Größen T und C sind zwei stochastisch unabhängige Zufallsvariablen. Unabhängigkeit gilt zumindest bedingt auf eine Menge erklärender Variablen (bedingte Unabhängigkeit gegeben X).

Unabhängige Zensierung (engl. “independent censoring”)

Die Patienten/Patientinnen, die zum Zeitpunkt t zensiert sind, sind für alle Patienten/Patientinnen, die zum Zeitpunkt t noch kein Ereignis erlebt haben und weiter unter Beobachtung stehen, repräsentativ. Dies gilt für jede Untergruppe bezüglich der erklärenden Variablen X . Die zufällige Zensierung stellt einen Spezialfall der unabhängigen Zensierung dar (Kalbfleisch und Prentice, 2002).

Nicht-informative Zensierung (engl. “non-informative censoring”)

Die Verteilung der Zensierungszeiten C hängt von keinem Parameter ab, der für die Modellierung der Ereigniszeiten T herangezogen wurde. Der Zensierungsmechanismus enthält somit keinerlei Informationen über die Parameter, die die Verteilung von T bestimmen (Kalbfleisch und Prentice, 2002). Im Fall unabhängiger Zensierung gilt in der Regel auch nicht-informative Zensierung. Als hypothetisches Beispiel mit unabhängiger, aber informativer Zensierung nennen Kalbfleisch und Prentice (2002) eine Studie, in der die Zensierungszeiten C von Ereigniszeiten ähnlicher Patienten/Patientinnen abhängen, die jedoch nicht in die Studie eingeschlossen wurden.

In dieser Arbeit werden ausschließlich Analysemethoden vorgestellt, die voraussetzen, dass die Ereigniszeiten die Annahmen von zufälliger und nicht-informativer Zensierung erfüllen.

1.2 Diskrete Ereigniszeiten

Der Fokus der Ereigniszeitanalyse liegt darin, die Verteilung der Ereigniszeiten, d.h. den dynamischen Verlauf über die Zeit, möglichst genau zu erfassen. Für die Modellierung ist dabei entscheidend, welche Zeitskala in den Daten zugrunde liegt. Viele Methoden zur Modellierung von Ereigniszeitdaten nehmen an, dass die Zeit auf einer stetigen Skala gemessen wurde, d.h., dass T Werte aus den positiven reellen Zahlen annehmen kann ($T \in \mathbb{R}_0^+$). Eine ausführliche Darstellung der Methodik für stetige Ereigniszeiten findet sich unter anderem in Kalbfleisch und Prentice (2002), Klein und Moeschberger (2003), Lawless (2003) und Kleinbaum und Klein (2012).

In praktischen Anwendungen, wie z.B. in vielen longitudinalen, klinischen Studien, wird die Zeit entgegen der obigen Annahme jedoch in der Regel auf einer diskreten Skala gemessen, d.h. T nimmt lediglich Werte aus den natürlichen Zahlen an ($T \in \mathbb{N}$). Für diskrete Ereigniszeiten lassen sich grundsätzlich zwei Fälle unterscheiden, die nachfolgend beschrieben werden, siehe Tutz und Schmid (2016).

Diskretisierte Ereigniszeiten

In klinischen Studien, bei denen Follow-up-Untersuchungen zu festgelegten Zeitpunkten stattfinden, entsprechen die (diskreten) Ereigniszeiten $t = 1, \dots, k$, den Intervallen $[q_{t-1}, q_t)$ mit den stetigen Grenzen $0 = q_0 < q_1 < \dots < q_k = \infty$. Die eigentlich zugrunde liegenden, stetigen Ereigniszeiten sind somit gruppiert und in feste Zeitintervalle unterteilt. Liegen gruppierte Daten vor, spricht man auch von intervallzensierten Daten, wobei Beobachtungen mit einer identischen Ereigniszeit als Bindungen (engl. “ties”) bezeichnet werden (Sun, 2007).

Diskretisierte Daten finden sich auch in Studien zur Hospitalisierung (siehe Kapitel 2.2) oder intensivmedizinischen Behandlung (z.B. bezüglich des Auftretens von nosokomialen Infektionen, siehe Kapitel 2.4) von Patienten/Patientinnen.

Immanent diskrete Ereigniszeiten

Ereigniszeiten können natürlicherweise von diskreter Struktur sein und Werten der natürlichen Zahlen entsprechen. Ein oft genanntes Beispiel sind Studien zur Fertilität, in denen die Zeit von der Pubertät bis zur Geburt des ersten Kindes analysiert wird. Für diese Fragestellung ist es sinnvoll, nicht die Zeit selbst,

sondern die Anzahl an Menstruationszyklen als natürliche Zeiteinheit heranzuziehen, da die Länge eines Zyklus zwischen jungen Frauen variieren kann (vgl., Scheike und Keiding, 2006).

Diskrete Hazard- und Überlebensfunktion

Die wichtigste Größe, um die Verteilung des interessierenden Ereignisses über die Zeit zu beschreiben, ist die diskrete Hazardfunktion

$$\lambda(t|X) = P(T = t|T \geq t, X), \quad t = 1, \dots, k, \quad (1)$$

die der Wahrscheinlichkeit entspricht, dass das Ereignis zum Zeitpunkt t eintritt, unter der Bedingung, dass bis zum Zeitpunkt t noch kein Ereignis aufgetreten ist (Tutz und Schmid, 2016). Die Hazardfunktion in Gleichung (1) ist spezifisch für bestimmte Patienten/Patientinnen, da sie zusätzlich auf die erklärenden Variablen X bedingt. Die zugehörige diskrete Überlebensfunktion ist gegeben durch

$$S(t|X) = P(T > t|X) = \prod_{s=1}^t (1 - \lambda(s|X)), \quad t = 1, \dots, k, \quad (2)$$

und entspricht der Wahrscheinlichkeit, dass das Ereignis erst nach dem Zeitpunkt t eintritt (Tutz und Schmid, 2016). Damit berechnet sich die unbedingte Wahrscheinlichkeit für ein Ereignis zum Zeitpunkt $t \in \{1, \dots, k\}$ durch

$$P(T = t|X) = \lambda(t|X) \prod_{s=1}^{t-1} (1 - \lambda(s|X)) = \lambda(t|X)S(t-1|X), \quad (3)$$

wobei $S(0|X) := 1$. Im Fall gruppierter Daten entspricht $P(T = t|X)$ der Wahrscheinlichkeit, dass das Ereignis im Intervall $[q_{t-1}, q_t)$ auftritt.

Methoden zur Modellierung der diskreten Hazardfunktion sind Gegenstand der Kapitel 2.1 und 2.2. Wie bereits in Tutz und Schmid (2016) dargestellt, lassen sich die vorgeschlagenen Modelle in die Klasse der binären Regressionsmodelle einbetten. Daraus ergeben sich zahlreiche mögliche Erweiterungen, um komplexere Zusammenhänge zwischen der Hazardfunktion und den erklärenden Variablen flexibel abzubilden.

1.3 Konkurrierende Ereignisse

Oftmals ist in klinischen Studien nicht nur ein bestimmtes, einzelnes Ereignis von Interesse, sondern das Auftreten mehrerer möglicher Ereignisse. Kann dabei jedes Ereignis zu jedem Zeitpunkt als Erstes eintreten und ist jedes Ereignis ein absorbierendes Ereignis, sodass ein/eine Patient/in im Studienzeitraum höchstens eines erlebt, spricht man von konkurrierenden Ereignissen (Beyersmann et al., 2011). Konkurrierende Ereignisse können unterschiedliche Todesursachen oder verschiedene Arten einer Erkrankung sein. Ziel der MODIAMD-Studie (Steinberg et al., 2016) ist es beispielsweise, das Auftreten zweier unterschiedlicher, sich gegenseitig ausschließender Ausprägungen der altersbedingten Makuladegeneration im Spätstadium (der feuchten oder trockenen Form) zu analysieren. In der Studie von Wolkewitz et al. (2008) können Patienten/Patientinnen entweder während des Aufenthalts auf der Intensivstation eine Lungenentzündung erleiden, gesund aus dem Krankenhaus entlassen werden oder vorzeitig versterben. Der eventuelle Tod oder eine mögliche Entlassung stellen daher zwei konkurrierende Ereignisse für das Auftreten einer Lungenentzündung dar. Für eine Einführung zur Analyse von konkurrierenden Ereignissen in stetiger Zeit sei verwiesen auf Putter et al. (2007) und Andersen et al. (2012).

Ereignis-spezifische Hazardfunktion und kumulative Inzidenzfunktion

Bei der Analyse von diskreten Ereigniszeiten mit konkurrierenden Ereignissen betrachtet man nicht nur eine diskrete Hazardfunktion, wie in (1) definiert, sondern eine ereignis-spezifische Hazardfunktion für jedes der möglichen Ereignisse. Bezeichne $\varepsilon \in \{1, \dots, J\}$ die unterschiedlichen Ereignisse, so ist die diskrete, ereignis-spezifische Hazardfunktion für ein Ereignis vom Typ j definiert durch

$$\lambda_j(t|X) = P(T = t, \varepsilon = j | T \geq t, X), \quad j = 1, \dots, J, \quad t = 1, \dots, k. \quad (4)$$

Die Hazardfunktion (4) entspricht der Wahrscheinlichkeit, dass das Ereignis j zum Zeitpunkt t eintritt, unter der Bedingung, dass bis zum Zeitpunkt t noch kein Ereignis aufgetreten ist, und bedingt auf erklärende Variablen X (Tutz und Schmid, 2016). Kombiniert man alle ereignis-spezifischen Hazardfunktionen, kann man insgesamt den dynamischen Verlauf über die Zeit (unabhängig von der

Art des Ereignisses) über die aggregierte Hazardfunktion

$$\lambda(t|X) = \sum_{j=1}^J \lambda_j(t) = P(T = t|T \geq t, X), \quad t = 1, \dots, k, \quad (5)$$

beschreiben. Die Funktion $\lambda(t|X)$ entspricht der bedingten Wahrscheinlichkeit, dass ein Ereignis jeglicher Art zum Zeitpunkt t eintritt (Tutz und Schmid, 2016). Die bedingte Wahrscheinlichkeit, dass ein Ereignis erst nach Zeitpunkt t eintritt, ergibt sich damit durch $P(T > t|T \geq t, X) = 1 - \lambda(t)$. Die zugehörige diskrete Überlebensfunktion ergibt sich analog zu (2) als

$$S(t|X) = P(T > t|X) = \prod_{s=1}^t (1 - \lambda(s|X)) \quad t = 1, \dots, t, \quad (6)$$

und entspricht hier der (unbedingten) Wahrscheinlichkeit, dass ein Ereignis jeglicher Art erst nach dem Zeitpunkt t eintritt (Tutz und Schmid, 2016). Die unbedingte Wahrscheinlichkeit für das Auftreten eines Ereignisses jeglicher Art berechnet sich analog zu (3) durch $P(T = t|X) = \lambda(t|X)S(t-1|X)$, wobei auch hier $S(0|X) := 1$ gilt. Unter Verwendung der Überlebensfunktion (6) lautet eine gängige Darstellung der ereignis-spezifischen Hazardfunktion

$$\lambda_j(t|X) = \frac{f_j(t|X)}{S(t-1|X)}, \quad j = 1, \dots, J, \quad t = 1, \dots, k, \quad (7)$$

wobei $f_j(t|X)$ der Wahrscheinlichkeit $P(T = t, \varepsilon = j|X)$ entspricht (Lee, 2017).

Eine weitere, wichtige Größe, die zur Beschreibung von Ereigniszeitdaten mit konkurrierenden Ereignissen herangezogen wird, ist die kumulative Inzidenzfunktion für ein Ereignis vom Typ j , definiert durch

$$F_j(t|X) = P(T \leq t, \varepsilon = j|X), \quad j = 1, \dots, J, \quad t = 1, \dots, k. \quad (8)$$

Die kumulative Inzidenzfunktion $F_j(t|X)$ entspricht der Wahrscheinlichkeit, dass das Ereignis j vor oder zum Zeitpunkt t eintritt, bedingt auf erklärende Variablen X (Fine und Gray, 1999; Klein und Andersen, 2005). Per Definition liegt der Wertebereich von F_j zwischen 0 und $F_j(k|X) = P(\varepsilon = j|X) \leq 1$. Kombiniert man die Gleichungen (4) bis (6), lässt sich die kumulative Inzidenzfunktion für

ein Ereignis vom Typ j zum Zeitpunkt t darstellen als

$$F_j(t|X) = \sum_{s=1}^t \left(\lambda_j(s|X) \prod_{q=1}^{s-1} (1 - \lambda(q|X)) \right) = \sum_{s=1}^t \left(\lambda_j(s|X) S(s-1|X) \right), \quad (9)$$

siehe Lee (2017). Aus Gleichung (9) folgt, dass die kumulative Inzidenzfunktion für ein Ereignis vom Typ j von den ereignis-spezifischen Hazardfunktionen aller möglichen Ereignisse $\lambda_1, \dots, \lambda_J$ abhängt. Insbesondere gibt es keine unmittelbare Verknüpfung zwischen λ_j und F_j (Beyersmann et al., 2011).

Im Allgemeinen kann man für die Analyse von Ereigniszeiten mit konkurrierenden Ereignissen zwei Vorgehensweisen unterscheiden:

- Modellierung jedes Ereignisses $j = 1, \dots, J$, über ereignis-spezifische Hazardfunktionen. Im Falle stetiger Ereigniszeitdaten wird dabei ein separates Regressionsmodell für jedes Ereignis j spezifiziert, wobei alle Patienten/Patientinnen, die ein konkurrierendes Ereignis erleben, jeweils wie zensierte Beobachtungen behandelt werden (Prentice et al., 1978). Die Modellierung der ereignis-spezifischen Hazardfunktionen im Fall diskreter Ereigniszeiten ist Gegenstand von Kapitel 2.3. Insbesondere wird dargestellt, wie sich das vorgeschlagene Modell für diskrete Ereigniszeiten auf etablierte Verfahren für kategoriale Datenanalyse zurückführen lässt.
- Modellierung eines einzelnen, interessierenden Ereignisses vom Typ j über die kumulative Inzidenzfunktion unter Berücksichtigung der anderen, konkurrierenden Ereignisse. Für stetige Ereigniszeitdaten wurden Ansätze dieser Art von Fine und Gray (1999) und Klein und Andersen (2005) vorgeschlagen. Die Modellierung der kumulativen Inzidenzfunktion im Fall diskreter Ereigniszeiten ist Gegenstand der Kapitel 2.4 und 2.5. Es wird gezeigt, wie sich das vorgeschlagene Modell in die Klasse der binären Regressionsmodelle einbetten lässt.

1.4 Parametrische und nicht-parametrische Regressionsmodelle

Grundprinzip aller in dieser Arbeit vorgeschlagenen Methoden ist es, ein Regressionsmodell aufzustellen, das den Zusammenhang zwischen der Hazardfunktion, der ereignis-spezifischen Hazardfunktionen oder der kumulativen Inzidenzfunkti-

on und einer Menge erklärender Variablen quantitativ beschreibt, und es ermöglicht, Vorhersagen der Ereigniszeiten für zukünftige Patienten/Patientinnen zu treffen. Wie in den Kapiteln 2.1 bis 2.5 dargestellt, lassen sich die Modelle für diskrete Ereigniszeiten auf die Struktur von klassischen generalisierten Regressionsmodellen (Fahrmeir et al., 2013) zurückführen. Die Konzepte grundlegender, regressionsanalytischer Verfahren, auf die später Bezug genommen wird, werden im Folgenden kurz skizziert.

Betrachte man allgemein den Zusammenhang zwischen einer interessierenden, abhängigen Variablen Y (Zielvariable) und einer Vielzahl an erklärenden Variablen X . In einem generalisierten linearen Regressionsmodell (McCullagh und Nelder, 2019) wird typischerweise der bedingte Erwartungswert $\mu = \mathbb{E}(Y|X)$ einer Zielvariable als Funktion der erklärenden Variablen in der Form

$$\mu = h(\eta(X)), \quad (10)$$

modelliert. Dabei verknüpft die Antwortfunktion $h(\cdot)$ den Erwartungswert mit einer linearen Vorhersagefunktion der Form

$$\eta(X) = \beta_0 + X\boldsymbol{\beta} = \beta_0 + X_1\beta_1 + \dots + X_p\beta_p. \quad (11)$$

Die Parameter β_0 und $\boldsymbol{\beta}^\top = (\beta_1, \dots, \beta_p)$ stellen reellwertige Regressionskoeffizienten dar. In einem generalisierten additiven Regressionsmodell (Wood, 2017) werden (für einen Teil der erklärenden Variablen) die linearen Terme durch glatte Funktionen ersetzt und man erhält eine Vorhersagefunktion der Form

$$\eta(X) = \beta_0 + f_1(X_1) + \dots + f_q(X_q) + X_{q+1}\beta_{q+1} + \dots + X_p\beta_p, \quad (12)$$

wobei die Effekte der erklärenden Variablen X_1, \dots, X_q den unspezifizierten, glatten Funktionen f_1, \dots, f_q entsprechen und die erklärenden Variablen X_{q+1}, \dots, X_p mit linearen Effekten in die Modellgleichung eingehen (Fahrmeir et al., 2013). Zur Modellierung der glatten Funktionen wird üblicherweise angenommen, dass sich jede Funktion als gewichtete Summe von Basisfunktionen darstellen lässt, nämlich als

$$f_j(X_j) = \sum_{m=1}^M \phi_m(X_j)\beta_{jm}, \quad (13)$$

mit M festen Basisfunktionen ϕ_1, \dots, ϕ_M und zugehörigen, zu schätzenden Koeffizienten $\beta_{j1}, \dots, \beta_{jM}$. Eine gängige Wahl der Basisfunktionen sind B-Splines (De Boor, 1978). Eine flexible Schätzung von f_j erhält man, wenn man eine große Zahl an B-Spline-Basisfunktionen verwendet (z.B., 20 bis 40) und gleichzeitig einen Penalisierungsterm einführt, der zu großer Variabilität der Schätzung vorbeugt (Eilers und Marx, 1996). Die Modellierung mithilfe von penalisierten Splines (P-Splines) wird in Kapitel 2.1 und 2.2 im Detail betrachtet.

Für eine binäre Zielvariable $Y \in \{0, 1\}$ mit der Verteilungsannahme $Y \sim B(1, \pi)$ ergibt sich Modellgleichung (10) mit Antwortfunktion $h(\cdot) \in [0, 1]$ zu

$$\mu = P(Y = 1|X) = \pi = h(\eta(X)). \quad (14)$$

Das populärste binäre Regressionsmodell ist das logistische Modell (Logit-Modell) mit Antwortfunktion $h(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$. Eine Alternative, die für die Modellierung der kumulativen Inzidenzfunktion in Kapitel 2.4 und 2.5 herangezogen wird, ist das Gompertz-Modell, das über die inverse komplementäre log-log-Funktion $h(\cdot) = 1 - \exp(-\exp(\cdot))$ definiert ist (Fahrmeir et al., 2013). In Kapitel 2.1 und 2.4 wird erläutert, wie sich die Methoden zur Modellierung der diskreten Hazardfunktion und der kumulativen Inzidenzfunktion in die Klasse der (gewichteten) binären Regressionsmodelle einbetten lassen und, wie damit etablierte Softwareprogramme zur Schätzung der Regressionskoeffizienten herangezogen werden können.

Für eine kategoriale Zielvariable $Y \in \{1, \dots, K\}$ mit K Kategorien lässt sich das Logit-Modell zum multinomialen Logit-Modell erweitern (Tutz, 2012). Mit Referenzkategorie K lauten die Modellgleichungen

$$P(Y = r|X) = \pi_r = \frac{\exp(\eta_r)}{1 + \sum_{s=1}^{K-1} \exp(\eta_s)}, \quad r = 1, \dots, K-1, \quad (15)$$

mit kategoriespezifischen Vorhersagefunktionen $\eta_r = \beta_{0r} + X\beta_r$. Für die Wahrscheinlichkeit der Referenzkategorie gilt dann $\pi_K = 1/(1 + \sum_{s=1}^{K-1} \exp(\eta_s))$. Die in Kapitel 2.3 vorgeschlagene Methode zur Modellierung der diskreten, ereignisspezifischen Hazardfunktionen lässt sich in die Klasse der kategorialen Regressionsmodelle einbetten.

Generalisierte additive Modelle sind wesentlich flexibler als generalisierte lineare Modelle und behalten aufgrund der additiven Struktur der Vorhersagefunk-

tion gleichzeitig ihre einfache Interpretierbarkeit. Nachteil von Modellen der Form (12) ist jedoch, dass die Vorhersagefunktion nur Haupteffekte der erklärenden Variablen berücksichtigt. Es ist insbesondere schwierig, mögliche Interaktionen, vor allem höherer Ordnung, ohne konkretes (klinisches) Vorwissen mit einzu beziehen. Abhilfe dafür schafft die nicht-parametrische Modellierung mithilfe von rekursiver Partitionierung, sogenannten Baum-basierten Verfahren. Die bekanntesten Methoden sind classification and regression trees, abgekürzt durch CART (Breiman et al., 1984), conditional inference trees (Hothorn et al., 2006) und der C4.5-Algorithmus (Quinlan, 1993). Das Grundkonzept von Baum-basierten Verfahren ist es, den Variablenraum durch sequentielles Aufteilen in disjunkte Unterräume (Knoten) zu zerlegen und in jedem Unterraum ein einfaches Modell (z.B. eine Konstante) anzupassen. CART zerlegt in jedem Schritt der Baumkonstruktion einen Knoten U durch eine binäre Aufteilungsregel in zwei Unterräume U_1 und U_2 . Nach Abschluss der Baumkonstruktion erhält man eine Menge Q disjunkter Endknoten U_1, \dots, U_Q und eine nicht-parametrische Vorhersagefunktion der Form

$$\mu = f(X) = \sum_{q=1}^Q c_q I(X \in U_q), \quad (16)$$

wobei $I(\cdot)$ die Indikatorfunktion bezeichnet und c_1, \dots, c_Q den Vorhersagen der Zielvariable in den Endknoten entsprechen (Hastie et al., 2009). Baum-basierte Verfahren zur Modellierung der Hazardfunktion und der ereignis-spezifischen Hazardfunktionen werden in den Kapiteln 2.1 bis 2.3 betrachtet. In Kapitel 2.2 wird eine Baum-basierte Methode vorgeschlagen, um erklärende Variablen zu identifizieren, deren Effekte auf die Ereigniszeit über den Beobachtungszeitraum der Studie variieren.

1.5 Anwendungsbeispiele

Im Rahmen dieser Arbeit werden fünf verschiedene Anwendungsbeispiele vorgestellt. Dabei handelt es sich um Daten der drei bereits genannten klinischen Studien, Daten einer Studie zur Dauer der Arbeitslosigkeit US-amerikanischer Bürger/innen, und Daten einer sozialwissenschaftlichen Studie zur Erforschung partnerschaftlicher und familialer Lebensformen. Im Folgenden werden die einzelnen Studien in chronologischer Reihenfolge jeweils kurz vorgestellt.

Dauer der Arbeitslosigkeit

In Kapitel 2.1 wurden Daten des Current Population Survey's der Jahre 1986, 1988, 1990 und 1992 analysiert (Croissant, 2016). Modelliert wurde die Dauer der Arbeitslosigkeit von 3.343 US-amerikanischen Bürger/innen, die in Zwei-Wochen-Intervallen erhoben wurde. Entsprechend der Analyse von Cameron und Trivedi (2005) wurde die Zielvariable definiert als die Zeit bis zum Antritt einer Arbeitsstelle jeglicher Art (unter anderem einer Vollzeit- oder Teilzeitstelle). Für die Analyse wurden die Daten über einen Zeitraum von 40 Wochen betrachtet. Dies resultierte in 21 diskreten Ereigniszeitpunkten, wobei die Ereigniszeit $t = 21$ dem Antritt einer Arbeitsstelle zu einem späteren Zeitpunkt nach 40 Wochen entsprach. Erklärende Variablen, die in die Analyse mit einbezogen wurden, sind unter anderem die Ersatzrate des Arbeitslosengeldes und das wöchentliche Gehalt der vorherigen Arbeit.

Akute odontogene Infektionen

In Kapitel 2.2 wurden Daten einer retrospektiven Studie der Klinik und Poliklinik für Mund-, Kiefer- und Plastische Gesichtschirurgie des Universitätsklinikums Bonn untersucht. Gegenstand der Betrachtung waren 303 Patienten/Patientinnen, die im Zeitraum von 2012 bis 2017 mit einer akuten odontogenen Infektion vorstellig wurden (Heim et al., 2019). Hauptziel der Analyse war die Identifizierung von Risikofaktoren, die den notwendigen Krankenhausaufenthalt der Patienten/Patientinnen nach dem operativen Eingriff entscheidend verlängern. Eine genaue Vorhersage der Liegedauer kann zur Transparenz der Kosten und zur Erleichterung des Managements dieser Patienten/Patientinnen beitragen. Zielvariable war die Zeit bis zur Entlassung aus dem Krankenhaus, die zwischen einem Tag und 18 Tagen lag. Potentielle Risikofaktoren, die in die Analyse mit einbezogen wurden, sind unter anderem das Alter und die Erkrankung an Diabetes Typ 2.

Familiäre Entwicklungen

Als zweites Anwendungsbeispiel dienen in Kapitel 2.2 die Daten des Beziehungs- und Familienpanels (pairfam), das die Entwicklung von Partnerschafts- und Generationenbeziehungen untersucht (Huinink et al., 2011). Zu Beginn der Studie im Jahre 2008 wurden 12.000 Teilnehmer/innen der Geburtsjahrgänge 1971–1973, 1981–1983 und 1991–1993 und deren Familien rekrutiert. Diese werden seitdem

jährlich zu Themen der Familienplanung und Einstellungen zur Elternschaft befragt. Basierend auf einer Stichprobe von 861 Frauen wurde in Kapitel 2.2 die Zeit bis zur Geburt des ersten Kindes modelliert. Entscheidende Faktoren, die für die Analyse von Interesse waren, sind unter anderem die Anzahl der Geschwister und die Art der Freizeitgestaltung.

Altersbedingte Makuladegeneration

Das klinische Anwendungsbeispiel in Kapitel 2.3 befasst sich mit Daten der MODIAMD-Studie (Molecular Diagnostics of Age-related Makular Degeneration). Die MODIAMD-Studie ist eine laufende Beobachtungsstudie der Augenklinik des Universitätsklinikums Bonn an Patienten/Patientinnen mit einem hohen Risiko, an altersbedingter Makuladegeneration (AMD) im Spätstadium zu erkranken (Steinberg et al., 2016). AMD gilt als Hauptursache für Erblindung im Alter und äußert sich entweder in trockener Form, der sogenannten geographischen Atrophie (GA), oder in feuchter Form, der sogenannten choroidalen Neovaskularisation (CNV). Ziel der Analyse war die Bestimmung von Risikofaktoren für die Entwicklung beider Ausprägungen der AMD. Dies ermöglicht die Entwicklung von frühzeitigen Maßnahmen für Risikopatienten. Einbezogen wurden die Daten von 98 Patienten/Patientinnen, die zwischen November 2010 und September 2011 in die Studie rekrutiert wurden und seitdem jährlich zu Follow-up-Untersuchungen vorstellig werden. Potentielle Risikofaktoren für die Entwicklung von AMD, die in die Analyse mit einbezogen wurden, sind unter anderem die Sehkraft und das Vorhandensein von refraktilen Drusen.

Nosokomiale Lungenentzündungen

In Kapitel 2.4 und 2.5 werden Daten einer Kohortenstudie von Februar 2000 bis Juli 2001 betrachtet (Beyersmann et al., 2006; Wolkewitz et al., 2008). Untersucht wurden Daten von 1.876 Patienten/Patientinnen aus fünf Universitätskliniken, die mindestens zwei Tage intensivmedizinisch behandelt werden mussten. Hauptziel der Analyse war die Bestimmung von Risikofaktoren für die Erkrankung an nosokomialer Lungenentzündung, die das Risiko der Patienten/Patientinnen, während oder nach dem Krankenhausaufenthalt zu versterben, deutlich erhöht. Einbezogen wurden die Daten über einen Zeitraum von 60 Tagen. Dies resultierte in 61 diskreten Ereigniszeitpunkten, wobei die Ereigniszeit $t = 61$ einer Infektion zu einem späteren Zeitpunkt nach 60 Tagen entsprach. Wichtige Risi-

kofaktoren, die für die Analyse von Interesse waren, sind unter anderem die Art der Operation (geplant oder als Notfall) und die Notwendigkeit einer Intubation.

1.6 Ziele der vorliegenden Arbeit

Die primäre Fragestellung vieler klinischer und epidemiologischer Beobachtungsstudien bezieht sich auf die Analyse von Ereigniszeiten. Dabei steht insbesondere die frühzeitige, individualisierte Vorhersage möglicher Ereignisse für Patienten/Patientinnen mit einem hohen Risiko im Vordergrund. Vorhersagemodelle stellen dafür ein unverzichtbares Instrument für die klinische Entscheidungsfindung dar und leisten einen wichtigen Beitrag für den Einsatz von individuellen Behandlungs- und Therapiestrategien (Moons et al., 2012b; Steyerberg, 2019).

Wie bereits an einigen Anwendungsbeispielen illustriert, werden in der Praxis die Ereigniszeiten, oftmals bedingt durch den Aufbau der Studien, auf einer diskreten Skala gemessen oder liegen nur in gruppierter Form vor. Dies macht die Anwendung klassischer Regressionsmodelle für stetige Ereigniszeiten, wie das Cox-Modell (Cox, 1972), problematisch, und bedarf geeigneter Methoden, die auf die diskrete Datenstruktur zugeschnitten sind. Aufbauend auf den grundlegenden Methoden der Arbeit von Tutz und Schmid (2016) wird in dieser Arbeit ein erweitertes Instrumentarium für die Analyse diskreter Ereigniszeiten entwickelt, das eine breite Anwendbarkeit sowohl in der klinischen Forschung als auch in anderen Bereichen der angewandten Forschung ermöglicht. Die Beiträge der Kapitel 2.1 bis 2.5 lassen sich wie folgt zusammenfassen:

- Semi-parametrische und nicht-parametrische Modellierung der diskreten Hazardfunktion mithilfe von penalisierten Splines und Baum-basierten Verfahren.
- Modellierung der diskreten Hazardfunktion über zeit-variierende Koeffizienten mithilfe von Baum-basierten Verfahren.
- Baum-basierte Modellierung der diskreten, ereignis-spezifischen Hazardfunktionen.
- Entwicklung des Subdistribution Hazard-Modells zur Modellierung der diskreten, kumulativen Inzidenzfunktion.
- Explorative und formale Validierung der Kalibrierung von diskreten Subdistribution Hazard-Modellen.

1.7 Software

Sofern nicht anders angegeben, wurden die Berechnungen in dieser Arbeit mit dem statistischen Softwareprogramm R (R Core Team, 2021) durchgeführt. Folgende Zusatzpakete wurden für die Durchführung der Analysen herangezogen²:

- **discSurv**, Version 1.4.1 (Welchowski und Schmid, 2019) in den Kapiteln 2.1 bis 2.5.
- **mgcv**, Version 1.8-36 (Wood, 2021) in den Kapiteln 2.1, 2.2, 2.4 und 2.5.
- **rpart**, Version 4.1-15 (Therneau und Atkinson, 2019) in Kapitel 2.1.
- **TSVC**, Version 1.2.1 (Berger, 2020) in Kapitel 2.2.
- **VGAM**, Version 1.1-5 (Yee, 2021) in Kapitel 2.3.
- **MRSP**, Version 0.4.3 (Pößnecker, 2014) in Kapitel 2.3.
- **cmprsk**, Version 2.2-10 (Gray, 2020) in Kapitel 2.4.

Neuimplementierungen von Softwareprogrammen und deren Verfügbarkeit sind jeweils in den entsprechenden Ergebnisteilen angezeigt.

²Aktuellste auf CRAN verfügbare Version

2 Ergebnisse

2.1 Berger et al., *Statistical Modelling* 18, 322-345

Das gängigste Regressionsmodell zur Modellierung der diskreten Hazardfunktion (1) ist das logistische Hazard-Modell der Form

$$\lambda(t|X) = \frac{\exp(\eta(t, X))}{1 + \exp(\eta(t, X))}, \quad t = 1, \dots, k - 1, \quad (17)$$

wobei die Vorhersagefunktion $\eta(\cdot)$ von den erklärenden Variablen X und der Zeit t abhängt, siehe Tutz und Schmid (2016) für eine Einführung der grundlegenden Konzepte. In diesem Kapitel wird die Form des logistischen Hazard-Modells mit klassischer parametrischer, linearer Vorhersagefunktion rekapituliert und erläutert, wie die Vorhersagefunktion durch Spezifizierung von glatten, nicht-linearen Funktionen noch flexibler gestaltet werden kann. Es wird gezeigt, wie die resultierenden semi-parametrischen, additiven Modelle mithilfe von P-Splines geschätzt und die zugehörige Likelihood-Funktion hergeleitet werden können. Insbesondere weist die Likelihood-Funktion die Form eines klassischen Regressionsmodells für binäre Zielvariablen $Y \in \{0, 1\}$ auf, was es erlaubt, etablierte Softwareprogramme, die für diese Klasse von Modellen entwickelt wurden, für die Schätzung der Regressionskoeffizienten heranzuziehen.

Anhand der Daten des Current Population Survey's zur Dauer der Arbeitslosigkeit US-amerikanischer Bürger/innen wird Schritt für Schritt illustriert, wie die Datenaufbereitung in R mithilfe des Zusatzpaketes **discSurv** (Welchowski und Schmid, 2019) erfolgen kann und die Schätzung diskreter, semi-parametrischer Hazard-Modelle mithilfe des Zusatzpaketes **mgcv** (Wood, 2021) durchgeführt werden kann. Des Weiteren wird zur Modellierung der Arbeitslosendaten das Baum-basierte Verfahren von Schmid et al. (2016) herangezogen, das als selbst-implementierte R Funktion **survivalTree()** zur Verfügung gestellt wurde.

Zur Evaluierung der Anpassungsgüte parametrischer, semi-parametrischer und Baum-basierter Modelle werden (i) ein Diagramm zur grafischen Beurteilung der Kalibrierung, und (ii) Martingal-Residuen zur Beurteilung des Einflusses einzelner, erklärender Variablen vorgeschlagen. Zuletzt werden die vier vorgestellten Hazard-Modelle bezüglich ihrer Vorhersagegenauigkeit verglichen, wobei das Baum-basierte Modell bei der Modellierung der Arbeitslosendaten am besten abschneidet.

Semiparametric regression for discrete time-to-event data

Moritz Berger¹ and Matthias Schmid¹

¹Department of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn Germany.

Abstract: Time-to-event models are a popular tool to analyse data where the outcome variable is the time to the occurrence of a specific event of interest. Here, we focus on the analysis of time-to-event outcomes that are either intrinsically discrete or grouped versions of continuous event times. In the literature, there exists a variety of regression methods for such data. This tutorial provides an introduction to how these models can be applied using open source statistical software. In particular, we consider semiparametric extensions comprising the use of smooth nonlinear functions and tree-based methods. All methods are illustrated by data on the duration of unemployment of US citizens.

Key words: Discrete time-to-event data, hazard models, semiparametric regression, survival analysis

Received April 2017; revised August 2017; accepted September 2017

1 Introduction

The objective of many statistical analyses is to model a duration time until a specific event occurs. This is usually referred to as *time-to-event* or *survival analysis*. In biostatistics, for example, one often examines the time to death or the progression of a disease. In economics and the social sciences, popular examples include the modelling of the duration of unemployment or the time to retirement. Generally, in regression models for time-to-event data, the event time itself is the response variable, and one wants to investigate the association of the response with several explanatory variables. Most often it is assumed in these analyses that the survival time is given by a random variable measured on a *continuous scale*. This case has been studied extensively in the literature; see, for example, Kalbfleisch and Prentice (2002) and Klein and Moeschberger (2003). However, in practice, measurements of time are often discrete. Durations, for example, are often measured in days, years or months. Moreover, there are situations where the exact event time may not be known, but only an interval during which the event of interest took place.

Here, we consider the application of regression models for *discrete* time-to-event data, which are characterized by an ordinal response variable taking the numbers $1, 2, \dots, k$, with k equal to the number of event times. These numbers either refer to

Address for correspondence: Moritz Berger, Department of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Siegmund-Freud-Straße 25, 53127 Bonn, Germany.
E-mail: Moritz.Berger@imbie.uni-bonn.de

a situation where event times are intrinsically discrete (such as the time to pregnancy, which in clinical applications is usually measured by the number of menstrual cycles) or when continuous event times have been grouped. In the latter case, the numbers $t = 1, 2, \dots, k$ refer to mutually exclusive time intervals $[0, a_1)$, $[a_1, a_2)$, \dots , $[a_{k-1}, \infty)$, with fixed boundaries a_1, \dots, a_{k-1} . Generally, a great advantage of discrete time-to-event models is that they can be viewed as regression models with binary response, giving rise, for example, to the application of logistic regression or probit regression (Willett and Singer, 1993).

A comprehensive treatment of the statistical methodology for discrete time-to-event data has recently been given by Tutz and Schmid (2016). Similar to Gaussian regression, a large part of this methodology has been designed to estimate predictor-response relationships using a *linear* combination of the explanatory variables. In addition, Tutz and Schmid (2016) discuss several (less well known) approaches for *semiparametric* discrete time-to-event modelling. The aim of this tutorial is to provide an in-depth explanation of how these semiparametric models can be fitted and implemented using the R software for statistical computing (R Core Team, 2017). In particular, we will explain how smooth nonlinear functions and tree-based methods can be incorporated into discrete time-to-event models.

A frequently observed phenomenon in time-to-event analysis is censoring. Generally, a duration time is termed ‘censored’, if its total length has not been fully observed. In this article, we consider the most common type of censoring, the so-called *type-I* or *right censoring*, which means that the beginnings of the duration times are observed for all individuals in a study, whereas the respective ends are only observed for part of the individuals. Hence, for some of the individuals, it is only known that the event occurred later than the observed time.

All models discussed in this article will be illustrated by means of a publicly available dataset on the duration of unemployment. The data comprise observations obtained from $n = 3\,343$ US citizens and were collected between 1986 and 1992 as part of the January Current Population Surveys Displaced Workers Supplements (DWS). The original dataset is available as part of the R add-on package `Ecdat` (Croissant, 2016). The response variable that will be considered here is the time to re-employment in any kind of job, which includes full-time, part-time or other kind of jobs. Due to the study design, the observed unemployment durations are discrete, as they were measured in two-week intervals. In this article, we will analyse the data over a period of 40 weeks comprising 21 possible event times $t = 1, 2, \dots, 21$, where $t = 21$ refers to event times > 40 weeks. Explanatory variables that will be included in our analyses are the age of the US citizen in years (`age`), an indicator on whether an unemployment insurance claim was submitted (`ui`), the eligible replacement rate (`reprate`, defined by the weekly benefit amount divided by the amount of weekly earnings in the lost job), the eligible disregard rate (`disrate`, defined as the amount up to which recipients of unemployment insurance who accept part-time work can earn without any reduction in unemployment benefits divided by the weekly earnings in the lost job), the log weekly earnings in the lost job in US\$ (`logwage`) and the tenure in the lost job in years (`tenure`). The summary statistics of the six explanatory

Table 1 Summary statistics of the six explanatory variables used in the modelling of the US unemployment data ($n = 3\,210$)

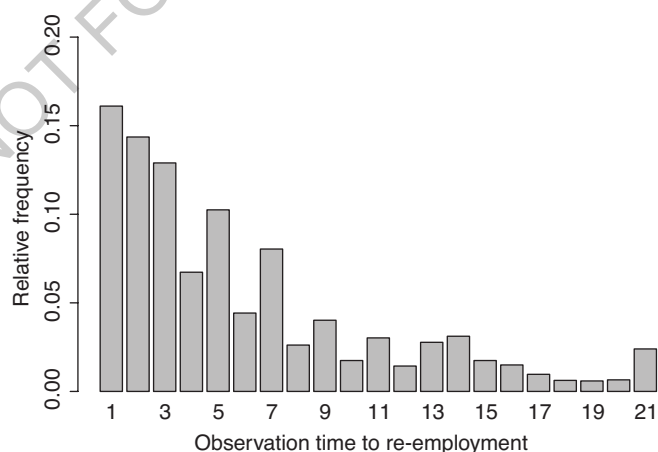
Variable	Summary Statistics					
	X_{min}	$X_{0.25}$	X_{med}	\bar{x}	$X_{0.75}$	X_{max}
age	20	27	34	35.45	43	61
reprate	0.06	0.39	0.50	0.45	0.52	2.05
disrate	0.01	0.05	0.10	0.11	0.15	1.02
logwage	2.70	5.29	5.68	5.69	6.05	7.60
tenure	0	0	2	4.11	5	40
ui	no: 1 437 (44.8%)			yes: 1 773 (55.2%)		

variables are presented in Table 1. Due to missing values in the variables, some observations were excluded from the data, arriving at a sample containing the data of 3 210 citizens.

The observed times to re-employment are visualized in Figure 1. If an individual is still jobless at the end of the survey (i.e., after 40 weeks) or dropped out of the study before finding a job, it is subject to right censoring. In this case, its observation time corresponds to the *censoring time*, otherwise to the true time of re-employment.

The analysis in this tutorial shows how the relationship between the chance of re-employment and the explanatory variables can be estimated in a flexible way, using tailored semiparametric models.

The rest of this tutorial is organized as follows: Section 2 provides the basic theoretical framework and an introduction to parametric as well as semiparametric discrete time-to-event modelling. Details on model fitting and data preparation are given in Section 3. Section 4 presents measures that are useful for assessing the

**Figure 1** Observed time to re-employment (measured in two-week intervals) in the US unemployment data. The median observation time in the data is 4, corresponding to a time period of eight weeks

goodness of fit of discrete time-to-event models. In Section 5, we illustrate the methods by presenting a detailed analysis of the US unemployment data, showing how the various regression models can be applied by use of the R language for statistical computing. Furthermore, we provide guidance on model choice and compare the models in terms of their prediction accuracy. Section 6 discusses additional aspects related to discrete time-to-event modelling and puts the methods considered in this article into perspective.

The R code to reproduce all the numerical results is provided as electronic supplement to this tutorial.

2 Notation and basic concepts

Given n observations, let, in the following, T_i denote the event time and C_i the censoring time of individual i , $i = 1, \dots, n$. T_i and C_i are assumed to be independent random variables taking discrete values in $\{1, \dots, k\}$. In addition, one observes a vector of p time-constant explanatory variables $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$. For right-censored data, the observation time is defined by $\tilde{T}_i = \min(T_i, C_i)$, that is, \tilde{T}_i corresponds to the true event time, if $T_i < C_i$, and to the censoring time otherwise. If originally continuous data have been grouped, the discrete event times $1, \dots, k$ refer to time intervals $[0, a_1), [a_1, a_2), \dots, [a_{k-1}, \infty)$, where $T_i = t$ means that the event occurred in time interval $[a_{t-1}, a_t)$. For example, in our application on unemployment durations, where time was measured in two-week intervals, $T_i = 3$ implies that re-employment of individual i took place between four and six weeks after the start of the study.

The main tool to model discrete time-to-event data is the *hazard function*, which captures the dynamics of the survival process at each time point. For a given vector of time-constant explanatory variables \mathbf{x}_i , the hazard function is defined by

$$\lambda(t|\mathbf{x}_i) = P(T_i = t | T_i \geq t, \mathbf{x}_i), \quad t = 1, \dots, k, \quad (2.1)$$

describing the conditional probability of an event at time t , given that the individual survived until t . The corresponding *survival function* is given by

$$S(t|\mathbf{x}_i) = P(T_i > t|\mathbf{x}_i) = \prod_{s=1}^t (1 - \lambda(s|\mathbf{x}_i)), \quad t = 1, \dots, k, \quad (2.2)$$

denoting the probability that an event occurs later than at time t , or, alternatively, the probability of surviving interval $[a_{t-1}, a_t)$.

An important consequence of the definition of the hazard function in (2.1) is that for a fixed time t , the hazard $\lambda(t|\mathbf{x}_i)$ drives a binary variable that distinguishes between the event taking place at time t or not, conditional on $T_i \geq t$. Therefore, a model for the discrete hazard function can be derived from regression modelling strategies for

a binary response data. This allows to use established tools and reliable software packages that have been developed for this class of models.

2.1 Parametric discrete hazard models

A general class of binary response models applied to the discrete hazard function is defined by

$$\lambda(t|\mathbf{x}_i) = h(\gamma_{0t} + \mathbf{x}_i^\top \boldsymbol{\gamma}), \quad (2.3)$$

where $h(\cdot)$ is a strictly monotone increasing distribution function. A common assumption is that the model contains a time-varying intercept and a set of covariate effects that are fixed over time. Hence, the linear predictor of the model, $\eta_{it} = \gamma_{0t} + \mathbf{x}_i^\top \boldsymbol{\gamma}$, comprises the intercepts γ_{0t} , $t = 1, \dots, k-1$, and a vector of regression coefficients $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)^\top$ independent of t . Note that there is no intercept parameter for $t = k$, as the hazard function in (2.1) is fully determined by $h(\cdot)$ and the coefficients $\gamma_{01}, \dots, \gamma_{0,k-1}$, $\boldsymbol{\gamma}^\top$. The hazard of the last interval is not explicitly modelled, but by definition is given by $\lambda(k|\mathbf{x}_i) = 1$.

The most popular version of model (2.3) is the *logistic discrete hazard model* or *proportional continuation ratio model*, which is specified by the equation

$$\lambda(t|\mathbf{x}_i) = \frac{\exp(\gamma_{0t} + \mathbf{x}_i^\top \boldsymbol{\gamma})}{1 + \exp(\gamma_{0t} + \mathbf{x}_i^\top \boldsymbol{\gamma})}. \quad (2.4)$$

By definition, the proportional continuation ratio model uses the logistic distribution function for $h(\cdot)$. It can be shown that an alternative representation of the model is

$$\log \left(\frac{P(T_i = t|\mathbf{x}_i)}{P(T_i > t|\mathbf{x}_i)} \right) = \gamma_{0t} + \mathbf{x}_i^\top \boldsymbol{\gamma}. \quad (2.5)$$

The ratio $P(T_i = t|\mathbf{x}_i)/P(T_i > t|\mathbf{x}_i)$ compares the probability of an event at time t to the probability of an event later than t . It is also known as *continuation ratio*; see, for example, Agresti (2013). One has to note, that the odds in (2.5) are equivalent to the odds under the condition $T_i \geq t$. This representation of the model allows for an easy interpretation of the effects see the application in Section 5.2.

Generally, the number of parameters in model (2.4) depends on the number of time points, as there is a separate intercept for each t . The set of intercepts $\gamma_{01}, \dots, \gamma_{0,k-1}$ defines the hazard that is always present for any given set of covariates. This hazard is usually referred to as *baseline hazard*, and the intercepts γ_{0t} correspond to the log continuation ratio when all covariates are zero.

2.2 Semiparametric extensions

The parametric model introduced in the previous section is linear in $\boldsymbol{\gamma}$, implying that each covariate has a linear effect on the transformed hazard. In practice, this

linearity assumption may be too restrictive, as predictor-response relationships are often characterized by nonlinear functional forms. Furthermore, so far it has been assumed that the baseline hazard is represented by a separate intercept coefficient for each t . This can lead to numerical problems, especially when the number of time points (and, hence, the number of intercept parameters) is large relative to the sample size, implying that the event counts at some of these time points may become small. In the following, we will consider popular semiparametric alternatives for the definition of η_{it} that address these issues. We first consider *additive hazard models* and subsequently *tree-based methods*, which can also be embedded into the framework of binary response models.

To avoid numerical problems in the estimation of the baseline hazard, it is often convenient to consider an additive model with predictor

$$\eta_{it} = f_0(t) + \mathbf{x}_i^\top \boldsymbol{\gamma}, \quad (2.6)$$

where $f_0(t)$ is a smooth (possibly nonlinear) function of time. By relating the values of the baseline hazard at neighbouring time points via $f_0(t)$, the number of parameters involved in model fitting effectively reduces, and low event counts at some time points become less problematic. A common way to specify the smooth function in t is to use splines, which are represented by a weighted sum of M basis functions. One possible representation of $f_0(t)$ is by B -spline basis functions. These are polynomials of fixed degree d differing from zero in $d + 1$ adjoining intervals. For a comprehensive introduction to B -splines, see De Boor (1978). Very flexible spline functions can be obtained by choosing a relatively large number of basis functions M and at the same time using a penalty term to prevent estimates becoming too rough ('wiggly'). This approach, on which we will focus in this article, is called *P-splines* and was first proposed by Eilers and Marx (1996).

An extension of the semiparametric model (2.6) that weakens the linearity assumption on the effects of the covariates is given by the additive model

$$\eta_{it} = f_0(t) + \sum_{j=1}^p f_j(\mathbf{x}_{ij}), \quad (2.7)$$

where the $f_j(\cdot)$ are unknown smooth functions. That is, the effects of the covariates (or subsets of the covariates) are determined by smooth, possibly nonlinear, functions. A common approach is again to use P -splines and to expand each function separately by a weighted sum of B -spline basis functions depending on the covariates.

Possible further extensions of model (2.7) are, for example, the use of smooth time-varying effects of the form $f_j(\mathbf{x}_{ij}) \cdot t$ or $f_j(\mathbf{x}_{ij}, t)$. This kind of models is extensively discussed in Bender et al. (2018). The tutorial by Bender et al. (2018) illustrates the use of generalized additive mixed models for semiparametric continuous time-to-event modelling and is also part of this special issue.

In the semiparametric models with predictors (2.6) and (2.7), it is assumed that the predictor is given by an additive function of time and a linear (or additive)

function of the covariates. Although these models are very flexible, they may not capture the structure of the data very well, if interactions between covariates are present. For example, it is quite conceivable that the effect of a covariate on the hazard depends on the values of a second covariate, implying the presence of an interaction between the two covariates. The problem when incorporating interactions in parametric or additive models is that the relevant interactions have to be known and specified before model fitting. Furthermore, parametric and additive models are hard to handle if the interaction terms involve more than two covariates. An alternative regression approach that addresses these problems is *recursive partitioning*, which is also known as *tree modelling*. The most popular tree method is *classification and regression trees* (CART), as proposed and described in detail by Breiman et al. (1984). The basic CART method is conceptually very simple: The covariate space is partitioned recursively into a set of rectangles, and in each rectangle a simple model (for example, a covariate-free model) is fitted. A user-friendly introduction to the basic concepts of tree modelling is found in Hastie et al. (2009). Recently, Schmid et al. (2016) proposed a recursive partitioning method that is specifically designed to model discrete time-to-event data. The main principle is to fit a discrete hazard model of the form

$$\lambda(t|\mathbf{x}_i) = f(t, \mathbf{x}_i), \quad (2.8)$$

where $f(t, \mathbf{x}_i)$ is represented by a classification tree with binary outcome. Each split of this tree is determined by either t (treated as an ordinal variable) or one of the covariates. As a result, each terminal node of the tree refers to an estimate of the hazard function for a specific covariate combination and a specific time interval $[t_1, t_2] \subset [1, k]$. For details on the calculation of the corresponding estimates, see Section 5.3.

3 Estimation and data preparation for additive hazard models

To derive the log-likelihood function for discrete hazard models, it is useful to introduce a binary variable indicating whether the target event was observed or not:

$$\Delta_i = \begin{cases} 1, & \text{if } T_i \leq C_i, \\ 0, & \text{if } T_i > C_i. \end{cases} \quad (3.1)$$

Thus, Δ_i becomes 1, if the exact true time is observed; otherwise, $\Delta_i = 0$. In the case where continuous time-to-event data are grouped, $\Delta_i = 1$ and $\tilde{T}_i = t$ implies an event in interval $[a_{t-1}, a_t)$ and $T_i = \tilde{T}_i = t$. Similarly, $\Delta_i = 0$ and $\tilde{T}_i = t$ implies $C_i = \tilde{T}_i = t$ and survival beyond a_t , that is, $T_i > \tilde{T}_i = t$.

Note that when continuous time-to-event data are grouped or rounded, additional assumptions are implicitly imposed on the censoring mechanism. To see this, consider the case where both the continuous event time $T_{\text{cont},i}$ and the continuous censoring

time $C_{\text{cont},i}$ are within the same interval, say, $[a_{t-1}, a_t)$. Then, by definition, $T_i = C_i = t$, $\tilde{T}_i = t$, and $\Delta_i = 1$, leading to the usual interpretation that an event was observed in interval $[a_{t-1}, a_t)$. At the same time, however, this interpretation implicitly assumes $T_{\text{cont},i} \leq C_{\text{cont},i}$, that is, the continuous event time $T_{\text{cont},i} \in [a_{t-1}, a_t)$ is not allowed to be larger than the continuous censoring time $C_{\text{cont},i} \in [a_{t-1}, a_t)$. Without this assumption, the scenario where both $T_{\text{cont},i}, C_{\text{cont},i} \in [a_{t-1}, a_t)$ and $T_{\text{cont},i} > C_{\text{cont},i}$ would result in $\tilde{T}_i = t$ and $\Delta_i = 1$ but *no* observed event in $[a_{t-1}, a_t)$. This assumption on the nature of the censoring mechanism is often referred to as ‘censoring at the end of the interval’.

With data $(\tilde{T}_i, \Delta_i, \mathbf{x}_i)$, $i = 1, \dots, n$, the contribution of the i -th observation to the likelihood function is given by

$$L_i = P(T_i = \tilde{T}_i)^{\Delta_i} P(T_i > \tilde{T}_i)^{1-\Delta_i} P(C_i \geq \tilde{T}_i)^{\Delta_i} P(C_i = \tilde{T}_i)^{1-\Delta_i}. \tag{3.2}$$

A crucial assumption that is usually made to simplify the likelihood function (3.2) is that the censoring process does not depend on the parameters determining the event times T_i . A consequence of this assumption is that the terms involving the censoring times can be ignored in the maximization of the likelihood function for the time-to-event process. Omitting the terms involving C_i in (3.2) and inserting the definitions of the hazard function (2.1) and the survival function (2.2), one obtains (expect for some constants)

$$L_i \propto \lambda(\tilde{T}_i|\mathbf{x}_i)^{\Delta_i} (1 - \lambda(\tilde{T}_i|\mathbf{x}_i))^{1-\Delta_i} \prod_{j=1}^{\tilde{T}_i-1} (1 - \lambda(j|\mathbf{x}_i)). \tag{3.3}$$

Note that, by definition of Δ_i in (3.1), one always obtains $\Delta_i = 1$ and $\lambda(\tilde{T}_i|\mathbf{x}_i) = 1$, if \tilde{T}_i is equal to the last time point k . For maximum likelihood estimation, it is therefore convenient to re-code observations with $\tilde{T}_i = k$ as follows:

$$\tilde{T}_i = k, \Delta_i = 1, \mathbf{x}_i \mapsto \tilde{T}_i = k - 1, \Delta_i = 0, \mathbf{x}_i, \tag{3.4}$$

making use of the fact that the value of the likelihood contribution in (3.3) will not be altered by this transformation.

With some algebra, it can be shown that the likelihood function (3.3) is equal to the likelihood of a binary response model with outcome variables

$$(y_{i1}, \dots, y_{i\tilde{T}_i}) = \begin{cases} (0, \dots, 0, 1), & \text{if } \Delta_i = 1 \\ (0, \dots, 0, 0), & \text{if } \Delta_i = 0. \end{cases} \tag{3.5}$$

For individuals where the exact event time is observed, one defines the observation vector $(0, \dots, 0, 1)$ of length \tilde{T}_i . For censored individuals, the observation vector contains only zeros. According to this definition, one has \tilde{T}_i binary observations for each individual i , resulting in a total of $\tilde{T}_1 + \dots + \tilde{T}_n$ observations.

Using these definitions, the log-likelihood of the proportional continuation ratio model becomes

$$\ell \propto \sum_{i=1}^n \sum_{s=1}^{\tilde{T}_i} y_{is} \log(\lambda(s|\mathbf{x}_i)) + (1 - y_{is}) \log(1 - \lambda(s|\mathbf{x}_i)). \quad (3.6)$$

The main advantage of this representation of the log-likelihood is that it allows to use software for fitting binary response models. For example, it follows from (3.6) that fitting a continuation ratio model is equivalent to fitting a logistic regression model with predictor (2.6) or (2.7). In this model, the values of the binary responses y_{is} can be interpreted as binary decisions for the transition from interval $[a_{s-1}, a_s)$ to $[a_s, a_{s+1})$. For instance, in the application on unemployment duration, one observes $y_{is} = 0$ for each two-week interval as long as the individual i is not re-employed yet.

The models with smooth components (2.6) and (2.7) can be fitted by maximizing a penalized likelihood of the form

$$\ell_p = \ell - \delta J, \quad (3.7)$$

where $\delta \in \mathbb{R}^+$ is a penalty parameter and $J \in \mathbb{R}^+$ is the penalty term already mentioned in Section 2.2, putting restrictions on the weights of the B -spline basis functions and preventing estimates from becoming too rough. When using P -splines, J is a difference penalty on adjacent B -spline coefficients. A common procedure is to use cubic B -splines ($d = 3$) with second order differences. Then, for example, the penalty term for the estimation of the smooth baseline hazard $f_0(t)$ in model (2.6) contains only the parameters $\gamma_{01}, \dots, \gamma_{0M}$, corresponding to $f_0(t)$ and has the form

$$J = \sum_{m=3}^M (\Delta^2 \gamma_{0m})^2 = \sum_{m=3}^M (\gamma_{0m} - 2\gamma_{0,m-1} + \gamma_{0,m-2})^2. \quad (3.8)$$

For further details, see Eilers and Marx (1996).

The degree of smoothness is determined by the tuning parameter δ . The larger the value of δ , the smoother is the resulting function, and vice versa. When several smooth functions are included in the model, one uses a difference penalty for each spline effect, based on the differences of adjacent B -spline coefficients for the corresponding covariate. The smoothness of the individual spline estimates can be determined by the same or separate penalty parameters δ_j .

Before fitting proportional continuation ratio models with software for binary outcome data, one has to generate the required binary observations presented in (3.5). This is done by the generation of an *augmented data matrix*. For the setup of the matrix, one has to distinguish between censored and non-censored individuals. For an individual whose event was observed ($\Delta_i = 1$) at time \tilde{T}_i , the augmented data matrix is given by

$$\begin{pmatrix} 0 & 1 & x_{i1} & \dots & x_{ip} \\ 0 & 2 & x_{i1} & \dots & x_{ip} \\ 0 & 3 & x_{i1} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \tilde{T}_i & x_{i1} & \dots & x_{ip} \end{pmatrix}. \quad (3.9)$$

For an individual that is censored ($\Delta_i = 0$) at time \tilde{T}_i , the augmented data matrix is given by

$$\begin{pmatrix} 0 & 1 & x_{i1} & \dots & x_{ip} \\ 0 & 2 & x_{i1} & \dots & x_{ip} \\ 0 & 3 & x_{i1} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & \tilde{T}_i & x_{i1} & \dots & x_{ip} \end{pmatrix}. \quad (3.10)$$

The first column in the augmented data matrices corresponds to the binary responses $y_{i1}, \dots, y_{i\tilde{T}_i}$. The second column is the time interval running from 1 to \tilde{T}_i . When fitting a model with fixed intercept parameters γ_{0t} , this column has necessarily to be coded as a nominal factor, for example, via dummy variables. The remaining part of the data contains the covariates. When the covariates are constant over time, the values in each row of columns 3 to $(p+2)$ are the same, that is, covariate vector x_i is repeated row-wise. This is also the case in the US unemployment data. Otherwise, when time-varying covariates are considered, the observed time series are entered in the respective columns of the augmented data matrices. For each individual, the augmented data matrix has \tilde{T}_i rows, and the whole data matrix, which is obtained by ‘glueing’ the individual augmented matrices together, has $\sum_{i=1}^n \tilde{T}_i$ rows.

In R, the augmented data matrix can be generated by applying the function `dataLong()` in the R package **discSurv** (Welchowski and Schmid, 2017). The general interface of the function is

```
> dataLong(dataSet, timeColumn, censColumn, timeAsFactor = TRUE).
```

The function requires the original data of class `data.frame` in ‘non-augmented’ short format (argument `dataSet`), the column name of the observed discrete event times (argument `timeColumn`) and the column name of the binary event indicator as defined in Equation (3.1) (argument `censColumn`). The variable required by `timeColumn` can either be numeric or coded as an ordinal or nominal factor. If `timeAsFactor = TRUE`, the time column in the augmented data matrix will be returned as a nominal factor. The variable required by `censColumn` can either be a numerically coded 0/1 vector or a labelled factor variable. Note that `dataLong()`

assumes that the covariates are constant over time. If this is not the case, the function `dataLongTimeDep()` should be used instead to generate the augmented data matrix. For details on the required format of the raw data matrix, we refer to the documentation in `discSurv`.

The augmented data matrix returned by `dataLong()` contains the binary responses as defined in Equation (3.5) in the form of a numerically coded 0/1 vector named y . Further details on the output are given by the application in Section 5.1.

4 Goodness-of-fit measures

In this tutorial, we consider two diagnostic tools that are useful to investigate discrete hazard models in terms of their goodness of fit. By appropriate visualizations, both tools can be used to check whether a model is well *calibrated*, that is, to check how well the fitted probabilities agree with their corresponding observed proportions. Note that these graphical checks do not constitute ‘formal’ calibration tests, in the sense that they neither rely on asymptotics nor on distributional results.

First, one can generate a *calibration plot*. The idea is to compare the estimated hazards $\hat{\lambda}(t|\mathbf{x}_i)$, $i = 1, \dots, n$, $t = 1, \dots, \tilde{T}_i$, of the model to the relative frequencies of observed events ($y_{it} = 1$) in predefined subsets of the augmented set of observations. More specifically, one splits the data into subsets D_k , $k = 1, \dots, K$, defined by the percentiles of the estimated hazards. Common choices for K are $K = 10$ or $K = 20$. Then the relative frequency of observed events (‘empirical hazard’) is calculated in each subset by

$$\sum_{i,t:\hat{\lambda}(t|\mathbf{x}_i) \in D_k} \frac{y_{it}}{|D_k|}, \quad (4.1)$$

where $|D_k|$ corresponds to the number of observations in subset D_k . If the fit of the model is satisfactory, the empirical hazard measure in (4.1) should be close to the average of the estimated hazards in D_k for all k , that is, close to the mean of $\hat{\lambda}(t|\mathbf{x}_i) \in D_k$, $k = 1, \dots, K$. Examples of calibration plots are shown in Figure 3 in the application.

Second, we consider *martingale residuals*, which allow for assessing the importance of single covariates x_j . The idea of the martingale residuals is to compare for each individual the observed number of events with the expected number of events up to \tilde{T}_i . Using the binary response variables $y_{i1}, \dots, y_{i\tilde{T}_i}$ the residuals are defined as

$$r_i = \sum_{t=1}^{\tilde{T}_i} (y_{it} - \hat{\lambda}(t|\mathbf{x}_i)), \quad i = 1, \dots, n. \quad (4.2)$$

For a well-fitting model that includes all relevant predictors, the difference between y_{it} and $\hat{\lambda}(t|\mathbf{x}_i)$ should be ‘random’ and therefore uncorrelated with the covariate

values. To assess the importance of a covariate graphically, one can plot the residuals against the covariate values. Martingale residuals can be computed by the function `martingaleResid()` contained in the `discSurv` package. Examples are shown in Figure 3 in the application.

5 Application: Duration of unemployment

In the following, discrete hazard regression modelling is illustrated by means of a step-by-step analysis of the US unemployment data. Throughout this section, we use the logistic link function, that is, we consider the fitting of a proportional continuation ratio model.

5.1 Preprocessing of the data

To fit a logistic discrete hazard model of the form (2.4), the original data matrix first has to be transformed to an augmented data matrix, as described earlier. The dataset `UnempDur`, which (after application of the preprocessing steps outlined in the Introduction) is a slightly modified version of the data frame available in the R package `Ecdat`, has the following form:

```
> head(UnempDur)
  spell age  ui reprice disrate logwage tenure status
1     5  41 no   0.179   0.045 6.89568      3      1
2    13  30 yes  0.520   0.130 5.28827      6      1
4     3  26 yes  0.448   0.112 5.97889      3      1
5     9  22 yes  0.320   0.080 6.31536      0      1
6    11  43 yes  0.187   0.047 6.85435      9      0
8     3  32 no   0.373   0.093 6.16121      0      1
```

The first column named `spell` is the observed time to re-employment of individual i and contains the values of \tilde{T}_i , $i = 1, \dots, n$. As mentioned earlier, these values correspond to the lengths of the spells (measured in two week intervals), whose distribution is displayed in Figure 1. The last column named `status` indicates whether the exact event time of individual i has been observed (`status = 1`) or if the individual is subject to right censoring (`status = 0`); it corresponds to the random variable Δ_i defined in equation (3.1). Summarizing the `status` column yields a censoring rate of 0.391:

```
> table(UnempDur$status) / nrow(UnempDur)
      0      1
0.3909657 0.6090343
```

Columns two to seven of the data frame `UnempDur` contain the explanatory variables described in Table 1. All covariates are constant over the time of the survey.

When using `dataLong()` to obtain the augmented data matrix, one has to pass the column names `spell` and `status` to the arguments `timeColumn` and `censColumn`, respectively:

```
> library(discSurv)
> UnempDurLong <- dataLong(UnempDur, timeColumn = "spell",
+                           censColumn = "status").
```

The augmented data matrix `UnempDurLong` has 10 columns with the following names:

```
> names(UnempDurLong)

 [1] "obj"      "timeInt" "y"        "spell"    "age"      "ui"
 [7] "reprate" "disrate" "logwage" "tenure"   "status"
```

The new columns are `obj`, which is an identifier of the individuals, `timeInt`, which contains the discrete time values (i.e., the second column of the augmented matrices in (3.9) and (3.10), stored as a nominal factor) and `y`, which contains the binary response variables $y_{i1}, \dots, y_{i\tilde{T}_i} \in \{0, 1\}$. The head of the augmented data matrix is given by

```
> UnempDurLong[UnempDurLong$obj==1, ]

   obj timeInt y spell age ui reprate disrate logwage tenure status
1     1         1 0     5 41 no  0.179  0.045 6.89568      3      1
1.1   1         2 0     5 41 no  0.179  0.045 6.89568      3      1
1.2   1         3 0     5 41 no  0.179  0.045 6.89568      3      1
1.3   1         4 0     5 41 no  0.179  0.045 6.89568      3      1
1.4   1         5 1     5 41 no  0.179  0.045 6.89568      3      1,
```

showing that the first individual (`obj = 1`) had an event after 10 weeks (`spell = 5` and `status = 1`). Accordingly, the augmented data matrix for the first individual has five rows, where each row corresponds to one time interval (`timeInt = 1, \dots, 5`). The corresponding vector of responses is $y = (0, 0, 0, 0, 1)$. The values of the covariates remain constant over time and are therefore the same in each row.

As a second example, consider the augmented data matrix of the 12th individual (`obj = 12`). This individual is censored after six weeks (`spell = 3` and `status = 0`). Hence, the corresponding data matrix has three rows with response $y = (0, 0, 0)$:

```
> UnempDurLong[UnempDurLong$obj==12, ]

   obj timeInt y spell age ui reprate disrate logwage tenure status
14    12         1 0     3 40 yes  0.52   0.13 4.95583      0      0
14.1  12         2 0     3 40 yes  0.52   0.13 4.95583      0      0
14.2  12         3 0     3 40 yes  0.52   0.13 4.95583      0      0.
```


5.2 Regression modelling

A parametric proportional continuation ratio model with a linear predictor is estimated in R by passing the augmented data matrix `UnempDurLong` to `glm()` with the usual specifications:

```
> model1 <- glm(formula = y ~ timeInt - 1 +
+               age + rebrate + disrate + logwage + tenure + ui,
+               data = UnempDurLong, family = binomial(link = "logit")).
```

The left-hand side of the `formula` argument contains the binary response vector `y`. In addition to the names of the six covariates, the right-hand side contains the discrete time variable as a nominal factor without intercept (`timeInt - 1`). The family argument `binomial(link = "logit")` is the same as for ‘usual’ logistic regression models with binary outcome.

The more complex model with smooth baseline hazard (Equation (2.6)) is estimated by use of the R package `mgcv` (Wood, 2017). A detailed introduction to the estimation procedures is found in Wood (2011). The corresponding function `gam()` essentially has the same interface as `glm()`:

```
> library("mgcv")
> UnempDurLong$timeIntNum <- as.numeric(UnempDurLong$timeInt)
> model2 <- gam(formula = y ~ s(timeIntNum, bs = "ps", k = 5, m = 2) +
+               age + rebrate + disrate + logwage + tenure + ui,
+               data = UnempDurLong, family = binomial(link = "logit")).
```

Before passing the nominal factor `timeInt` to `gam()`, it has to be transformed to a continuous variable (`timeIntNum`). Then a smooth baseline hazard is specified by use of the function `s()` on the right-hand side of the model formula. Required arguments are the type of spline smoother `bs`, the dimension of the basis `k` (which corresponds to the number of basis functions $M = k - 1$) and the order of the penalty `m`. Here, we use *P*-splines with cubic basis functions (`bs = "ps"`) and a second order difference penalty (`m = 2`). The chosen dimension (`k = 5`) results in four cubic basis functions. Based on these specifications, the optimal smoothing parameter δ , see Equation (3.7), is computed by generalized cross-validation (see Wood, 2006). Its estimate is stored in the argument `sp`. In case of the US unemployment data, `sp` is estimated as:

```
> model2$sp
s(timeIntNum)
  0.08562284.
```

Generally, the `mgcv` package implements a large variety of alternative spline estimators and methods for smoothing parameter optimization. In principle, all of these methods may be used for discrete hazard modelling, in the same way as they

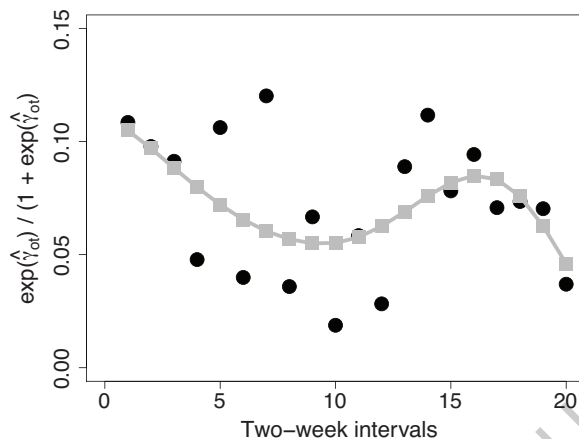


Figure 2 Analysis of the US unemployment data. The figure shows the estimated discrete baseline hazard of `model1` (black dots) and the smooth baseline hazard of `model2` (grey squares)

would be used in logistic regression (or, more generally, in additive models with a binary response).

The estimated baseline hazards of `model1` and `model2` are shown in Figure 2. The discrete baseline hazard obtained for `model1` is visualized by black dots. From these estimates, a reasonable interpretation is hard to derive. On the other hand, the smooth baseline hazard obtained for `model2`, visualized by grey squares, is more meaningful. It is seen that the conditional probability of re-employment decreases until week 20 and subsequently increases up to week 32 before it diminishes again. The reason for this might be that in many US states, workers are eligible for up to 26 weeks of benefits from the state-funded unemployment compensation programme.

Table 2 shows the estimates of the coefficients γ , the corresponding estimated standard errors, and the p -values of the covariate effects obtained for `model1` and `model2`. Apart from the eligible replacement rate (`reprate`) and the tenure in

Table 2 Analysis of the US unemployment data. The table contains coefficient estimates (`coef`), estimated standard errors (`se`) and p -values of the covariate effects obtained for `model1` and `model2` (bh = baseline hazard)

	model1 (discrete bh)			model2 (smooth bh)		
	coef	se	p -value	coef	se	p -value
age	-0.012	0.003	0.000	-0.012	0.003	0.000
reprate	0.285	0.342	0.406	0.301	0.338	0.373
disrate	-0.764	0.383	0.046	-0.755	0.379	0.047
logwage	0.231	0.072	0.001	0.236	0.071	0.001
tenure	-0.005	0.005	0.280	-0.006	0.005	0.266
ui	-1.151	0.052	0.000	-1.175	0.051	0.000

the lost job, (tenure), the covariates are significantly associated with the time to re-employment. According to the signs of the estimates of both models, the chance of getting re-employed decreases with increasing values of age, increasing disregard rate and with the filing of an unemployment claim. On the other hand, the higher the earnings in the lost job, the better the chance of re-employment. Table 2 also shows that the differences in coefficient estimates between the two models are small. By use of Equation (2.5), the effects can be interpreted in an easy way. For example, let us compare citizens who submitted an unemployment insurance claim ($ui = 1$) to those who did not ($ui = 0$). Based on the estimate of `model1` ($\hat{\gamma}_{ui} = -1.151$), one obtains that the probability of re-employment at (any) time t , compared to the probability of re-employment later than t , decreases for citizens who submitted an unemployment insurance claim by the factor $\exp(\hat{\gamma}_{ui}) = 0.316$. The chance of re-employment is therefore much smaller in this group. One might speculate that due to benefits from the state-funded unemployment, the motivation to search for a new job is lower.

The goodness-of-fit measures for `model1` and `model2` are presented in Figure 3. The left figures show the calibration plots (average fitted hazards against the relative frequencies of events). It is seen that the values of `model1` are closer to the 45 degree line than those of `model2`, indicating a better model fit. The right figures show the martingale residuals, defined in (4.2), against the values of the covariate age for the fit of `model1` and `model2` without age. The black lines correspond to the estimated trend obtained by a local polynomial regression using the R function `loess()`. The functional form of the trend lines (compared to the zero line) shows a nonlinear effect on the martingale residuals for both models. This indicates that the covariate age is an influential variable (cf. Table 2) with a possibly nonlinear effect on the response.

Therefore, as a possible extension, we consider a model where the baseline hazard as well as the covariate age are both modelled as smooth P -spline functions:

```
> model3 <- gam(formula = y ~ s(timeIntNum, bs = "ps", k = 5, m = 2) +
+               s(age, bs = "ps", k = 25, m = 2) +
+               rebrate + disrate + logwage + tenure + ui,
+               data = UnempDurLong, family = binomial(link = "logit")).
```

As seen from the R code, the estimation of the smooth function of covariate age in `model3` is based on 24 cubic basis functions (dimension $k = 25$). The estimated penalty parameter $\hat{\delta}$ for age, stored in `sp`, is:

```
> model3$sp["s(age)"]
s(age)
56 860.2.
```

From the resulting function shown in Figure 4, it is seen that the association between the time to re-employment and age is definitely not linear. Its form is very similar to

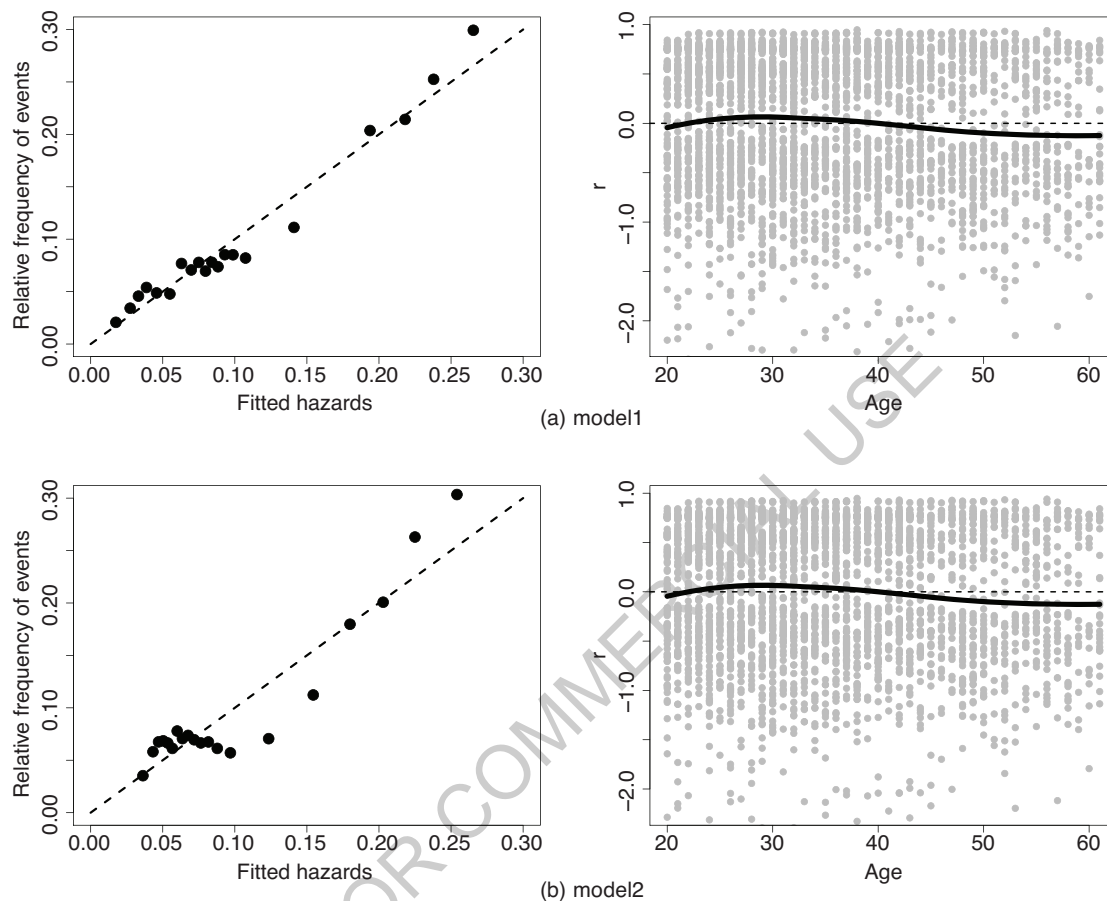


Figure 3 Analysis of the US unemployment data. The two panels show the calibration plot (left) and the martingale residuals against the values of age (right) obtained for `model1` (a) and `model2` (b) without age, respectively. The trend lines were obtained by local polynomial regression

the loess trend shown in Figure 3. The value on the y-axis of the figure corresponds to the contribution of age to the predictor η_{it} of the model. The chance of re-employment has a peak between 20 years and 30 years, and subsequently decreases.

5.3 Tree-based modelling

Finally, we fit a recursive partitioning model of the form (2.8). Again, we consider a procedure that is based on the augmented data matrix with binary outcomes $y_{i1}, \dots, y_{i\bar{T}_i}$.

When growing trees, one has to take two main decisions: First, one has to choose an appropriate criterion for performing the splits. Criteria that have already been used in

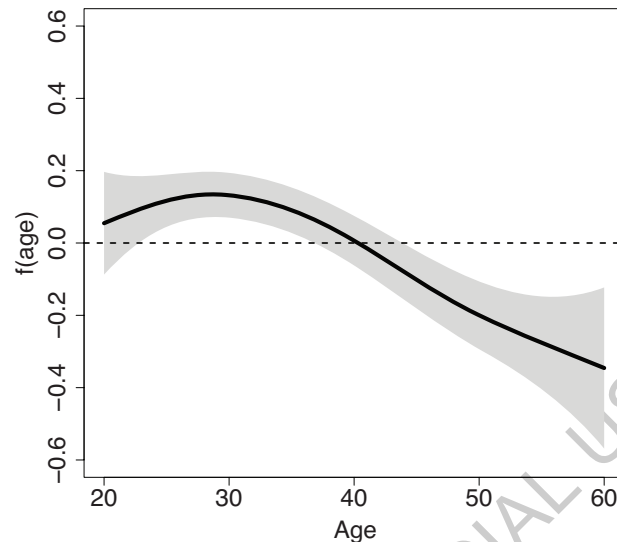


Figure 4 Analysis of the US unemployment data. The plot shows the estimated P -spline function for the covariate age in `model13`

the early days of tree construction are impurity measures. For discrete survival trees, a natural measure of node impurity is the *Brier score*, which evaluates the average squared difference between the binary outcome values y_{it} and the respective hazard estimate $\hat{\lambda}(t|\mathbf{x}_i)$ in each node (see Schmid et al., 2016). It can be shown that using the Brier score is equivalent to the traditional Gini impurity measure. For a single node m , the Gini impurity is given by

$$G_m = 2\pi_m(1 - \pi_m), \quad (5.1)$$

where π_m is the proportion of ones in node m (see Breiman, 1996). This equivalence implies that the traditional CART algorithm based on the Gini criterion can be used for the construction of the tree. The latter is done by using the function `rpart()` of the eponymous R package `rpart` (Therneau et al., 2015).

Second, one has to determine the optimal size of the tree. For the discrete survival tree, an appropriate tuning parameter controlling tree size is the minimal number of observations in each terminal node ('minimal node size'). Optimizing this number avoids overfitting, as the number of terminal nodes is prevented from becoming too large and, at the same time, the node sizes are prevented from becoming too small. Growing the largest possible tree with exactly one observation in each terminal node, for example, is not desirable, as the estimated hazards would all be either 0 or 1 in this case. Accordingly, splitting is stopped when further splitting in any of the current nodes would result in an additional node containing less observations than the minimal node size. Given a sequence of tree estimates depending on the minimal node size, the optimal tree (i.e., the tree with 'optimal' minimal node size) is determined by either minimization of an information criterion (such as AIC or BIC, see in

the following) or maximization of the predictive log-likelihood. The latter strategy means to repeatedly draw subsamples from the original non-augmented data (for example, by cross-validation, bootstrapping or subsampling without replacement) and to calculate the log-likelihood for the omitted observations. One determines the optimal tree as the one for which the predictive log-likelihood (averaged across the subsamples) becomes maximal. The R function `survivalTree()` automatically generates the augmented data matrix by `dataLong()`, estimates the discrete survival tree by `rpart()` and returns the optimal one according to the specified performance criterion. The function is part of the electronic supplement of this article.

Once the optimal minimal node size has been determined, the estimate of $\lambda(t|\mathbf{x}_i)$ is given by the relative frequency of events (proportion of ones) in each node, possibly after applying some sort of correction procedure like the Laplace correction (see in the following).

To fit a tree model to the US unemployment data, we call the `survivalTree` function using the following arguments:

```
> source("survivalTree.R")
> model4 <- survivalTree(formula = y ~ timeInt + age +
+                         replate + disrate + logwage + tenure + ui,
+                         data = UnempDur, tuning = "BIC",
+                         timeColumn = "spell", censColumn = "status",
+                         minimal_ns = seq(100, 1500, by = 10),
+                         trace = TRUE).
```

The formula required for the tree model is analogous to the one specified for a model with linear predictor. Note that internally the time variable `timeInt` is coded as a numeric vector. This is in analogy to `model2` and `model3` with smooth baseline hazards. The original data frame `UnempDur` (in non-augmented format) is passed to the `data` argument. In addition, one has to specify the `timeColumn` and `CensColumn` arguments used in `dataLong()`. The performance criterion is specified by the argument `tuning`. For tuning, we use the Bayesian information criterion (BIC) defined by

$$\text{BIC} := -2\ell + \log(\tilde{n}) n_s,$$

where ℓ is the log-likelihood (3.6), \tilde{n} is the number of rows of the augmented data matrix and n_s denotes the number of splits as a measure of the complexity of the tree. Other possible arguments for tuning are "AIC" (Akaike's information criterion) and "ll" (predictive log-likelihood method). When using "ll", `survivalTree()` performs a five-fold cross-validation based on subsamples without replacement. To ensure that each subsample contains a sufficient number of observations per observed event time, the subsamples are stratified by `spell`. The `survivalTree` function searches for the best model among the sequence of models with minimal node sizes `minimal_ns`. If `minimal_ns` is not specified, the sequence of minimal node sizes is set to $1, \dots, \lfloor \tilde{n}/2 \rfloor$.

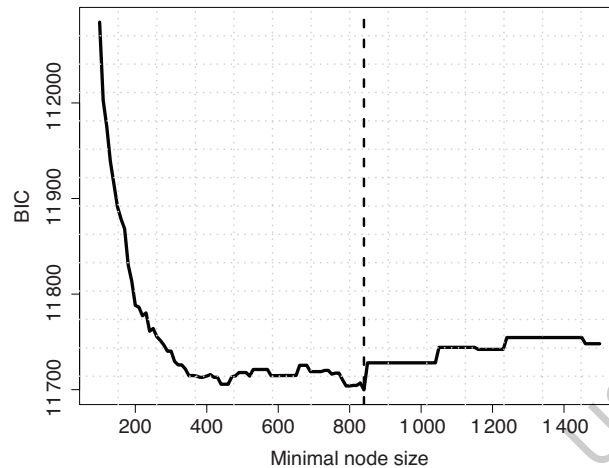


Figure 5 Analysis of the US unemployment data. The plot shows the BIC values for the sequence of survival trees that was obtained by fitting `model4` with minimal node sizes ranging from 100 to 1500. The minimal BIC value (obtained for node size 840) is marked by the vertical dashed line

The BICs obtained for `model4` with minimal node sizes 100, ..., 1500 (in steps of 10) are shown in Figure 5. If an increase of the minimal node size does not change the number of splits and therefore does not influence the resulting tree, the BIC remains the same. This is the case, for example, between minimal node sizes 900 and 1000. According to the BIC, the optimal tree model has minimal node size 840, marked by the dashed line in Figure 5. This results in a tree with 11 splits or 12 terminal nodes. The estimated tree is shown in Figure 6.

The most important covariate, which was chosen in the first split of the tree, is `ui`. As already derived from the parametric models, the submission of an unemployment insurance claim (`ui = 'yes'`) has a negative effect on the 'chance' of re-employment. Within the group of citizens who submitted an insurance claim, the chance is lowest for citizens aged 43 years, or older and with a tenure in the lost job of at least 6 years (leftmost node in Figure 6). For citizens younger than 43 years, all further splits are performed with regard to the discrete time variable (`timeInt`). This confirms the results from Figure 2 in that the chance of re-employment is highly time-dependent. With the high hazard estimate of 0.110 after 26 weeks (`timeInt >= 13`) of unemployment in this group, the tree estimate also indicates similar tendencies that were already seen in Figure 2 after 20 weeks. The best opportunities of re-employment are observed for citizens without an unemployment insurance claim, within the first six weeks (`timeInt < 4`) of unemployment, and with log weekly earnings of at least 5.52\$ (rightmost node in Figure 6). For this subgroup, the estimated hazard rate is 0.302. The two covariates `reprate` and `disrate` were not selected in any of the splits and are therefore excluded from the model. This is in contrast to `model1` and `model2` (see Table 2), where `disrate` showed a significant effect on the hazard.

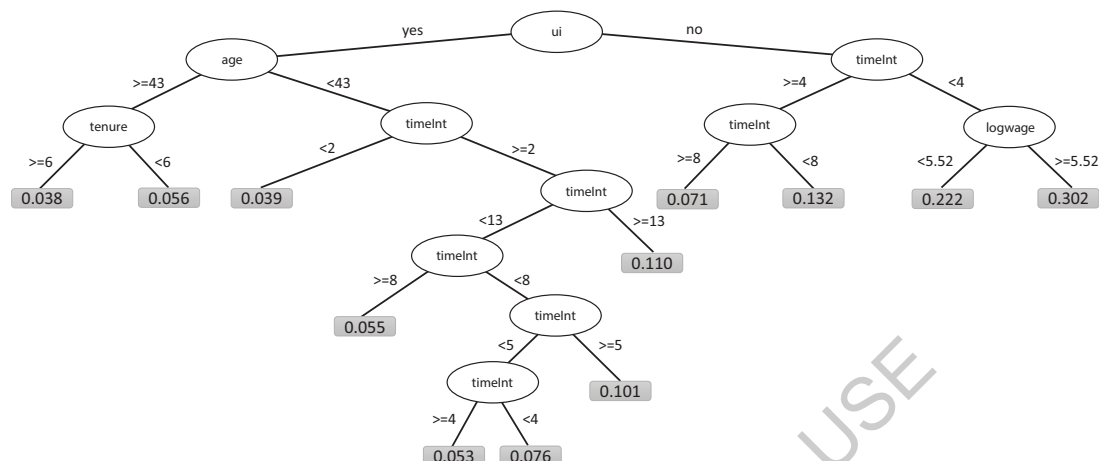


Figure 6 Analysis of the US unemployment data. The graph visualizes the survival tree obtained from fitting `model4` with BIC-optimal minimal node size 840. The numbers at the terminal nodes refer to the estimated hazards. All estimated hazards were additionally post-processed by application of the *Laplace correction*, which was suggested by Ferri et al. (2003) to correct for estimates near the boundaries 0 and 1 in nodes with very few observations. The Laplace correction is automatically performed by `survivalTree()`

5.4 Comparison and model choice

In the previous sections, we presented the results of the analysis of the US unemployment data, having used three different parametric models and a tree-based model. In addition to the goodness-of-fit measures defined in Section 4, the interpretability of the model coefficients and the performance with regard to predicting events of future observations can be used as criteria for model choice.

Regarding the interpretation of the covariate effects, there is an important difference between the parametric models and the tree-based model. After the first split in the tree, all nodes represent interactions between either the covariates or between the covariates and time. For example, the second split in the left node (`ui = 'yes'`) in the covariate `age` implies an interaction between the covariates `ui` and `age` (see Figure 6). In contrast to the interaction effects, it is usually difficult to detect and quantify main effects using tree modelling (e.g., by `age`). On the other hand, in the parametric models, the main covariate effects can easily be interpreted. For example, according to our analysis, covariate `age` has a negative linear effect (see Table 2) or smooth effect (see Figure 4) on the chance of re-employment. Higher order interactions (e.g., between `ui`, `age` and `tenure`) are usually hard to model by the parametric models, as they—unlike trees—do not include a data-driven selection of interaction terms during model fitting. This implies that interactions either need to be known or that a large number of parametric models (including/excluding the various interaction terms) need to be fitted and compared.

A common measure for the prediction accuracy of the models is the deviance $D = -2\ell$, evaluated on $(y_{is}, \hat{\lambda}(s|x_j))$, $i = 1, \dots, n'$, of an independent test dataset

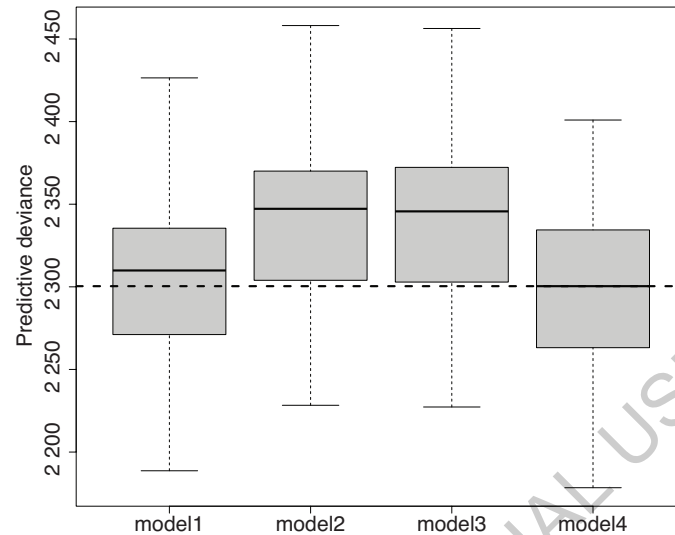


Figure 7 Analysis of the US unemployment data. The boxplots show the predictive deviance values of the four models based on 100 subsamples without replacement of size $n = 2568$ each. The models were evaluated on the remaining 100 test sets of $n' = 642$ observations each. For a better comparison, the median of the tree-based model is marked by a dashed line

comprising n' observations. Because D is small if the log-likelihood ℓ is large, one should prefer the model with minimal D . To further compare the models, we drew 100 subsamples without replacement of size $n = 2568$ (i.e., 80% of the original sample), estimated the four models on each of the 100 subsamples and computed the predictive deviances from the remaining 100 test sets of $n' = 642$ observations each. Subsampling was stratified by `spell` to ensure a sufficient number of observations per observed event. From the results in Figure 7, it is seen that `model1` with discrete baseline hazard performed better than `model2` and `model3` with smooth baseline hazard. This was already indicated by the calibration plots in Figure 3. The tree-based model (rightmost boxplot) had a smaller predictive deviance on average than the parametric models and may hence be considered the best-performing model for the US unemployment data.

6 Concluding remarks

In this tutorial, we have described a basic set of tools to fit semiparametric regression models with a discrete time-to-event outcome. All presented models are very general, in that they are applicable to any type of censored discrete response, regardless of whether the data-generating process is defined by an intrinsically discrete process or by the rounding/grouping of continuous event times. Furthermore, the presented methods are applicable in basically any field of research, as for example, in the social sciences, biostatistics, epidemiology and many more. The US unemployment data

considered in this article is therefore only one of many possible examples. Further applications are presented in Tutz and Schmid (2016).

It is important to realize that all models considered in this tutorial can be fitted easily by use of standard software for binary regression modelling. This has the great advantage that established tools for estimation and inference can be used, that are already available. The most important functions in R are `glm()`, `gam()` (of the **mgcv** package) and `rpart()` (of the eponymous package). In addition to the *P*-spline and CART methodologies considered here, many other options for semiparametric discrete time-to-event modelling exist in R. For example, **mgcv** provides a variety of alternative spline modelling tools such as cardinal splines and smoothing splines, which can be used for discrete hazard modelling by specifying the `bs` argument in `gam()` accordingly. Further extensions also include time-varying coefficient models as considered in Bender et al. (2018) for continuous data. Similarly, there is an alternative tree modelling approach developed by Bou-Hamad et al. (2009) that operates directly on the non-augmented time-to-event data. This procedure is implemented in the R package **DStree** (Mayer et al., 2014).

The basic functionalities required for applying the aforementioned software packages are all implemented in the **discSurv** package. Next to the functions used in this tutorial, **discSurv** provides additional functions to calculate, for example, measures for model evaluation like the concordance index (Schmid et al., 2017), and alternative tools for residual analysis.

We finally note that there exists a number of additional modelling options that are beyond the scope of this tutorial. These include, among many others, (a) regularized estimation via boosting or penalized optimization of the log-likelihood, which is useful for variable selection in higher-dimensional settings, (b) random effects and finite mixture modelling, which account for unobserved heterogeneity in the data and (c) competing-risks models, which extend the models considered in this tutorial by allowing for more than one target event. For details on further methodology, including semiparametric extensions, see Tutz and Schmid (2016). In particular, a basic introduction into boosting for regression modelling is given by Mayr and Hofner (2018) as part of this special issue.

Acknowledgements

The work of MS was supported by the German Research Foundation (DFG), grant SCHM 2966/1-2.

References

- Agresti A (2013) *Categorical Data Analysis, 3rd edition*. New York, NY: John Wiley & Sons.
- Bender A, Groll A and Scheipl F (2018) A tutorial on the estimation of time-varying coefficient models in event data analysis. *Statistical Modelling*, 18, 299–321.
- Bou-Hamad I, Larocque D, Ben-Ameur H, Mâsse LC, Vitaro F and Tremblay RE (2009)

- Discrete-time survival trees. *Canadian Journal of Statistics*, **37**, 17–32.
- Breiman L (1996) Technical note: Some properties of splitting criteria. *Machine Learning*, **24**, 41–7.
- Breiman L, Friedman JH, Olshen RA and Stone JC (1984) *Classification and Regression Trees*. Monterey, CA: Wadsworth.
- Croissant Y (2016) *Ecdat: Data sets for econometrics*. R package version 0.3-1. URL <http://CRAN.R-project.org/package=Ecdat>
- De Boor C (1978) *A Practical Guide to Splines*. New York, NY: Springer.
- Eilers PH and Marx BD (1996) Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–102.
- Ferri C, Flach PA and Hernández-Orallo J (2003) Improving the AUC of probabilistic estimation trees. In Nada Lavrač, Dragan Gamberger, Hendrik Blockeel and Ljupčo Todorovski, eds. *European Conference on Machine Learning*, pages 121–32. Berlin Heidelberg: Springer.
- Hastie T, Tibshirani R and Friedman JH (2009) *The Elements of Statistical Learning, 2nd edition*. New York, NY: Springer.
- Kalbfleisch J and Prentice R (2002) *The Survival Analysis of Failure Time Data, 2nd edition*. New Jersey: Wiley Inter-Science.
- Klein J and Moeschberger M (2003) *Survival Analysis: Statistical Methods for Censored and Truncated Data*. New York, NY: Springer.
- Mayer P, Larocque D and Schmid M (2014) *DStree: Recursive partitioning for discrete-time survival trees*. R package version 1.0. URL <https://CRAN.R-project.org/package=DStree>
- Mayr A and Hofner B (2018) Boosting for statistical modelling: A non-technical introduction. *Statistical Modelling*, **18**, 365–84.
- R Core Team (2017) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. URL <https://www.R-project.org/>
- Schmid M, Küchenhoff H, Hoerauf A and Tutz G (2016) A survival tree method for the analysis of discrete event times in clinical and epidemiological studies. *Statistics in Medicine*, **35**, 734–51.
- Schmid M, Tutz G and Welchowski T (2017) Discrimination measures for discrete time-to-event predictions. *Econometrics and Statistics*.
- Therneau T, Atkinson B and Ripley B (2015) *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-10. URL <https://CRAN.R-project.org/package=rpart>
- Tutz G and Schmid M (2016) *Modeling Discrete Time-to-event Data*. New York, NY: Springer.
- Welchowski T and Schmid M (2017) *discSurv: Discrete Time Survival Analysis*. R package version 1.1.7 URL <http://CRAN.R-project.org/package=discSurv>
- Willett JB and Singer JD (1993) Investigating onset, cessation, relapse, and recovery: Why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *Journal of Consulting and Clinical Psychology*, **61**, 952–65.
- Wood S (2006) *Generalized Additive Models: An Introduction with R*. Florida: CRC press.
- (2017) *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*. R package version 1.8-15. URL <https://CRAN.R-project.org/package=mgcv>
- (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 3–36.

2.2 Puth et al., Lifetime Data Analysis 26, 545-572

In parametrischen und semi-parametrischen diskreten Hazard-Modellen, wie in Kapitel 2.1 beschrieben, setzt sich die Vorhersagefunktion aus der Baseline-Hazardfunktion über die Zeit t und einer Funktion der erklärenden Variablen X zusammen. Dabei wird üblicherweise angenommen, dass die Effekte der erklärenden Variablen über die Zeit konstant sind. Diese Annahme kann jedoch oftmals zu einschränkend sein, beispielsweise wenn Risikofaktoren zu Beginn einer Studie einen stärkeren Effekt aufweisen als im späteren Verlauf der Studie. Eine flexiblere Methode, die erlaubt, dass Effekte über die Zeit variieren, ist ein diskretes Hazard-Modell mit additiver Vorhersagefunktion der Form


$$\eta(t, X) = \gamma_{0t} + X_1\gamma_1(t) + \dots + X_p\gamma_p(t), \quad t = 1, \dots, k-1, \quad (18)$$

wobei die Funktionen $\gamma_j(t)$, $j = 1, \dots, p$, den Koeffizienten der erklärenden Variablen über die Zeit entsprechen. Zur Modellierung dieser Funktionen können, wie für die semi-parametrischen Modelle in Kapitel 2.1, P-Splines verwendet werden. Dies impliziert, dass sich die Effekte über die Zeit gleichmäßig verändern. Im Fall diskreter Ereigniszeiten, ist es jedoch plausibler, dass sich die Effekte nur zu bestimmten Zeitpunkten verändern und in den Zeitintervallen dazwischen konstant bleiben. Damit ergeben sich stückweise konstante Effekte über die Zeit. Zur Modellierung stückweiser konstanter Funktionen $\gamma_j(t)$ wird in diesem Kapitel ein Baum-basiertes Verfahren eingeführt. Der vorgeschlagene Algorithmus ermöglicht eine automatische Selektion der erklärenden Variablen, deren Effekte über die Zeit variieren, deren Effekte über die Zeit konstant sind und, welche gar keinen Effekt auf das interessierende Ereignis haben. Die Schätzung kann in R mithilfe des Zusatzpaketes **TSVC** (Berger, 2020) durchgeführt werden. Eine Simulationsstudie veranschaulicht den Mehrwert des Baum-basierten Modells gegenüber dem parametrischen Modell ohne zeit-variierende Effekte und dem semi-parametrischen Modell mit gleichmäßigen Effekten über die Zeit.

Das Baum-basierte Verfahren wird angewendet, um ein Vorhersagemodell für die Liegedauer von Patienten/Patientinnen mit einer akuten odontogenen Infektion zu erstellen und, um die Daten der pairfam-Studie zur Geburt des ersten Kindes zu analysieren. Die erstere Analyse deckt insbesondere auf, dass die Erkrankung an Diabetes Typ 2 einen relevanten Risikofaktor darstellt, der die Wahrscheinlichkeit einer Entlassung innerhalb der ersten vier Tage signifikant verringert.



Tree-based modeling of time-varying coefficients in discrete time-to-event models

Marie-Therese Puth^{1,2}  · Gerhard Tutz³ · Nils Heim⁴ · Eva Münster² · Matthias Schmid¹ · Moritz Berger¹

Received: 17 January 2019 / Accepted: 30 October 2019 / Published online: 11 November 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

Hazard models are popular tools for the modeling of discrete time-to-event data. In particular two approaches for modeling time dependent effects are in common use. The more traditional one assumes a linear predictor with effects of explanatory variables being constant over time. The more flexible approach uses the class of semiparametric models that allow the effects of the explanatory variables to vary smoothly over time. The approach considered here is in between these modeling strategies. It assumes that the effects of the explanatory variables are piecewise constant. It allows, in particular, to evaluate at which time points the effect strength changes and is able to approximate quite complex variations of the change of effects in a simple way. A tree-based method is proposed for modeling the piecewise constant time-varying coefficients, which is embedded into the framework of varying-coefficient models. One important feature of the approach is that it automatically selects the relevant explanatory variables and no separate variable selection procedure is needed. The properties of the method are investigated in several simulation studies and its usefulness is demonstrated by considering two real-world applications.

Keywords Discrete time-to-event data · Time-varying coefficients · Recursive partitioning · Semiparametric regression · Survival analysis

Marie-Therese Puth
puth@imbie.uni-bonn.de

- ¹ Department of Medical Biometry, Informatics and Epidemiology, Faculty of Medicine, University of Bonn, Venusberg-Campus 1, 53127 Bonn, Germany
- ² Institute of General Practice and Family Medicine, Faculty of Medicine, University of Bonn, Venusberg-Campus 1, 53127 Bonn, Germany
- ³ Department of Statistics, Ludwig-Maximilians-University Munich, Ludwigstrasse 33, 80539 Munich, Germany
- ⁴ Department of Oral and Cranio-Maxillo and Facial Plastic Surgery, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

1 Introduction

Time-to-event models, also referred to as survival models, are a popular tool to analyze data where the outcome variable describes the time to the occurrence of a specific event of interest. In clinical research, for example, one often examines the time to death, the progression of a specific disease, the onset of an infection or the length of stay in hospital (Klein et al. 2016). Further examples from the field of social sciences are the time to re-employment and family developments, like the time to pregnancy or relationship durations (Van den Berg 2001).

The objective of statistical analyses typically is the modeling of the hazard function $\xi(t) = \lim_{\Delta t \rightarrow 0} \{P(t < T \leq t + \Delta t | T > t, \mathbf{x}) / \Delta t\}$, where T denotes the event time, and to relate ξ to a set of explanatory variables $\mathbf{x}^\top = (x_1, \dots, x_p)$. Traditional methods, like the Cox proportional hazards model (Cox 1972), usually assume that the event times T are measured on a *continuous scale*. This case has been studied extensively in the literature, see, for example, Kalbfleisch and Prentice (2002) and Klein and Möscherberger (2003). Yet, in practice, measurements of time are often intrinsically discrete or the exact (continuous) event times are not recorded, but it is only known that the event occurred between pairs of consecutive points in time, i.e. within pre-specified follow-up visits. Thus, time is measured on a discrete scale $t = 1, 2, \dots, k$. In the latter case, the event times t refer to mutually exclusive time intervals $[0, a_1)$, $[a_1, a_2)$, \dots , $[a_{k-1}, \infty)$, with fixed boundaries a_1, \dots, a_{k-1} . A comprehensive treatment of the statistical methodology for discrete time-to-event data has recently been given by Tutz and Schmid (2016) and Berger and Schmid (2018). Generally, a great advantage of discrete time-to-event models is that they can be viewed as regression models with binary outcome variable. This allows to use established tools and standard software packages that have been developed for the analysis of binary outcome data, e.g. logistic regression or probit regression (Willett and Singer 1993).

In parametric discrete time-to-event models one usually uses simple linear combinations of the explanatory variables, that is, one assumes that the effects of the explanatory variables on the outcome are linear. Moreover, it is often assumed that the effects of the explanatory variables on the outcome are constant over the entire observation time. In many applications, however, this assumption is too restrictive and may produce artefacts, see, for example, Tutz and Binder (2004). An important example constitutes the case where the explanatory variables describe an initial condition like the type of treatment at the beginning of a study. Then, the effect on the hazard at earlier times is expected to be stronger than at later times during the study.

This phenomenon can be addressed by semiparametric regression models that incorporate interactions between the explanatory variables and time. In this class of models one allows the effects of the explanatory variables to vary smoothly over time. A common way to specify smooth functions in t is to use splines, which are represented by a weighted sum of basis functions. In the continuous-time case, smooth time-varying effects, inter alia, have been considered by Sargent (1997), Cai and Sun (2003), Tian et al. (2005), Lambert and Eilers (2005), Groll and Tutz (2017) and Ruhe (2018). In discrete time, the modeling of smooth time-varying has been considered by Fahrmeir and Wagenpfeil (1996), Tutz and Binder (2004) and Groll and Tutz (2017), and, for

example, employed by Adebayo and Fahrmeir (2005), Kandala and Ghilagaber (2006) and Djeundje and Crook (2018) in specific applications.

Smoothly time-varying effects are a quite flexible tool but are typically unable to model adequately the effects of explanatory variables that can be constant over a wide range of time though not being constant over the whole range. In particular, if one is interested in the time points where the strength of effects changes, it is more appropriate to model the variation of effects over time by using *piecewise constant* functions. One should also keep in mind that in discrete survival time points refer to intervals. Thus, smooth variation of effects on the underlying continuous time scale may show jumps of the effect strength on the discrete scale. In our approach the ranges where the effects are constant are identified by the use of *recursive partitioning techniques* or *tree-based modeling*. A tree-based approach for modeling time-varying coefficients in continuous time has been proposed by Xu and Adak (2002). To the best of our knowledge for discrete-time models, no tree-based modeling strategy exists so far.

We propose a tree-based approach for modeling piecewise constant time-varying effects in discrete time-to-event models. Specifically, our method is based on the tree-structured varying coefficients (TSVC) approach that was recently proposed by Berger et al. (2018b). Here we use this approach to allow the effects to be modified by the time t , the so-called *effect modifier*, making use of the fact that regression models incorporating time-varying effects may be seen as varying-coefficient models (Hastie and Tibshirani 1993). By iterative splitting in one of the explanatory variables the method yields a tree for each variable that shows time-varying coefficients. For each explanatory variable the proposed algorithm determines whether the effect varies across t , is constant over the whole range of t , or if the variable has no effect on the outcome and should therefore be excluded from the model.

The remainder of the article is organized as follows: In Sect. 2 we give the notation and definitions focusing on right censored data. Details on modeling smooth time-varying coefficients and the proposed tree-structured time-varying coefficient model are introduced in Sect. 3. Section 4 presents the results of several simulation studies. In these studies we investigated the properties of the TSVC model and compared it to a simple model without time-varying coefficients and a model with smooth time-varying coefficients. In Sect. 5 we consider two real-world applications dealing with data collected by the Department of Oral and Cranio-Maxillo and Facial Plastic Surgery at the University Hospital Bonn and data from the German Family Panel (pairfam; Brüderl et al. 2018). Section 6 summarizes the main findings of the article.

2 Notation and methodology

Let in the following T_i denote the event time and C_i the censoring time of an individual i , $i = 1, \dots, n$, with n individuals given. The times T_i and C_i are assumed to be independent random variables taking discrete values in $\{1, \dots, k\}$. For right censored data, the time period during which an individual is under observation is denoted by $\tilde{T}_i = \min(T_i, C_i)$, i.e., \tilde{T}_i corresponds to the true event time if $T_i \leq C_i$ and to the censoring time otherwise. The random variable $\Delta_i := I(T_i \leq C_i)$

indicates whether \tilde{T}_i is right-censored ($\Delta_i = 0$) or not ($\Delta_i = 1$). If originally continuous data have been grouped, the discrete event times $1, \dots, k$ refer to time intervals $[0, a_1), [a_1, a_2), \dots, [a_{k-1}, \infty)$, where $T_i = t$ means that the event occurred in time interval $[a_{t-1}, a_t)$.

The main tool to describe the stochastic behavior of the discrete random variable T_i is the *hazard function*. For given values of p time-constant explanatory variables $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$, the discrete hazard function is defined by

$$\lambda(t|\mathbf{x}_i) = P(T_i = t | T_i \geq t, \mathbf{x}_i), \quad t = 1, \dots, k, \quad (1)$$

which is the conditional probability of an event at time t given that the individual reaches time t . An alternative way to describe the stochastic behavior is to consider the *survival function* given by

$$S(t|\mathbf{x}_i) = P(T_i > t | \mathbf{x}_i) = \prod_{s=1}^t (1 - \lambda(s|\mathbf{x}_i)), \quad (2)$$

denoting the probability that an event occurs later than at time t . For further details on the basic concept of discrete time-to-event data, see Tutz and Schmid (2016), Chapter 1. In the following we consider parametric as well as semiparametric regression models for the discrete hazard $\lambda(t|\mathbf{x}_i)$.

A class of regression models that relates the discrete hazard function (1) to the explanatory variables \mathbf{x}_i is defined by

$$\lambda(t|\mathbf{x}_i) = h(\eta(t, \mathbf{x}_i)), \quad t = 1, \dots, k - 1, \quad (3)$$

where $h(\cdot)$ is a strictly monotone increasing distribution function. Usually it is assumed that the predictor function has the form

$$\eta(t, \mathbf{x}_i) = \gamma_{0t} + \mathbf{x}_i^\top \boldsymbol{\gamma}, \quad (4)$$

which is composed of time-varying intercepts $\gamma_{01}, \dots, \gamma_{0,k-1}$, referred to as *baseline coefficients*, and a linear function of the explanatory variables with coefficients $\boldsymbol{\gamma} \in \mathbb{R}^p$ that do not depend on t . Using the logistic distribution function for $h(\cdot)$ yields the widely applied *logistic discrete hazard model*, specified by

$$\lambda(t|\mathbf{x}_i) = \frac{\exp(\eta(t, \mathbf{x}_i))}{1 + \exp(\eta(t, \mathbf{x}_i))}, \quad (5)$$

which is also known as *proportional continuation ratio model*. The continuation ratio compares the probability of an event at time t to the probability later than t , see, for example, Agresti (2013). As it is the most common model and as the results presented in this paper can easily be extended to other link functions $h(\cdot)$ we reduce our considerations to the logistic model throughout the rest of this article.

By definition, the discrete hazard model (3) has the form of a regression model for binary response data. Therefore, standard estimation techniques for binary regression

can be used for deriving estimates of the model parameters. With data $(\tilde{T}_i, \Delta_i, \mathbf{x}_i)$, $i = 1, \dots, n$, the log-likelihood of model (5) is given by

$$\ell = \sum_{i=1}^n \sum_{t=1}^{\tilde{T}_i} y_{it} \log(\lambda(t|\mathbf{x}_i)) + (1 - y_{it}) \log(1 - \lambda(t|\mathbf{x}_i)), \quad (6)$$

with binary outcome values

$$(y_{i1}, \dots, y_{i\tilde{T}_i}) = \begin{cases} (0, \dots, 0, 1), & \text{if } \Delta_i = 1, \\ (0, \dots, 0, 0), & \text{if } \Delta_i = 0, \end{cases} \quad (7)$$

see, for example, Berger and Schmid (2018). To construct the log-likelihood (6) and to fit the model with software for binary outcomes, the original data has to be converted into an *augmented data matrix* comprising the binary outcome values (7) beforehand. This results in an augmented design matrix with \tilde{T}_i rows for each individual. The whole data matrix, which is obtained by concatenating the individual augmented matrices together, has $\tilde{n} = \sum_{i=1}^n \tilde{T}_i$ rows. For further details on data preparation and the estimation procedure for discrete hazard models, see Tutz and Schmid (2016) and Berger and Schmid (2018).

3 Modeling time-varying coefficients

In common models with predictor (4) it is supposed that the coefficients $\boldsymbol{\gamma}$ do not depend on t . That is, one assumes that the effects of the explanatory variables are constant over the entire observation time. This assumption is typically too restrictive, as, for example, the effect of an explanatory variable on the hazard might be stronger at the beginning of the study than at later times.

3.1 Smooth and piecewise constant time-varying coefficients

A general approach that allows the effects to vary over time is a model with predictor

$$\eta(t, \mathbf{x}_i) = \gamma_{0t} + \mathbf{x}_i^T \boldsymbol{\gamma}(t). \quad (8)$$

The predictor of model (8) contains the vector-valued function $\boldsymbol{\gamma}(t) = (\gamma_1(t), \dots, \gamma_p(t))$. Each component $\gamma_j(t)$ represents the coefficients of the j th explanatory variable depending on the time t . The modeling of discrete event times including smooth time-varying coefficients was, for example, considered by Fahrmeir and Wagenpfeil (1996) and Tutz and Binder (2004). A conventional way to specify such a smooth function in t is to use splines, represented by a weighted sum of M basis functions (e.g. Wood 2017). Then each component $\gamma_j(t)$ has the form

$$\gamma_j(t) = \sum_{m=1}^M \phi_m(t) \beta_{jm}, \quad (9)$$

where $\phi_1(t), \dots, \phi_M(t)$ are M fixed basis functions and $\beta_{j1}, \dots, \beta_{jM}$ are the corresponding parameters to be estimated. Since the explanatory variables all refer to the same period of time (i.e. are measured on the same time scale), the basis functions $\phi_m(t)$ are the same for all variables. With this assumption, the expansion in basis functions yields a model with predictor

$$\eta(t, \mathbf{x}_i) = \gamma_{0t} + \sum_{j=1}^p \sum_{m=1}^M x_{ij} \phi_m(t) \beta_{jm}, \quad (10)$$

which constitutes a linear predictor in the parameters $\gamma_{01}, \dots, \gamma_{0,k-1}, \beta_{11}, \dots, \beta_{pM}$. A popular class of basis functions are B-splines, which are defined as polynomials of fixed degree d differing from zero in $d + 1$ adjacent intervals, see De Boor (1978). In practice one typically uses *P-splines* (Eilers and Marx 1996), i.e., a relatively large number of B-spline basis functions and an additional penalty term that penalizes differences of adjacent coefficients. Fitting of the model can be done by maximization of the corresponding penalized log-likelihood

$$\ell_p = \ell - \epsilon J, \quad (11)$$

where J is the penalty term preventing estimates becoming too wiggly and $\epsilon \in \mathbb{R}^+$ is a penalty parameter that determines the degree of smoothness of the fitted functions $\gamma_j(t)$. For details on spline fitting, see Wood (2011, 2017).

When using P-splines, a smooth variation of the effect strength tends to miss the points where the effect strength changes strongly. Although abrupt changes seem implausible for continuous time data, for discrete time data that refer to intervals of continuous time abrupt changes are to be expected. Therefore, in the following it is assumed that the effects of an explanatory variable do not vary over the whole range of t , but are constant over a certain period of time (or within several time intervals). That is, one assumes that the time-varying coefficients for the j th variable are *piecewise constant* and have the form

$$\gamma_j(t) = \sum_{q=1}^{Q_j} \gamma_{jq} I(t \in T_{jq}), \quad (12)$$

where T_{j1}, \dots, T_{jQ_j} are Q_j time intervals, $\gamma_{j1}, \dots, \gamma_{jQ_j}$ are the corresponding coefficients, and $I(\cdot)$ denotes the indicator function with $I(a) = 1$ if a is true and $I(a) = 0$ otherwise. More specifically, the observation times are divided by the thresholds $1 = t_{j0} \leq t_{j1} \leq \dots \leq t_{j,Q_j-1} \leq t_{jQ_j} = k$, and one obtains a partitioning into the time intervals $T_{j1} = \{t_{j0}, \dots, t_{j1}\}$, $T_{jq} = \{t_{j,q-1} + 1, \dots, t_{jq}\}$, $q = 2, \dots, Q_j$. Accordingly, the coefficients γ_{jq} are constant over the adjacent time points collected

in T_{jq} . The simplest case, a partition of the coefficients of x_j into two time intervals with regard to threshold t_{j1} , yields the function

$$\gamma_j(t) = \gamma_{j1}I(t \in \{1, \dots, t_{j1}\}) + \gamma_{j2}I(t \in \{t_{j1} + 1, \dots, k\}), \quad (13)$$

where the parameter γ_{j1} denotes the effect of x_j in the first interval until time t_{j1} and γ_{j2} denotes the effect of x_j in the second interval between time $t_{j1} + 1$ and k .

For each explanatory variable, the partitioning into the time intervals T_{jq} can be determined by using recursive partitioning techniques. We propose to adapt the tree-based approach that was recently proposed by Berger et al. (2018b). By iterative splitting in one of the explanatory variables the method yields a tree for each variable that shows time-varying coefficients (see Sect. 3.2). Thereby, the algorithm itself identifies the coefficients (corresponding to an explanatory variable) that deviate from a constant, and the corresponding thresholds.

Importantly, the use of the tree-based approach by Berger et al. (2018b) (described in detail in Sects. 3.2 and 3.3) not only achieves the selection of varying and non-varying coefficients, but additionally enforces the selection of variables. More specifically, for each explanatory variable x_j the algorithm determines whether the effect varies across t (by a piecewise constant function), is constant over the whole range of t , or if the variable is influential at all.

3.2 Modeling piecewise constant coefficients by tree-based splits

Assume that we start with the discrete hazard model without time-varying coefficients (4). Then the first split in x_j of a common tree yields a model with predictor

$$\eta(t, \mathbf{x}_i) = \gamma_{0t} + x_{ij} \left[\gamma_{j1}^{[1]} I(t \leq t_{j1}^*) + \gamma_{j2}^{[1]} I(t > t_{j1}^*) \right] + \sum_{s \neq j} x_{is} \gamma_s. \quad (14)$$

This model just uses an alternative representation of the function in (13), but the two intervals regarding x_j are constructed by a split at split point t_{j1}^* with the two parameters $\gamma_{j1}^{[1]}$ (left interval) and $\gamma_{j2}^{[1]}$ (right interval). For given t_{j1}^* , estimates of the parameters in model (14) can still be obtained by maximizing the log-likelihood function (6), plugging in an augmented data matrix, where the column associated with the j th explanatory variable is replaced by two new columns containing the values $x_{ij}I(t \leq t_{j1}^*)$ and $x_{ij}I(t > t_{j1}^*)$, see ‘‘Appendix 1’’.

If the effects of x_j are further modified, a second split (for example in the left interval) with regard to split point t_{j2}^* yields two new intervals

$$I(t \leq t_{j1}^*)I(t \leq t_{j2}^*) \quad \text{and} \quad I(t \leq t_{j1}^*)I(t > t_{j2}^*),$$

and the model with predictor

$$\eta(t, \mathbf{x}_i) = \gamma_{0t} + x_{ij} \left[\gamma_{j1}^{[2]} I(t \leq t_{j1}^*) I(t \leq t_{j2}^*) + \gamma_{j2}^{[2]} I(t \leq t_{j1}^*) I(t > t_{j2}^*) + \gamma_{j3}^{[2]} I(t > t_{j1}^*) \right] + \sum_{s \neq j} x_{is} \gamma_s, \quad (15)$$

where $\gamma_{j1}^{[2]}, \gamma_{j2}^{[2]}, \gamma_{j3}^{[2]}$ are the new effects in the intervals after the second split. Several splits in the coefficients of x_j , result in a sequence of $Q_j - 1$ selected split points $t_{j1}^*, \dots, t_{j, Q_j-1}^*$ and coefficients $\gamma_{j1}^{[Q_j-1]}, \dots, \gamma_{j, Q_j}^{[Q_j-1]}$. Ordering the selected split points, such that $1 \leq t_{(j1)}^* < t_{(j2)}^* < \dots < t_{(j, Q_j-1)}^* < k$, yields the partitioning into the Q_j time intervals

$$T_{j1} = \{1, \dots, t_{(j1)}^*\}, T_{j2} = \{t_{(j1)}^* + 1, \dots, t_{(j2)}^*\}, \dots, T_{j, Q_j} = \{t_{(j, Q_j-1)}^* + 1, \dots, k\},$$

with the corresponding coefficients $\gamma_{j1}^{[Q_j-1]}, \dots, \gamma_{j, Q_j}^{[Q_j-1]}$ representing the piecewise constant function $\gamma_j(t)$.

In general, the effects of all explanatory variables x_1, \dots, x_p in model (4) are allowed to vary over time. This results in several tree components $\gamma_j(t)$, i.e. piecewise constant functions, in the predictor $\eta(t, \mathbf{x}_i)$ of the model. When fitting the model, the first split is determined by selecting the best model among all the explanatory variables x_j and possible split points $t = 1, \dots, k - 1$ (see Sect. 3.3 for details on the selection procedure). The second split is either in the coefficients of the same or another explanatory variable. As in later steps the search is the same but for variables that have already been split, one starts from already built time intervals (corresponding to the current nodes of the tree) which are possibly further split in disjoint intervals. If an explanatory variable is never selected for splitting during iteration, it is assumed to simply have a constant effect γ_j on the hazard over time.

After termination of the algorithm (see Sect. 3.3 for details on stopping criteria), let $V \subseteq \{x_1, \dots, x_p\}$ denote the subset of explanatory variables that have been selected for splitting and $L \subseteq \{x_1, \dots, x_p\} \setminus V$ denote the subset of explanatory variables with a constant effect on the hazard (not selected for splitting). If no split is performed at all, V is an empty set and the result is the simple time-constant model with predictor (4). In the other extreme case, where all explanatory variables are selected for splitting at least once, L is an empty set. With this notation, the tree-structured discrete hazard model has the form

$$\eta(t, \mathbf{x}_i) = \gamma_{0t} + \sum_{x_j \in V} x_{ij} \gamma_j(t) + \sum_{x_\ell \in L} x_{i\ell} \gamma_\ell. \quad (16)$$

In the last step of the algorithm, again following the TSVC approach by Berger et al. (2018b), the time-constant effects γ_ℓ of variables that were not chosen for splitting during iteration are tested for inclusion in the model by using a stepwise elimination scheme. Accordingly the variables are removed from L or kept in the model. If none

of the variables is influential at all, the predictor of the model reduces to the baseline coefficients γ_{0t} only.

3.3 Fitting procedure

In each step of the TSVC algorithm, one selects the best split among all the explanatory variables and possible split points $t = 1, \dots, k - 1$. This is done by testing the equivalence of the two coefficients γ_{jq} and $\gamma_{j,q+1}$ that are associated with the new intervals after splitting. More specifically, one examines all the null hypotheses $H_0 : \gamma_{jq} = \gamma_{j,q+1}$ against the alternatives $H_1 : \gamma_{jq} \neq \gamma_{j,q+1}$ and chooses the combination of x_j and t with the smallest p value of the corresponding likelihood ratio (LR) test.

To decide whether the selected split should be performed, the distribution of the maximally selected LR test statistic, i.e. the maximum of the LR test statistics of the selected variable x_j with regard to t , is investigated. The corresponding p value provides a measure of the dependence between the outcome values and t at a global level and already takes the number of observation times (i.e., the number of possible split points) into account. Therefore, one explicitly accounts for the involved multiple testing problem. To derive a decision on the null hypothesis we propose to use a permutation test. That means one permutes the values of t in the relevant part of the augmented data matrix, which breaks the relation of t and the outcome values in the selected time interval, and computes the corresponding value of the maximally selected LR test statistic (Berger et al. 2018b). For a large number of permutations, one obtains an approximation of the distribution under the null hypothesis and a corresponding p value.

To summarize, the following steps are carried out during the fitting procedure:

1. (*Initial Model*) Fit the model without time-varying coefficients (4), yielding the estimates $\hat{\gamma}_{01}, \dots, \hat{\gamma}_{0,k-1}$ and $\hat{\gamma}_1, \dots, \hat{\gamma}_p$.
2. (*Tree Building*)
 - (a) For all explanatory variables x_j , $j = 1, \dots, p$, fit all the candidate models with one additional split in one of the already built time intervals.
 - (b) Select the best model using the p values of the LR test statistics.
 - (c) Carry out the permutation test for the selected node (defined by a combination of x_j and t) with significance level α . If significant, fit the selected model and continue with Step 2(a), else continue with Step 3.
3. (*Time-Constant Effects*) For all explanatory variables $x_\ell \in L$, examine the null hypotheses $H_0 : \beta_\ell = 0$ by a stepwise backward elimination scheme. Iteratively, the variable with the largest p value, obtained from LR permutation tests with significance level α , is excluded from L . Stop, if none of the p values exceeds α anymore.
4. (*Selected Model*) Fit the final model with components $\hat{\gamma}_{0t}$, $\hat{\gamma}_j(t)$ and $\hat{\gamma}_\ell$.

The resulting discrete hazard model is a specific version of a TSVC model as proposed by Berger et al. (2018b) with the time t (treated as an ordinal variable) being the only permitted effect modifier. The main tuning parameter of the algorithm is the error level α which is used as significance level of the permutation tests. As outlined in

Berger et al. (2018b) the error level α constitutes an upper bound for the proportion of falsely identified variables with time-varying coefficients.

In R, the augmented data matrix for fitting discrete time-to-event models can be generated by using the function `dataLong()` of the add-on package **discSurv** (Welchowski and Schmid 2018). The proposed TSVC model can be fitted by applying the function `TSVC()` of the eponymous add-on package (Berger 2018) with the time t (the only considered effect modifier) specified in the two arguments `effmod` and `only_effmod`.

4 Simulation study

We considered different simulation scenarios in order to evaluate the performance of the proposed TSVC model and to compare the model to alternative approaches. The different scenarios are described in detail in the following subsections: we assessed the performance in terms of a true model without time-varying effects (Sect. 4.1), a true model with smooth time-varying effects (Sect. 4.2) and a true model with piecewise constant time-varying effects (Sect. 4.3).

Particularly, we compared the fit of the TSVC model to the fit of a simple discrete hazard model given by Eq. (4) that did not account for possible time-varying effects (referred to as *NVC model*). In R, simple discrete hazard models without time-varying coefficients were fitted by running `glm()` with family argument `binomial()`. Further, we considered a discrete hazard model allowing for smooth time-varying effects as defined in Eq. (9) (referred to as *SVC model*), using a P-spline for each component $\gamma_j(t)$. In R, discrete hazard models with smooth time-varying coefficients can be fitted by applying the function `gam()` in the add-on package **mgcv** (Wood 2018). The modeling of smooth time-varying coefficients was done by using the function `s()` with the explanatory variables specified in the `by`-argument. The number of basis functions M was set to the default value of **mgcv** with fixed degree $d = 2$. We used a first-order difference penalty J with the optimal smoothing parameter ϵ , see Eq. (11), computed by generalized cross-validation (see Wood 2017).

In all the scenarios we simulated data with constant baseline coefficients $\gamma_{0t} = -2$, $t = 1, \dots, k - 1$, two independent binary explanatory variables, $x_1, x_2 \sim B(1, 0.5)$ and two independent standard normally distributed explanatory variables, $x_3, x_4 \sim N(0, 1)$. The definitions of the respective coefficients of the explanatory variables $\gamma_1, \dots, \gamma_4$ differed in each scenario, hence they are given in the following subsections.

Each scenario was based on 100 independent samples of size $n = 500$ each and the number of discrete time points was set to $k = 11$. During the estimation procedure, for each permutation test we used 1000 permutations with error level $\alpha = 0.05$ throughout all scenarios. Following a strategy already used in Schmid et al. (2017), the censoring times C_i were sampled independently of T by drawing from a discrete distribution with probability density function $P(C_i = t) = b^{(k+1)-t} / \sum_{j=1}^k b^j$, $t = 1, \dots, k$. Three different censoring rates were considered: the value $b = 0.1$ resulted in low censoring ($\sim 30\%$), a value of $b = 1.0$ was used for a medium level of censoring ($\sim 50\%$), and for strong censoring ($\sim 70\%$) the value b was set to 1.5. This resulted in

augmented data matrices with an average number of about $\tilde{n} = 3000$ (low censoring), $\tilde{n} = 2000$ (medium censoring) and $\tilde{n} = 1200$ (strong censoring) rows.

Using the same data-generating process, the performance of the three approaches was assessed by computing the predictive log-likelihood based on new samples with $n = 500$ observations each. Further, to evaluate the performance of the proposed TSVC model, we generated the true positive rate on the covariate level (TPR) and the false positive rate on the covariate level (FPR) that have been introduced in Berger et al. (2018b). The true positive rate describes the amount of all predefined explanatory variables with time-varying effects that have been correctly identified to have an effect that changes over time. Specifically, it is given by

$$TPR = \frac{1}{\#\{j : \vartheta_j = 1\}} \sum_{j:\vartheta_j=1} I(j : \hat{\vartheta}_j = 1),$$

where $\vartheta_j = 1$ if the explanatory variable x_j , $j = 1, \dots, 4$, varies over time. In contrast, the false positive rate displays the amount of all prescribed explanatory variables with time constant effects that have falsely been determined to have an effect that varies in the course of time. It is given by

$$FPR = \frac{1}{\#\{j : \vartheta_j = 0\}} \sum_{j:\vartheta_j=0} I(j : \hat{\vartheta}_j = 1),$$

where $\vartheta_j = 0$ if the explanatory variable x_j , $j = 1, \dots, 4$, has time-constant effects only.

4.1 Model without time-varying effects

In the first scenario, the predictor was given by a linear function of the form

$$\eta(t, \mathbf{x}_i) = \gamma_{0t} + x_{i1}\gamma_1 + x_{i2}\gamma_2 + x_{i3}\gamma_3 + x_{i4}\gamma_4$$

with fixed coefficients $\gamma_1 = 0.4$, $\gamma_2 = -0.4$, $\gamma_3 = -0.2$ and $\gamma_4 = 0.2$. Hence, only samples with time-constant coefficients for all explanatory variables were generated. We used this scenario to examine whether the algorithm of the more complex TSVC model was able to identify the simple model with time-constant coefficients only. This was evaluated by the false positive rate which is anticipated to meet the error level α .

In our simulation study, the TSVC model (on average over the 100 replications) yielded false positive rates that approximately met the intended level of $\alpha = 0.05$ regardless of the censoring rate. In detail, the three different settings resulted in false positive rates of 0.050 (low), 0.058 (medium) and 0.073 (strong), respectively. For low censoring, in 82% of all replications none of the four explanatory variables had been selected for splitting during the fitting procedure. For medium and strong censoring, this rate slightly reduced to 79% and 74%, respectively.

Comparing the performance of the three competing approaches with respect to the predictive log-likelihood (see Fig. 1), the log-likelihood values of the TSVC model

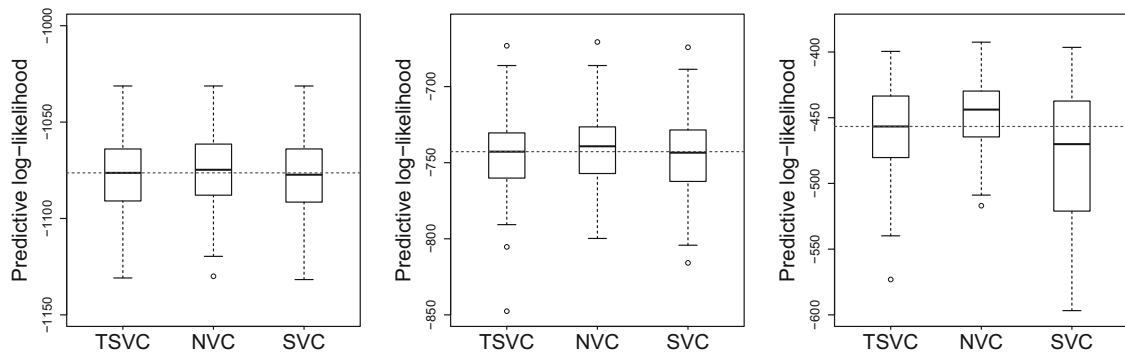


Fig. 1 Results of the simulation study (scenario 1). The figure shows boxplots of the predictive log-likelihood of the TSVC, NVC and SVC model for low (left), medium (center) and strong (right) censoring. The reference line represents the median log-likelihood value of the TSVC model, respectively

were comparable to the ones of the true model (NVC model) with linear predictor (which was expected to perform best). Further, the TSVC model exhibited higher log-likelihood values than the model allowing for smooth time-varying effects (SVC model), in particular for strong censoring the values of the SVC model showed strong variability. The TSVC algorithm showed a rather good performance, which may partly be due to the fact that it was a simple time-constant discrete hazard model if none of the explanatory variables was selected for splitting (see Sect. 3.2).

4.2 Model with smooth time-varying effects

In the second simulation scenario, the underlying model included two explanatory variables with a smooth time-varying effect each while the other effects were kept time-constant. The predictor function had the form

$$\eta(t, \mathbf{x}_i) = \gamma_{0t} + x_{i1}\gamma_1 + x_{i2}\gamma_2 + x_{i3}\gamma_3(t) + x_{i4}\gamma_4(t)$$

with fixed coefficients $\gamma_1 = -0.3$ and $\gamma_2 = 0.3$. For the time-varying coefficients of x_3 and x_4 we used two different sigmoid functions given by

$$\gamma_3(t) = (1 + \exp(5 - t))^{-1}, \quad t = 1, \dots, k,$$

and

$$\gamma_4(t) = \left(1 + \exp(5 - t)^{-1}\right)^{-1}, \quad t = 1, \dots, k.$$

Accordingly, the true data-generating model was a discrete hazard model with smooth time-varying effects. Nevertheless, the TSVC model should still be capable of approximating the functional form of the coefficients of x_3 and x_4 by piecewise constant functions. Figure 2 visualizes the true functions $\gamma_3(t)$ (left panel) and $\gamma_4(t)$ (right panel) and the estimated functions $\hat{\gamma}_3(t)$ and $\hat{\gamma}_4(t)$ obtained by the three approaches for one randomly chosen sample with low censoring. In this example, both the TSVC model and SVC model were well able to approximate the true smooth functions.

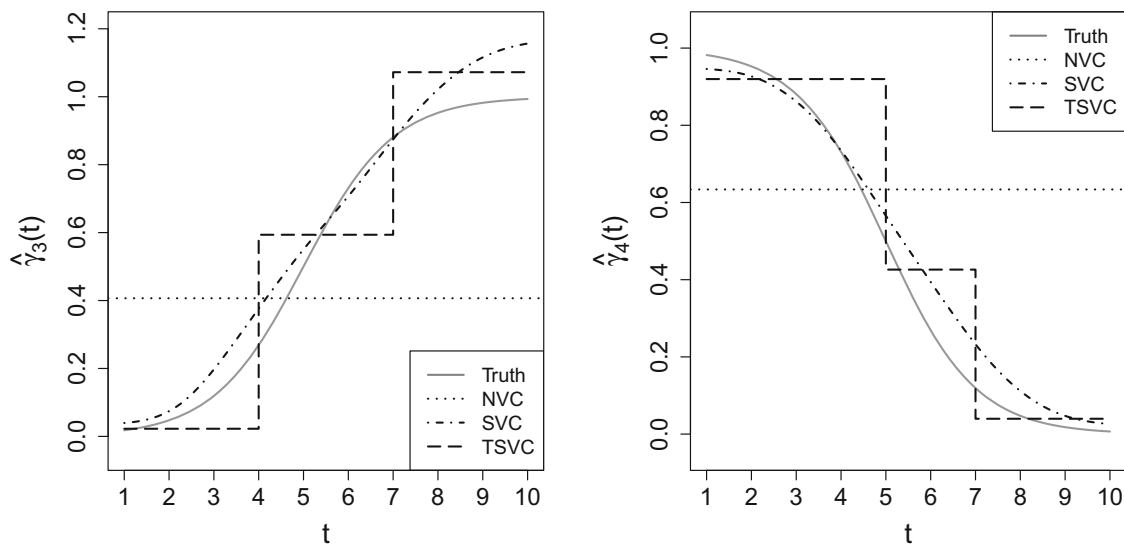


Fig. 2 Results of the simulation study (scenario 2). Estimated coefficients $\hat{\gamma}_3(t)$ of explanatory variable x_3 (left) and $\hat{\gamma}_4(t)$ of explanatory variable x_4 (right) obtained by the three approaches for one randomly chosen sample with low censoring. The true functions are represented by solid lines

Table 1 Results of the simulation study (scenario 2)

Scenario 2	Censoring		
	Low	Medium	Strong
FPR	0.060	0.075	0.070
TPR	1.000	0.945	0.395

Average false positive rates (FPR) and true positive rates (TPR)

Table 1 shows that with increasing level of censoring, the average false positive rate of the TSVC model remained stable while the true positive rate decreased in size. More precisely, for low and medium censoring, the results were similar, with true positive rates higher than 90%. However, for strong censoring the performance of the TSVC model considerably deteriorated. Given all the splits that were generated by the algorithm in any explanatory variable, the proportion of splits in the third or fourth explanatory variable made up 95% for low censoring, 93% for medium censoring and 84% for strong censoring.

For low and medium censoring, the TSVC model and the SVC model provided similar median log-likelihood values whereas the NVC model performed worst (Fig. 3). With increasing level of censoring, the NVC model achieved considerably better results and showed a higher median log-likelihood than the SVC model for strong censoring. In this setting, the TSVC model performed best. The SVC model performed worse as there may be only few observations at later points in time, which made it difficult to correctly identify the time-varying effects of an explanatory variable during the fitting procedure. Further of note, when fitting the SVC model, all four explanatory variables were modeled by smooth model terms using the function $s(\cdot)$. Thus the effects of x_1 and x_2 were not forced to be time-constant, but were also allowed to be fitted as time-varying. This might also be a reason for the high variance of the performance of the SVC model in the strong censoring case.

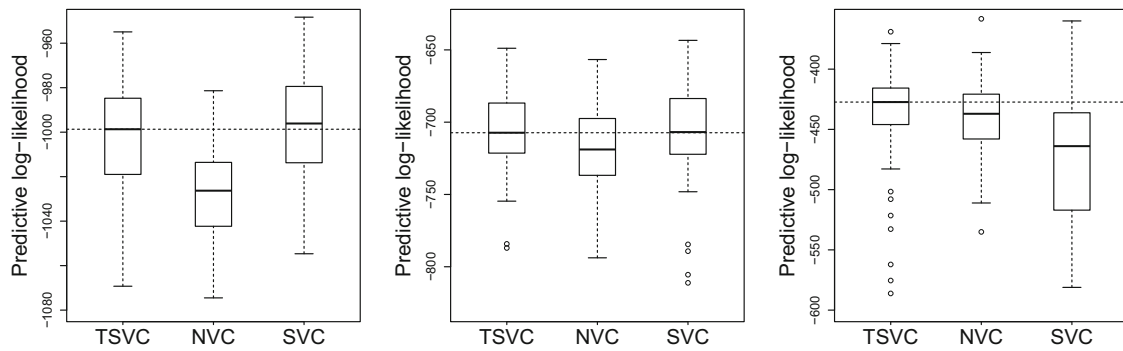


Fig. 3 Results of the simulation study (scenario 2). The figure shows boxplots of the predictive log-likelihood of the TSVC, NVC and SVC model for low (left), medium (center) and strong (right) censoring. The reference line represents the median log-likelihood value of the TSVC model, respectively

4.3 Model with piecewise constant time-varying effects

The third scenario was based on samples with piecewise constant time-varying effects in two explanatory variables, whereas the effects of the other two explanatory variables were kept constant over time. We defined two splits at different event times: for the binary time-varying explanatory variable x_2 , one split was defined at event time $t = 2$. For the standard normally distributed explanatory variable x_4 , one split was defined at event time $t = 5$. Hence, the predictor function of the true model was specified by

$$\eta(t, \mathbf{x}_i) = \gamma_{0t} + x_{i1}\gamma_1 + x_{i2} \left[\gamma_{21}I(t \leq 2) + \gamma_{22}I(t > 2) \right] \\ + x_{i3}\gamma_3 + x_{i4} \left[\gamma_{41}I(t \leq 5) + \gamma_{42}I(t > 5) \right]$$

with fixed coefficients $\gamma_1 = 0.3$, $\gamma_3 = -0.3$. The time-varying effects were generated by $\gamma_{21} = -0.3$, $\gamma_{22} = \gamma_{21} - \delta$ and $\gamma_{41} = 0.5$, $\gamma_{42} = \gamma_{41} + \delta$, respectively. In order to analyze how the amount of change in the effect of the explanatory variable over the course of time affects the performance of the approaches, we considered different values of δ . A value of $\delta = 0$ corresponds to a model with time-constant coefficients only, so δ was set to 0.5, 0.8 or 1.0.

The effect of δ is illustrated in Fig. 4. Without time-varying effects in the coefficients, the number of events consistently decreased over time (see Fig. 4, left panel). Increasing the negative effect of the second explanatory variable resulted in a greater decline of the number of events after time $t = 2$ (see Fig. 4, right panel). In the further course, the number of events increased once the positive effect of the fourth explanatory variable at time $t = 5$ had been modified (see Fig. 4, right panel).

A summary of the results for the different settings with varying δ and varying censoring rate is given in Table 2. Overall, the false positive rates showed values that were close to the anticipated value of 0.05. Irrespective of the level of censoring, the true positive rate increased for higher values of δ . Further we analyzed the rate of correctly splitting at the predefined event times. For low censoring, in 68% of all splits (averaged over the three settings with varying δ) the algorithm correctly generated a split at $t = 2$ in x_2 and at $t = 5$ in x_4 . This rate reduced to average values of 59% and

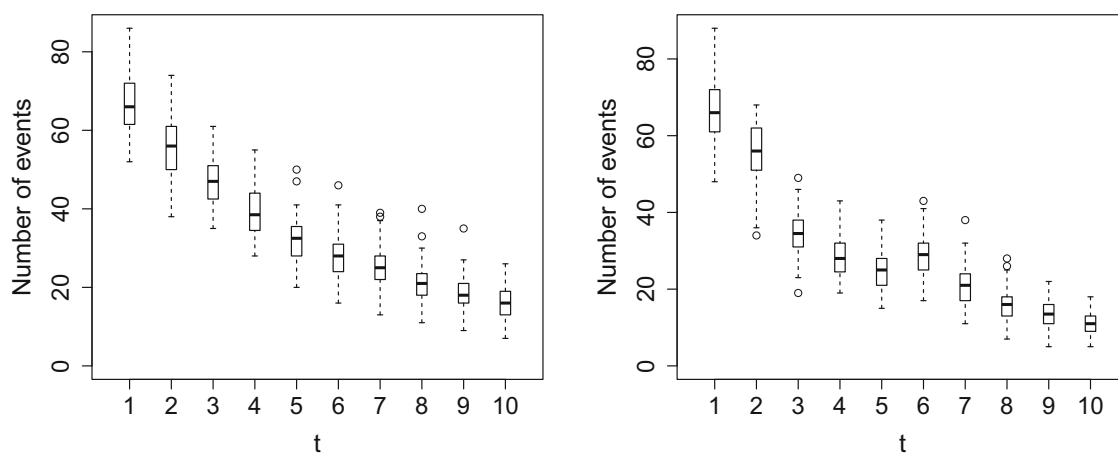


Fig. 4 Results of the simulation study (scenario 3). Illustration of the effect of δ on the number of events. The figure shows boxplots of the number of events over time (observations with $\Delta_i = 1$) for the 100 samples in the setting with low censoring for $\delta = 0.0$ (left) and $\delta = 1.0$ (right). Note the steep decline after $t = 2$ and the increase after $t = 5$ for $\delta = 1.0$

Table 2 Results of the simulation study (scenario 3)

Scenario 3	δ	Censoring		
		Low	Medium	Strong
FPR	0.5	0.070	0.050	0.045
	0.8	0.070	0.085	0.075
	1.0	0.055	0.075	0.050
TPR	0.5	0.550	0.330	0.110
	0.8	0.880	0.655	0.295
	1.0	0.950	0.825	0.340

Average false positive rates (FPR) and true positive rates (TPR) for different values of δ

45% for medium and strong censoring, respectively. The resulting trees for a randomly chosen sample obtained by the TSVC model showing the estimates for the effects of x_2 and x_4 are presented in Fig. 5. The true coefficients in this setting were $\gamma_{21} = -0.3$, $\gamma_{22} = -1.1$ (left panel), and $\gamma_{41} = 0.5$, $\gamma_{42} = 1.3$ (right panel), which were very close to the estimated effects of $\hat{\gamma}_{21} = -0.281$, $\hat{\gamma}_{22} = -1.273$ (left panel) and $\hat{\gamma}_{41} = 0.446$, $\hat{\gamma}_{42} = 1.274$ (right panel), respectively.

Throughout all settings, the median log-likelihood values of the TSVC model were among the highest (Fig. 6), whereas the performance of the NVC model and SVC model strongly varied. For medium censoring, the NVC model and the SVC model showed values comparable to those of the TSVC model. However, for low censoring the performance of the NVC model suffered considerably for higher values of δ . Further, the SVC model (as in the previous scenarios) performed very poorly for strong censoring.

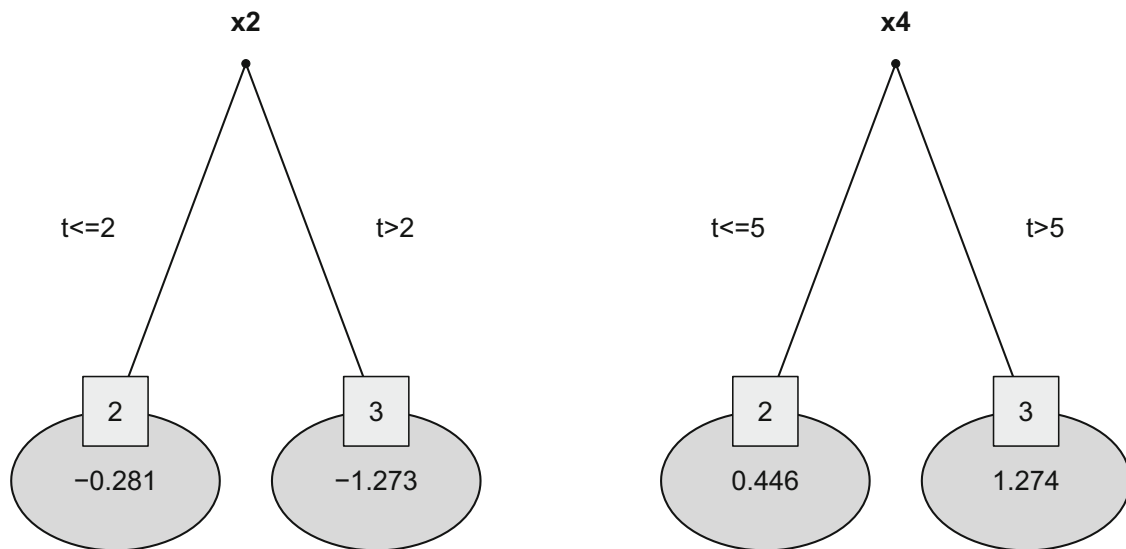


Fig. 5 Results of the simulation study (scenario 3). Estimated trees by the TSVC model for the explanatory variables x_2 (left) and x_4 (right). The results refer to a randomly chosen sample for the setting with $\delta = 0.8$ and low censoring. The estimated time-varying coefficients are given in the leaves of the trees. The true coefficients were $\gamma_{21} = -0.3$, $\gamma_{22} = -1.1$ (left), $\gamma_{41} = 0.5$ and $\gamma_{42} = 1.3$ (right)

4.4 Computational complexity of the fitting procedure

The computational complexity of the TSVC algorithm (in particular the tree building step) is mainly determined by the number of explanatory variables p , the number of time points k and the sample size n . To evaluate this property, we measured the computation time of the fitting procedure in simulation scenarios with time-constant effects in all explanatory variables and a low censoring rate. The specification of the predictor was analogous to the first scenario in Sect. 4.1. We considered scenarios with a varying number of explanatory variables $p = \{4, 8\}$, a varying number of discrete time points $k = \{6, 11, 21\}$ and varying sample size $n = \{100, 500, 1000\}$. In the scenario with $p = 8$ we added two independent binary explanatory variables x_5, x_6 , and two independent standard normally distributed explanatory variables x_7, x_8 with coefficients $\gamma_5 = 0.3$, $\gamma_6 = -0.5$, $\gamma_7 = -0.3$ and $\gamma_8 = 0.5$. Depending on k and n , this also led to a varying number of rows in the augmented data matrices. For the lowest dimensional scenario with $p = 4$, $k = 6$ and $n = 100$, the average number of rows was $\tilde{n} = 275$. With increasing number of discrete time points ($p = 4$, $n = 100$), the average number of rows increased to $\tilde{n} = 405$ for $k = 11$ and $\tilde{n} = 620$ for $k = 21$. With increasing sample sizes ($p = 4$, $k = 6$), the average number of rows increased to $\tilde{n} = 1390$ for $n = 500$ and $\tilde{n} = 2775$ for $n = 1000$.

Figure 7 shows the results (computation time in seconds) based on 100 independent samples each. It is seen that a higher number of explanatory variables (left panel), a higher number of discrete time points (middle panel) as well as a larger sample size (right panel) affected the computation time of the fitting procedure. As the permutation test in each iteration evaluates the candidate models with an additional split at all possible time points, the value of k caused the largest rise in time whereas p and n had a smaller influence.

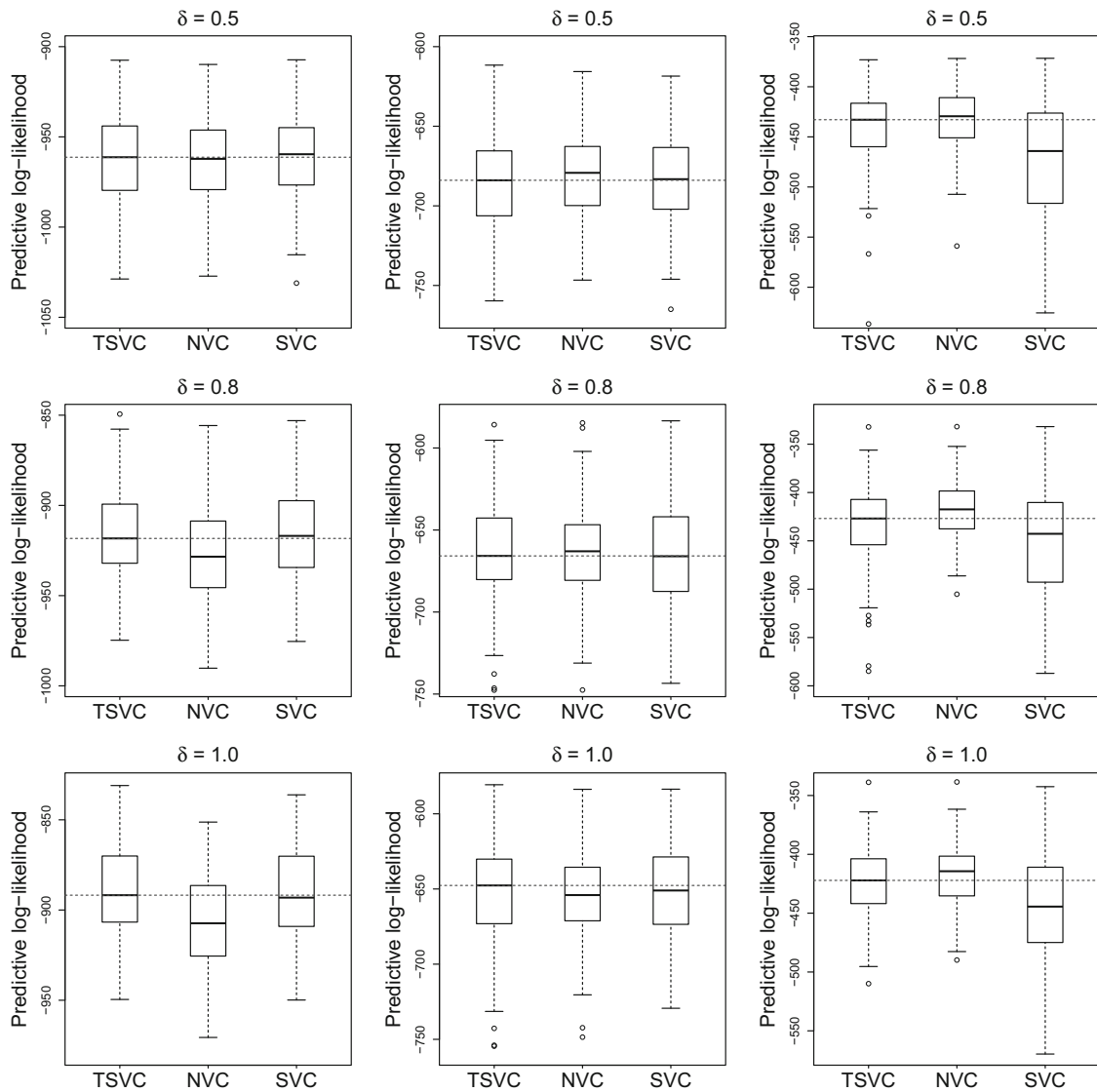


Fig. 6 Results of the simulation study (scenario 3). The figure shows boxplots of the predictive log-likelihood of the TSVC, NVC and SVC model values for low (left), medium (center) and strong (right) censoring. The reference line represents the median log-likelihood value of the TSVC model, respectively

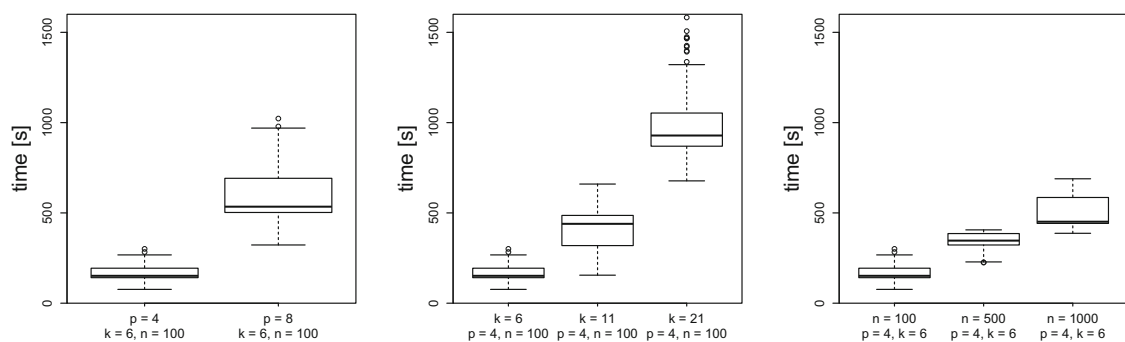


Fig. 7 Results of the simulation study (computational complexity). The figure shows boxplots of the computation time (Hardware: 504 Cores; Opteron 8431 2.4 GHz, Xeon X5650 2.67 GHz, 2.9TB RAM) when running the TSVC algorithm for scenarios with low censoring that differ with regard to the number of explanatory variables p (left), the number of discrete time points k (center) and the sample size n (right)

5 Applications

To further illustrate the use of the TSVC model we considered two real-world applications. In those examples, the use of the TSVC model appeared to be appropriate as the model was able to detect important effects that were easy to interpret but were, for example, not found by a more simple model. As in the previous sections we compared the TSVC model to a simple discrete hazard model (NVC model) and to a discrete hazard model allowing for smooth time-varying effects (SVC model).

5.1 Patients with acute odontogenic infection

We considered data of a 5-year retrospective study investigating in-hospital patients with abscess of odontogenic origin conducted between 2012 and 2017 by the Department of Oral and Cranio-Maxillo and Facial Plastic Surgery at the University Hospital Bonn. An acute odontogenic infection is a major burden for patients' health and public health care systems in western countries (Burnham et al. 2011). Practically, every patient suffers from pain, swelling, erythema and hyperthermia. If not treated at an early stage, odontogenic infections are likely to spread into deep neck spaces and cause perilous complications by menacing anatomical structures, such as major blood vessels, the upper airway and even the mediastinum (Biasotto et al. 2004). The main objective of this study was to investigate risk factors (like age, gender, presence of diabetes mellitus type 2) that tend to prolong the length of stay (LOS) in the treatment of severe odontogenic infections. Predicting the LOS may promote transparency to costs and management of patients under inpatient treatment. For this purpose a discrete time-to-event model was considered, where the event of interest was the discharge from the hospital with the hospitalization measured in days ($t = 1, \dots, 18$).

Here we focused on the data of 303 patients that underwent surgical treatment in terms of incision and drainage of the abscess. Intravenous antibiotics were administered during the operation and for the length of inpatient treatment. For further details on the study we refer to Heim et al. (2018). The LOS of the patients and the patients characteristics considered as explanatory variables in the analysis are summarized in Table 3. These were: age in years (centered around 48), gender (0: female, 1: male), an indicator if the infection spread into other facial spaces (0: no, 1: yes), the location of the infection focus (0: mandible, 1: maxilla), the administered antibiotics (0: ampicillin, 1: clindamycin), the presence of diabetes mellitus type 2 (0: no, 1: yes), and an indicator if the infection was already removed at admission (0: no, 1: yes).

The results of the NVC model and the proposed TSVC model are given in Table 4. The simple NVC model that was recently applied for statistical analysis by Heim et al. (2018) indicated that age and spreading of the infection focus significantly increase the LOS (at the 5% type I error level), while all the other variables showed no evidence for an effect. In particular, diabetes mellitus type 2 revealed no significant increase of the LOS in the present study ($\hat{\gamma} = -0.429$, z value = -1.699), although diabetes stands out as a well investigated cause for an increased LOS (Rao et al. 2010).

As seen from the right part of Table 4, the picture changes when fitting the TSVC model. The algorithm performed one split with respect to the risk factor diabetes at

Table 3 Analysis of the odontogenic infection data

Variable	Summary statistics					
	x_{min}	$x_{0.25}$	x_{med}	\bar{x}	$x_{0.75}$	x_{max}
LOS	1	4	5	5.9	7	18
Age	6	31	48	48.6	64	92
Gender		0: 146	(48.2%)		1: 157	(51.8%)
Spreading		0: 268	(88.4%)		1: 35	(11.6%)
Location		0: 263	(86.8%)		1: 40	(13.2%)
Antibiosis		0: 263	(86.8%)		1: 40	(13.2%)
Diabetes		0: 278	(91.7%)		1: 25	(8.3%)
Remaining focus		0: 118	(38.9%)		1: 185	(61.1%)

Summary statistics of the LOS and the patients characteristics incorporated in the analysis ($n = 303$)

Table 4 Analysis of the odontogenic infection data

Variable	NVC model			TSVC model	
	Coefficient	SE	z value	Estimation	Coefficients
Age	-0.007	0.003	-2.032	Time-constant	-0.008
Gender	-0.222	0.139	-1.592	-	-
Spreading	-0.970	0.212	-4.566	Time-constant	-0.939
Location	0.069	0.208	0.332	-	-
Antibiosis	-0.057	0.203	-0.285	-	-
Diabetes	-0.429	0.252	-1.699	Time-varying	-2.437 0.002
Remaining focus	-0.185	0.148	-1.247	-	-

Overview of the results of the NVC (left) and the TSVC model (right). The algorithm performed one split regarding diabetes at $t = 4$

split point $t = 4$. According to the estimates, there was a strong negative effect ($\hat{\gamma} = -2.437$) at the beginning of the hospitalization ($t \leq 4$), but the effect vanished for later time points ($t > 4$). This result suggested that patients suffering from diabetes mellitus type 2 will hardly be released from the hospital before day 4, an important finding that could not be uncovered by the simple NVC model.

The resulting smooth functions $\gamma_j(t)$ when fitting the SVC model, are shown in Fig. 8. As in the simulation study we used penalized B-spline basis functions with degree $d = 2$ and a first-order difference penalty. In line with the results of the NVC and the TSVC model, the fitted functions and corresponding confidence intervals showed no evidence for an effect of gender, the location of the infection focus and the administered antibiotics. In contrast to the previous results the SVC model revealed linear time-varying effects for the two risk factors spreading of the infection focus and diabetes. However, the confidence intervals for later time points were very wide. This was also the case for γ_{age} and $\gamma_{remaining\ focus}$, which made the effects rather difficult to interpret and strongly suggested that the more parsimonious TSVC model is more appropriate in this analysis.

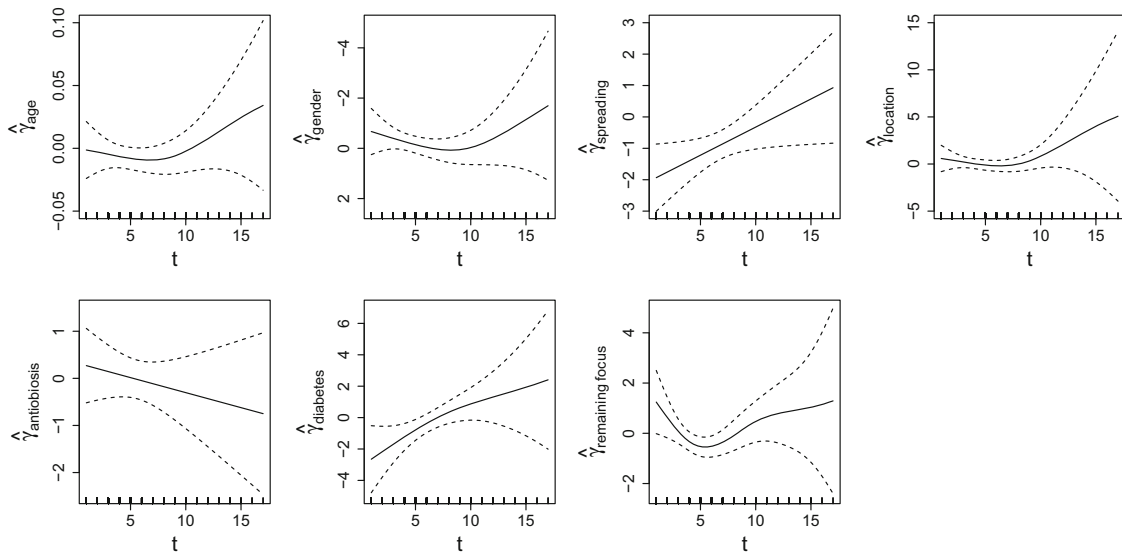


Fig. 8 Analysis of the odontogenic infection data. Estimated effects of the explanatory variables in the SVC model varying over t . Pointwise confidence intervals are drawn by dashed lines, respectively

5.2 Family developments

In a second application, we evaluated data from the first nine waves of the German Family Panel (*pairfam*: Panel Analysis of Intimate Relationships and Family Dynamics), which provides data on family processes in Germany (Brüderl et al. 2018). The first survey in 2008 collected data from a nationwide random sample comprising more than 12,000 respondents of the birth cohorts 1971–1973, 1981–1983 and 1991–1993 and their families. In the multi-cohort approach the main focus is on so-called *anchor persons* of a certain birth cohort, who were annually interviewed to get detailed information on topics like the development of partnership, family plans and formation as well as attitudes regarding parenting in general. In addition, information from parents, partner and children of the anchor person was gathered as well. For further details on the study we refer to Huinink et al. (2011).

As all the information was gathered in one-year intervals, the observed duration times of the *pairfam* study are discrete. The event of interest was defined by the binary outcome whether an anchor woman gave birth to her first child or not. In line with Groll and Tutz (2017), we restricted our consideration to women of the birth cohorts 1971–1973 or 1981–1983 and considered age measured in years as the unit of the discrete hazard model starting with women of at least 25 years. The analysis data set comprised 4077 observations of 861 anchor women who stated to have no children in the initial wave.

As explanatory variables, we included the educational level of the anchor woman measured in years (*yeduc*), the educational levels of the parents of the anchor woman in years (*myeduc* and *fyeduc*), the degree of life satisfaction of the anchor woman (*sat6*, with higher values indicating a higher life satisfaction), the status of relationship of the anchor woman (*relstat*, 0: single, 1: married and/or cohabitation), the employment status of the anchor woman (*casprim*, 0: not employed, 1: employed), the number of siblings of the anchor woman (*siblings*), the amount of leisure time spent for going to

Table 5 Analysis of the pairfam data

Variable	Summary statistics					
	x_{min}	$x_{0.25}$	x_{med}	\bar{x}	$x_{0.75}$	x_{max}
Yeduc	8	11.5	13	13.99	17	20
Myeduc	8	10.5	11.5	12.18	13	20
Fyeduc	8	10.5	11.5	12.76	14.5	20
Sat6	0	7	8	7.47	9	10
Siblings	0	1	1	1.69	2	16
Leisure	1	2	2	1.96	2	4
Relstat		0: 460 (53.4%)			1: 401 (46.6%)	
Casprim		0: 283 (32.9%)			1: 578 (67.1%)	

Summary statistics of the explanatory variables at the first wave in 2008 ($n = 861$)

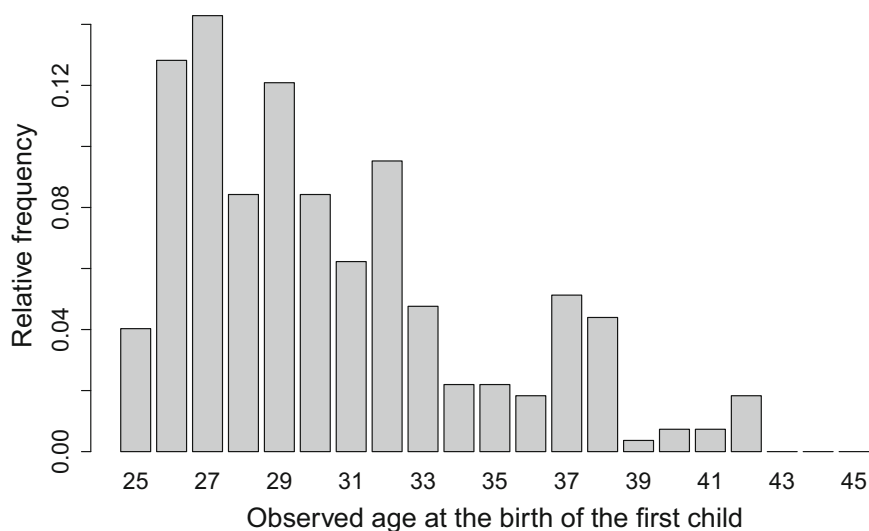


Fig. 9 Analysis of the pairfam data. Distribution of the observed age of the anchor women at the birth of the first child

bars/cafés/restaurants, doing sport, meeting with friends and/or going to a discotheque of the anchor woman (*leisure*, 1: daily, 2: at least once a week, 3: at least once a month, 4: less often, 5: never). A descriptive overview of all explanatory variables at the first wave in 2008 is summarized in Table 5.

In total, there were 273 observed births in our sample and the amount of censoring was 40%. The distribution of the observed age of the anchor women at the birth of the first child is presented in Fig. 9. The median age was 29 years.

The baseline coefficients, which correspond to the effect of age, were fitted by a smooth function as defined in Eq. (9), using P-splines of degree $d = 2$ and a second-order difference penalty. The resulting *baseline hazards* when fitting the TSVc, NVC and SVC model are presented in Fig. 10. The baseline hazard was respectively obtained by transforming the estimated baseline coefficients using the distribution function $\exp(\hat{\gamma}_{0,age}) / (1 + \exp(\hat{\gamma}_{0,age}))$ of the logistic model. It can be seen that the baseline hazards were very similar for the NVC and TSVc model. They were found to show a

Fig. 10 Analysis of the pairfam data. The figure shows the estimated smooth baseline hazard depending on age for the three models TSVC, NVC and SVC

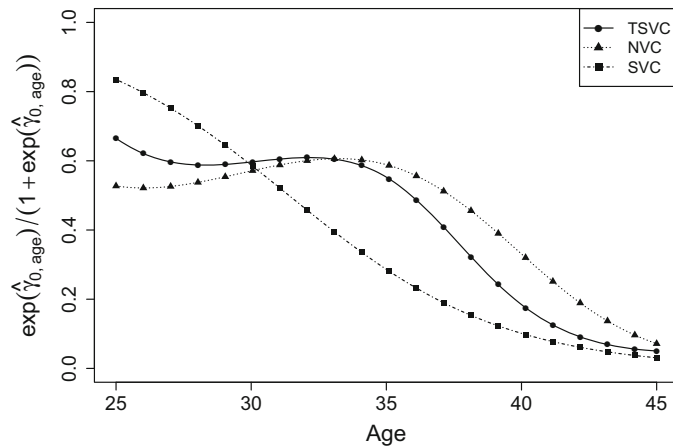


Table 6 Analysis of the pairfam data

Variable	NVC model			TSVC model		
	Coefficient	SE	z value	Estimation	Coefficients	
Yeduc	0.006	0.026	0.249	Time-varying	-0.005	0.044
Myeduc	-0.026	0.034	-0.760	-	-	-
Fyeduc	0.011	0.030	0.370	-	-	-
Sat6	0.204	0.047	4.333	Time-varying	0.172	0.216
Siblings	0.117	0.043	2.697	Time-constant	0.123	-
Leisure	0.250	0.111	2.249	Time-constant	0.251	-
Relstat	1.696	0.172	9.849	Time-constant	1.699	-
Casprim	-0.135	0.156	-0.865	-	-	-

Overview of the results of the NVC (left) and the TSVC model (right). The algorithm performed one split regarding *yeduc* at *age* > 36 and one split with respect to *sat6* at *age* > 28

high hazard up to age 35 followed by a strong decline beyond age 35. In contrast the SVC model yielded a steady decline across time.

The estimated coefficients, standard errors and *z* values obtained by the NVC model are given in Table 6 (left part). There were significant effects for all variables except for the years of education (of the anchor woman and her parents) and the employment status (*casprim*). The estimates indicate that the chance to have a child increased with having a relationship, with the number of siblings, with a higher degree of life satisfaction and a lower amount of leisure time.

In the right part of Table 6, the results when fitting the TSVC model are presented. As can be seen there, the algorithm performed two splits with respect to the explanatory variables *yeduc* and *sat6*. Further, there were time-constant effects of the relationship status, the number of siblings and the amount of leisure time. The employment status as well as the educational achievements of the parents were excluded from the model. Figure 11 shows the estimated trees for *yeduc* and *sat6*. In general, the degree of a woman’s life satisfaction had a positive effect on the chance of having a child (as already indicated by the NVC model) but got even stronger with age (*age* > 28 years). The effect of the educational level measured in years of a woman was opposing: while a higher educational level had a positive effect for relatively old women (*age* > 36

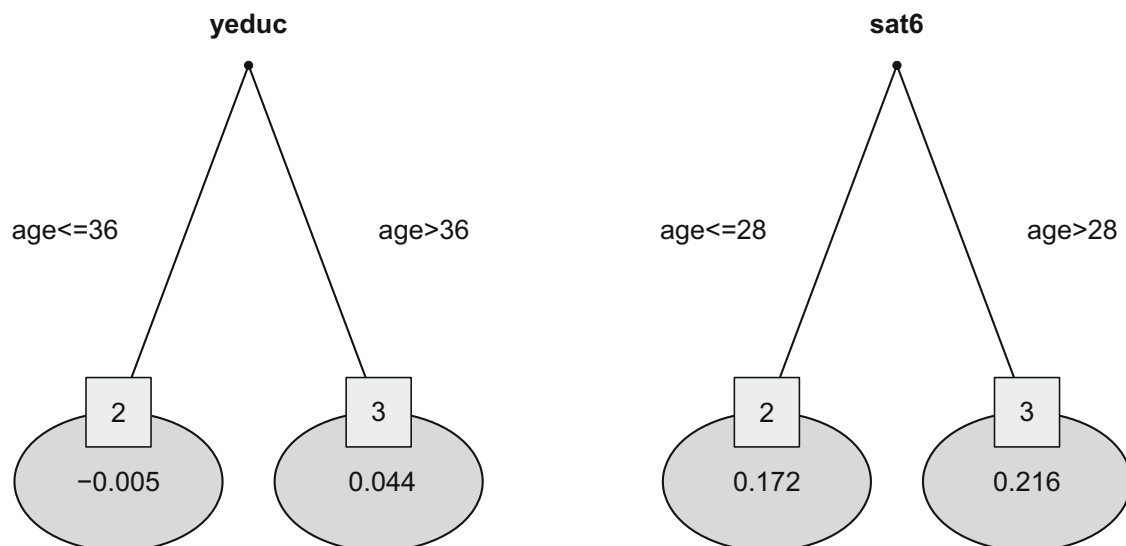


Fig. 11 Analysis of the pairfam data. The estimated time-varying coefficients by the TSVC model for explanatory variables *yeduc* (left) and *sat6* (right). The varying coefficients are given in the leaves of the trees

years), the effect was close to zero for younger women ($\text{age} \leq 36$ years). This finding is in line with a previous analysis of the first four waves of the pairfam study which found that half of the women with an academic degree were at least 35 years at the birth of the first child (Huininik 2014).

Comparing the results of the TSVC model to the NVC model, the TSVC model was able to detect a relevant time-varying effect of *yeduc* which remained undetected by the simple discrete hazard model. For the explanatory variable *sat6*, the NVC model also detected a positive effect, but not the difference over the course of time.

The resulting coefficients when fitting the SVC model allowing for smooth time-varying effects in all explanatory variables are shown in Fig. 12. As seen from the fitted functions and the confidence intervals, there was evidence for (i) time-constant effects of the number of siblings and the amount of leisure time, (ii) time-varying effects of a woman's educational level, the degree of life satisfaction and the relationship status, and (iii) no effects of the parent's educational level and the employment status.

Both, the TSVC and the SVC model showed similar effects for variable *yeduc* on the chances of starting a family, although the function fitted by the SVC model was much more complex. The function also indicated that the effect was close to zero for young women, increased and turned into a constant positive effect for relatively old women ($\text{age} > 35$ years). For the degree of life satisfaction, both models showed a positive effect on the chance for having a child which became slightly stronger with age. Time-constant effects that were similar in magnitude were found for the explanatory variables *leisure* and *siblings*. A difference between the models was obtained for the explanatory variable *relstat*, which was estimated to have a time-constant effect by the TSVC model but a time-varying (decreasing) effect by the SVC model. The confidence intervals of the SVC model are very wide making it dubious that the effect is truly time-varying, which favors the more parsimonious TSVC model.

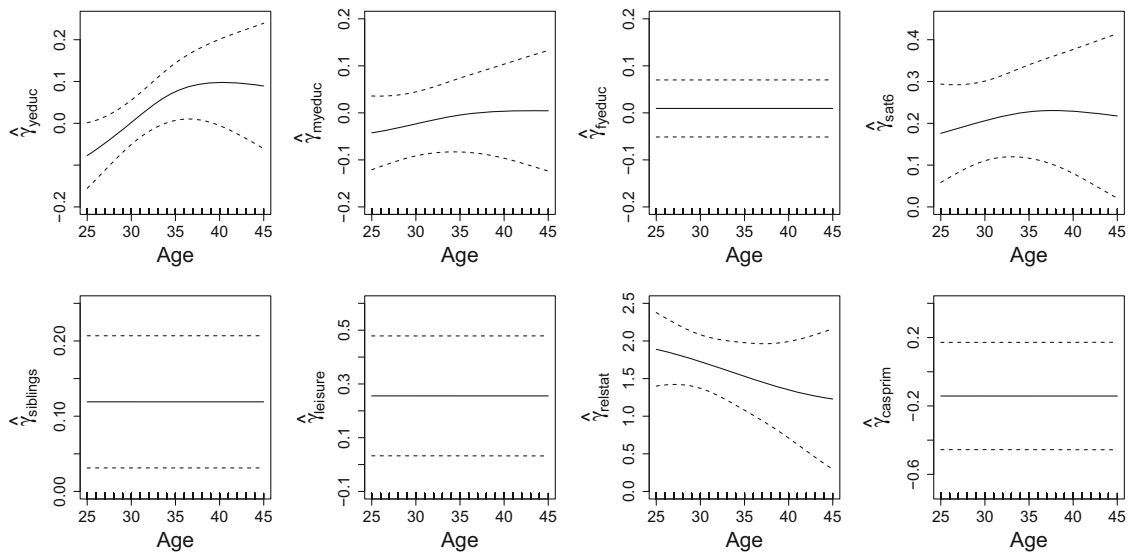
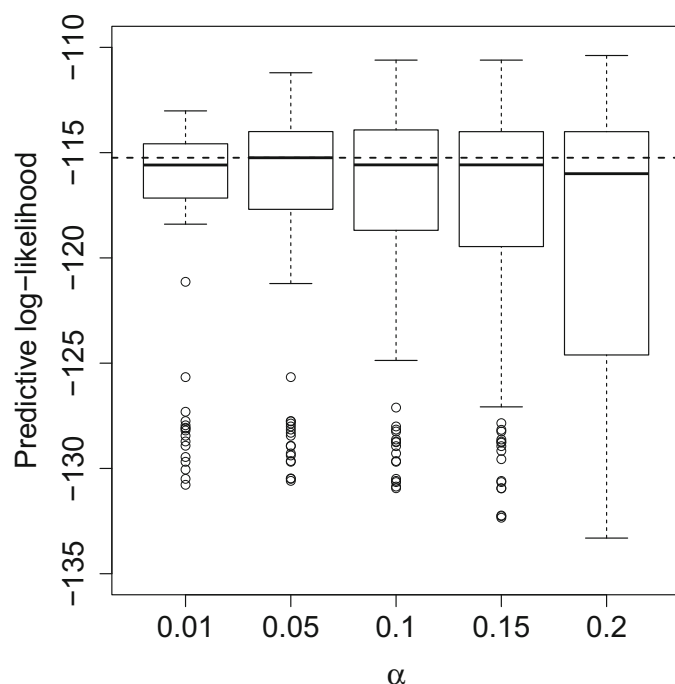


Fig. 12 Analysis of the pairfam data. Estimated effects of the explanatory variables in the SVC model varying over age. Pointwise confidence intervals are drawn by dashed lines, respectively

5.3 Choice of the tuning parameter α

As described in Sect. 3.3, the main tuning parameter of the algorithm is the error level α , which was set to $\alpha = 0.05$ in all the previous simulations and the applications. To investigate the dependence of the proposed TSVC model on α , we compared the prediction accuracy for different values of α using the odontogenic infection data analyzed in Sect. 5.1. We drew 100 subsamples without replacement of size $n_{\text{train}} = 242$ (i.e., 80% of the original sample), fitted the TSVC model using the grid $\alpha = (0.01, 0.05, 0.10, 0.15, 0.20)$ in each of the 100 subsamples and computed

Fig. 13 Analysis of the odontogenic infection data. The boxplots show the predictive log-likelihood values of the TSVC model using different values of α (on the x-axis) based on 100 subsamples of size $n_{\text{train}} = 242$ each. The models were evaluated on the remaining 100 test sets of size $n_{\text{test}} = 61$ each. The reference line represents the median log-likelihood value of the best-performing model



the predictive log-likelihood values from the remaining 100 test sets of $n_{\text{test}} = 61$. Subsampling was stratified by t to ensure a sufficient number of observations per observed event time. It is seen from the boxplots in Fig. 13 that the median log-likelihood value was highest for $\alpha = 0.05$, but did only slightly vary for the other values of α . The variance, however, strongly increased for a high error level ($\alpha = 0.20$), which was caused by the fitting of too large trees in some of the replications. These results underline that the algorithm shows the desired behavior and that the use of $\alpha = 0.05$ is a reasonable choice.

6 Concluding remarks

We propose the use of a tree-based algorithm for the modeling of time-varying coefficients in discrete time-to-event models. The output of the method is a set of piecewise constant functions that are visualized in small trees and are therefore easily accessible. The method constitutes a flexible alternative to models with smooth time-varying coefficients. One of the main features of the algorithm is simultaneous variable selection (of the explanatory variables to be split and corresponding split points) and model fitting, because all the model parameters are refitted in each iteration.

The simulation study essentially showed that the proposed TSVC model (i) performed well in terms of true positive and false positive rates, (ii) was competitive to the simple NVC model in scenarios without time-varying effects, and (iii) was robust against high censoring rates, where the performance of the SVC model strongly suffered. Obviously, a small number of observations at later time points impedes the reliable detection of time-varying effects fitted by smooth functions. Both applications demonstrated the usefulness of the TSVC model, as the model (i) was well able to identify relevant time-varying effects that could not be detected by the simple NVC model, and (ii) was more parsimonious than the SVC model, which yielded easier interpretations of the model fits.

It is important to note that in the representations of the models in Sects. 2 and 3.2 the explanatory variables for simplicity are considered as being constant over time. This restriction is easily removed by allowing time-dependent values $\mathbf{x}_{it}^{\top} = (x_{i1t}, \dots, x_{ip_t})$, $t = 1, \dots, \tilde{T}_i$, as was already done in the pairfam data. The vectors \mathbf{x}_{it} simply need to be inserted in the rows of the augmented data matrices (see also “Appendix 1”) and the analysis can be run in the usual way.

Finally, we restricted our consideration to time-to-event data with a single type of event. An obvious direction for future research is to extend the TSVC model to the competing-risks case with more than one target event that could be realized by embedding the R-package **VGAM** (Yee 2010, 2017) for fitting vector generalized additive models into the fitting algorithm. Recent extensions of the discrete hazard modeling framework to competing-risks models, allowing for more than one target event, were inter alia considered by Möst et al. (2016), Berger et al. (2018a, c) and Heyard et al. (2018).

Acknowledgements This paper uses data from the German Family Panel pairfam, coordinated by Josef Brüderl, Karsten Hank, Johannes Huinink, Bernhard Nauck, Franz Neyer, and Sabine Walper. Pairfam is funded as long-term project by the German Research Foundation (DFG).

Funding The work was supported by the German Research Foundation (DFG), Grant SCHM 2966/2-1.

Appendix 1: Augmented data matrices of the TSVC model given in Eq. (14)

For an individual whose event was observed ($\Delta_i = 1$) at time \tilde{T}_i the augmented data matrix after a split in x_j at split point t_{j1}^* is given by

$$\begin{array}{c} y_i \quad t \quad X_i \\ \left(\begin{array}{ccccccc} 0 & 1 & x_{i1} & \dots & x_{ij} & 0 & \dots & x_{ip} \\ 0 & 2 & x_{i1} & \dots & x_{ij} & 0 & \dots & x_{ip} \\ 0 & 3 & x_{i1} & \dots & x_{ij} & 0 & \dots & x_{ip} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & t_{j1}^* & x_{i1} & \dots & x_{ij} & 0 & \dots & x_{ip} \\ 0 & t_{j1}^* + 1 & x_{i1} & \dots & 0 & x_{ij} & \dots & x_{ip} \\ 0 & t_{j1}^* + 2 & x_{i1} & \dots & 0 & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & \tilde{T}_i & x_{i1} & \dots & 0 & x_{ij} & \dots & x_{ip} \end{array} \right). \end{array} \quad (17)$$

For an individual that is censored ($\Delta_i = 0$) at time \tilde{T}_i the augmented data matrix after a split in x_j at split point t_{j1}^* is given by

$$\begin{array}{c} y_i \quad t \quad X_i \\ \left(\begin{array}{ccccccc} 0 & 1 & x_{i1} & \dots & x_{ij} & 0 & \dots & x_{ip} \\ 0 & 2 & x_{i1} & \dots & x_{ij} & 0 & \dots & x_{ip} \\ 0 & 3 & x_{i1} & \dots & x_{ij} & 0 & \dots & x_{ip} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & t_{j1}^* & x_{i1} & \dots & x_{ij} & 0 & \dots & x_{ip} \\ 0 & t_{j1}^* + 1 & x_{i1} & \dots & 0 & x_{ij} & \dots & x_{ip} \\ 0 & t_{j1}^* + 2 & x_{i1} & \dots & 0 & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & \tilde{T}_i & x_{i1} & \dots & 0 & x_{ij} & \dots & x_{ip} \end{array} \right). \end{array} \quad (18)$$

The matrices (17) and (18) contain two columns associated with the j th explanatory variable including the values $\mathbf{x}_{ij}^\top I(t \leq t_{j1}^*)$ and $\mathbf{x}_{ij}^\top I(t > t_{j1}^*)$.

References

- Adebayo SB, Fahrmeir L (2005) Analysing child mortality in Nigeria with geoaddivitive discrete-time survival models. *Stat Med* 24:709–728
- Agresti A (2013) *Categorical data analysis*, 3rd edn. Wiley, New York
- Berger M (2018) TSVC: tree-structured modelling of varying coefficients. R package version 1.2.0. <https://CRAN.R-project.org/package=TSVC>
- Berger M, Schmid M (2018) Semiparametric regression for discrete time-to-event data. *Stat Model* 18:322–345
- Berger M, Schmid M, Welchowski T, Schmitz-Valckenberg S, Beyersmann J (2018a) Subdistribution hazard models for competing risks in discrete time. *Biostatistics*. <https://doi.org/10.1093/biostatistics/kxy069>
- Berger M, Tutz G, Schmid M (2018b) Tree-structured modelling of varying coefficients. *Stat Comput*. <https://doi.org/10.1007/s11222-018-9804-8>
- Berger M, Welchowski T, Schmitz-Valckenberg S, Schmid M (2018c) A classification tree approach for the modeling of competing risks in discrete time. *Adv Data Anal Classif*. <https://doi.org/10.1007/s11634-018-0345-y>
- Biasotto M, Pellis T, Cadenaro M, Bevilacqua L, Berlot G, Lenarda RD (2004) Odontogenic infections and descending necrotising mediastinitis: case report and review of the literature. *Int Dental J* 54:97–102
- Brüderl J, Drobnič S, Hank K, Huinink J, Nauck B, Neyer F, Walper S, Alt P, Borschel E, Bozoyan C, Buhr P, Finn C, Garrett M, Greischel H, Hajek K, Herzig M, Huyer-May B, Lenke R, Müller B, Peter T, Schmiedeberg C, Schütze P, Schumann N, Thönnissen C, Wetzel M, Wilhelm B (2018) The German family panel (pairfam). GESIS Data Archive, Cologne. ZA5678 Data file Version 9.1.0. <https://doi.org/10.4232/pairfam.5678.9.1.0>
- Burnham R, Rishi RB, Bridle C (2011) Changes in admission rates for spreading odontogenic infection resulting from changes in government policy about the dental schedule and remunerations. *Br J Oral Maxillofac Surg* 49:26–28
- Cai Z, Sun Y (2003) Local linear estimation for time-dependent coefficients in Cox's regression models. *Scand J Stat* 30:93–111
- Cox DR (1972) Regression models and life-tables. *J R Stat Soc, Ser B (Stat Methodol)* 34:187–220 (with discussion)
- De Boor C (1978) *A practical guide to splines*. Springer, New York
- Djeundje VB, Crook J (2018) Dynamic survival models with varying coefficients for credit risks. *Eur J Oper Res* 275:319–333. <https://doi.org/10.1016/j.ejor.2018.11.029>
- Eilers PH, Marx BD (1996) Flexible smoothing with B-splines and penalties. *Stat Sci* 11:89–102
- Fahrmeir L, Wagenpfeil S (1996) Smoothing hazard functions and time-varying effects in discrete duration and competing risks models. *J Am Stat Assoc* 91:1584–1594
- Groll A, Tutz G (2017) Variable selection in discrete survival models including heterogeneity. *Lifetime Data Anal* 23:305–338
- Hastie T, Tibshirani R (1993) Varying-coefficient models. *J R Stat Soc, Ser B (Stat Methodol)* 55:757–796
- Heim N, Berger M, Wiedemeyer V, Reich RH, Martini M (2018) A mathematical approach improves the predictability of length of hospitalization due to acute odontogenic infection. A retrospective investigation of 303 patients. *J Cranio-Maxillofac Surg* 47:334–340. <https://doi.org/10.1016/j.jcems.2018.11.002>
- Heyard R, Timsit JF, Essaïed W, Held L (2018) Dynamic clinical prediction models for discrete time-to-event data with competing risks—a case study on the OUTCOMEREA database. *Biom J*. <https://doi.org/10.1002/bimj.201700259>
- Huinink J (2014) Alter der Mütter bei Geburt des ersten und der nachfolgenden Kinder - europäischer Vergleich. In: Deutsche Familienstiftung (Hrsg) Wenn Kinder - wann Kinder? Ergebnisse der ersten Welle des Beziehungs- und Familienpanels. Parzellers Buchverlag, Fulda, pp 13–26
- Huinink J, Brüderl J, Nauck B, Walper S, Castiglioni L, Feldhaus M (2011) Panel analysis of intimate relationships and family dynamics (pairfam): conceptual framework and design. *J Fam Res* 23:77–101
- Kalbfleisch JD, Prentice R (2002) *The survival analysis of failure time data*, 2nd edn. Hoboken, Wiley

- Kandala NB, Ghilagaber G (2006) A geo-additive Bayesian discrete-time survival model and its application to spatial analysis of childhood mortality in Malawi. *Qual Quant* 40:935–957
- Klein J, Möscherberger M (2003) *Survival analysis: statistical methods for censored and truncated data*. Springer, New York
- Klein JP, Houwelingen HCV, Ibrahim JG, Scheike TH (2016) *Handbook of survival analysis*. Chapman & Hall, Boca Raton
- Lambert P, Eilers P (2005) Bayesian proportional hazards model with time-varying regression coefficients: a penalized Poisson regression approach. *Stat Med* 24:3977–3989
- Möst S, Pöbnecker W, Tutz G (2016) Variable selection for discrete competing risks models. *Qual Quant* 50:1589–1610
- Rao D, Desai A, Kulkarni R, Gopalkrishnan K, Rao C (2010) Comparison of maxillofacial space infection in diabetic and nondiabetic patients. *Oral Surg, Oral Med, Oral Pathol, Oral Radiol, Endod* 110:e7–e12
- Ruhe C (2018) Quantifying change over time: interpreting time-varying effects in duration analyses. *Polit Anal* 26:90–111
- Sargent DJ (1997) A flexible approach to time-varying coefficients in the Cox regression setting. *Lifetime Data Anal* 3:13
- Schmid M, Tutz G, Welchowski T (2017) Discrimination measures for discrete time-to-event predictions. *Econom Stat* 7:153–164
- Tian L, Zucker D, Wei L (2005) On the Cox model with time-varying regression coefficients. *J Am Stat Assoc* 100:172–183
- Tutz G, Binder H (2004) Flexible modelling of discrete failure time including time-varying smooth effects. *Stat Med* 23:2445–2461
- Tutz G, Schmid M (2016) *Modeling discrete time-to-event data*. Springer, New York
- Van den Berg GJ (2001) Duration models: specification, identification and multiple durations. In: Heckman JJ, Leamer E (eds) *Handbook of econometrics*. North Holland, Amsterdam
- Welchowski T, Schmid M (2018) *discSurv: discrete time survival analysis*. R package version 1.3.4. <http://CRAN.R-project.org/package=discSurv>
- Willett JB, Singer JD (1993) Investigating onset, cessation, relapse, and recovery: why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *J Consult Clin Psychol* 61:952–965
- Wood SN (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models. *J R Stat Soc: Ser B (Stat Methodol)* 73:3–36
- Wood SN (2017) *Generalized additive models: an introduction with R*, 2nd edn. Chapman & Hall, Boca Raton
- Wood SN (2018) *mgcv: mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation*. R package version 1.8-15. <https://CRAN.R-project.org/package=mgcv>
- Xu R, Adak S (2002) Survival analysis with time-varying regression effects using a tree-based approach. *Biometrics* 58:305–315
- Yee TW (2010) The VGAM package for categorical data analysis. *J Stat Softw* 32:1–34
- Yee TW (2017) *VGAM: vector generalized linear and additive models*. R package version 1.0-4. <https://CRAN.R-project.org/package=VGAM>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

2.3 Berger et al., Advances in Data Analysis and Classification 13, 965-990

Das gängigste Regressionsmodell zur Modellierung der diskreten ereignis-spezifischen Hazardfunktionen (4) ist das multinomiale logistische Hazard-Modell der Form

$$\lambda_j(t|X) = \frac{\exp(\eta_j(t, X))}{1 + \sum_{j=1}^J \exp(\eta_j(t, X))}, \quad j = 1, \dots, J, \quad t = 1, \dots, k-1, \quad (19)$$

wobei die Vorhersagefunktionen $\eta_j(\cdot)$ jeweils von den erklärenden Variablen X und der Zeit t abhängen, siehe Tutz und Schmid (2016) für eine Einführung der grundlegenden Konzepte. Parametrische Vorhersagefunktionen, die, wie in Kapitel 2.1 und 2.2 betrachtet, über eine Linearkombination der erklärenden Variablen definiert sind, haben im multinomialen Hazard-Modell (19) den Nachteil, dass sie eine sehr große Anzahl an Parametern beinhalten. Insbesondere wenn die Anzahl der zu schätzenden Koeffizienten im Vergleich zur Anzahl der Beobachtungen in den Daten sehr groß ist, kann dies zu numerischen Problemen führen. Des Weiteren können parametrische Modelle oftmals zu einschränkend sein, wenn Interaktionen höherer Ordnung zwischen den erklärenden Variablen vorhanden sind (siehe auch Kapitel 1.4). Um diesen Problemen zu begegnen, wird in diesem Kapitel ein Baum-basiertes Modell der Form

$$\lambda_j(t|X) = f_j(t, X), \quad j = 1, \dots, J, \quad t = 1, \dots, k-1, \quad (20)$$

vorgeschlagen. Die Funktionen $f_j(\cdot)$ sind dabei durch einen CART (Breiman et al., 1984) mit kategorialer Zielvariable bestimmt. Das Verfahren stellt eine Erweiterung der Methode von Schmid et al. (2016) dar (siehe auch Kapitel 2.1). Insbesondere wird neben dem klassischen Gini-Koeffizienten (Breiman, 1996) die sogenannte Hellinger-Distanz (Cieslak et al., 2012) als Kriterium zur Selektion der optimalen Aufteilungsregeln bei der Baumkonstruktion betrachtet. Die Schätzung des Modells kann mithilfe eines selbst-implementierten R Programms durchgeführt werden, das auf GitHub zur Verfügung gestellt wurde.

Der Nutzen der Baum-basierten Methode wird an den Daten der MODIAMD-Studie illustriert. Das entwickelte Vorhersagemodell für das Auftreten von GA oder CNV zeigt auf, dass das Vorhandensein von refraktilen Drusen und das Alter der Patienten/Patientinnen die wichtigsten Risikofaktoren darstellen.



A classification tree approach for the modeling of competing risks in discrete time

Moritz Berger¹  · Thomas Welchowski¹ · Steffen Schmitz-Valckenberg² · Matthias Schmid¹

Received: 6 April 2018 / Revised: 12 July 2018 / Accepted: 20 September 2018 /

Published online: 28 September 2018

© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

Cause-specific hazard models are a popular tool for the analysis of competing risks data. The classical modeling approach in discrete time consists of fitting parametric multinomial logit models. A drawback of this method is that the focus is on main effects only, and that higher order interactions are hard to handle. Moreover, the resulting models contain a large number of parameters, which may cause numerical problems when estimating coefficients. To overcome these problems, a tree-based model is proposed that extends the survival tree methodology developed previously for time-to-event models with one single type of event. The performance of the method, compared with several competitors, is investigated in simulations. The usefulness of the proposed approach is demonstrated by an analysis of age-related macular degeneration among elderly people that were monitored by annual study visits.

Keywords Discrete time-to-event data · Competing risks · Recursive partitioning · Cause-specific hazards · Regression modeling

Mathematics Subject Classification 62N01 · 62N02 · 62P10 · 62-07

1 Introduction

There are many clinical and epidemiological studies where individuals may experience events of various types. The analysis of this kind of data requires a time-to-event model describing the progression to each of the *competing events*. Typical examples are the development of different kinds of diseases or the occurrence of specific causes of

Moritz Berger
Moritz.Berger@imbie.uni-bonn.de

¹ Department of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Sigmund-Freud-Strasse 25, 53105 Bonn, Germany

² University Eye Hospital Bonn, Sigmund-Freud-Strasse 25, 53127 Bonn, Germany

death that are analyzed in clinical research (Lau et al. 2009; Austin et al. 2016). Detailed introductions to state-of-the-art techniques for competing risks analysis were given by Putter et al. (2007) and Beyersmann et al. (2011). The objective typically is the modeling of the cause-specific hazard functions $\xi_j(t) = \lim_{\Delta t \rightarrow 0} \{P(t < T \leq t + \Delta t, \epsilon = j | T > t, \mathbf{x}) / \Delta t\}$, $j = 1, \dots, J$, where $\epsilon \in \{1, \dots, J\}$ is a random variable indicating the type of event at T (Prentice et al. 1978), and to relate ξ_j to a set of explanatory variables $\mathbf{x}^\top = (x_1, \dots, x_p)$.

Traditional methods, like the Cox proportional hazards model (Cox 1972), which are readily applied to modeling cause-specific hazards, usually assume that the event times are measured on a *continuous* scale. In practice, however, the exact (continuous) event times are often not observed, but only intervals (i.e., pairs of fixed consecutive points in time) at which the events of interest took place. Thus, time is measured on a *discrete* scale. An example, which will be considered in this article, is the development of age-related macular degeneration (AMD) among elderly people that were monitored by annual study visits (Steinberg et al. 2016).

There exist several established approaches for the modeling of cause-specific hazards in discrete time. A comprehensive treatment of the statistical methodology has been given by Tutz and Schmid (2016). Recent extensions, among others, have been proposed by Möst et al. (2016), Luo et al. (2016), Vallejos and Steel (2017) and Meggiolaro et al. (2017). A large part of this methodology refers to parametric regression models using *linear* combinations of the explanatory variables for modeling ξ_j . In many applications, however, parametric models are too restrictive, for example, when higher-order interactions between the explanatory variables are present. Also, the specification of a parametric model, like the multinomial logit model, results in a very large number of parameters relative to the sample size.

These issues can be addressed by the use of *recursive partitioning* or *tree-based modeling*. Tree-based approaches for ordinary discrete time-to-event data with one single type of event have been proposed by Bou-Hamad et al. (2009) and Schmid et al. (2016) and have also been referred to as *survival trees*. In this article, we propose a novel extension of the approach by Schmid et al. (2016) to discrete time-to-event data with competing events. The principle is to model the cause-specific hazards by the use of a classification tree with multi-categorical outcome.

The underlying concept of recursive partitioning has its roots in automatic interaction detection. The most popular version, which the proposed approach is based on, is due to Breiman et al. (1984) and is known by the name *classification and regression trees* (CART). The basic method is conceptually very simple: The predictor space defined by the explanatory variables (containing the complete set of observations) is partitioned into a set of disjoint rectangles (i.e., subgroups of observations) by sequentially applying binary splits. On each rectangle, a simple model (e.g., a constant) is fitted. An easily accessible introduction into the basic concepts is found in Hastie et al. (2009). A comparison of several recent developments of recursive partitioning methods has been given by Doove et al. (2014).

In the presence of competing events the outcome variable of a discrete time-to-event model is a categorical variable, with the outcome categories denoting the type of the observed event or the censoring event. A tailored classification tree (originally designed for multi-categorical outcomes) results in a partition composed of disjoint

subgroups of observations and associated cause-specific hazard estimates that can be used for the prediction of events of future observations.

The rest of the article is organized as follows: in Sect. 2 we give basic notations and definitions. The class of parametric cause-specific discrete hazard models is specified in Sect. 3.1, and the proposed tree-based model is introduced in Sect. 3.2. The results of several simulation studies will be presented in Sect. 4. Specifically, we will compare the performance of the tree-based model to several parametric approaches in terms of predicting future events. Section 5 contains an application on the aforementioned analysis of AMD development. Finally, Sect. 6 summarizes the main findings of the article.

2 The discrete cause-specific hazard function

Let T_i be the event time and C_i the censoring time of individual i , $i = 1, \dots, n$. Both T_i and C_i are assumed to be independent random variables taking discrete values in $\{1, \dots, k\}$. In situations where originally continuous data have been grouped, the discrete event times $1, \dots, k$ refer to time intervals $[0, a_1), [a_1, a_2), \dots, [a_{k-1}, \infty)$, where $T_i = t$ means that the event has occurred in time interval $[a_{t-1}, a_t)$. For right-censored data, the time period during which an individual is under observation is denoted by $\tilde{T}_i = \min(T_i, C_i)$, i.e., \tilde{T}_i corresponds to the true event time if $T_i \leq C_i$ and to the censoring time otherwise. The random variable $\Delta_i := I(T_i \leq C_i)$ indicates whether \tilde{T}_i is right-censored ($\Delta_i = 0$) or not ($\Delta_i = 1$). It is assumed that there are J competing events and that the event type is denoted by $\epsilon_i \in \{1, \dots, J\}$. Throughout this article, the focus is on modeling the occurrence of one of the J competing events by also taking into account the censoring event ($\Delta_i = 0$).

For given values of p explanatory variables $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$ that are constant over time, the discrete *cause-specific hazard function* for an event of type j is defined by

$$\lambda_j(t|\mathbf{x}_i) = P(T_i = t, \epsilon_i = j | T_i \geq t, \mathbf{x}_i), \quad j = 1, \dots, J, \quad t = 1, \dots, k, \quad (1)$$

which is the conditional probability of an event of type j at time t given that the individual reaches time t . To describe the whole dynamics of the survival process one can combine all the cause-specific hazard functions $\lambda_1, \dots, \lambda_J$ to obtain the *overall hazard function* given by

$$\lambda(t|\mathbf{x}_i) = \sum_{j=1}^J \lambda_j(t|\mathbf{x}_i) = P(T_i = t | T_i \geq t, \mathbf{x}_i), \quad t = 1, \dots, k, \quad (2)$$

which is the probability of experiencing any event at time t given that t has been reached. The conditional probability of experiencing no event at t , i.e., $P(T_i > t | T_i \geq t, \mathbf{x}_i)$, is then given by $1 - \lambda(t|\mathbf{x}_i)$. The corresponding *survival function* derived from Equation (2) has the form

$$S(t|\mathbf{x}_i) = P(T_i > t | \mathbf{x}_i) = \prod_{s=1}^t (1 - \lambda(s|\mathbf{x}_i)), \quad (3)$$

which denotes the probability that an event (of any type) occurs later than at time t . For an overview of the basic concepts for modeling competing risks, see Tutz and Schmid (2016), Chapter 8. The focus of the next sections will be on parametric and tree-based models for the cause-specific hazards $\lambda_j(t|\mathbf{x}_i)$.

3 Methods

The proposed tree-based model (presented in Sect. 3.2) is rooted in the concepts of classical discrete cause-specific hazard modeling, which will be summarized briefly in the following.

3.1 Parametric models for discrete cause-specific hazards

To model the cause-specific hazard functions (1) one usually considers the general class of multi-categorical response models of the form

$$\lambda_j(t|\mathbf{x}_i) = h(\eta_j(t, \mathbf{x}_i)), \quad j = 1, \dots, J, \quad t = 1, \dots, k - 1, \quad (4)$$

where $h(\cdot)$ is a response function and $\eta_j(\cdot) \in \mathbb{R}$ are predictor functions. Usually the predictor functions are specified by $\eta_j(t, \mathbf{x}_i) = \gamma_{0j}(t) + \mathbf{x}_i^\top \boldsymbol{\gamma}_j$, which contain time-dependent intercepts $\gamma_{0j}(t)$ (referred to as *baseline coefficients*) and linear functions $\mathbf{x}_i^\top \boldsymbol{\gamma}_j$ of the explanatory variables with coefficients $\boldsymbol{\gamma}_j \in \mathbb{R}^p$ that do not depend on t . The baseline coefficients $\gamma_{0j}(t)$ can either be specified by separate intercepts for each t using dummy variables, or by smooth (possibly non-linear) functions of unspecified form using P-splines or smoothing splines. For details on semiparametric approaches for discrete time-to-event models, see Berger and Schmid (2018).

The most popular model for multi-categorical outcomes is the *multinomial logistic regression model*, see Tutz (2012). The associated cause-specific hazard model is specified by setting $h(\eta_j(t, \mathbf{x}_i))$ equal to the logistic response function (Tutz 1995), yielding

$$\lambda_j(t|\mathbf{x}_i) = \frac{\exp(\eta_j(t, \mathbf{x}_i))}{1 + \sum_{j=1}^J \exp(\eta_j(t, \mathbf{x}_i))}. \quad (5)$$

Accordingly, the overall hazard function is obtained by

$$\lambda(t|\mathbf{x}_i) = 1 - \frac{1}{1 + \sum_{j=1}^J \exp(\eta_j(t, \mathbf{x}_i))}. \quad (6)$$

The model based on the logistic response function will be used for comparison purposes in the simulation study and the application in Sects. 4 and 5.

Estimates of the parameters γ_{0j} , $\boldsymbol{\gamma}_j$ are obtained by maximizing the log-likelihood of Model (4). With data $(\tilde{T}_i, \Delta_i, \epsilon_i, \mathbf{x}_i)$, $i = 1, \dots, n$, the likelihood of the model for one individual is given by

$$\begin{aligned}
 L_i &= P(T_i = \tilde{T}_i, \epsilon_i = j | \mathbf{x}_i)^{\Delta_i} P(T_i > \tilde{T}_i | \mathbf{x}_i)^{1-\Delta_i} \\
 &= \lambda_j(\tilde{T}_i | \mathbf{x}_i)^{\Delta_i} (1 - \lambda(\tilde{T}_i | \mathbf{x}_i))^{1-\Delta_i} \prod_{t=1}^{\tilde{T}_i-1} (1 - \lambda(t | \mathbf{x}_i)). \tag{7}
 \end{aligned}$$

For the optimization of the likelihood one can exploit the property that L_i is the same as the likelihood of a multi-categorical response model with outcome categories $j \in \{0, 1, \dots, J\}$, where $j = 0$ denotes the reference category (Tutz 2012). For each t , the corresponding binary outcome variables are defined by

$$\mathbf{y}_{it}^\top = (y_{it0}, y_{it1}, \dots, y_{it\epsilon_i}, \dots, y_{itJ}) = \begin{cases} (1, 0, \dots, 0, \dots, 0), & \text{if } t < \tilde{T}_i, \\ (0, 0, \dots, 1, \dots, 0), & \text{if } t = \tilde{T}_i, \Delta_i = 1, \\ (1, 0, \dots, 0, \dots, 0), & \text{if } t = \tilde{T}_i, \Delta_i = 0. \end{cases} \tag{8}$$

It is assumed that the binary indicator variables are multinomially distributed with $\mathbf{y}_{it}^\top = (y_{it0}, y_{it1}, \dots, y_{itJ}) \sim M(1, 1 - \lambda(t | \mathbf{x}_i), \lambda_1(t | \mathbf{x}_i), \dots, \lambda_J(t | \mathbf{x}_i))$. Using this definitions, the total log-likelihood becomes

$$\ell = \sum_{i=1}^n \sum_{t=1}^{\tilde{T}_i} \left(\sum_{j=1}^J y_{itj} \log(\lambda_j(t | \mathbf{x}_i)) + y_{it0} \log(1 - \lambda(t | \mathbf{x}_i)) \right). \tag{9}$$

The cause-specific hazards $\lambda_j(t | \mathbf{x}_i)$ can be estimated by fitting a multi-categorical regression model with outcome values \mathbf{y}_{it} . This is done by the generation of an *augmented data matrix*, which is composed of smaller (augmented) data matrices for each individual. More specifically, for an individual that experienced an event of type j ($\Delta_i = 1, \epsilon_i = j$) at time \tilde{T}_i the augmented data matrix is given by

$$\begin{pmatrix}
 \mathbf{y}_{i0} & \mathbf{y}_{i1} & \dots & \mathbf{y}_{ij} & \dots & \mathbf{y}_{iJ} & \mathbf{t} & \mathbf{X}_i \\
 1 & 0 & \dots & 0 & \dots & 0 & 1 & x_{i1} \dots x_{ip} \\
 1 & 0 & \dots & 0 & \dots & 0 & 2 & x_{i1} \dots x_{ip} \\
 1 & 0 & \dots & 0 & \dots & 0 & 3 & x_{i1} \dots x_{ip} \\
 \vdots & \vdots & & \vdots & & \vdots & \vdots & \vdots \\
 1 & 0 & \dots & 0 & \dots & 0 & \tilde{T}_i - 1 & x_{i1} \dots x_{ip} \\
 0 & 0 & \dots & 1 & \dots & 0 & \tilde{T}_i & x_{i1} \dots x_{ip}
 \end{pmatrix}. \tag{10}$$

In line with the definition of \mathbf{y}_{it} in Equation (8), the outcome values with $y_{i0} = 1$ in the first column indicate that no event has been experienced yet.

For an individual that is censored ($\Delta_i = 0$) at time \tilde{T}_i the augmented data matrix is given by

$$\begin{pmatrix}
 y_{i0} & y_{i1} & \dots & y_{ij} & \dots & y_{iJ} & t & \mathbf{X}_i \\
 1 & 0 & \dots & 0 & \dots & 0 & 1 & x_{i1} \dots x_{ip} \\
 1 & 0 & \dots & 0 & \dots & 0 & 2 & x_{i1} \dots x_{ip} \\
 1 & 0 & \dots & 0 & \dots & 0 & 3 & x_{i1} \dots x_{ip} \\
 \vdots & \vdots & & \vdots & & \vdots & \vdots & \vdots \\
 1 & 0 & \dots & 0 & \dots & 0 & \tilde{T}_i - 1 & x_{i1} \dots x_{ip} \\
 1 & 0 & \dots & 0 & \dots & 0 & \tilde{T}_i & x_{i1} \dots x_{ip}
 \end{pmatrix}. \tag{11}$$

The overall augmented data matrix, which is obtained by “glueing” the individual augmented data matrices together, has $\tilde{n} = \sum_{i=1}^n \tilde{T}_i$ rows and $J + 2 + p$ columns. The first $J + 1$ columns of the augmented data matrices correspond to the multinomial outcomes y_{it0}, \dots, y_{itJ} . The $(J + 2)$ th column is the time running from $1, \dots, \tilde{T}_i$, which is used as an additional explanatory variable to estimate the baseline coefficients. The values of the explanatory variables are contained in columns $(J + 3)$ to $(J + 2 + p)$.

Instead of maximizing the “pure” log-likelihood (9), it is often more appropriate to use a penalized likelihood of the form $\ell_p(\boldsymbol{\gamma}_{0j}, \boldsymbol{\gamma}_j) = \ell(\boldsymbol{\gamma}_{0j}, \boldsymbol{\gamma}_j) - \vartheta P(\boldsymbol{\gamma}_j)$, where P is a functional that penalizes the magnitude of the coefficients $\boldsymbol{\gamma}_j$ and ϑ is a tuning parameter. A penalized likelihood approach that accounts for the special structure of multi-categorical response models, called *CATS Lasso*, was proposed by Tutz et al. (2015). The corresponding functional P additionally enforces variable selection among the explanatory variables \mathbf{x}_i and thus reduces the complexity of the model. An alternative approach that enforces variable selection in multinomial logistic regression models by likelihood-based boosting was proposed by Zahid and Tutz (2013). An application of the *CATS Lasso* penalty to discrete competing risks models was more recently provided by Möst et al. (2016). As the approach is a competitor to the tree-based method proposed here, it will be compared by means of the application and the simulations in Sects. 4 and 5.

It is important to note that by the representation of \mathbf{X}_i in (10) and (11) one assumes that the explanatory variables are constant over time. This restriction can easily be removed by inserting time-dependent values $\mathbf{x}_{it}^\top = (x_{i1t}, \dots, x_{ipt})$, $t = 1, \dots, \tilde{T}_i$, of the explanatory variables in the rows of the individual augmented data matrices. However, for notational simplicity we reduce our considerations to the case of time-constant explanatory variables throughout the rest of the article.

3.2 Recursive partitioning for discrete cause-specific hazards

The principle of the tree-based method by Schmid et al. (2016), which was designed for time to-event models with one single type of event ($J = 1$), is to fit a discrete hazard model of the form $\lambda(t|\mathbf{x}_i) = f(t, \mathbf{x}_i)$, where the function $f(\cdot)$ is represented by a classification tree with binary outcome. For the construction of the tree, the explanatory variables x_1, \dots, x_p as well as the time t (represented by an ordinal variable) are considered as candidates for splitting. Schmid et al. (2016) propose to apply the CART algorithm based on the Gini impurity measure with minimal node

size pruning. This strategy results in a set of terminal nodes that are represented by sets of binary outcome values and are used to obtain an estimate of the hazard function.

Building a tree means to successively find a partition of the predictor space defined by the explanatory variables. The *root* of the tree is the top node representing the whole predictor space and the resulting *terminal nodes* refer to a disjoint partition. When growing a tree, one successively splits one node M into two subsets M_1 and M_2 . In each step, a single explanatory variable x_s , or the time t , and a corresponding splitting rule are selected for splitting. The splitting rule is applied depending on the scale of the variable. For a *metrically scaled* or *ordinal* variable x_s , the partition into two subsets has the form ' $M_1 = M \cap \{x_s \leq c\}$ and $M_2 = M \cap \{x_s > c\}$ ', with regard to split point c . For a *multi-categorical* variable without ordering $x_s \in \{1, \dots, r\}$, the partition has the form ' $M_1 = M \cap C_1$ and $M_2 = M \cap C_2$ ', where C_1 and C_2 are disjoint, non-empty subsets $C_1 \subset \{1, \dots, r\}$ and $C_2 = \{1, \dots, r\} \setminus C_1$.

In the presence of competing events ($J > 1$), we propose to extend the survival tree method by Schmid et al. (2016) by defining a cause-specific hazard model of the form

$$\lambda_j(t|\mathbf{x}_i) = f_j(t, \mathbf{x}_i), \tag{12}$$

where the functions $f_j(\cdot)$ are determined by a classification tree with multi-categorical outcome. Similar to the single-event case, the building blocks of the proposed tree-based method, which will be explained in the following, are: (i) the specification of the data structure, (ii) the choice of an appropriate splitting criterion, (iii) the choice of tuning parameters for splitting and pruning of the built tree, and (iv) the estimation of the cause-specific hazard functions.

Specification of the data structure

The proposed algorithm is based on the multi-categorical representation of the likelihood function in (9). The corresponding values of the multinomially distributed outcomes are given by the first $J + 1$ columns of the augmented data matrices (10) and (11). To account for the multi-categorical structure of the outcomes, the indicator variables y_{i0}, \dots, y_{iJ} are replaced by one factor variable $y_i \in \{0, \dots, J\}$. Due to its ordered structure, the time t , which is additionally considered as a candidate for splitting during tree building, has to be treated as an ordinal or numerical variable. Thus, the data structure for one individual is given by

$$\begin{pmatrix} y_i & t & \mathbf{X}_i \\ 0 & 1 & \mathbf{x}_i^\top \\ 0 & 2 & \mathbf{x}_i^\top \\ 0 & 3 & \mathbf{x}_i^\top \\ \vdots & \vdots & \vdots \\ \epsilon_i & \tilde{T}_i & \mathbf{x}_i^\top \end{pmatrix} \text{ if } \Delta_i = 1, \epsilon_i = j \text{ and } \begin{pmatrix} y_i & t & \mathbf{X}_i \\ 0 & 1 & \mathbf{x}_i^\top \\ 0 & 2 & \mathbf{x}_i^\top \\ 0 & 3 & \mathbf{x}_i^\top \\ \vdots & \vdots & \vdots \\ 0 & \tilde{T}_i & \mathbf{x}_i^\top \end{pmatrix} \text{ if } \Delta_i = 0. \tag{13}$$

The resulting concatenated data matrix has $\tilde{n} = \sum_{i=1}^n \tilde{T}_i$ rows and $p + 2$ columns.

Choice of the splitting criterion

For the construction of a classification tree with a multi-categorical outcome variable, popular splitting strategies are based on impurity measures. Important examples are the entropy (Quinlan 1986) and the Gini impurity (Breiman et al. 1984). Given the multi-categorical outcome variable y_i , the Gini impurity measure in one node M of a built tree is defined by

$$GI_M = \sum_{j=0}^J \sum_{j \neq k} \pi_j(M) \pi_k(M), \tag{14}$$

where $\pi_j(M)$ is the proportion of observations with outcome value j in node M , see Breiman (1996). In each step of the tree-building algorithm, one chooses the split (among all explanatory variables and corresponding splitting rules) that minimizes the pooled Gini impurity

$$GI(M_1, M_2) = |M_1|GI_{M_1} + |M_2|GI_{M_2}, \tag{15}$$

where $|\cdot|$ denotes the cardinality of the node. The proportions of the outcome values in each node M yield estimates of the cause-specific hazards $\hat{\lambda}_1(M) = \pi_1(M), \dots, \hat{\lambda}_J(M) = \pi_J(M)$ for the subset of individuals and a time interval determined by the node. An estimate of the overall hazard is obtained by the proportion of zero values, namely $\hat{\lambda}(M) = 1 - \pi_0(M)$.

An alternative splitting criterion considered here is the Hellinger distance. Hellinger distance decision tree (HDDT) algorithms have been proposed for binary classification by Cieslak and Chawla (2008) and Cieslak et al. (2012). Their focus was on tree-based methods for unbalanced datasets, that is datasets where one of the two outcome classes is particularly rare. Cieslak and Chawla (2008) showed that HDDT outperforms the classical CART algorithm in terms of the Area under the curve (AUC) in the presence of substantial class imbalance. An extension to multi-categorical classification problems was proposed by Hoens et al. (2012). For the tree-based method proposed here, the Hellinger distance is an attractive choice, as the augmented data matrix used for tree building comprises a disproportionately high number of zero values.

Given a single pair of outcome values, e.g. $\{0, 1\}$, and one node M that is splitted into the two subsets M_1 and M_2 , the Hellinger distance is defined by

$$HD(M_1, M_2) = \sqrt{\left(\sqrt{\pi_1(M_1)} - \sqrt{\pi_0(M_1)}\right)^2 + \left(\sqrt{\pi_1(M_2)} - \sqrt{\pi_0(M_2)}\right)^2}, \tag{16}$$

where $\pi_0(\cdot)$ and $\pi_1(\cdot)$ denote the proportions of zeros and ones in the respective nodes. To account for the multi-categorical structure of the outcome variable, in accordance with Hoens et al. (2012), we propose to consider pairs of subsets of outcome values $j_1 \subset \{0, \dots, J\}$ and $j_2 = \{0, \dots, J\} \setminus j_1$ and to assign the value zero to all categories

in j_1 and the value one to all categories in j_2 . In each step of the tree construction $HD(M_1, M_2)$ is then calculated for all possible pairs of subsets. To determine the best split, one chooses the split with the minimal distance among all pairs of outcome values, all explanatory variables and corresponding splitting rules.

Specification of the tuning parameters

When building trees, the most important tuning parameter is the number of splits that controls the depth and hence the size of the trees. There exist several strategies to determine the adequate size of classification trees. As in traditional approaches we propose to grow large trees and to prune them to an adequate size afterward (Breiman et al. 1984; Ripley 1996). In discrete time-to-event analysis, pruning is essential to ensure sensible estimates of the cause-specific hazards, as the variance of the estimators $\hat{\lambda}_1, \dots, \hat{\lambda}_J$ (defined in detail in the next subsection) is inversely related to the terminal node size. Hence the accuracy of these estimates highly depends on the number of observations in the terminal nodes.

With traditional approaches, tuning is usually achieved by using *cost-complexity pruning*. Starting with the terminal nodes of the grown tree, nodes that result in the smallest decrease in classification accuracy are successively collapsed. One obtains a sequence of nested subtrees, where each subtree minimizes a cost-complexity criterion among all subtrees with the same size. The optimal tree out of this sequence is then given by the subtree with the minimal value of the cost-complexity criterion. This controls the trade-off between classification accuracy and tree size (Mingers 1989). However, it has been shown in several studies that this pruning strategy is not optimal for probability estimation from the terminal nodes (which is the main objective here since the cause-specific hazards (1) are defined in terms of conditional probabilities), because trees that optimize classification accuracy are usually too small (e.g., Provost and Domingos 2003).

For these reasons, we propose to optimize the size of the tree by using the minimal number of observations in the nodes (*minimal node size*) as the main pruning parameter for tree building. The latter is specified as the number of observations that has to be necessarily contained in the current nodes to perform further splits. Thus, if the number of observations in a current node falls below a (predefined) minimal node size, the node is flagged as terminal node. The algorithm terminates when all current nodes are flagged as terminal nodes. For a given sequence of minimal node sizes, the result is again a sequence of nested subtrees, where each subtree is defined by a specific minimal node size.

To determine the tree with the optimal minimal node size we propose to use the log-likelihood of the model. One can either minimize an information criterion (such as AIC and BIC) or maximize the predictive log-likelihood. In analogy to Schmid et al. (2016), we define the information criteria by $-2\ell + \xi(Q - 1)$, where ℓ is the log-likelihood (9), Q is the number of terminal nodes (serving as a measure for the complexity of the tree) and $\xi \in \{2, \log(\tilde{n})\}$. When the predictive log-likelihood is used for tuning, the algorithm performs five-fold cross validation based on subsamples without replacement of size 80% (drawn from the original non-augmented data). To

ensure that each subsample contains a sufficient number of observations per observed event time, the subsamples are stratified by \tilde{T}_i .

The pruning strategy described above is designed such that it results in a compromise between “too small” trees (optimizing classification accuracy but being suboptimal for probability estimation) and “too large” trees (with too few observations in the terminal nodes, resulting in cause-specific hazard estimates with an overly large variance). Analogous to cost-complexity pruning, it generates a sequence of nested trees. In contrast to the former, however, it is not guaranteed that minimal node size pruning will always produce a tree that is optimal among all probability estimation trees of the same size.

Estimation of the cause-specific hazard functions

When a tree has been constructed based on the augmented data matrix, the result is a set of Q terminal nodes that contain a set of n_q multi-categorical outcome values $\mathbf{y}_q = (y_{it1}, \dots, y_{itn_q})^\top \in \{0, \dots, J\}$, $i \in \{1, \dots, n\}$, $t \in \{1, \dots, \tilde{T}_i\}$, $q = 1, \dots, Q$. Estimates of the cause-specific hazards $\hat{\lambda}_{1q}, \dots, \hat{\lambda}_{Jq}$ are derived by the proportions of the outcome values in the respective terminal node. Because the time t is a candidate splitting variable, each terminal node of the tree corresponds to a subset defined by the explanatory variables x_1, \dots, x_p and to a time interval $T_q = [a_q, b_q]$, $1 \leq a_q \leq b_q \leq k$. Splitting in t , which causes the observations from one individual to be allocated to different nodes of the fitted tree, indicates an interaction between the involved explanatory variables and time. This implies the presence of time-varying effects on the cause-specific hazards, which are captured by the tree structure in a very flexible way. If the time t has never been selected for splitting during tree building, it implies that the resulting cause-specific hazards are constant over time. In the other extreme case where only t has been selected for splitting, it implies that the cause-specific hazards depend on time but that the explanatory variables are not influential. As each terminal node is directly interpretable in terms of cause-specific hazard estimates within a specific time interval T_q , estimates of the cause-specific hazard functions are obtained for each individual by concatenating the terminal nodes to which the observations of the individual have been allocated to.

If the number of observations in a terminal node is relatively small or one of the competing events is quite rare, the raw estimate of the associated cause-specific hazard might be close to zero, or even become exactly zero. This is not desirable, as one might observe the same hazard estimates in several nodes independent of the size of the nodes. Therefore, we propose to apply probability smoothing, which was suggested by Ferri et al. (2003) to correct for probability estimates near the boundaries zero and one. For category j in node q , the *Laplace-corrected* cause-specific hazard estimate is defined by

$$\hat{\lambda}_{jq} = \frac{n_j(q) + 1}{|q| + J + 1}, \quad (17)$$

where $n_j(q)$ is the number of observations with outcome value j in node q . Although pruning of the built tree is already designed to avoid extreme estimates close to zero and one, we will use the Laplace correction in both the simulations and application.

For a (new) individual with explanatory variables \tilde{x}_i one obtains the estimated hazard functions $\hat{\lambda}_j(t|\tilde{x}_i)$ by dropping the set of vectors $(\tilde{x}_i^\top, 1), \dots, (\tilde{x}_i^\top, \tilde{T}_i)$ down the final tree. The estimated conditional survival function $\hat{S}(t|\tilde{x}_i)$ can then be derived by applying Eqs. (2) and (3).

3.3 Implementation and available software

In R, the augmented data matrix can be generated by applying the function `dataLongCompRisk()` of the add-on package **discSurv** (Welchowski and Schmid 2017). Parametric cause-specific hazard models can be fitted by the use of the add-on package **VGAM** (Yee 2010, 2017); the function `vglm()` with family argument `multinomial()` allows to fit traditional multinomial logit models. Penalized maximum likelihood estimation can be performed by using the function `MRSP()` of the eponymous add-on package (Pöbnecker 2014). The proposed tree-based method is implemented in a R program that is available from GitHub (<https://github.com/jmober/CompetingRisksTreeDiscSurvival.git>).

4 Simulation study

In this section we present the results of numerical experiments to demonstrate the performance of the tree-based cause-specific hazard model. The aims of the study were: (i) to compare the two splitting criteria *GI* and *HD*, as well as to compare various pruning strategies for determining the optimal minimal node size, and (ii) to compare the tree-based model to parametric models in the presence of interactions between the explanatory variables, and in higher dimensional settings with a large number of non-influential variables.

4.1 Experimental design

We considered simulation scenarios with two competing events $\epsilon_i \in \{1, 2\}$ and discrete event times $\tilde{T}_i = 1, \dots, 11$. In all scenarios we simulated data with four independent binary explanatory variables $x_1, \dots, x_4 \in \{1, 2\}$. The cause-specific hazards were generated by use of the logistic response function and had the form

$$\lambda_j(t|x_1, \dots, x_4) = \frac{\exp(\eta_j(t, x_1, \dots, x_4))}{1 + \sum_{j=1}^2 \exp(\eta_j(t, x_1, \dots, x_4))}, \quad j = 1, 2, \quad (18)$$

with predictors

$$\begin{aligned} \eta_j(t, x_1, \dots, x_4) &= \gamma_{0j}(t) + \eta_j(x_1, \dots, x_4) \\ &= \gamma_{0j}(t) + \gamma_{j1} \cdot x_1x_2 + \gamma_{j2} \cdot x_2x_3 + \gamma_{j3} \cdot x_3x_4, \end{aligned}$$

where the baseline coefficients were randomly drawn from a uniform distribution: $\gamma_{01}(t) \sim U[-5, -4.5]$, $\gamma_{02}(t) \sim U[-4.5, -4]$. The effects of the explanatory variables were set to $\boldsymbol{\gamma}_1 = (0.4, 0.2, 0.3)^\top$ and $\boldsymbol{\gamma}_2 = (0.3, 0.1, 0.5)^\top$. By definition, $\eta_1(x_1, \dots, x_4)$ and $\eta_2(x_1, \dots, x_4)$ were determined by three interaction terms and could take 16 distinct values in the interval $[0.9, \dots, 3.6]$. Following a strategy already used in Schmid et al. (2018), the censoring times were obtained by drawing random numbers from a distribution with probability density function

$$P(C_i = t) = b^{11-t+1} / \sum_{s=1}^{11} b^s, \quad t = 1, \dots, 11, \tag{19}$$

where the degree of censoring was determined by the parameter $b \in \mathbb{R}^+$.

The following steps were carried out for data generation:

- (a) Compute the cause-specific hazards $\lambda_j(t|\mathbf{x}_i)$, $t = 1, \dots, 10$, $j \in \{1, 2\}$.
- (b) Obtain the associated overall hazard functions $\lambda(t|\mathbf{x}_i)$ and survival functions $S(t|\mathbf{x}_i)$. Set $\lambda(11|\mathbf{x}_i) = 1$ for all individuals i .
- (c) Generate the discrete event times T_i from the discrete distribution with probability density function $P(T_i = t|\mathbf{x}_i) = \lambda(t|\mathbf{x}_i)S(t-1|\mathbf{x}_i)$, $t = 1, \dots, 11$.
- (d) Generate the discrete censoring times C_i according to (19).
- (e) Compute the observed event times $\tilde{T}_i = \min(T_i, C_i)$ and the status indicators $\Delta_i = I(T_i \leq C_i)$.
- (f) If $\Delta_i = 1$, determine the type of the event by drawing from a binomial distribution with probabilities $P(\epsilon_i = j|T_i = \tilde{T}_i, \mathbf{x}_i) = \lambda_j(\tilde{T}_i|\mathbf{x}_i)/\lambda(\tilde{T}_i|\mathbf{x}_i)$.

We simulated data comprising $n \in \{200, 500\}$ individuals. Furthermore, we considered a scenario with four additional non-influential variables $x_5, \dots, x_8 \in \{1, 2\}$ (*low-dimensional*) and a scenario with 50 additional non-influential variables $x_5, \dots, x_{54} \in \{1, 2\}$ (*noisy*). The degree of censoring was determined by the parameter b of the censoring distribution (19). We used the values $b = 0.5$ (*weak*), $b = 1$ (*medium*) and $b = 1.5$ (*strong*), yielding the approximate censoring rates shown in Fig. 1. In total this resulted in $2 \times 2 \times 3 = 12$ different scenarios. In each of the scenarios we performed 100 replications, respectively.

Figure 1 shows the relative frequency of observed events for the three low-dimensional scenarios with $n = 200$. It is seen that the number of censoring events increased with increasing value of b . As the true simulated hazards for a type 2 event were higher than for a type 1 event across all t , one observed more events of type 2 than of type 1. For varying b , the ratio of observed type 1 and type 2 events remained approximately the same ($\sim 6/10$). For the scenarios with $n = 500$ and the noisy scenarios, the observed frequencies were almost the same and are thus not shown.

In all 12 scenarios the following modeling approaches were considered: (i) the tree-based approaches introduced in Sect. 3.2, differing with regard to the splitting criterion and the pruning strategy, (ii) a fully specified parametric model (referred to as *Full*) with linear predictors $\eta_j(t, \mathbf{x}_i) = \gamma_{0j}(t) + \mathbf{x}_i^\top \boldsymbol{\gamma}_j$, $j \in \{1, 2\}$, (iii) a parametric model without any explanatory variable (referred to as *Null*), and (iv) a parametric model with linear predictors based on the penalized likelihood with CATS Lasso penalty, where

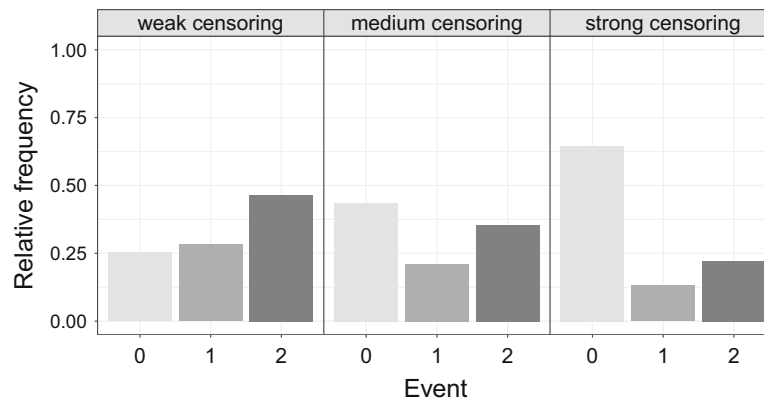


Fig. 1 Illustration of the experimental design of the simulation study. The bars display the average relative frequencies of observed events (0 = censoring event, 1 = event of type 1, 2 = event of type 2) depending on the degree of censoring that were obtained from 100 simulated data sets ($n = 200$)

the effects of all explanatory variables were allowed to be event-specific (referred to as *Lasso*; Tutz et al. 2015). The tuning parameter ϑ was chosen by ten-fold cross-validation based on the predictive deviance. We evaluated the performance of the modeling approaches with regard to predicting events of future observations. In order to do so, we computed the log-likelihood on an independently drawn test data set with equal sample size n in each replication.

4.2 An illustrative example

First we consider in detail the results obtained from fitting one tree-based model. We used an exemplary data set of the low-dimensional scenario with medium censoring and sample size $n = 200$. The fitted trees with tuning by BIC and splitting by Gini impurity (GI) and Hellinger distance (HD) are shown in Fig. 2. It is seen that during tree building both approaches only selected variables from the pool of influential explanatory variables x_1, \dots, x_4 , but none of the additional variables x_5, \dots, x_8 . Both approaches also selected the time t for splitting. Overall, the two approaches yielded quite different splits, which resulted in trees with five (GI) and seven (HD) terminal nodes. High estimated hazards ($\hat{\lambda}_1 = 0.116$, $\hat{\lambda}_2 = 0.197$) were observed for the subset of individuals $\{x_2 = 2 \cap x_3 = 2\}$, which coincided in both trees. The theoretical (simulated) hazards of this subset were $\lambda_1 = 0.117$ and $\lambda_2 = 0.189$ (averaged over all individuals and time points) and thus were very close to the estimated ones. Rather small estimates were obtained for the subset $\{x_2 = 1 \cap x_3 = 1\}$, which was also in line with the true data-generating process. In this scenario, the two approaches yielded the same fitted tree (and thus the same estimated hazards for all the observations) in 10 of 100 replications.

Figure 3 shows the predicted survival functions when applying the fitted models to two exemplary groups of observations in the corresponding test data set containing $n = 200$ observations. It is seen from Fig. 3 that the predicted survival functions only slightly deviated from the true one. Thus the two tree-based models were well able to describe the true underlying survival mechanism.

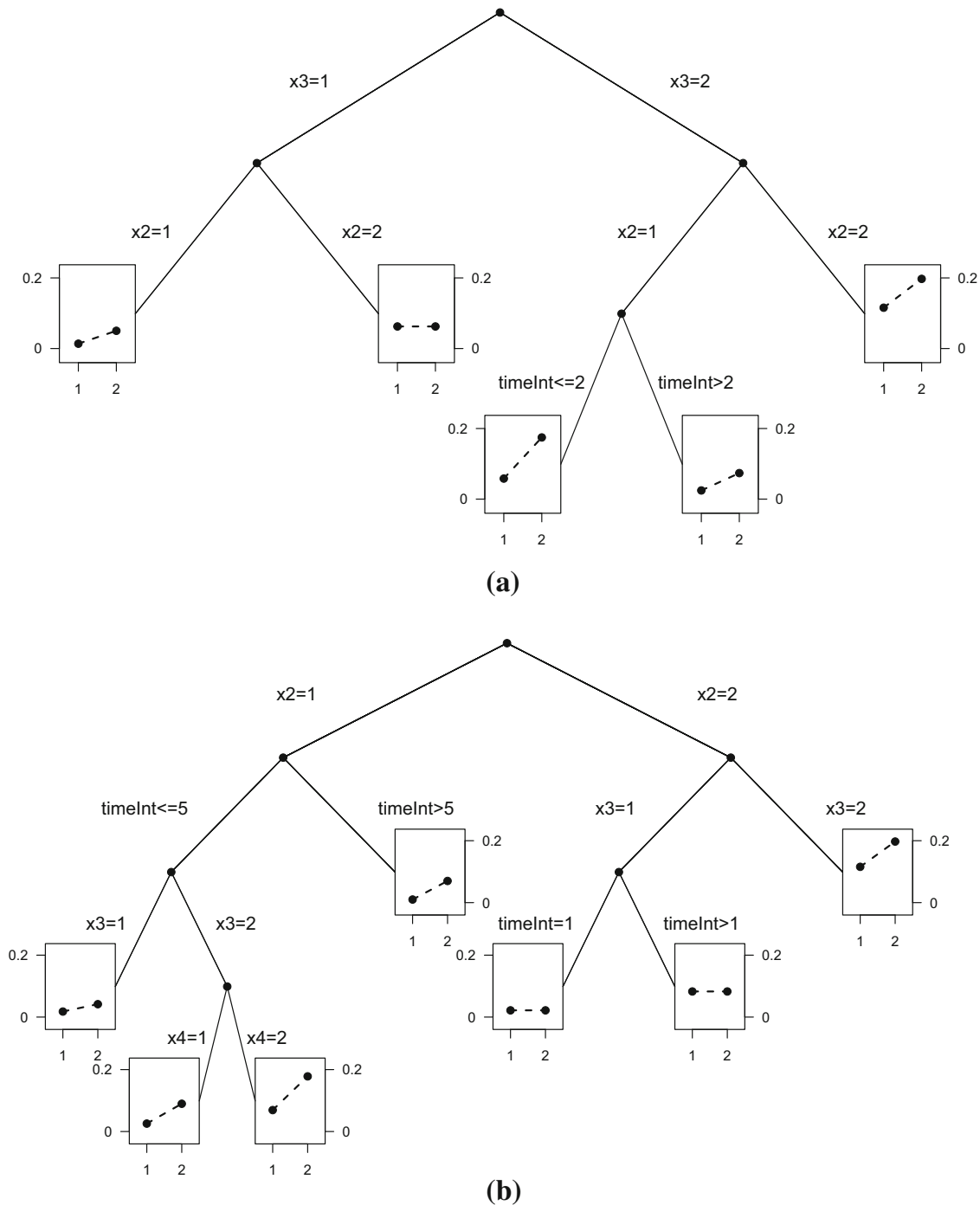


Fig. 2 Fitted trees obtained from one data set ($n = 200$, low-dimensional setting, medium censoring). The trees were built by tuning with BIC and splitting with Gini impurity (GI, upper panel) and Hellinger distance (HD, lower panel). The splitting variable *timeInt* refers to the time column t . At each terminal node the estimated hazards $\hat{\lambda}_1$ and $\hat{\lambda}_2$ (post-processed by application of the Laplace correction) are depicted in the small subfigure

4.3 Comparison of tree-based approaches

Figure 4 shows the prediction accuracy (i.e. the predictive log-likelihood values on the test samples) for the six scenarios with sample size $n = 200$, as obtained from the various tree-based approaches. Specifically, we compared splitting by Gini impurity

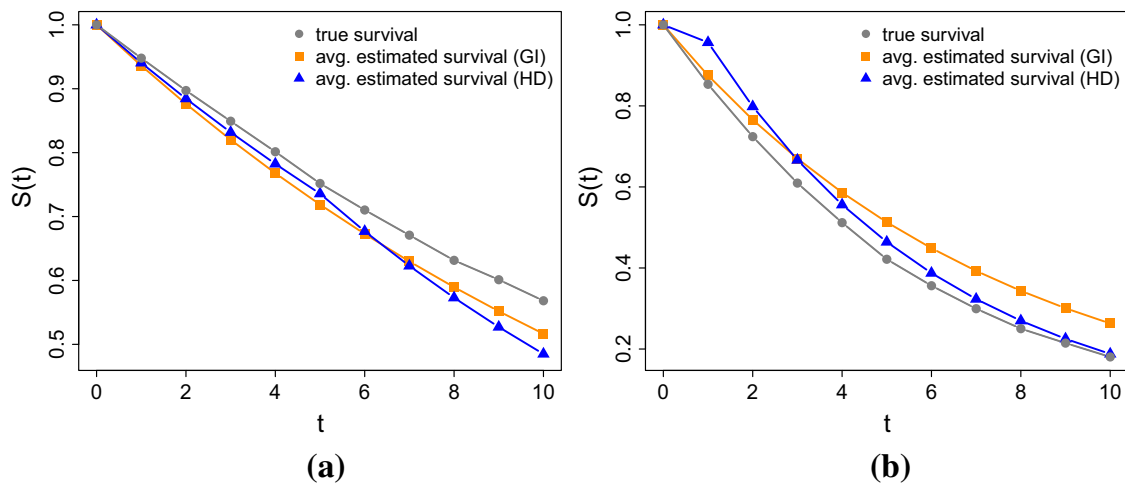


Fig. 3 Predicted survival functions for two groups of observations from one test data set ($n = 200$, low-dimensional setting, medium censoring). Estimates were obtained using the tree-based approach with tuning by BIC and splitting with Gini impurity (GI, orange squares) and Hellinger distance (HD, blue triangles). The respective true survival functions are marked by gray circles (color figure online)

(GI) and Hellinger distance (HD) and pruning based on AIC, BIC and five-fold cross-validated log-likelihood (referred to as ll). This resulted in the $2 \times 3 = 6$ different approaches shown in each panel of Fig. 4. It is seen that the predictive log-likelihood values were smallest for AIC-based pruning throughout all scenarios, yielding the worst performance. BIC-based pruning and ll -based pruning resulted in very similar predictions in the low-dimensional scenarios (left panels). Differences between the two pruning strategies occurred in the noisy scenarios (right panels), where ll was clearly the best-performing pruning method.

Regarding the two splitting criteria, great differences between GI and HD were observed with AIC-based pruning only (where HD-based splitting resulted in a better prediction accuracy). With BIC-based pruning and ll -based pruning, GI performed slightly better than HD in the low-dimensional setting with weak censoring (upper left panel of Fig. 4). In all the other scenarios the performance of both splitting criteria was largely the same. This was particularly the case for the scenarios with strong censoring (lower panels of Fig. 4). Our results thus confirmed the findings reported by Hoens et al. (2012) that HD does not perform significantly better than classical impurity measures in a multi-categorical classification problem. Overall, the simulations suggested that both splitting criteria (GI and HD) are equivalent alternatives for data analysis.

The results for sample size $n = 500$ are shown in Fig. 9 in “Appendix”. There are only slight differences to the previous results across all the six scenarios. We also compared the two splitting criteria in the application (see Sect. 5), where HD-based splitting turned out to be the better choice.

4.4 Comparison to alternative models

The predictive log-likelihood values for the six scenarios with sample size $n = 200$ obtained from the different modeling approaches are presented in Fig. 5. It is seen that the tree-based models outperformed the parametric models in all of the six scenarios.

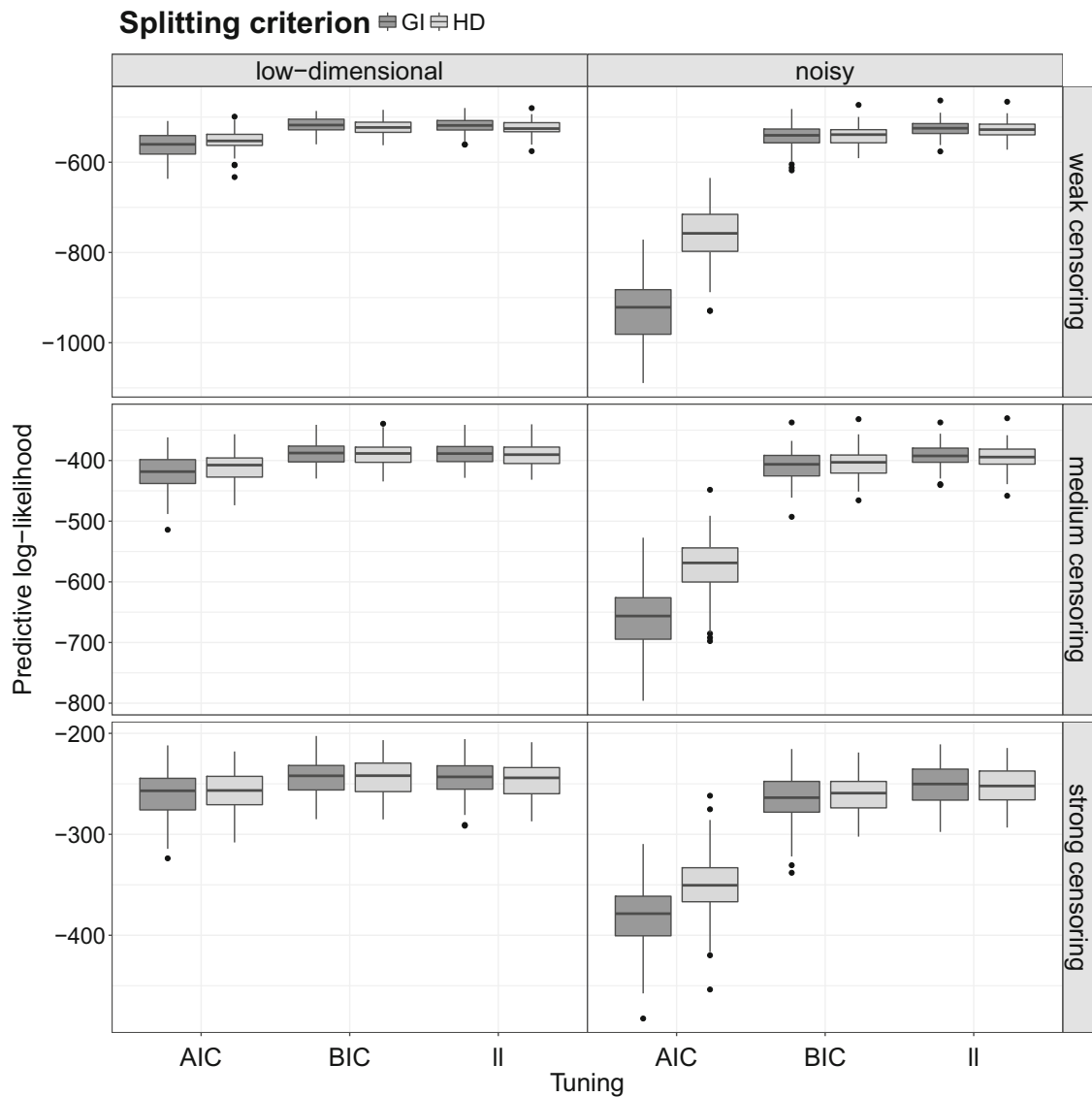


Fig. 4 Results of the simulation study. The boxplots visualize the predictive log-likelihood values obtained from the various tree-based approaches for the six scenarios with $n = 200$. Dark gray boxplots refer to the results with splitting by Gini impurity (GI), light gray boxplots refer to the results with splitting by Hellinger distance (HD). High values of the predictive log-likelihood correspond to good model fits, and vice versa

As expected, the fully specified parametric model (Full) was only competitive in the low-dimensional scenarios (left panels), but strongly suffered in the noisy scenarios (right panels). The penalized likelihood approach with CATS Lasso penalty conspicuously showed a very poor performance in the scenarios with weak censoring (upper panels).

The superiority of the tree-based models might be explained in that the linear modeling approaches were not able to adequately account for the interaction terms contained in the data-generating model (18). Moreover, the results stress the added value of variable selection in the noisy scenarios, which was also enforced by the tailored CATS Lasso penalty.

The results obtained for the scenarios with sample size $n = 500$ are shown in Fig. 10 in “Appendix”. For larger sample size the Full model performed best in the

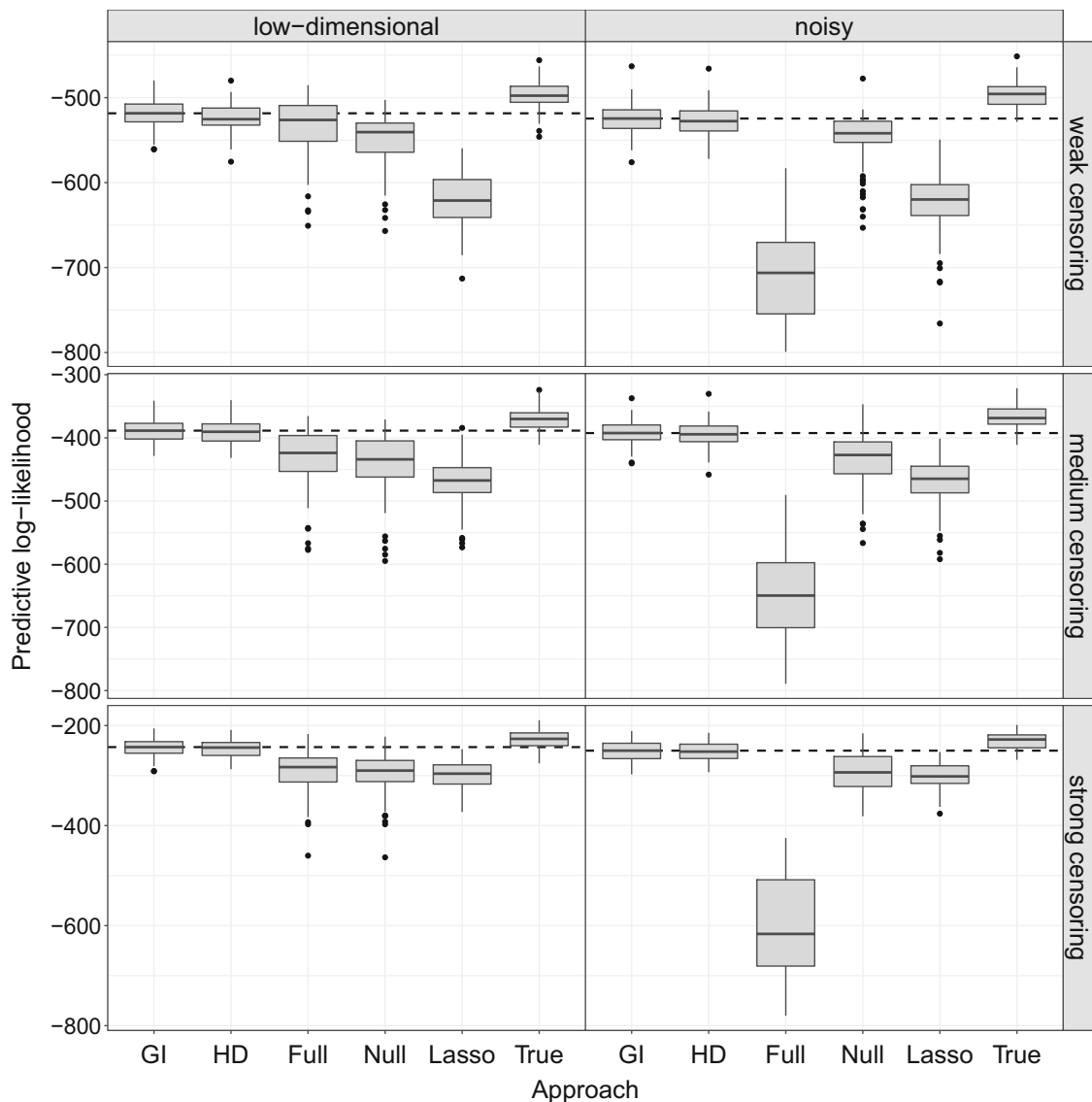


Fig. 5 Results of the simulation study. The boxplots visualize the predictive log-likelihood values obtained from various modeling approaches for the six scenarios with $n = 200$. The first two boxplots (GI and HD) obtained from the tree-based models refer to the results with tuning by the predictive log-likelihood (II). The sixth boxplot in each of the six panels contains the true log-likelihood values of the 100 test data sets (*True*), based on the true hazards defined in (18). Dashed lines refer to the median values of the best-performing tree-based model

low-dimensional scenarios with weak censoring, but again deteriorated in the other scenarios. Surprisingly, CATS Lasso performed worst in all the scenarios. This can again be explained by the misspecification of the predictor of the model, but might also be due to a strong shrinkage of the (influential) parameters by the penalty towards zero.

5 Application: age-related macular degeneration

To illustrate the application of the tree-based cause-specific hazard model, we analyzed the database of the MODIAMD (Molecular Diagnostics of Age-related Macular

Degeneration) study, which is an ongoing non-interventional study in patients at high risk for developing late-stage age-related macular degeneration (AMD, Steinberg et al. 2016). Late-stage AMD is the leading cause of blindness among elderly people in industrialized countries. It either manifests by geographic atrophy (GA) or by choroidal neovascularization (CNV). GA is an advanced stage of AMD with irreversible loss of photoreceptors and severe loss of vision. CNV, also called the “wet” form of advanced AMD, causes vision loss due to abnormal blood vessel growth. Currently, there is no therapeutic intervention to delay or stop this progression to late stage AMD. Although the more common CNV can be treated with anti-vascular endothelial growth factor therapy, it remains a vision threatening disease. Therefore, it is of high interest to develop intervention strategies for high-risk patients. For this purpose, we analyzed the effects of potential risk factors on the development of the two “competing” disease outcomes GA and CNV.

Patients were enrolled between November 2010 and September 2011 at the Department of Ophthalmology, University of Bonn, Germany. There was one study eye per patient. Criteria for the inclusion in the study were: (i) to be older than 50 years of age, (ii) to be AREDS stage 3 or 4, according to the Age-Related Eye Disease (AREDS) classification, and (iii) to have no advanced AMD (GA or CNV) at baseline in the study eye (Steinberg et al. 2016). All patients were monitored at the time of their inclusion in the study (baseline visit) and subsequently monitored by annual study visits. Hence, observing $T_i = 1$ means that an event occurred during the first year of the study. For our analysis, the data up to and including the fifth annual study was available ($t = 1, \dots, 5$).

In total, 98 patients were enrolled in the study. Exclusion of one patient with missing values in the analyzed risk factors resulted in an analysis data set of size $n = 97$. On completion of the fifth visit, 16 study eyes had developed GA and 25 study eyes had developed CNV; 26 patients were still in the study while 30 patients were censored (i.e., had dropped out at earlier visits). Only one of the 30 censored patients died before the fifth visit. Due to this very small number we did not consider death as an additional competing event. The processing of the analysis data resulted in an augmented data matrix with $\tilde{n} = 344$ lines.

Summary statistics of the risk factors incorporated in our analysis (all measured at baseline) are given in Table 1. They included visual acuity (measured as the total number of correctly identified letters on the Snellen chart), drusen volume (mm^3), the presence of the natural crystalline lens of the eye (phakia), smoking, the presence of refractile drusen (ref_drusen) and the disease status of the fellow eye.

Table 2 shows the estimates of the coefficients $\boldsymbol{\gamma}_j$, $j \in \{\text{GA}, \text{CNV}\}$, the corresponding estimated standard errors and the p -values of the explanatory variables obtained from fitting a parametric cause-specific hazard model with linear predictors $\eta_j(t, \mathbf{x}_i) = \gamma_{0j}(t) + \mathbf{x}_i^\top \boldsymbol{\gamma}_j$, $j \in \{\text{GA}, \text{CNV}\}$ and logistic response function (for details, see Tutz and Schmid, 2016, Chapter 8). The p -values in Table 2 indicate that there were risk factors (e.g. visual acuity, drusen volume) with a significant effect on the development of GA or CNV. However, the specification of one parameter vector for each event resulted in a very large number of parameters compared to the observed number of events in the data. This lead to partly unreliable estimates and numerical problems (cf. the effect of a GA in the fellow eye on CNV; last row

Table 1 Analysis of the AMD data

Variable	Summary statistics						
	Min	Q1	Median	Q3	Max	Mean	SD
Age (years)	51	70	73	77	89	73.09	7.10
Visual acuity (snellen)	56	75	79	83	91	77.94	6.94
Drusen volume (mm ³)	0.09	0.15	0.19	0.25	0.66	0.21	0.10
Phakia	No	30 (30.93%)			Yes	67 (69.07%)	
Smoking	No	60 (61.85%)			Yes	37 (38.15%)	
Ref_drusen	No	78 (79.59%)			Yes	19 (20.41%)	
Fellow eye	Healthy	23 (23.71%)			GA	5 (5.15%)	
	CNV	69 (71.14%)					

Description and summary statistics of the variables used for the analysis (Q1 = first quartile, Q3 = third quartile)

Table 2 Analysis of the AMD data

	GA			CNV		
	$\hat{\gamma}$	se($\hat{\gamma}$)	<i>p</i> -value	$\hat{\gamma}$	se($\hat{\gamma}$)	<i>p</i> -value
Intercept (t = 1)	- 3.8228	5.0247	0.4468	1.8512	4.8802	0.7044
Intercept (t = 2)	- 3.3412	5.0468	0.5079	3.6576	4.8472	0.4505
Intercept (t = 3)	- 3.8007	5.1116	0.4572	3.0804	4.8682	0.5269
Intercept (t = 4)	- 2.6469	5.0855	0.6027	3.4437	4.8784	0.4802
Intercept (t = 5)	- 2.8738	5.0559	0.5698	4.3565	4.8307	0.3671
Age (years)	0.0257	0.0483	0.5939	- 0.0117	0.0433	0.7869
Visual acuity (snellen)	- 0.0629	0.0462	0.1729	- 0.0749	0.0325	0.0212
Drusen volume (mm ³)	6.7520	2.3100	0.0035	1.4389	2.1350	0.5003
Phakia	1.0522	0.8500	0.2158	- 0.0836	0.5438	0.8778
Smoking	0.1120	0.5941	0.8505	- 0.5985	0.4961	0.2277
Ref_drusen	1.5514	0.7482	0.0381	1.0211	0.6588	0.1212
Fellow eye (CNV)	0.3518	0.7524	0.6401	0.0874	0.5471	0.8731
Fellow eye (GA)	2.3432	0.9986	0.0190	- 13.3469	1043.4122	0.9898

The table contains the coefficient estimates ($\hat{\gamma}$), the estimated standard errors [se($\hat{\gamma}$)] and the *p*-values (based on Wald test statistics) that were obtained from fitting a parametric cause-specific hazard model with logistic response function

in Table 2). Moreover, the additional incorporation of possibly relevant interactions became numerically infeasible. These results suggested the use of a regularized estimation approach that enforces variable selection but still enables an easy interpretation of effects and accounts for possible non-linear effects as well as interactions between the explanatory variables.

To compare the approaches introduced in the previous sections, we generated 100 subsamples without replacement of size $n = 65$ (i.e. 2/3 of the original sample) each and computed the predictive log-likelihood from the remaining 100 test data sets of

Fig. 6 Analysis of the AMD data. The boxplots show the predictive log-likelihood values of the tree-based models (GI and HD with ll-based pruning) and the penalized likelihood approach (CATS Lasso). The models were evaluated on 100 test sets of $n' = 32$ each. The median value of the best-performing approach is marked by a dashed line

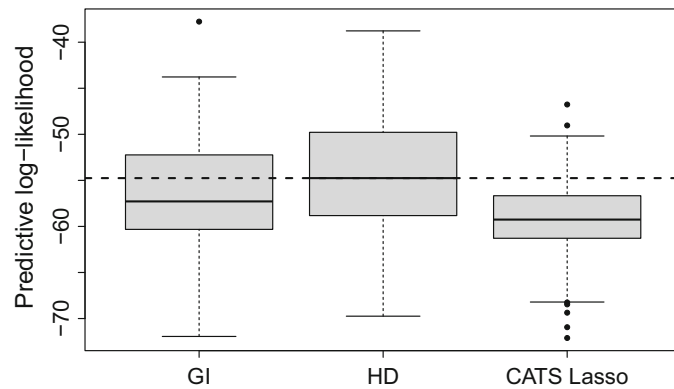
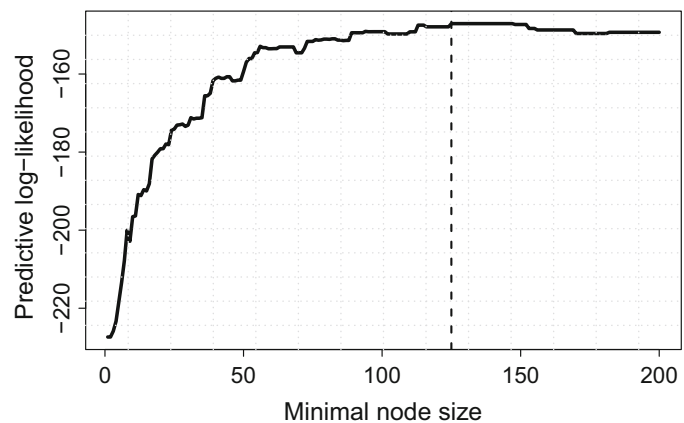


Fig. 7 Analysis of the AMD data. The figure shows the predictive log-likelihood values that were obtained by five-fold cross-validation when fitting the tree-based model with ll-based pruning and splitting by HD to the entire data. The maximal value (obtained for the minimal node size 125) is marked by the vertical dashed line



$n' = 32$ observations each. Figure 6 shows the results obtained from fitting (i) the proposed tree-based model with ll-based pruning and splitting by either Gini impurity (GI) or Hellinger distance (HD), and (ii) a parametric model based on the penalized likelihood with CATS Lasso penalty. It is seen that the tree with splitting by HD was the best-performing method. Furthermore, both tree-based models (on average) yielded higher values than CATS Lasso. This superiority was already present in the simulations in Sect. 4 and suggested that there might be interaction effects or non-linear relations that were not accounted for by the simple linear predictors.

Finally, we performed an analysis on the entire data by applying the tree-based model with ll-based pruning and splitting by HD. The values of the predictive log-likelihood obtained by five-fold cross validation are shown in Fig. 7. If an increase of the minimal node size did not change the number of splits and therefore did not influence the resulting tree, the value of the log-likelihood accordingly remained the same. The optimal minimal node size (with the maximal log-likelihood) was found to be 125 (referred to the augmented data matrix) resulting in a tree with five terminal nodes, see Fig. 8. By definition, the number of observations in each of the terminal nodes of the tree is below 125 and hence no further split is performed.

As seen from Fig. 8, the most important risk factor, which was chosen in the first split during tree building, was the presence of refractile drusen. Patients with refractile drusen had a particularly high risk for the development of GA. The corresponding estimated hazard ($\hat{\lambda}_{GA} = 0.145$) pictured in the rightmost node in Fig. 8 represented the highest value among all the estimated hazards. Within the group of patients without refractile drusen the risk for the development of CNV was very high in patients older

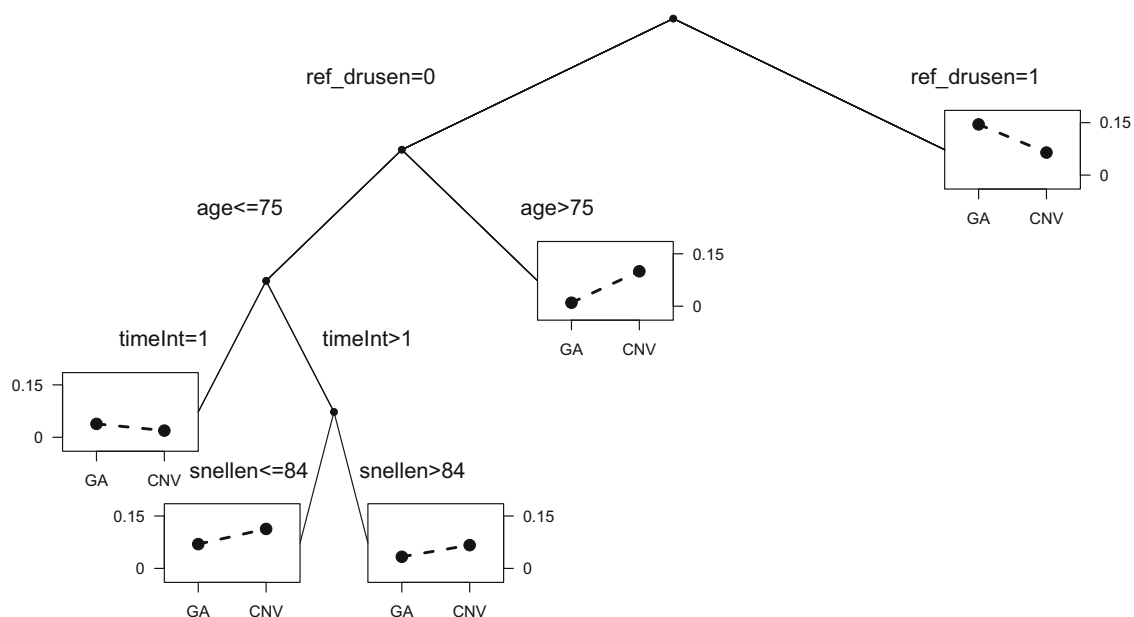


Fig. 8 Analysis of the AMD data. The figure visualizes the tree that was obtained from fitting the tree-based model with ll-based pruning and splitting by HD to the entire data. At each terminal node the estimated hazards $\hat{\lambda}_{GA}$ and $\hat{\lambda}_{CNV}$ (post-processed by application of the Laplace correction) are depicted in the small subfigure

than 75 ($\hat{\lambda}_{CNV} = 0.10$). For the subgroup of younger patients, the risk for the development of a GA or CNV was high in patients with a low or moderate visual acuity (snellen ≤ 84) after the first year of the study (timeInt > 1), yielding $\hat{\lambda}_{GA} = 0.07$ and $\hat{\lambda}_{CNV} = 0.11$ in the respective terminal node of the tree.

The risk factors drusen volume, phakia, smoking and disease status of the fellow eye were not selected for splitting during tree building. Thus they were effectively excluded from the model. In contrast, drusen volume and a GA in the fellow eye yielded significant effects on the development of a GA in the simple parametric model (see Table 2), where the higher-order interactions captured by the tree (e.g. between the presence of refractile drusen and age) were not taken into account.

6 Concluding remarks

This article proposes a tree-based method for the modeling of discrete competing risks data. For models in continuous time, similar approaches based on the Cox proportional hazards model have been proposed before by Ibrahim et al. (2008) and Xu et al. (2016). The attractive feature of the method proposed here is that it is directly based on the likelihood of a multinomial logit model. Therefore, the algorithm can be implemented using established tools for multi-categorical response models.

The results show that our method performs well in both simulations and the analysis of real-world data. In particular, it outperformed parametric models in the presence of interactions between the explanatory variables, and in higher dimensional situations

with a huge number of noisy variables. Both, Gini impurity and Hellinger distance, turned out to be attractive splitting criteria for tree building. The analysis of the MODIAMD study suggests that our method yields plausible results and is an appropriate modeling strategy for competing risks outcomes in clinical and epidemiological studies.

As described in Sect. 3.1, it is possible to insert time-dependent values of the explanatory variables into the augmented data matrices (10) and (11), enabling the tree-based method to additionally deal with time-varying information. Splitting in a time-varying explanatory variable then implies that not all the observations belonging to one individual must necessarily be allocated to the same node of the fitted tree. This strategy is in line with an alternative tree-based approach for time-varying explanatory variables in the single-event case, proposed by Bou-Hamad et al. (2011).

A disadvantage of tree-based methods is that the resulting trees are often affected by a large variance. This means that even a small variation in the data may result in different trees. Therefore, when the focus is on prediction accuracy, it might be worth stabilizing the results obtained from single trees by applying ensemble methods, such as bagging or random forests. Methods of this kind have been suggested by Ishwaran et al. (2014) for the continuous-time case and investigated by Janitza and Tutz (2015) in the discrete-time case.

Further extensions of parametric discrete competing risks models include, for example, the use of regression splines to model non-linear effects (Luo et al. 2016) and the incorporation of heterogeneity components in hierarchical settings (Meggiolaro et al. 2017). Finally, an attractive modeling approach to enforce variable selection in high-dimensional settings might also be boosting techniques, as examined by Binder et al. (2009) and Tapak et al. (2015) for competing risks data in continuous time.

Acknowledgements Support by the German Research Foundation (DFG), Grant SCHM 2966/1-2 and SCHM 2966/2-1, is gratefully acknowledged. The MODIAMD study is funded by the German Ministry of Education and Research (BMBF), Funding Number 13N10349.

Appendix: Further simulation results

See Figs. 9 and 10.

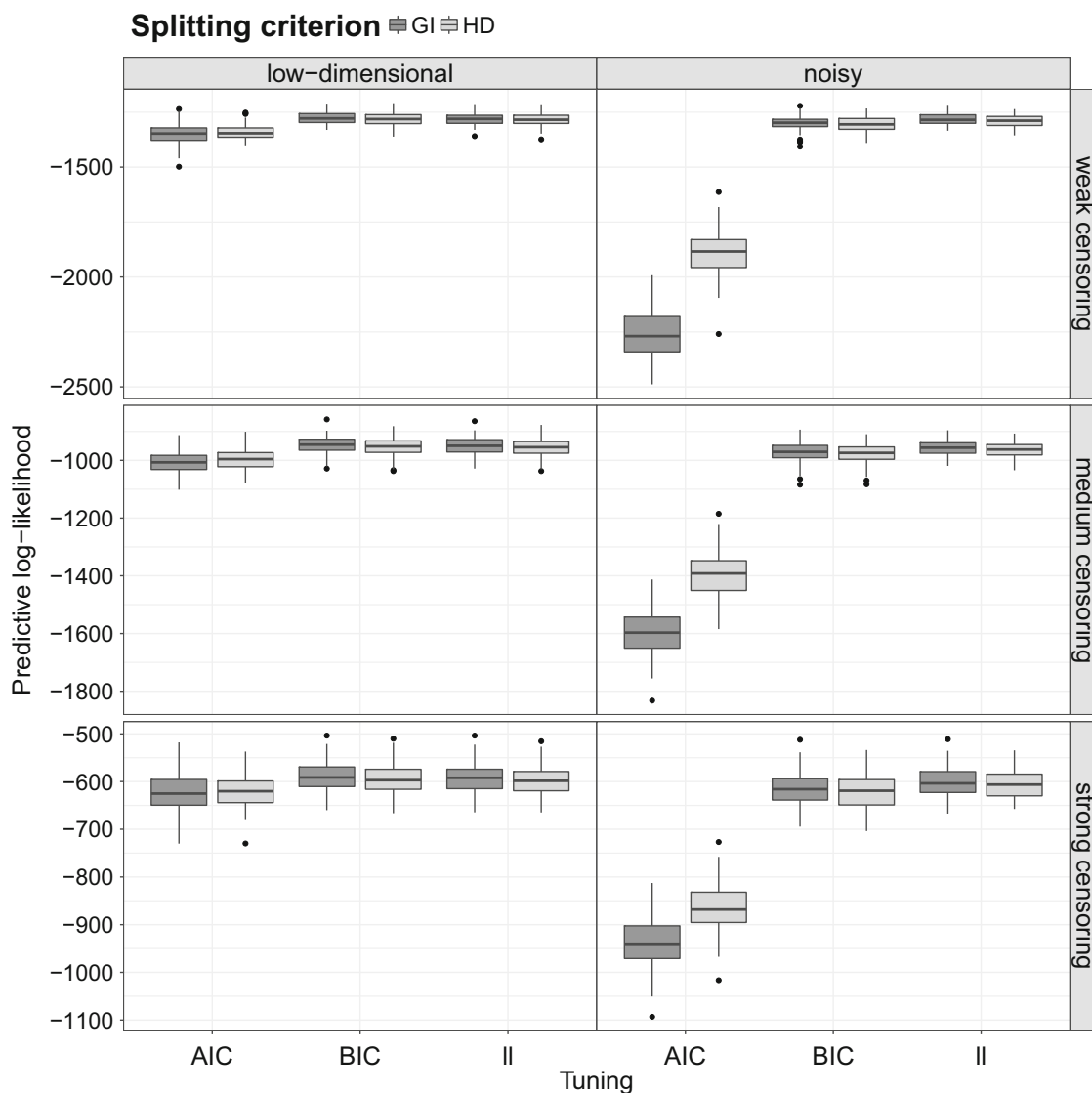


Fig. 9 Results of the simulation study. The boxplots visualize the predictive log-likelihood values obtained from the various tree-based approaches for the six scenarios with $n = 500$. Dark gray boxplots refer to the results with splitting by Gini impurity (GI), light gray boxplots refer to the results with splitting by Hellinger distance (HD). High values of the predictive log-likelihood correspond to good model fits, and vice versa

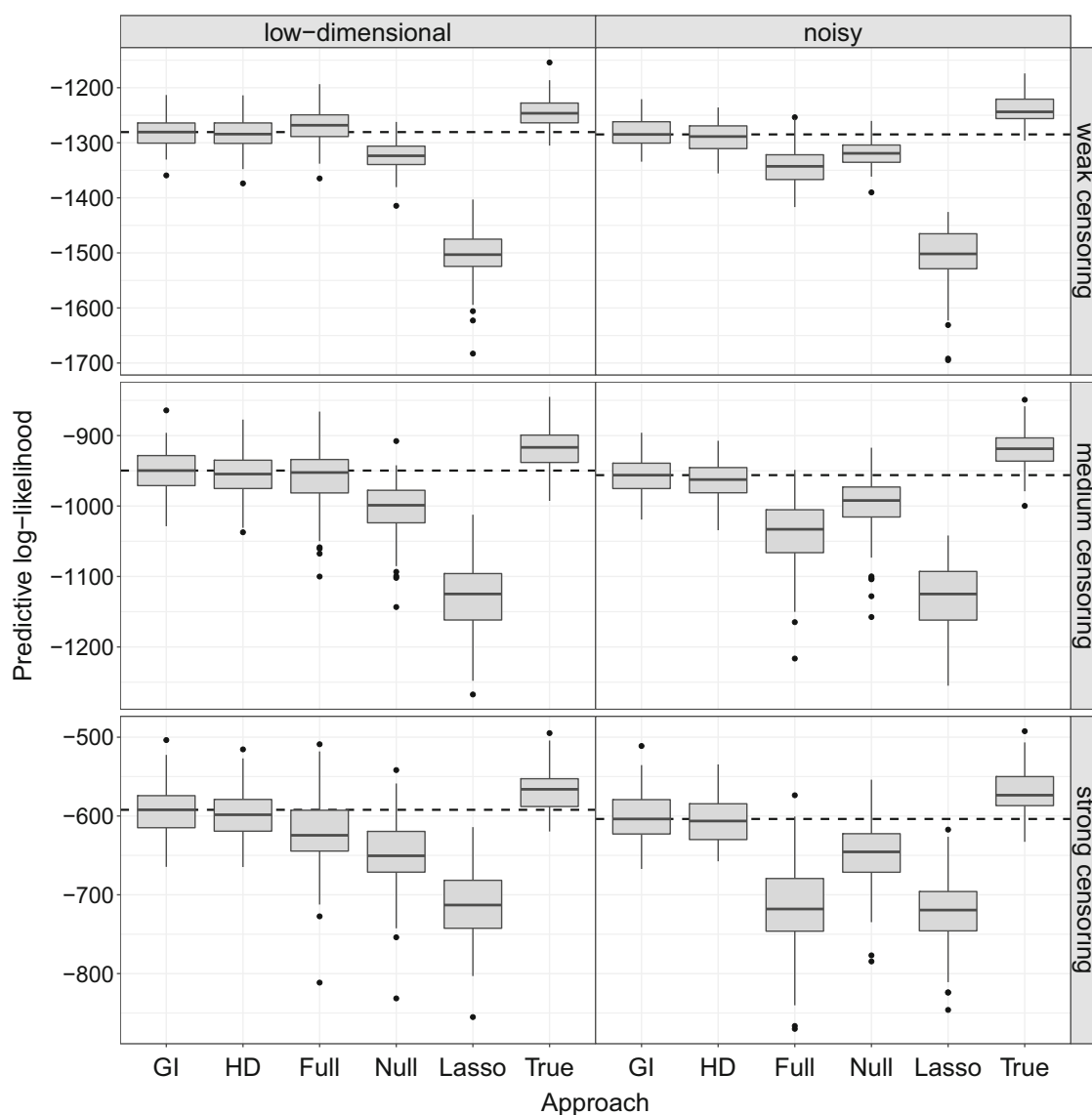


Fig. 10 Results of the simulation study. The boxplots visualize the predictive log-likelihood values obtained from various modeling approaches for the six scenarios with $n = 500$. The first two boxplots (GI and HD) obtained from the tree-based models refer to the results with tuning by the predictive log-likelihood (II), respectively. The sixth boxplot in each of the six panels contains the true log-likelihood values of the 100 test data sets (*True*), based on the true hazards defined in (18). Dashed lines refer to the median values of the best-performing tree-based model

References

- Austin PC, Lee DS, Fine JP (2016) Introduction to the analysis of survival data in the presence of competing risks. *Circulation* 133:601–609
- Berger M, Schmid M (2018) Semiparametric regression for discrete time-to-event data. *Stat Model* 18:1–24
- Beyersmann J, Allignol A, Schumacher M (2011) *Competing risks and multistate models with R*. Springer, New York
- Binder H, Allignol A, Schumacher M, Beyersmann J (2009) Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics* 25:890–896
- Bou-Hamad I, Larocque D, Ben-Ameur H, Mâsse LC, Vitaro F, Tremblay RE (2009) Discrete-time survival trees. *Can J Stat* 37:17–32
- Bou-Hamad I, Larocque D, Ben-Ameur H (2011) Discrete-time survival trees and forests with time-varying covariates: application to bankruptcy data. *Stat Model* 11:429–446

- Breiman L (1996) Technical note: some properties of splitting criteria. *Mach Learn* 24:41–47
- Breiman L, Friedman JH, Olshen RA, Stone JC (1984) *Classification and regression trees*. Wadsworth, Monterey
- Cieslak DA, Chawla NV (2008) Learning decision trees for unbalanced data. In: Daelemans W, Goethals B, Morik K (eds) *Machine learning and knowledge discovery in databases*. Springer, Berlin, pp 241–256
- Cieslak DA, Hoens TR, Chawla NV, Kegelmeyer WP (2012) Hellinger distance decision trees are robust and skew-insensitive. *Data Min Knowl Discov* 24:136–158
- Cox DR (1972) Regression models and life-tables (with discussion). *J R Stat Soc Series B* 34:187–220
- Doove LL, Dusseldorp E, Deun KV, Mechelen IV (2014) A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions. *Adv Data Anal Classif* 8:403–425
- Ferri C, Flach PA, Hernández-Orallo J (2003) Improving the AUC of probabilistic estimation trees. In: Lavrač N, Blockeel DGH, Todorovski L (eds) *European conference on machine learning*. Springer, Berlin, pp 121–132
- Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning*, 2nd edn. Springer, New York
- Hoens TR, Qian Q, Chawla NV, Zhou ZH (2012) Building decision trees for the multi-class imbalance problem. In: Tan P, Chawla S, Ho C, Bailey J (eds) *Advances in knowledge discovery and data mining*. Springer, Berlin, pp 122–134
- Ibrahim NA, Kudus A, Daud I, Bakar MRA (2008) Decision tree for competing risks survival probability in breast cancer study. *Int J Biol Med Sci* 3:25–29
- Ishwaran H, Gerds TA, Kogalur UB, Moore RD, Gange SJ, Lau BM (2014) Random survival forests for competing risks. *Biostatistics* 15:757–773
- Janitza S, Tutz G (2015) Prediction models for time discrete competing risks. Ludwig-Maximilians-Universität München, Department of Statistics Technical Report, p 177
- Lau B, Cole SR, Gange SJ (2009) Competing risk regression models for epidemiologic data. *Am J Epidemiol* 170:244–256
- Luo S, Kong X, Nie T (2016) Spline based survival model for credit risk modeling. *Eur J Oper Res* 253:869–879
- Meggiolaro S, Giraldo A, Clerici R (2017) A multilevel competing risks model for analysis of university students' careers in Italy. *Stud High Educ* 42:1259–1274
- Mingers J (1989) An empirical comparison of pruning methods for decision tree induction. *Mach Learn* 4:227–243
- Möst S, Pöbnecker W, Tutz G (2016) Variable selection for discrete competing risks models. *Qual Quant* 50:1589–1610
- Pöbnecker W (2014) MRSP: multinomial response models with structured penalties. R package version 0.4.3. <http://CRAN.R-project.org/package=MRSP>
- Prentice RL, Kalbfleisch JD, Peterson AV Jr, Flournoy N, Farewell VT, Breslow NE (1978) The analysis of failure times in the presence of competing risks. *Biometrics* 34:541–554
- Provost F, Domingos P (2003) Tree induction for probability-based ranking. *Mach Learn* 52:199–215
- Putter H, Fiocco M, Geskus RB (2007) Tutorial in biostatistics: competing risks and multi-state models. *Stat Med* 26:2389–2430
- Quinlan JR (1986) Induction of decision trees. *Mach Learn* 1:81–106
- Ripley BD (1996) *Pattern recognition and neural networks*. University Press, Cambridge
- Schmid M, Küchenhoff H, Hörauf A, Tutz G (2016) A survival tree method for the analysis of discrete event times in clinical and epidemiological studies. *Stat Med* 35:734–751
- Schmid M, Tutz G, Welchowski T (2018) Discrimination measures for discrete time-to-event predictions. *Econ Stat* 7:153–164
- Steinberg JS, Göbel AP, Thiele S, Fleckenstein M, Holz FG, Schmitz-Valckenberg S (2016) Development of intraretinal cystoid lesions in eyes with intermediate age-related macular degeneration. *Retina* 36:1548–1556
- Tapak L, Saidijam M, Sadeghifar M, Poorolajal J, Mahjub H (2015) Competing risks data analysis with high-dimensional covariates: an application in bladder cancer. *Genomics Proteomics Bioinformatics* 13:169–176
- Tutz G (1995) Competing risks models in discrete time with nominal or ordinal categories of response. *Qual Quant* 29:405–420
- Tutz G (2012) *Regression for categorical data*. University Press, Cambridge
- Tutz G, Schmid M (2016) *Modeling discrete time-to-event data*. Springer, New York

- Tutz G, Pöbnecker W, Uhlmann L (2015) Variable selection in general multinomial logit models. *Comput Stat Data Anal* 82:207–222
- Vallejos CA, Steel MFJ (2017) Bayesian survival modelling of university outcomes. *J R Stat Soc Series A Stat Soc* 180:613–631
- Welchowski T, Schmid M (2017) discSurv: discrete time survival analysis. R package version 1.1.7. <http://CRAN.R-project.org/package=discSurv>
- Xu W, Che J, Kong Q (2016) Recursive partitioning method on competing risk outcomes. *Cancer Inform* 15:CIN–S39364
- Yee TW (2010) The VGAM package for categorical data analysis. *J Stat Softw* 32:1–34
- Yee TW (2017) VGAM: vector generalized linear and additive models. R package version 1.0-4. <https://CRAN.R-project.org/package=VGAM>
- Zahid FM, Tutz G (2013) Multinomial logit models with implicit variable selection. *Adv Data Anal Classif* 7:393–416

2.4 Berger, Schmid et al., Biostatistics 21, 449-466

Liegen Ereigniszeitdaten mit konkurrierenden Ereignissen vor, ist oftmals nur das Eintreten eines der Ereignisse von besonderem Interesse. In diesem Fall ist es sinnvoll, nicht die ereignis-spezifischen Hazardfunktionen, wie in Kapitel 2.3 betrachtet, sondern direkt die kumulative Inzidenzfunktion (8) für dieses Ereignis zu modellieren. Diese Vorgehensweise wird bei Vorliegen von konkurrierenden Ereignissen empfohlen, wann immer der Schwerpunkt auf der Schätzung der Inzidenz oder auf der Bestimmung von Vorhersagen liegt (Austin et al., 2016).

In diesem Kapitel wird eine neuartige Methode zur Modellierung der kumulativen Inzidenzfunktion für diskrete Ereigniszeiten vorgeschlagen. Diese fußt auf dem Subdistribution Hazard-Modell, das von Fine und Gray (1999) für stetige Ereigniszeitdaten eingeführt wurde und das die Proportionalitätsannahme des Cox-Modells (Cox, 1972) voraussetzt. Analog zu Fine und Gray (1999) wird o.B.d.A. die diskrete Subdistribution Hazardfunktion für ein Ereignis vom Typ $\varepsilon = 1$ zum Zeitpunkt $t \in \{1, \dots, k\}$ definiert durch

$$\lambda_1(t|X) = P(T = t, \varepsilon = 1 | (T \geq t) \cup (T \leq t - 1, \varepsilon \neq 1), X). \quad (21)$$

Es wird gezeigt, dass sich eine direkte Verknüpfung der Funktion $\lambda_1(t|X)$ zur kumulativen Inzidenzfunktion $F_1(t|X)$ ergibt. Wie in Kapitel 2.1 und 2.2 wird zur Modellierung der diskreten Subdistribution Hazardfunktion ein parametrisches Regressionsmodell der Form

$$\lambda_1(t|X) = h(\gamma_{0t} + X_1\gamma_1 + \dots + X_p\gamma_p), \quad t = 1, \dots, k - 1, \quad (22)$$

vorgeschlagen, wobei in den Simulationen und der praktischen Anwendung die inverse komplementäre log-log-Funktion $h(\cdot) = 1 - \exp(-\exp(\cdot))$ herangezogen wird. Die Regressionskoeffizienten γ_{0t} , $t = 1, \dots, k - 1$, und $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)^\top$ können über gewichtete Maximum-Likelihood-Schätzung berechnet werden und in R mithilfe des Zusatzpaketes **mgcv** (Wood, 2021) konsistent geschätzt werden.

Umfangreiche Simulationen veranschaulichen die Unverzerrtheit der gewichteten Schätzung und den Mehrwert der neuen Methode gegenüber dem stetigen Modell nach Fine und Gray (1999), insbesondere wenn die Anzahl an diskreten Zeitpunkten klein ist. Zur Illustration werden die Daten der Studie zum Auftreten von nosokomialen Lungenentzündungen aus Wolkewitz et al. (2008) neu analysiert.

Subdistribution Hazard Models for Competing Risks in Discrete Time

MORITZ BERGER^{1,†}, MATTHIAS SCHMID^{1,†,*}, THOMAS WELCHOWSKI¹,
STEFFEN SCHMITZ-VALCKENBERG², JAN BEYERSMANN³

¹*Department of Medical Biometry, Informatics and Epidemiology, University of Bonn*

²*University Eye Hospital Bonn, Sigmund-Freud-Strasse 25, D-53127 Germany*

³*Institute of Statistics, Ulm University, Helmholtzstrasse 20, D-89081 Ulm*

matthias.schmid@imbie.uni-bonn.de

SUMMARY

A popular modeling approach for competing risks analysis in longitudinal studies is the proportional subdistribution hazards model by Fine & Gray (1999). This model is widely used for the analysis of continuous event times in clinical and epidemiological studies. However, it does not apply when event times are measured on a discrete time scale, which is a likely scenario when events occur between pairs of consecutive points in time (e.g., between two follow-up visits of an epidemiological study) and when the exact lengths of the continuous time spans are not known. To adapt the Fine & Gray approach to this situation, we propose a technique for modeling subdistribution hazards in discrete time. Our method, which results in consistent and asymptotically normal estimators of the model parameters, is based on a weighted maximum likelihood estimation scheme for binary regression. We illustrate the modeling approach by an analysis of nosocomial pneumonia in patients treated in hospitals.

*To whom correspondence should be addressed. †Contributed equally to this work.

Key words: Competing risks; Discrete time-to-event data; Regression modeling; Subdistribution hazard; Survival analysis.

1. INTRODUCTION

The purpose of time-to-event analysis is to model the time span T until the occurrence of an event of interest. In many studies there is not only one single type of event but $J > 1$ possible events, called *competing events*. A multitude of examples are found in clinical and epidemiological studies (e.g., Lau *and others* 2009; Austin *and others* 2016), where competing events such as cause-specific death, the progression of a disease, or the occurrence of an infection are often strongly related and therefore need to be analyzed together. In these situations, the use of suitable techniques for competing risks analysis is of increasing importance (Andersen *and others*, 2012). The objective is often to build a regression model that links the occurrence of the event(s) of interest to a set of predictors $\mathbf{x} = (x_1, \dots, x_p)^\top$. For introductions to competing risks analysis, see in particular Andersen and Keiding (2002) and Putter *and others* (2007). Recent extensions, among many others, have been suggested by Bartlett and Taylor (2016) and Cederkvist *and others* (2018).

Commonly used approaches for competing risks analysis are: (i) To focus on one specific event $j \in \{1, \dots, J\}$ and to consider individuals experiencing a competing event as random drop-outs. It has been shown in several studies that this incomplete approach yields biased estimates of the cumulative event probabilities, cf. Wolbers *and others* (2009). (ii) The modeling of the cause-specific hazard function $\xi_j(t) = \lim_{\Delta t \rightarrow 0} \{P(t < T \leq t + \Delta t, \epsilon = j | T > t, \mathbf{x}) / \Delta t\}$, where T is the time to the first event and $\epsilon \in \{1, \dots, J\}$ is a random variable indicating the type of event at T (Prentice *and others*, 1978). In this approach each type of event is analyzed separately, and all individuals experiencing a competing event may technically be treated as censored observations in the modeling of ξ_j . While fitting a separate cause-specific hazard model for each j is technically simple, the derivation and interpretation of cumulative event probabilities in these models is involved. This is because the cumulative incidence function, defined as $F_j(t) = P(T \leq t, \epsilon = j | \mathbf{x})$,

depends on a combination of all cause-specific hazard functions ξ_j , $j = 1, \dots, J$. Hence it is necessary to fit separate cause-specific hazard models for all J events even if the interest is solely in the modeling of F_j for one specific event $j \in \{1, \dots, J\}$. Also, interpretation of the cumulative incidence function may become difficult in this case, as all cause-specific hazards need to be considered together (Beyersmann *and others*, 2011). (iii) The modeling of the cumulative incidence function of one specific event of interest. Direct modeling approaches for $F_j(t)$ that account for the effects of competing events were proposed by Fine and Gray (1999) and Klein and Andersen (2005). The method by Klein and Andersen (2005), which uses pseudo values from a jackknife statistic derived from the cumulative incidence function, will not be dealt with here. Instead, we focus on the approach by Fine and Gray (1999), which is based on the modeling of a *subdistribution hazard* function for the event of interest. The subdistribution hazard, which is directly linked to $F_j(t)$, is defined in terms of the probability of experiencing j at time t , given that either no event has occurred yet or that a competing event occurred prior to t . In contrast to cause-specific hazard modeling, this approach has the advantage that only one model needs to be considered for interpretation; on the other hand, it does not provide insight in the characteristics of the cause-specific hazard functions (being the driving forces of competing risks data).

The subdistribution hazard model by Fine and Gray (1999) is based on the proportional hazards specification by Cox (1972). In particular, it is assumed that survival times are given by random variables measured on a *continuous* scale. In this setting, the approach by Fine and Gray (1999) has become hugely popular, and it has been recommended that analysts use subdistribution hazard models whenever “the focus is on estimating incidence or predicting prognosis in the presence of competing risks” (Austin *and others*, 2016). A remaining issue, however, is that in some clinical and epidemiological studies the exact (continuous) event times are not recorded. Instead, it may only be known that the events occurred between pairs of consecutive points in time (i.e. within pre-specified follow-up intervals). In these cases, time is measured on a discrete

scale $t = 1, 2, \dots$, where each t refers to a specific time interval. The subdistribution hazard model in continuous time may not apply to these situations. An example, which will also be considered in this paper, is the duration to the development of nosocomial pneumonia (NP) in intensive care patients (measured on a daily basis, Wolkewitz *and others* 2008). In addition, there are also situations where event times are ‘intrinsically’ discrete, for example, time to pregnancy, which is usually measured by the number of menstrual cycles (Scheike and Keiding, 2006).

In the literature, there exists a variety of models for discrete time-to-event data, see e.g. Tutz and Schmid (2016). A common approach for discrete competing risks analysis is to model the cause-specific *discrete hazard* function $P(T = t, \epsilon = j | T \geq t, \mathbf{x})$, $t = 1, 2, \dots$, by use of a regression model for multi-categorical response (Tutz, 1995). Prominent examples are the multinomial logistic regression model and the proportional odds model (Tutz and Schmid 2016, Chapter 8). Again, a drawback of cause-specific discrete hazard modeling is the lack of a simple and interpretable model for the discrete cumulative incidence function $F_j(t) = P(T \leq t, \epsilon = j | \mathbf{x})$, $t = 1, 2, \dots$

To address these issues, we propose a novel approach for the direct modeling of discrete cumulative incidence functions with right-censored data. In accordance with the method by Fine and Gray (1999), our model is specified in terms of a discrete subdistribution hazard function for the event of interest. For the parameters of the model we will derive consistent and asymptotically normal estimators which are based on inverse probability (IP) weighting (van der Laan and Robins, 2003) and on a weighted maximum likelihood estimation scheme for binary regression. Of note, the consistency proof for the weighted estimators can be embedded in the framework of unbiased estimation equations (Carroll and Ruppert, 1988) and does not require use of counting process theory. In the absence of competing events, the discrete subdistribution hazard model reduces to the standard discrete hazard model presented in Tutz and Schmid (2016).

The rest of the paper is organized as follows: In Section 2.1 we will introduce notations and definitions. The basic class of subdistribution hazard models will be specified in Section 2.2, and

the weighted log-likelihood function for these models will be derived in Section 2.3. In addition, we will show in Section 2.3 that the weighted maximum likelihood estimator is consistent and asymptotically normal. Section 3 presents the results of several simulation studies. Specifically, we will compare our estimation approach to situations where all censoring times are assumed to be known and a standard unweighted estimator is available. Section 4 contains the application dealing with the analysis of NP infection. Section 5 summarizes the main findings of the paper.

2. METHODS

2.1 The Discrete Subdistribution Hazard Function

Let T_i be the event time and C_i the censoring time of individual i , $i = 1, \dots, n$. Both T_i and C_i are assumed to be independent random variables taking discrete values in $\{1, 2, \dots, k\}$, where k is a natural number. In situations where originally continuous data have been grouped, the discrete event times $1, \dots, k$ refer to time intervals $[0, a_1), [a_1, a_2), \dots, [a_{k-1}, \infty)$, where $T_i = t$ means that the event has occurred in time interval $[a_{t-1}, a_t)$ with $a_k = \infty$. For right-censored data, the time period during which an individual is under observation is denoted by $\tilde{T}_i = \min(T_i, C_i)$, i.e. \tilde{T}_i corresponds to the true event time if $T_i \leq C_i$ and to the censoring time otherwise. The random variable $\Delta_i := I(T_i \leq C_i)$ indicates whether \tilde{T}_i is right-censored ($\Delta_i = 0$) or not ($\Delta_i = 1$).

It is assumed that there are J competing events and that the event type of the i -th individual at T_i is denoted by $\epsilon_i \in \{1, \dots, J\}$. Throughout this paper, the focus is on modeling the occurrence of a type 1 event ($\epsilon_i = 1$), taking into account that there are $J - 1$ competing events and also the censoring event ($\Delta_i = 0$).

For given values of a set of time-constant predictor variables $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$, the aim is to estimate the cumulative incidence function for a type 1 event, given by

$$F_1(t|\mathbf{x}_i) = P(T_i \leq t, \epsilon_i = 1|\mathbf{x}_i). \quad (2.1)$$

By definition, F_1 is bounded between 0 and $F_1(k|\mathbf{x}_i) = P(\epsilon_i = 1|\mathbf{x}_i) < 1$. The subscript “1” in

Equation (2.1) indicates that F_1 refers to a type 1 event.

In accordance with Fine and Gray (1999), we propose to link the cumulative incidence function F_1 to a subdistribution time ϑ_i measuring the time to the occurrence of the type 1 event. In the presence of competing risks, ϑ_i needs to account for the possible occurrence of an event of type 2, \dots , J . The basic assumption made by Fine and Gray (1999) is that the type 1 event will never be the first event to be observed once a competing event has occurred, implying that there is no finite event time for the occurrence of a type 1 event if $\epsilon_i \neq 1$. Accordingly, the discrete subdistribution time for a type 1 event is defined by

$$\vartheta_i := \begin{cases} T_i, & \text{if } \epsilon_i = 1 \\ \infty, & \text{if } \epsilon_i \neq 1. \end{cases} \quad (2.2)$$

Analogous to Fine and Gray (1999), we define the *discrete subdistribution hazard function* by

$$\lambda_1(t|\mathbf{x}_i) = P(T_i = t, \epsilon_i = 1 | (T_i \geq t) \cup (T_i \leq t-1, \epsilon_i \neq 1), \mathbf{x}_i) = P(\vartheta_i = t | \vartheta_i \geq t, \mathbf{x}_i), \quad (2.3)$$

$t = 1, \dots, k$, which is the discrete hazard function of the ‘‘event time’’ ϑ_i defined in Equation (2.2).

The subdistribution hazard λ_1 is linked to the subdistribution function F_1 by

$$F_1(t|\mathbf{x}_i) = 1 - \prod_{s=1}^t (1 - \lambda_1(s|\mathbf{x}_i)) = 1 - S_1(t|\mathbf{x}_i), \quad (2.4)$$

where $S_1(t|\mathbf{x}_i) = P(\vartheta_i > t|\mathbf{x}_i)$ is the discrete survival function for a type 1 event. Equation (2.4) implies that a statistical model for the discrete subdistribution hazard has a direct interpretation in terms of the cumulative incidence function. The focus of the next sections will be on parametric regression models for the subdistribution hazard λ_1 .

2.2 Parametric Regression Models for the Discrete Subdistribution Hazard

To model the discrete subdistribution hazard for a type 1 event, we consider the class of parametric regression models

$$\lambda_1(t|\mathbf{x}_i) = h(\gamma_{0t} + \mathbf{x}_i^\top \boldsymbol{\gamma}), \quad (2.5)$$

where $h(\cdot)$ is a strictly monotone increasing distribution function. The linear predictor $\eta_{it} = \gamma_{0t} + \mathbf{x}_i^\top \boldsymbol{\gamma}$ contains the real-valued time-dependent intercepts γ_{0t} , $t = 1, \dots, k-1$ (referred to as *baseline coefficients*), and a vector of regression coefficients $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)^\top$ independent of t . Technically, the baseline coefficients can be treated as an additional factor variable in Model (2.5) (see below). When the number of event times is large relative to the sample size, it may also be useful to represent the baseline coefficients by a smooth (possibly nonlinear) function $f_0(t)$ of unspecified form. For example, estimation of $f_0(t)$ may be carried out using P-splines or smoothing splines (for details see Tutz and Schmid 2016, Chapter 5).

In classical discrete hazard modeling without competing events ($J = 1$), which is often based on models of the form (2.5), the most popular distribution functions are the logistic function $h(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$ defining the *proportional continuation ratio model* and the inverse complementary log-log function $h(\cdot) = 1 - \exp(-\exp(\cdot))$ defining the *Gompertz model*. An important property of the Gompertz model (also called *complementary log-log model*), is its equivalence to the Cox proportional hazards model in continuous time. As shown, for example, in Tutz and Schmid (2016), the Gompertz model holds when continuous time-to-event data satisfying the proportional hazards assumption have been grouped.

When the aim is to model discrete competing risks data, it is generally possible to use the same distribution functions in the subdistribution hazard model (2.5) as those used in discrete hazard modeling with only one event. For example, in the simulation study and the application (Sections 3 and 4), we will use the inverse complementary log-log function, which makes the parameters $\boldsymbol{\gamma}$ be the same as those of the continuous-time Fine & Gray model, provided that continuous-time data satisfying the proportional subdistribution hazards assumption have been grouped.

2.3 Estimation

To estimate the baseline and regression coefficients of Model (2.5), we propose a maximum likelihood (ML) estimation scheme that is based on discrete inverse probability (IP) weighting (Fine

and Gray, 1999; van der Laan and Robins, 2003). Our method is rooted in classical discrete hazard modeling without competing events (Tutz and Schmid, 2016), which will be summarized briefly in Section 2.3.1. The main theoretical results of the paper, including the consistency and the asymptotic normality of the weighted ML estimator, will be presented in Section 2.3.2.

2.3.1 Estimation without Competing Events In the simplified situation where each individual experiences either a type 1 event or a censoring event, the discrete subdistribution hazard reduces to the *discrete hazard function* $\lambda_1(t|\mathbf{x}_i) = P(T_i = t | T_i \geq t, \mathbf{x}_i)$. The subdistribution time equals $\nu_i = T_i$ in this case, and the likelihood per individual is given by

$$L_i = \lambda_1(\tilde{T}_i|\mathbf{x}_i)^{\Delta_i} (1 - \lambda_1(\tilde{T}_i|\mathbf{x}_i))^{1-\Delta_i} \prod_{t=1}^{\tilde{T}_i-1} (1 - \lambda_1(t|\mathbf{x}_i)), \quad (2.6)$$

see Tutz and Schmid (2016) for details. Estimates of the parameters of Model (2.5) are obtained by specifying $\lambda_1(t|\mathbf{x}_i) = h(\gamma_{0t} + \mathbf{x}_i^\top \boldsymbol{\gamma})$ and by maximizing the log-likelihood

$$\begin{aligned} l(\gamma_{01}, \dots, \gamma_{0,k-1}, \boldsymbol{\gamma}^\top) &= \sum_{i=1}^n \log(L_i(\gamma_{01}, \dots, \gamma_{0,k-1}, \boldsymbol{\gamma}^\top)) \\ &= \sum_{i=1}^n \log \left[h(\gamma_{0\tilde{T}_i} + \mathbf{x}_i^\top \boldsymbol{\gamma})^{\Delta_i} (1 - h(\gamma_{0\tilde{T}_i} + \mathbf{x}_i^\top \boldsymbol{\gamma}))^{1-\Delta_i} \prod_{t=1}^{\tilde{T}_i-1} (1 - h(\gamma_{0t} + \mathbf{x}_i^\top \boldsymbol{\gamma})) \right] \end{aligned} \quad (2.7)$$

over $\boldsymbol{\gamma}_0 := (\gamma_{01}, \dots, \gamma_{0,k-1})^\top$ and $\boldsymbol{\gamma}$.

Estimation of the model parameters by Equation (2.7) is greatly simplified by the fact that L_i is equivalent to the likelihood of a binary response model with values $y_{it} \in \{0, 1\}$, $t = 1, \dots, k-1$. The latter values indicate whether individual i experienced a type 1 event at time t ($y_{it} = 1$) or not ($y_{it} = 0$). Furthermore, y_{it} is only defined if $t \leq \tilde{T}_i$, i.e., as long as individual i is *at risk*. Accordingly, one obtains

$$L_i = \prod_{t=1}^{\tilde{T}_i} \lambda_1(t|\mathbf{x}_i)^{y_{it}} (1 - \lambda_1(t|\mathbf{x}_i))^{1-y_{it}}, \quad (2.8)$$

where $(y_{i1}, \dots, y_{i\tilde{T}_i}) = (0, \dots, 0, 1)$ if $\Delta_i = 1$ and $(y_{i1}, \dots, y_{i\tilde{T}_i}) = (0, \dots, 0)$ if $\Delta_i = 0$. The binomial likelihood in (2.8) implies that standard software for binary regression models can be

used to fit the discrete hazard model in (2.5). For details, see Berger and Schmid (2018).

An alternative representation of L_i , which will become useful when modeling the subdistribution hazard in Section 2.3.2, is given by

$$L_i = \prod_{t=1}^{k-1} \{ \lambda_1(t|\mathbf{x}_i)^{y_{it}} (1 - \lambda_1(t|\mathbf{x}_i))^{1-y_{it}} \}^{w_{it}}, \quad (2.9)$$

where $w_{it} = \mathbf{I}(t \leq \tilde{T}_i)$, $i = 1, \dots, n$, $t = 1, \dots, k-1$, is a set of weights indicating whether individual i is at risk at time t or not. The corresponding binary values y_{it} used in Equation (2.9) are defined by

$$(y_{i1}, \dots, y_{i, \tilde{T}_i}, \dots, y_{i, k-1}) = \begin{cases} (0, \dots, 0, 1, 0, \dots, 0), & \text{if } \Delta_i = 1, \\ (0, \dots, 0, 0, 0, \dots, 0), & \text{if } \Delta_i = 0. \end{cases} \quad (2.10)$$

With this representation, the total log-likelihood becomes

$$\ell = \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} \{ y_{it} \log(\lambda_1(t|\mathbf{x}_i)) + (1 - y_{it}) \log(1 - \lambda_1(t|\mathbf{x}_i)) \}. \quad (2.11)$$

2.3.2 Modeling the Subdistribution Hazard in the Presence of Competing Events In the presence of competing events of type $2, \dots, J$ (with λ_1 now denoting the subdistribution hazard for a type 1 event), the log-likelihood function (2.11) can be specified in a similar manner as in Section 2.3.1. To ensure consistency of the estimators of γ_{0t} and γ , the weights w_{it} need to be redefined appropriately. For this, we consider the *risk set* $r(t)$ at time t , defined as the set of individuals that neither experienced a type 1 event nor a censoring event prior to time t . If the i -th individual is known to be contained in this set, the idea is to define $w_{it} = 1$ as before. Conversely, we set $w_{it} = 0$ if individual i is known not to be a member of $r(t)$. A remaining problem is that $r(t)$ is not fully known if there are individuals that experienced a competing event before t . In fact, since $\vartheta_i = \infty$ for these individuals, they continue to be at risk beyond \tilde{T}_i until they eventually experience the censoring event. Consequently, as the censoring times C_i are unobserved if $C_i > \tilde{T}_i$, it cannot be determined whether an individual with $\epsilon_i > 1$ is still part of the risk set at $t > \tilde{T}_i$. In accordance with the continuous-time approach by Fine and Gray (1999), we therefore propose

to estimate the probability of each individual being part of the risk set $r(t)$, and to set the weights w_{it} in the log-likelihood function (2.11) equal to the estimated probabilities.

More specifically, we propose to define the weights w_{it} as follows:

- (i) For uncensored individuals that experience a type 1 event ($\Delta_i \epsilon_i = 1$), we propose to define $w_{it} := \mathbb{I}(t \leq \tilde{T}_i)$. This definition implies that

$$(w_{i1}, w_{i2}, \dots, w_{i\tilde{T}_i}, w_{i,(\tilde{T}_i+1)}, \dots, w_{i,(k-1)}) = (1, 1, \dots, 1, 0, \dots, 0), \quad (2.12)$$

accounting for the fact that the individuals cease to be at risk after their respective type 1 events have been observed.

- (ii) For individuals that experience the censoring event first ($\Delta_i \epsilon_i = 0$), we also propose to define $w_{it} := \mathbb{I}(t \leq \tilde{T}_i)$. This definition accounts for the fact that the individuals cease to be at risk after \tilde{T}_i .

- (iii) For uncensored individuals that experience a competing event first ($\Delta_i \epsilon_i > 1$), we propose to define $w_{it} := 1$ if $t \leq \tilde{T}_i$, accounting for the fact that the individuals are known to be at risk at least until \tilde{T}_i . For $t > \tilde{T}_i$ we estimate the probability of being a member of $r(t)$ by

$$w_{it} := \frac{\hat{G}(t-1)}{\hat{G}(\tilde{T}_i-1)}, \quad \tilde{T}_i < t \leq k-1, \quad (2.13)$$

where $\hat{G}(t)$ is an estimate of the censoring survival function $G(t) = P(C_i > t)$.

By combining (i) to (iii), the weights w_{it} can be expressed in the closed form

$$\begin{aligned} w_{it} &= \mathbb{I}(C_i \geq \min(T_i, t)) \cdot \frac{\hat{G}(t-1)}{\hat{G}(\min(T_i, C_i, t)-1)} \cdot \left(\mathbb{I}(t \leq T_i) + \mathbb{I}(T_i \leq t-1, \epsilon_i \neq 1) \right) \\ &= \frac{\hat{G}(t-1)}{\hat{G}(\min(\tilde{T}_i, t)-1)} \cdot \left(\mathbb{I}(t \leq \tilde{T}_i) + \mathbb{I}(\tilde{T}_i \leq t-1, \Delta_i \epsilon_i > 1) \right), \end{aligned} \quad (2.14)$$

which is analogous to the estimated risk sets given in Beyersmann *and others* (2011) for continuous time-to-event data. Just like in (2.10), the binary values y_{it} are defined by

$$(y_{i1}, \dots, y_{i,\tilde{T}_i}, \dots, y_{i,k-1}) = \begin{cases} (0, \dots, 0, 1, 0, \dots, 0), & \text{if } \Delta_i \epsilon_i = 1, \\ (0, \dots, 0, 0, 0, \dots, 0), & \text{if } \Delta_i \epsilon_i \neq 1. \end{cases} \quad (2.15)$$

Under these assumptions and definitions, it is possible to prove the main result of the paper:

Theorem 1. The solution to the optimization problem

$$\begin{aligned} \operatorname{argmax}_{\gamma_0, \gamma} \ell(\gamma_0, \gamma) &= \\ &= \operatorname{argmax}_{\gamma_0, \gamma} \left\{ \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} \{ y_{it} \log(h(\gamma_{0t} + \mathbf{x}_i^\top \gamma)) + (1 - y_{it}) \log(1 - h(\gamma_{0t} + \mathbf{x}_i^\top \gamma)) \} \right\} \end{aligned} \quad (2.16)$$

defines a consistent and asymptotically normal estimator ($n \rightarrow \infty$) of the parameters γ_0 and γ of the subdistribution hazard model (2.5).

Proof. The proof of Theorem 1, which is based on the theory of unbiased estimation equations (Carroll and Ruppert, 1988), is given in Appendix A of the supplementary material. The main step in the proof is to show that the solution to (2.16) solves an unbiased estimation equation conditional on the covariates using the true censoring survival function in the definition of the weights w_{it} . Stacking this estimation equation and another unbiased estimation equation for $G(t)$, the theory of unbiased estimation equations guarantees consistency and asymptotic normality when using the estimated censoring survival function in the weights. \square

Estimation Using Software for Binary Regression Theorem 1 implies that the parameters of the subdistribution hazard model (2.5) can be estimated consistently by fitting a weighted binary regression model with outcome values y_{it} and weights w_{it} . Similar to Geskus (2011) in the continuous-time case, we propose to use standard software for model fitting. For this it is necessary to set up an *augmented data matrix*, which is passed to the software routine for binary regression, and which is composed of a set of smaller (augmented) data matrices defined separately for each individual. More specifically, for uncensored individuals that experience a type 1 event ($\Delta_i \epsilon_i = 1$), the augmented data matrix and the vector of weights are defined by

$$\begin{array}{c}
\mathbf{y}_i \quad \mathbf{t} \quad \mathbf{X}_i \\
\left(\begin{array}{cccc}
0 & 1 & x_{i1} & \dots & x_{ip} \\
0 & 2 & x_{i1} & \dots & x_{ip} \\
\vdots & \vdots & \vdots & & \vdots \\
1 & \tilde{T}_i & x_{i1} & \dots & x_{ip} \\
0 & \tilde{T}_i + 1 & x_{i1} & \dots & x_{ip} \\
\vdots & \vdots & \vdots & & \vdots \\
0 & k-1 & x_{i1} & \dots & x_{ip}
\end{array} \right) \quad \text{and} \quad \left(\begin{array}{c}
\mathbf{w}_i \\
1 \\
1 \\
\vdots \\
1 \\
0 \\
\vdots \\
0
\end{array} \right), \text{ respectively,} \quad (2.17)
\end{array}$$

where \mathbf{t} is an additional predictor variable that refers to the baseline coefficients. For example, if the baseline coefficients are defined by a separate intercept γ_{0t} at each time point, \mathbf{t} corresponds to a factor variable that is converted to a set of $k-1$ dummy variables. If the baseline coefficients are modeled by a spline function, \mathbf{t} corresponds to a numeric variable that is converted to a set of variables representing the basis functions of the spline.

The augmented data matrix and the vector of weights for censored individuals ($\Delta_i \epsilon_i = 0$) are defined in the same way as in (2.17), except that $\mathbf{y}_i := (0, \dots, 0)^\top$ in these cases.

For uncensored individuals that experience a competing event first ($\Delta_i \epsilon_i > 1$), we define

$$\begin{array}{c}
\mathbf{y}_i \quad \mathbf{t} \quad \mathbf{X}_i \\
\left(\begin{array}{cccc}
0 & 1 & x_{i1} & \dots & x_{ip} \\
0 & 2 & x_{i1} & \dots & x_{ip} \\
\vdots & \vdots & \vdots & & \vdots \\
0 & \tilde{T}_i & x_{i1} & \dots & x_{ip} \\
0 & \tilde{T}_i + 1 & x_{i1} & \dots & x_{ip} \\
\vdots & \vdots & \vdots & & \vdots \\
0 & k-1 & x_{i1} & \dots & x_{ip}
\end{array} \right) \quad \text{and} \quad \left(\begin{array}{c}
\mathbf{w}_i \\
1 \\
1 \\
\vdots \\
1 \\
\frac{\hat{G}(\tilde{T}_i)}{\hat{G}(\tilde{T}_i-1)} \\
\vdots \\
\frac{\hat{G}(k-2)}{\hat{G}(\tilde{T}_i-1)}
\end{array} \right), \text{ respectively.} \quad (2.18)
\end{array}$$

As stated above, the full augmented data matrix is obtained by concatenating the individual augmented data matrices. The resulting matrix of dimension $(n \cdot (k-1)) \times (p+2)$ and the vector of weights of length $n \cdot (k-1)$ are subsequently passed to a software routine for binary regression in order to solve the optimization problem (2.16). In R, the augmented data matrix can be generated by applying the function `dataLongSubDist()` of the add-on package **discSurv**.

Parameter estimates can be obtained by using the functions `glm()` (for models with a separate baseline coefficient for each t) or `gam()` (contained in the R package `mgcv`, for models with a smooth baseline function).

Remark. The rows with $w_{it} = 0$, which have been included in (2.17) for notational convenience, are, in fact, not included in the output of the `dataLongSubDist()` function. This is because these rows are not needed for fitting the binary regression model.

3. SIMULATION STUDY

In this section we present the results of numerical experiments to investigate the performance of the proposed modeling approach. The aims of the study were (i) to analyze both the accuracy of the weighted ML estimates and the run-time of the method, (ii) to investigate the performance of IP weighting by comparing the proposed modeling approach to an analysis with known censoring times, and (iii) to compare the weighted ML estimates to the respective estimates obtained from application of the continuous-time approach by Fine and Gray (1999).

3.1 Experimental Design

In order to generate data from a given subdistribution hazard model for type 1 events, we used a scheme adopted from Fine and Gray (1999). This procedure is also described in Beyersmann *and others* (2011), where it was termed “indirect simulation”. In all simulation scenarios we considered data with two competing events, $\epsilon_i \in \{1, 2\}$, that was generated under the model specification of proportional subdistribution hazards. More specifically, our discrete subdistribution hazard model was based on the discretization of the continuous model

$$F_1(t|\mathbf{x}_i) = P(T_{cont,i} \leq t, \epsilon_i = 1|\mathbf{x}_i) = 1 - (1 - q + q \exp(-t))^{\exp(\mathbf{x}_i^\top \boldsymbol{\gamma}_1)}, \quad (3.19)$$

where $T_{cont,i} \in \mathbb{R}^+$ denotes the continuous time span of individual i and $\boldsymbol{\gamma}_1 = (\gamma_{11}, \dots, \gamma_{1p})^\top$ is a set of regression coefficients. The parameter $q \in (0, 1)$ affected the probability of a type 1

event which, according to (3.19), was given by $\pi_{i1} := P(\epsilon_i = 1|\mathbf{x}_i) = 1 - (1 - q)^{\exp(\mathbf{x}_i^\top \boldsymbol{\gamma}_1)}$. By definition, high values of q resulted in high probabilities of π_{i1} , and vice versa. The probability of a competing event was given by $\pi_{i2} := P(\epsilon_i = 2|\mathbf{x}_i) = 1 - \pi_{i1} = (1 - q)^{\exp(\mathbf{x}_i^\top \boldsymbol{\gamma}_1)}$.

Continuous time spans for type 2 events were drawn from the exponential model

$$T_{cont,i}|\epsilon_i = 2, \mathbf{x}_i \sim \text{Exp}(\xi_2 = \exp(\mathbf{x}_i^\top \boldsymbol{\gamma}_2)),$$

where $\boldsymbol{\gamma}_2 = (\gamma_{21}, \dots, \gamma_{2p})^\top$ denotes a set of regression coefficients linking the rate parameter ξ_2 with the values of the predictor variables \mathbf{x} .

In order to obtain discrete event times, we generated data according to the indirect simulation scheme described above and grouped the resulting continuous event times into $k = 20$ categories. The latter were defined by the time intervals $[0, q_5), [q_5, q_{10}), \dots, [q_{95}, \infty)$, where q_a denotes the $a\%$ quantile of the continuous event times. The values of q_a were pre-estimated from an independent sample with 1,000,000 observations. Accordingly, the same interval boundaries were used in each simulation run. Censoring times were generated from a discrete distribution with probability density function $P(C_{disc,i} = t) = b^{k-t+1} / \sum_{s=1}^k b^s$, $t = 1, \dots, k$, where the percentage of censored observations was determined by the parameter $b \in \mathbb{R}^+$.

Similar to Fine and Gray (1999), we considered two standard normally distributed predictor variables $x_{i1}, x_{i2} \sim N(0, 1)$ and two binary predictor variables $x_{i3}, x_{i4} \sim \text{Binomial}(1, 0.5)$. All predictor variables were independent. The true regression coefficients were set to $\boldsymbol{\gamma}_1 = (0.4, -0.4, 0.2, -0.2)^\top$ and $\boldsymbol{\gamma}_2 = (-0.4, 0.4, -0.2, 0.2)^\top$. Three sample sizes ($n \in \{100, 300, 500\}$) were considered. In addition, we specified three different censoring rates, denoted by *weak*, *medium* and *strong*. The degree of censoring was determined by the parameter b of the censoring distribution. We used the values $b = 0.85$ (weak), $b = 1$ (medium) and $b = 1.25$ (strong), resulting in the censoring rates shown in Supplementary Figure 1. We also considered three different probabilities of a type 1 event, specifying $q \in \{0.2, 0.4, 0.8\}$. In total, this resulted in $3 \times 3 \times 3 = 27$ different scenarios. All scenarios were analyzed using 1000 independent replications. For estimation

we used the inverse complementary log-log distribution function, which defines the same values of γ_1 as the Fine & Gray proportional subdistribution hazards model in continuous time.

The following estimation approaches were considered: (i) weighted ML estimation based on (2.11), and (ii) unweighted ML estimation using the complete censoring information. For the latter model, we did not estimate the risk sets $r(t)$ via IP weighting but used the complete censoring times $C_{disc,i}$ (which are unknown in practice if $\Delta_i \epsilon_i > 1$) and fitted an ordinary discrete hazard model for type 1 events, as described in Section 2.3.1. In Fine and Gray (1999) this condition was termed “censoring-complete”. A comparison of the estimates obtained from (i) and (ii) served to analyze the performance of the weighted ML estimation approach. Estimates of G were obtained by a life table estimator for the censoring event (cf. Schmid *and others* 2018).

In addition to the estimation approaches (i) and (ii), we applied the continuous-time Fine & Gray method to the grouped data, using the same discrete time scale ($t = 1, \dots, k$) and various numbers of categories ($k = 4, 8, 16, 32, 64$). For each k , the discretization procedure was based on intervals defined by quantiles of the continuous event times, as described above. This part of the simulation study served to evaluate the differences between the Fine & Gray method and the discrete-time subdistribution hazard modeling approach, which were to be expected due to the grouping of the (originally continuous) event times. We also compared the two estimation approaches with respect to their run-times.

Supplementary Figure 1 shows the relative frequencies of observed events for the nine scenarios with $n = 500$ and $k = 20$. It is seen that the rates of observed type 1 events increased with increasing value of q and that censoring rates increased with increasing value of b . For constant q and varying b , the ratio of observed type 1 and type 2 events remained approximately the same. For $q = 0.2$ and $q = 0.4$ we observed more events of type 2 than of type 1, and for $q = 0.8$ there were more events of type 1 than of type 2. For the scenarios with $n = 100$ and $n = 300$ the observed relative frequencies were almost the same and are thus not shown.

3.2 Results

The coefficient estimates $\hat{\gamma}_1$ obtained from the weighted ML estimation approach with $n = 500$ and $k = 20$ are presented in Figure 1. The boxplots show that on average the estimated coefficients were very close to the true ones, regardless of the degree of censoring and the rate of type 1 events. Figure 1 also shows that the variance of the estimates increased with increasing degree of censoring, in particular for the two binary predictors x_3 and x_4 . In contrast, the rate of type 1 events (determined by the value of q) had only a small impact on the variance of the estimates. This shows that weighted ML estimation of the subdistribution hazard λ_1 also worked well in the presence of only a relatively small number of type 1 events (note that, for strong censoring and $q = 0.2$ only about 5% type 1 events were observed).

An overview of several performance indicators for the nine scenarios with $n = 500$ and $k = 20$ is given in Table 1. For each element of $\hat{\gamma}_1 = (\gamma_{11}, \dots, \gamma_{14})^\top$ we computed (i) the mean squared error (MSE) and (ii) the empirical variance from the 1000 samples, and (iii) the average of the 1000 variance estimates obtained from the Fisher scoring algorithm applied for optimizing the weighted log-likelihood. In case of the weighted ML estimation approach (left part of Table 1), the three measures took almost identical values. This confirms the results presented in Figure 1 in that weighted ML estimation yielded nearly unbiased estimates. Importantly, the estimators exhibited only a small finite-sample bias in the scenarios with strong censoring, and the estimated variances obtained from Fisher scoring were close to the empirical variances obtained from the 1000 samples. In addition, the censoring-complete estimates (right part of Table 1) were nearly identical to the weighted ML estimates. This demonstrates that the weighted ML estimation scheme worked well in all analyzed scenarios and that using the estimated survival function \hat{G} instead of the true function G had a negligible effect on the variance of the weighted ML estimator.

The results obtained for the scenarios with $n = 100$ and $n = 300$ are shown in Supplementary Figures 2 and 3. As expected, estimation accuracy deteriorated for $n = 100$, as in this case

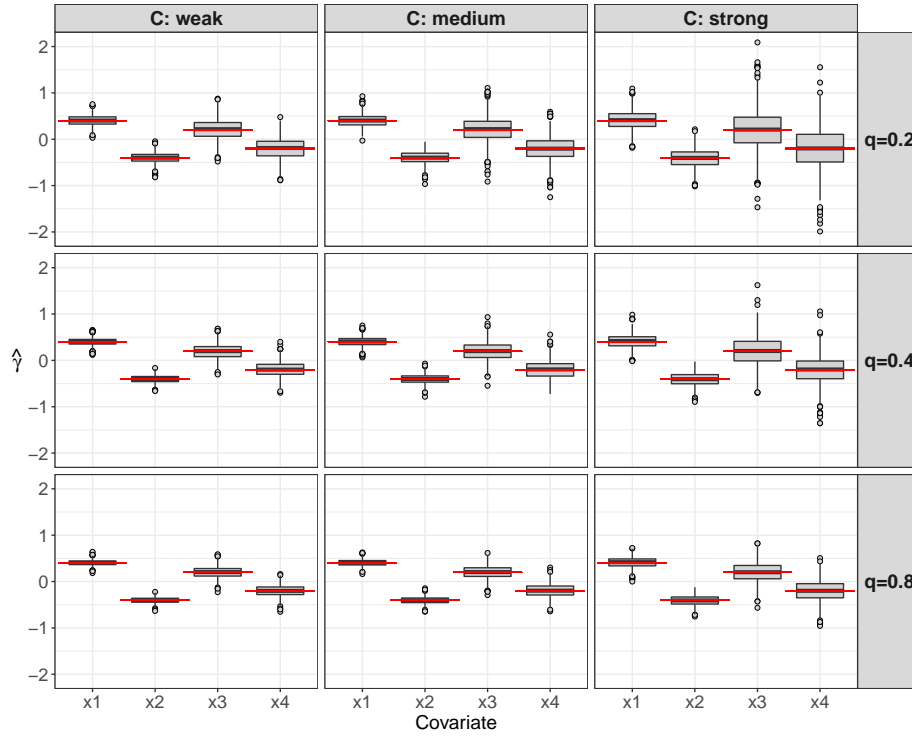


Fig. 1: Results of the simulation study. The boxplots visualize the estimates of the parameters $\gamma_1 = (0.4, -0.4, 0.2, -0.2)^\top$ that were obtained from fitting a discrete subdistribution hazard model using the proposed weighted ML estimation approach ($n = 500$). The horizontal lines refer to the true values of the parameters ($C =$ degree of censoring).

only very few type 1 events were observed. For example, in the most extreme scenario (strong censoring and $q = 0.2$, only five type 1 events on average), half of the estimates $\hat{\gamma}_{13}$ and $\hat{\gamma}_{14}$ did not take finite values. Not surprisingly, this result demonstrates that discrete subdistribution hazard modeling is not recommended when both event numbers and rates are small. On the other hand, when n was increased to 300, performance indicators were already very similar to those obtained in the scenario with $n = 500$ (cf. Figure 1).

The differences between the discrete-time subdistribution hazard model and the continuous-time Fine & Gray model are illustrated in Figure 2 and Supplementary Figures 4 to 7. It is seen

Table 1: Results of the simulation study. For each coefficient γ_{1j} , $j = 1, \dots, 4$, the table contains the following evaluation criteria: (i) $\text{MSE}(\hat{\gamma})$, as estimated by the mean squared error of $\hat{\gamma}_{1j}$ computed from the 1000 samples, (ii) $\text{var}(\hat{\gamma})$, as estimated by the empirical variance of $\hat{\gamma}_{1j}$ computed from the 1000 samples, and (iii) $\mathbb{E}(\widehat{\text{var}}(\gamma))$, which denotes the average of the 1000 variance estimates obtained from Fisher scoring. The left part contains the results obtained from the proposed weighted ML estimation approach, whereas the right part shows the respective results obtained from fitting a discrete hazard model under the censoring-complete condition.

C	q		weighted estimation				censoring-complete estimation			
			x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
weak	0.2	MSE($\hat{\gamma}$)	0.013	0.013	0.051	0.049	0.013	0.013	0.051	0.050
		var($\hat{\gamma}$)	0.013	0.013	0.051	0.050	0.013	0.013	0.051	0.050
		$\mathbb{E}(\widehat{\text{var}}(\gamma))$	0.012	0.012	0.049	0.049	0.012	0.012	0.049	0.049
	0.4	MSE($\hat{\gamma}$)	0.006	0.006	0.027	0.027	0.007	0.006	0.027	0.027
		var($\hat{\gamma}$)	0.006	0.006	0.027	0.027	0.007	0.006	0.027	0.027
		$\mathbb{E}(\widehat{\text{var}}(\gamma))$	0.007	0.007	0.026	0.026	0.007	0.007	0.026	0.026
	0.8	MSE($\hat{\gamma}$)	0.004	0.004	0.015	0.014	0.004	0.004	0.015	0.015
		var($\hat{\gamma}$)	0.004	0.004	0.015	0.014	0.004	0.004	0.015	0.015
		$\mathbb{E}(\widehat{\text{var}}(\gamma))$	0.004	0.004	0.014	0.014	0.004	0.004	0.014	0.014
medium	0.2	MSE($\hat{\gamma}$)	0.019	0.018	0.079	0.069	0.019	0.018	0.079	0.069
		var($\hat{\gamma}$)	0.019	0.018	0.079	0.069	0.019	0.018	0.079	0.069
		$\mathbb{E}(\widehat{\text{var}}(\gamma))$	0.018	0.019	0.073	0.073	0.018	0.019	0.074	0.073
	0.4	MSE($\hat{\gamma}$)	0.010	0.009	0.041	0.040	0.010	0.010	0.042	0.041
		var($\hat{\gamma}$)	0.010	0.009	0.041	0.040	0.010	0.010	0.042	0.041
		$\mathbb{E}(\widehat{\text{var}}(\gamma))$	0.010	0.010	0.038	0.038	0.010	0.010	0.038	0.038
	0.8	MSE($\hat{\gamma}$)	0.005	0.006	0.019	0.021	0.005	0.006	0.020	0.021
		var($\hat{\gamma}$)	0.005	0.006	0.019	0.021	0.005	0.006	0.020	0.021
		$\mathbb{E}(\widehat{\text{var}}(\gamma))$	0.006	0.006	0.021	0.021	0.006	0.006	0.021	0.021
strong	0.2	MSE($\hat{\gamma}$)	0.046	0.042	0.197	0.191	0.047	0.043	0.201	0.195
		var($\hat{\gamma}$)	0.045	0.042	0.197	0.191	0.047	0.043	0.202	0.195
		$\mathbb{E}(\widehat{\text{var}}(\gamma))$	0.044	0.044	0.182	0.181	0.045	0.045	0.182	0.182
	0.4	MSE($\hat{\gamma}$)	0.022	0.022	0.099	0.093	0.023	0.022	0.101	0.094
		var($\hat{\gamma}$)	0.022	0.022	0.099	0.093	0.023	0.022	0.101	0.094
		$\mathbb{E}(\widehat{\text{var}}(\gamma))$	0.023	0.023	0.091	0.091	0.023	0.023	0.091	0.091
	0.8	MSE($\hat{\gamma}$)	0.013	0.013	0.046	0.053	0.013	0.013	0.047	0.054
		var($\hat{\gamma}$)	0.013	0.013	0.046	0.053	0.013	0.013	0.047	0.054
		$\mathbb{E}(\widehat{\text{var}}(\gamma))$	0.013	0.013	0.048	0.048	0.013	0.013	0.049	0.049

that the Fine & Gray estimates showed a downward bias in absolute value when the number of intervals was small ($k \leq 16$). As expected, the bias became smaller when the number of intervals became larger and the number of ties decreased. The bias shown in Figure 2 may be attributed to the Breslow method for handling ties in Cox regression, which is the default (and only available) tie handling method in the R package **cmprsk** that was used for fitting the Fine & Gray models

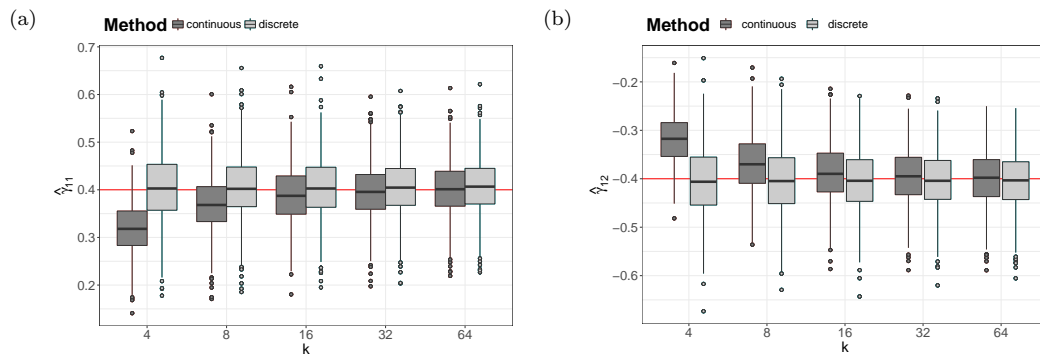


Fig. 2: Results of the simulation study. The boxplots visualize the estimates of the coefficients $\gamma_{11} = 0.4$ (panel a) and $\gamma_{12} = -0.4$ (panel b), as obtained from the discrete-time subdistribution hazard model and the continuous-time Fine & Gray model ($n = 500$, $q = 0.8$, $b = 0.85$). The horizontal lines refer to the true values of γ_{11} and γ_{12} . Interval numbers on the x-axes are presented on the \log_2 scale.

presented here. In Cox regression without competing events, the Breslow method has previously been shown to cause a bias in absolute value, similar to the bias observed in our simulation study (Hertz-Picciotto and Rockhill, 1997). Other tie handling methods will cause different types of bias (see, e.g., Supplementary Figures 8 and 9). Furthermore, the bias was also seen in the estimates of the cumulative incidence function F_1 (see Supplementary Figure 10).

Run-times of the discrete-time subdistribution hazard method and the continuous-time Fine & Gray method are illustrated in Supplementary Figure 11. It is seen that increasing the value of k resulted in a notable increase in the run-time of the discrete method. Although the differences between the two methods might be affected by different implementations in the R functions `crr()` (package `cmprsk`), `dataLongSubDist()`, and `glm()`, increases in the run-time of the discrete method are mainly attributed to the sizes of the augmented data matrices, which scale linearly with k . While the average run-time for a scenario with $q = 0.8$, $b = 0.85$, $n = 500$, and $k = 64$ intervals (19,800 data lines on average) was acceptable (~ 2.5 seconds), Supplementary Figure 11 clearly indicates the storage limitations of the discrete-time subdistribution hazard modeling approach when k becomes “too large”. It should be noted, however, that these storage issues are

inherent to many discrete hazard modeling approaches, regardless of the presence of competing events. Also, it appears difficult to derive a rule of thumb for the maximum feasible number of k , as this number does not only depend on computational resources, but also on the sample size, the rate of type 1 events, and the degree of censoring.

4. APPLICATION: NOSOCOMIAL PNEUMONIA INFECTION IN INTENSIVE CARE UNITS

To illustrate the application of the discrete subdistribution hazard model, we analyzed a data set on the development of pneumonia, which is a common nosocomial, i.e. hospital-acquired infection in intensive care units (ICUs). The data set was analyzed before in several publications (e.g., Beyersmann *and others* 2006; Wolkewitz *and others* 2008). As nosocomial pneumonia (NP) has a strong impact on the mortality of patients in ICUs, it is of high interest to determine the risk factors for the development of the disease.

The data were collected for a prospective cohort study at five ICUs in one university hospital, lasting 18 months from February 2000 to July 2001. We considered $n = 1,876$ patients with a duration of ICU stay of at least two days. The outcome of interest was the time to NP infection. Other possible events that were competing with the onset of NP (being the event of interest) were *death* and *discharge from hospital alive*. Due to the study design, the observed event times were discrete, as they were measured on a daily basis. We analyzed the data over a period of 60 days, resulting in 61 possible event times $t = 1, 2, \dots, 61$, where $t = k = 61$ refers to all individuals with event times ≥ 61 days. The observed event times are visualized in Supplementary Figure 12. At the observed times, each patient either acquired the NP infection ($n = 158$), was released from hospital (alive) or died ($n = 1,695$), or was administratively censored ($n = 23$).

In our analysis we investigated the impact of several baseline risk factors for NP acquisition by accounting for the competing event *death or discharge alive*. Descriptive summary statistics of the baseline risk factors considered in our analysis were presented in Table 1 of Wolkewitz

and others (2008). In addition to the age of the patients (centered at 60 years), the gender of the patients, and the simplified acute physiology score (SAPS) II, there were eleven binary risk factors characterizing the patients and their hospital stay. The binary variables either referred to the time of ICU admission (*on admission*) or the time prior to ICU admission (*before admission*).

Table 2 shows the estimates of the coefficients γ obtained from weighted ML fitting of a discrete subdistribution hazard model with complementary log-log link. We fitted a model with discrete baseline coefficients (*model 1*) and another model with a smooth baseline function represented by cubic P-splines with a second-order difference penalty (*model 2*). Model 2 was fitted using the R package **mgcv**. The proportional subdistribution hazards assumption was checked by fitting covariate-free subdistribution hazard models in various subgroups of the data and by comparing the resulting estimated cumulative incidence functions to the respective cumulative incidence functions estimated from the complementary log-log model (see Supplementary Figure 13, which does not indicate any major issues with the proportional subdistribution hazards assumption). For suggestions on how to analyze goodness-of-fit in the continuous-time case, see, in particular, Scheike and Zhang (2008) and Zhou and others (2013). The estimated baseline coefficients for the subdistribution hazard of NP acquisition are shown in Supplementary Figure 14.

Significant risk factors (on the subdistribution hazard scale and judged by the subdistribution hazard ratio) for the acquisition of NP at the 5% type I error level were (i) male gender, (ii) an intubation on admission, (iii) pneumonia on admission, (iv) another infection on admission, (v) an elective or emergency surgery before admission, and (vi) a cardiac/pulmonary or neurological underlying disease. According to the results in Table 2, there were only minor differences between models 1 and 2 regarding the magnitude of the coefficient estimates. Female gender reduced the subdistribution hazard of an NP acquisition ($\hat{\gamma} = -0.3432$, $\text{se}(\hat{\gamma}) = 0.1734$, model 1). Higher subdistribution hazards were, for instance, obtained for patients with an intubation on admission ($\hat{\gamma} = 0.6546$, $\text{se}(\hat{\gamma}) = 0.2432$) and for patients that underwent an elective surgery before admission

Table 2: Analysis of the NP infection data. The table contains the coefficient estimates ($\hat{\gamma}$), the estimated standard errors ($se(\hat{\gamma})$) and the p -values (based on Wald test statistics) obtained for models 1 and 2 (bc = baseline coefficients).

	model 1 (discrete bc)			model 2 (smooth bc)		
	$\hat{\gamma}$	$se(\hat{\gamma})$	p -value	$\hat{\gamma}$	$se(\hat{\gamma})$	p -value
Age (centered around 60)	0.0059	0.0052	0.2627	0.0059	0.0052	0.2608
SAPS II	0.0101	0.0061	0.0991	0.0100	0.0061	0.1008
Gender (female)	-0.3432	0.1734	0.0478	-0.3434	0.1734	0.0477
Intubation on admission	0.6546	0.2432	0.0071	0.6564	0.2431	0.0069
Pneumonia on admission	-3.6568	1.0099	0.0003	-3.6592	1.0099	0.0003
Urinary tract infection on admission	0.4628	0.5287	0.3814	0.4631	0.5291	0.3814
Other infections on admission	0.6191	0.2640	0.0190	0.6201	0.2638	0.0187
Hospitalization before admission	-0.3196	0.1769	0.0708	-0.3199	0.1769	0.0707
Elective surgery before admission	1.4017	0.1940	<0.0001	1.4030	0.1940	<0.0001
Emergency surgery before admission	0.5514	0.1827	0.0025	0.5516	0.1827	0.0025
Cardial/pulmonary underlying disease	0.6274	0.1957	0.0013	0.6304	0.1957	0.0013
Neurological underlying disease	0.4843	0.2087	0.0203	0.4867	0.2088	0.0197
Metabolic/renal underlying disease	-0.1198	0.3717	0.7471	-0.1204	0.3715	0.7458

($\hat{\gamma} = 1.4017$, $se(\hat{\gamma}) = 0.1940$). The estimated cumulative incidence function $F_1(t|\mathbf{x}_i)$ referring to the covariate profile of a randomly selected patient is displayed in Figure 3 (male gender, SAPS II score = 19, hospitalization before admission, absence of other risk factors). For this profile, there was only a very low probability of acquiring NP (< 1.5%). This probability would have increased to more than 5% after 30 days if the patient would have undergone elective surgery before admission, and to more than 2.5% if there would have been an intubation on admission.

Comparing the results of Table 2 with those of Wolkewitz *and others* (2008) who used time-continuous Cox models for all cause-specific hazards, *qualitative* identification of risk factors is similar. An interesting example are other infections on admission, which Table 2 identifies as significantly increasing the subdistribution hazard of nosocomial pneumonia incidence. Interestingly, Wolkewitz *and others* report a cause-specific hazard ratio of 1.08 (95% confidence interval [0.59, 1.98]) for pneumonia and a cause-specific hazard ratio of 0.72 [0.59, 0.89] for discharge. The interpretation is that patients with other infections on admission have a significantly reduced immediate ‘risk’ of discharge, while there appears to be no effect on the pneumonia hazard. Hence, patients stay in the intensive care unit longer while exposed to an essentially unchanged

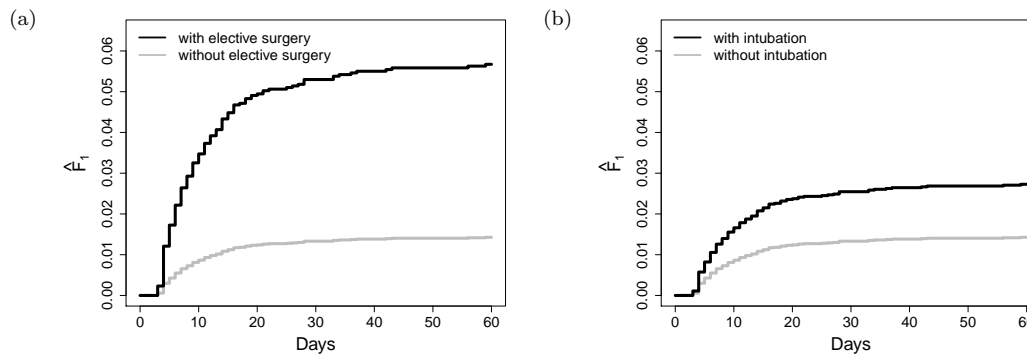


Fig. 3: Analysis of the NP infection data (model 1). The figure shows the estimated cumulative incidence functions for NP acquisition referring to the covariate profile of a randomly selected study participant (60 years of age, SAPS II score = 19, hospitalization before admission, absence of any of the other risk factors). Panel (a) refers to a situation where the risk factor *elective surgery before admission* would have been present in this patient, whereas panel (b) refers to a situation where the risk factor *intubation on admission* would have been present.

immediate risk of pneumonia. Over time, this leads to an increased pneumonia incidence. This is not an uncommon phenomenon in hospital epidemiology (Beyersmann *and others*, 2014), and our result in Table 2 directly identifies this increased incidence.

There is one covariate, intubation on admission, that Table 2 associates with an increased pneumonia incidence, but Wolkewitz *and others* (2008) do not. The reason is that Wolkewitz *and others* have also included time-dependent intubation status in their models, which was strongly associated with an increased pneumonia risk (cause-specific hazard ratio 5.90 [2.47, 14.09]). Intubation status is a highly time-dependent process in intensive care, possibly switched on and off multiple times. We have not included the time-dependent covariate information in our models, as the aim was regression for the cumulative incidence function (Cortese and Andersen, 2010).

Remark. In the cohort study there were rare instances where both an infection and death were recorded on the same day in hospital. In the original continuous-time analysis, Wolkewitz *and others* (2008) exploited the natural ordering of these events in that the infection was assumed to occur after half a day while death was assumed to have occurred by the end of the day. The discrete-time modeling approach presented here is also able to deal with the simultaneous occurrence of

competing events. In fact, since the method is designed to model the risk process for NP events only (implying that $w_{i\tilde{T}_i} = 1$ even if there has been a death event at $t = \tilde{T}_i$), no adjustments are necessary in this case. An alternative possibility, which could be used in the absence of a natural ordering, is to introduce an additional competing event modeling simultaneous occurrence.

5. DISCUSSION

We have suggested a discrete-time version of the subdistribution hazard model by Fine and Gray (1999). To fit the model, we proposed a weighted maximum likelihood estimation approach that can be implemented by using software for binary regression. An attractive feature of the approach is the relative ease with which the discrete-time version can be handled, see e.g. Geskus (2011); Mao and Lin (2017); Bellach *and others* (2018) for the challenges in the continuous-time case.

As shown in Section 2.3, the resulting weighted ML estimators are consistent and asymptotically normal as $n \rightarrow \infty$. Our method performed well in simulations, in particular when compared to the simpler situation of censoring-complete data.

Similar to the continuous-time case, the subdistribution hazard models considered here circumvent the difficulties associated with cause-specific discrete hazards modeling. Our framework also allows for the modeling of non-proportional subdistribution hazards. In fact, there are numerous options for specifying the link function of the binary regression model in (2.5), each yielding different characteristics and properties of the resulting discrete subdistribution hazard model, see e.g. Tutz and Schmid (2016), Chapter 3, and also Gerds *and others* (2012) for details.

In comparison to the continuous-time subdistribution hazard model by Fine & Gray, the method proposed here is particularly appropriate in applications where event times have been grouped or rounded (as in the example on nosocomial pneumonia infection). This situation corresponds to interval-censored data with fixed boundaries. As shown in the simulation study, differences between the Fine & Gray model and our method are largest when k is small and the

number of ties is large. In contrast, the Fine & Gray model may be preferred when the underlying time scale is truly continuous in the sense that ties rarely occur and patients/experimental units are monitored in such a way that the exact event or censoring times are practically known.

6. SUPPLEMENTARY MATERIAL

Supplementary material is available online at <http://biostatistics.oxfordjournals.org>. It contains the proof of Theorem 1 and additional numerical results.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

REFERENCES

- ANDERSEN, P. K., GESKUS, R. B., DE WITTE, T. AND PUTTER, H. (2012). Competing risks in epidemiology: Possibilities and pitfalls. *International Journal of Epidemiology* **41**, 861–870.
- ANDERSEN, P. K. AND KEIDING, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research* **11**, 91–115.
- AUSTIN, P. C., LEE, D. S. AND FINE, J. P. (2016). Introduction to the analysis of survival data in the presence of competing risks. *Circulation* **133**, 601–609.
- BARTLETT, J. W. AND TAYLOR, J. M. G. (2016). Missing covariates in competing risks analysis. *Biostatistics* **17**, 751–763.
- BELLACH, A., KOSOROK, M. R., RÜSCHENDORF, L. AND FINE, J. P. (2018). Weighted NPMLE for the subdistribution of a competing risk. *Journal of the American Statistical Association*. doi:10.1080/01621459.2017.1401540.

- BERGER, M. AND SCHMID, M. (2018). Semiparametric regression for discrete time-to-event data. *Statistical Modelling* **18**, 322–345.
- BEYERSMANN, J., ALLIGNOL, A. AND SCHUMACHER, M. (2011). *Competing Risks and Multi-state Models with R*. New York: Springer.
- BEYERSMANN, J., GASTMEIER, P., GRUNDMANN, H., BÄRWOLFF, S., GEFFERS, C., BEHNKE, M., RÜDEN, H. AND SCHUMACHER, M. (2006). Use of multistate models to assess prolongation of intensive care unit stay due to nosocomial infection. *Infection Control & Hospital Epidemiology* **27**, 493–499.
- BEYERSMANN, J., GASTMEIER, P. AND SCHUMACHER, M. (2014). Incidence in ICU populations: how to measure and report it? *Intensive Care Medicine* **40**, 871–876.
- CARROLL, R. J. AND RUPPERT, D. (1988). *Transformation and Weighting in Regression*. New York: Chapman and Hall.
- CEDERKVIST, L., HOLST, K. K., ANDERSEN, K. K. AND SCHEIKE, T. H. (2018). Modeling the cumulative incidence function of multivariate competing risks data allowing for within-cluster dependence of risk and timing. *Biostatistics*. doi:10.1093/biostatistics/kxx072.
- CORTESE, G. AND ANDERSEN, P.K. (2010). Competing risks and time-dependent covariates. *Biometrical Journal* **52**, 138–158.
- COX, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.
- FINE, J. P. AND GRAY, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94**, 496–509.
- GERDS, T. A., SCHEIKE, T. H. AND ANDERSEN, P. K. (2012). Absolute risk regression for competing risks. *Statistics in Medicine* **31**, 3921–3930.

- GESKUS, R. (2011). Cause-specific cumulative incidence estimation and the Fine and Gray model under both left truncation and right censoring. *Biometrics* **67**, 39–49.
- HERTZ-PICCIOTTO, I. AND ROCKHILL, B. (1997). Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics* **53**, 1151–1156.
- KLEIN, J. P. AND ANDERSEN, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics* **61**, 223–229.
- LAU, B., COLE, S. R. AND GANGE, S. J. (2009). Competing risk regression models for epidemiologic data. *American Journal of Epidemiology* **170**, 244–256.
- MAO, L. AND LIN, D. Y. (2017). Efficient estimation of semiparametric transformation models for the cumulative incidence of competing risks. *Journal of the Royal Statistical Society, Series B* **79**, 573–587.
- PRENTICE, R. L., KALBFLEISCH, J. D., PETERSON JR, A. V., FLOURNOY, N., FAREWELL, V. T. AND BRESLOW, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34**, 541–554.
- PUTTER, H., FIOCCO, M. AND GESKUS, R. B. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine* **26**, 2389–2430.
- SCHEIKE, T. H. AND KEIDING, N. (2006). Design and analysis of time-to-pregnancy. *Statistical Methods in Medical Research* **15**, 127–140.
- SCHEIKE, T. H. AND ZHANG, M.-J. (2008). Flexible competing risks regression modeling and goodness-of-fit. *Lifetime Data Analysis* **14**, 464–483.
- SCHMID, M., TUTZ, G. AND WELCHOWSKI, T. (2018). Discrimination measures for discrete time-to-event predictions. *Econometrics and Statistics* **7**, 153–164.

- TUTZ, G. (1995). Competing risks models in discrete time with nominal or ordinal categories of response. *Quality and Quantity* **29**, 405–420.
- TUTZ, G. AND SCHMID, M. (2016). *Modeling Discrete Time-to-Event Data*. New York: Springer.
- VAN DER LAAN, M. J. AND ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.
- WOLBERS, M., KOLLER, M. T., WITTEMANN, J. C. AND STEYERBERG, E. W. (2009). Prognostic models with competing risks: Methods and application to coronary risk prediction. *Epidemiology* **20**, 555–561.
- WOLKEWITZ, M., VONBERG, R. P., GRUNDMANN, H., BEYERSMANN, J., GASTMEIER, P., BÄRWOLFF, S., GEFFERS, C., BEHNKE, M., RÜDEN, H. AND SCHUMACHER, M. (2008). Risk factors for the development of nosocomial pneumonia and mortality on intensive care units: Application of competing risks models. *Critical Care* **12**, R44.
- ZHOU, B., FINE, J. AND LAIRD, G. (2013). Goodness-of-fit test for proportional subdistribution hazards model. *Statistics in Medicine* **32**, 3804–3811.

COPYRIGHT NOTICE

This is a pre-copyedited, author-produced version of an article accepted for publication in *Biostatistics* following peer review.

The version of record is available online at: <https://doi.org/10.1093/biostatistics/kxy069>.

[Received January 11, 2018; revised September 13, 2018; accepted for publication September 13, 2018]

Appendix to Subdistribution Hazard Models for Competing Risks in Discrete Time

MORITZ BERGER^{1,†}, MATTHIAS SCHMID^{1,†,*}, THOMAS WELCHOWSKI¹,
STEFFEN SCHMITZ-VALCKENBERG², JAN BEYERSMANN³

¹*Department of Medical Biometry, Informatics and Epidemiology, University of Bonn*

²*University Eye Hospital Bonn, Sigmund-Freud-Strasse 25, D-53127 Germany*

³*Institute of Statistics, Ulm University, Helmholtzstrasse 20, D-89081 Ulm*

matthias.schmid@imbie.uni-bonn.de

A PROOF OF THEOREM 1

As outlined in the main part of the paper, the key step in the proof is to show that the solution to (2.16) is the solution of a conditionally unbiased estimation equation. To this end, it is convenient to first consider the case where the censoring survival function needed for the definition of the weights is assumed to be known. Estimated weights can subsequently be handled by stacking estimation equations. To begin, we also first consider the case without competing risks in Appendix A.1 before allowing for multiple event types in Appendix A.2.

In order to solve the optimization problem in (2.16), it is necessary to determine the roots of

*To whom correspondence should be addressed. †Contributed equally to this work.

the weighted score function

$$s(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}) = \frac{\partial \ell}{\partial(\boldsymbol{\gamma}_0^\top, \boldsymbol{\gamma}^\top)^\top} = \frac{\partial \ell}{\partial \lambda_1} \cdot \frac{\partial \lambda_1}{\partial(\boldsymbol{\gamma}_0^\top, \boldsymbol{\gamma}^\top)^\top}. \quad (\text{A.1})$$

It can be shown (Carroll *and others* 2006, Appendix 6) that the estimator $(\hat{\boldsymbol{\gamma}}_0^\top, \hat{\boldsymbol{\gamma}}^\top)^\top$ solving the optimization problem (2.16) is a consistent and asymptotically normal estimator of $(\boldsymbol{\gamma}_0^\top, \boldsymbol{\gamma}^\top)^\top$ if the estimation equation $s(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}) = \mathbf{0}$ is *unbiased*, i.e., if

$$\mathbb{E} \{s(\boldsymbol{\gamma}_0, \boldsymbol{\gamma})\} = \mathbb{E} \left\{ \frac{\partial \ell}{\partial \lambda_1} \cdot \frac{\partial \lambda_1}{\partial(\boldsymbol{\gamma}_0^\top, \boldsymbol{\gamma}^\top)^\top} \right\} = \mathbb{E} \left\{ \frac{\partial \ell}{\partial \lambda_1} \right\} \cdot \frac{\partial \lambda_1}{\partial(\boldsymbol{\gamma}_0^\top, \boldsymbol{\gamma}^\top)^\top} = \mathbf{0}. \quad (\text{A.2})$$

To prove Theorem 1, we will therefore show that

$$\mathbb{E} \left\{ \frac{\partial \ell_i}{\partial \lambda_1} \right\} = 0, \quad i = 1, \dots, n, \quad (\text{A.3})$$

where ℓ_i is the contribution of the i -th individual to the weighted log-likelihood.

Remark: For the sake of simplicity and easier reading, the dependence of the subdistribution hazard $\lambda_1(t|\mathbf{x}_i)$ on the vector of explanatory variables \mathbf{x}_i will be omitted in the following.

A.1 Only Type 1 Events

We first consider the simplified situation where only type 1 events or censoring can occur (Section 2.3.1 of the paper). In this case, the contribution of individual i to the weighted log-likelihood (2.11) can be written as

$$\begin{aligned} \ell_i = & \sum_{t=1}^{k-1} \left\{ \begin{aligned} & \mathbb{I}(\tilde{T}_i = t \cap \Delta_i = 1) \log(\lambda_1(t)) && (\text{observed type 1 event at } t) \\ & + \mathbb{I}(\tilde{T}_i = t \cap \Delta_i = 0) \log(1 - \lambda_1(t)) && (\text{censoring at } t) \\ & + \mathbb{I}(\tilde{T}_i > t) \log(1 - \lambda_1(t)) \end{aligned} \right\}, && (\text{event after } t) \end{aligned} \quad (\text{A.4})$$

where $\mathbb{I}(\cdot)$ denotes the indicator function with $\mathbb{I}(a) = 1$ if a is true and $\mathbb{I}(a) = 0$ otherwise. Thus, in (A.4) the indicator function of the first addend is equal to y_{it} , whereas the indicator functions

in the second and third addends are equal to $1 - y_{it}$. The contribution of the i -th individual to the score function (ℓ_i differentiated with respect to $\lambda_1(t)$) is given by

$$\begin{aligned} s_i &= \frac{\partial \ell_i}{\partial \lambda_1(t)} = \frac{1}{\lambda_1(t)} \mathbb{I}(\tilde{T}_i = t \cap \Delta_i = 1) \\ &\quad - \frac{1}{1 - \lambda_1(t)} \mathbb{I}(\tilde{T}_i = t \cap \Delta_i = 0) \\ &\quad - \frac{1}{1 - \lambda_1(t)} \mathbb{I}(\tilde{T}_i > t). \end{aligned} \tag{A.5}$$

For the expectations one obtains

$$\mathbb{E}\{\mathbb{I}(\tilde{T}_i = t \cap \Delta_i = 1)\} = P(T_i = t)P(C_i \geq t), \tag{A.6}$$

$$\mathbb{E}\{\mathbb{I}(\tilde{T}_i = t \cap \Delta_i = 0)\} = P(T_i > t)P(C_i = t), \quad \text{and} \tag{A.7}$$

$$\mathbb{E}\{\mathbb{I}(\tilde{T}_i > t)\} = P(T_i > t)P(C_i > t). \tag{A.8}$$

Combining (A.5) to (A.8) yields

$$\begin{aligned} \mathbb{E}\{s_i\} &= \frac{1}{\lambda_1(t)} P(T_i = t)P(C_i \geq t) - \frac{1}{1 - \lambda_1(t)} \underbrace{[P(C_i = t) + P(C_i > t)]}_{=P(C_i \geq t)} P(T_i > t) \\ &= P(C_i \geq t) \left[\frac{P(T_i = t)}{\lambda_1(t)} - \frac{P(T_i > t)}{1 - \lambda_1(t)} \right] \\ &= P(C_i \geq t) \left[\frac{P(T_i = t)}{P(T_i = t|T_i \geq t)} - \frac{P(T_i > t)}{P(T_i > t|T_i \geq t)} \right] \\ &= P(C_i \geq t) \left[\frac{P(T_i = t)}{P(T_i = t)/P(T_i \geq t)} - \frac{P(T_i > t)}{P(T_i > t)/P(T_i \geq t)} \right] \\ &= 0. \end{aligned} \tag{A.9}$$

□

A.2 Competing Events

In the presence of the competing events of type $2, \dots, J$ (Section 2.3.2 of the paper), the contribution of the i -th individual to the weighted log-likelihood (2.11) can be written as

$$\begin{aligned}
\ell_i = & \sum_{t=1}^{k-1} \left\{ \mathbf{I}(\tilde{T}_i = t \cap \Delta_i = 1 \cap \epsilon_i = 1) \log(\lambda_1(t)) \right. && \text{(observed type 1 event at } t) \\
& + \mathbf{I}(\tilde{T}_i = t \cap \Delta_i = 0) \log(1 - \lambda_1(t)) && \text{(observed censoring event at } t) \\
& + \mathbf{I}(\tilde{T}_i > t) \log(1 - \lambda_1(t)) && \text{(event after } t) \\
& \left. + \mathbf{I}(\tilde{T}_i \leq t \cap \Delta_i = 1 \cap \epsilon_i \neq 1) \frac{G(t-1)}{G(\tilde{T}_i - 1)} \log(1 - \lambda_1(t)) \right\}, && \text{(event of type } \neq 1 \text{ at or before } t)
\end{aligned} \tag{A.10}$$

where, for the reasons mentioned above, we replaced the estimated function $\hat{G}(\cdot)$ by its true value $G(\cdot)$. The indicator function of the first addend in (A.10) corresponds to y_{it} , whereas the other indicator functions in (A.10) correspond to $1 - y_{it}$.

For the expectations one obtains

$$\mathbb{E}\{\mathbf{I}(\tilde{T}_i = t \cap \Delta_i = 1 \cap \epsilon_i = 1)\} = P(T_i = t, \epsilon_i = 1) P(C_i \geq t), \tag{A.11}$$

$$\mathbb{E}\{\mathbf{I}(\tilde{T}_i = t \cap \Delta_i = 0)\} = P(T_i > t) P(C_i = t), \tag{A.12}$$

$$\mathbb{E}\{\mathbf{I}(\tilde{T}_i > t)\} = P(T_i > t) P(C_i > t), \quad \text{and} \tag{A.13}$$

$$\begin{aligned}
& \mathbb{E}\left\{ \mathbf{I}(\tilde{T}_i \leq t \cap \Delta_i = 1 \cap \epsilon_i \neq 1) \frac{G(t-1)}{G(\tilde{T}_i - 1)} \right\} \\
& = \underbrace{G(t-1)}_{=P(C_i \geq t)} \sum_{u=1}^t \left[\frac{1}{G(u-1)} P(\tilde{T}_i = u \cap \Delta_i = 1 \cap \epsilon_i \neq 1) \right] \\
& = P(C_i \geq t) \sum_{u=1}^t \left[\frac{1}{G(u-1)} P(C_i \geq u \cap T_i = u \cap \epsilon_i \neq 1) \right] \\
& = P(C_i \geq t) \sum_{u=1}^t \left[\frac{1}{G(u-1)} \underbrace{P(C_i \geq u)}_{=G(u-1)} P(T_i = u \cap \epsilon_i \neq 1) \right] \\
& = P(C_i \geq t) \sum_{u=1}^t P(T_i = u \cap \epsilon_i \neq 1) \\
& = P(C_i \geq t) P(T_i \leq t, \epsilon_i \neq 1).
\end{aligned} \tag{A.14}$$

On the other hand, the subdistribution hazard $\lambda_1(t)$ it can be rewritten as

$$\begin{aligned}
 \lambda_1(t) &= P(T_i = t, \epsilon_i = 1 | (T_i \geq t) \cup (T_i \leq t-1, \epsilon_i \neq 1)) \\
 &= \frac{P(T_i = t, \epsilon_i = 1)}{\underbrace{P(T_i \geq t) + P(T_i < t, \epsilon_i \neq 1)}_{:=N_i}} \\
 &= \frac{P(T_i = t, \epsilon_i = 1)}{N_i}, \tag{A.15}
 \end{aligned}$$

implying that

$$\begin{aligned}
 1 - \lambda_1(t) &= \frac{P(T_i \geq t) + P(T_i < t, \epsilon_i \neq 1)}{P(T_i \geq t) + P(T_i < t, \epsilon_i \neq 1)} - \frac{P(T_i = t, \epsilon_i = 1)}{N_i} \\
 &= \frac{P(T_i > t) + P(T_i = t) - P(T_i = t, \epsilon_i = 1) + P(T_i < t, \epsilon_i \neq 1)}{N_i} \\
 &= \frac{P(T_i > t) + P(T_i = t, \epsilon_i \neq 1) + P(T_i < t, \epsilon_i \neq 1)}{N_i} \\
 &= \frac{P(T_i > t) + P(T_i \leq t, \epsilon_i \neq 1)}{N_i}. \tag{A.16}
 \end{aligned}$$

Combining (A.10) to (A.16) yields

$$\begin{aligned}
 \mathbb{E}\{s_i\} &\stackrel{(A.11)}{=} \frac{P(C_i \geq t)P(T_i = t, \epsilon_i = 1)}{\lambda_1(t)} \stackrel{(A.12),(A.13)}{=} \frac{P(C_i \geq t)P(T_i > t)}{1 - \lambda_1(t)} \\
 &\stackrel{(A.14)}{=} \frac{P(C_i \geq t)P(T_i \leq t, \epsilon_i \neq 1)}{1 - \lambda_1(t)} \\
 &\stackrel{(A.15),(A.16)}{=} P(C_i \geq t) \left[\frac{P(T_i = t, \epsilon_i = 1)}{P(T_i = t, \epsilon_i = 1)/N_i} - \frac{P(T_i > t) + P(T_i \leq t, \epsilon_i \neq 1)}{((P(T_i > t) + P(T_i \leq t, \epsilon_i \neq 1))/N_i)} \right] \\
 &= P(C_i \geq t) [N_i - N_i] \\
 &= 0. \tag{A.17}
 \end{aligned}$$

□

B FURTHER NUMERICAL RESULTS

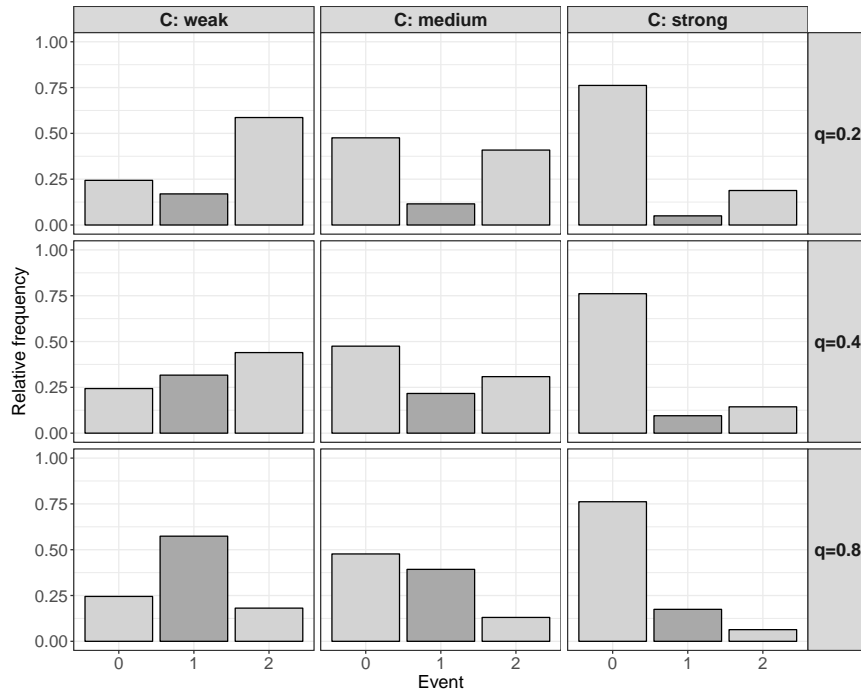


Fig. 1: Illustration of the experimental design of the simulation study. The bars display the average relative frequencies of observed events (0 = censoring event, 1 = event of interest, 2 = competing event) that were obtained from 1000 simulated data sets ($n = 500$). The ratio of type 1 and type 2 events was approximately the same in each row (C = degree of censoring).

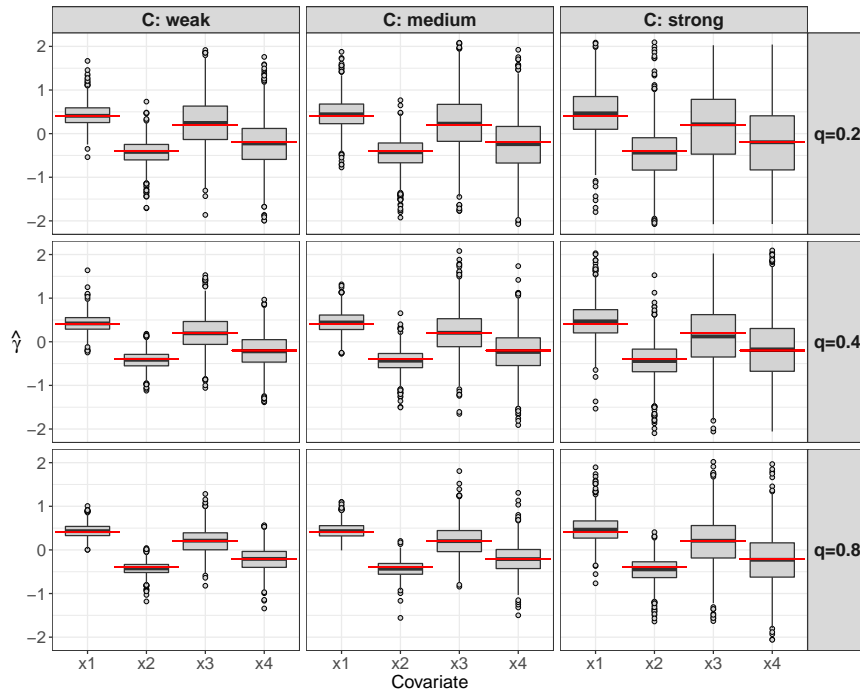


Fig. 2: Results of the simulation study. The boxplots visualize the estimates of the parameters $\gamma_1 = (0.4, -0.4, 0.2, -0.2)^\top$ that were obtained from fitting a discrete subdistribution hazard model using the proposed weighted ML estimation approach ($n = 100$). The horizontal lines refer to the true values of the parameters. Note that some of the boxplots have been truncated, as for reasons of comparability we used the same y -axis limits as in Figure 1 of the paper.

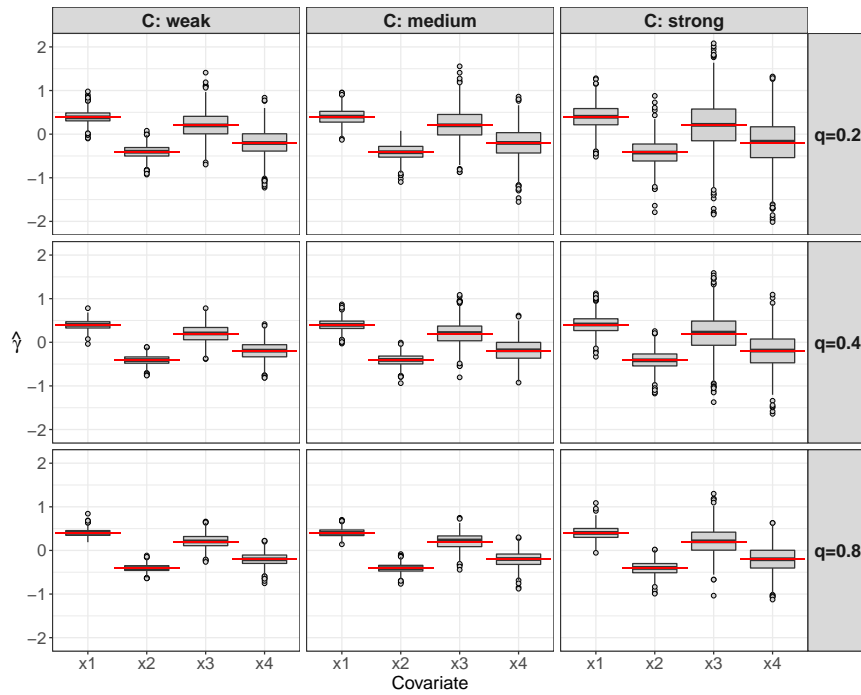


Fig. 3: Results of the simulation study. The boxplots visualize the estimates of the parameters $\gamma_1 = (0.4, -0.4, 0.2, -0.2)^\top$ that were obtained from fitting a discrete subdistribution hazard model using the proposed weighted ML estimation approach ($n = 300$). The horizontal lines refer to the true values of the parameters.

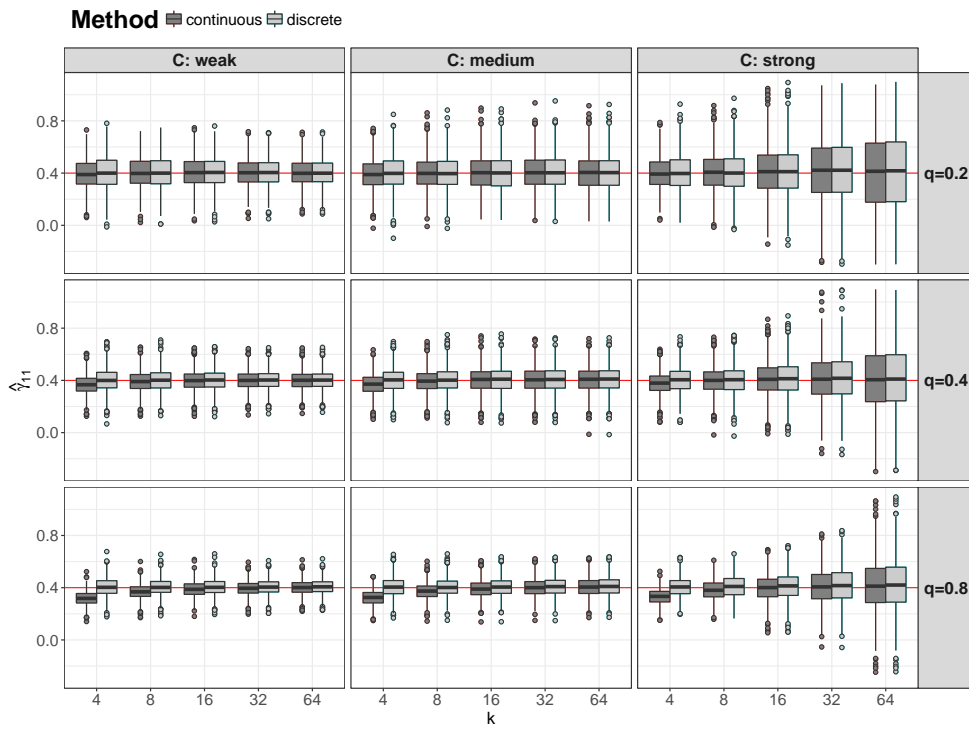


Fig. 4: Results of the simulation study. The boxplots visualize the estimates of the coefficient $\gamma_{11} = 0.4$, as obtained from the discrete-time subdistribution hazard model and the continuous-time Fine & Gray model ($n = 500$). The horizontal lines refer to the true values of γ_{11} . Interval numbers on the x-axes are presented on the \log_2 scale.

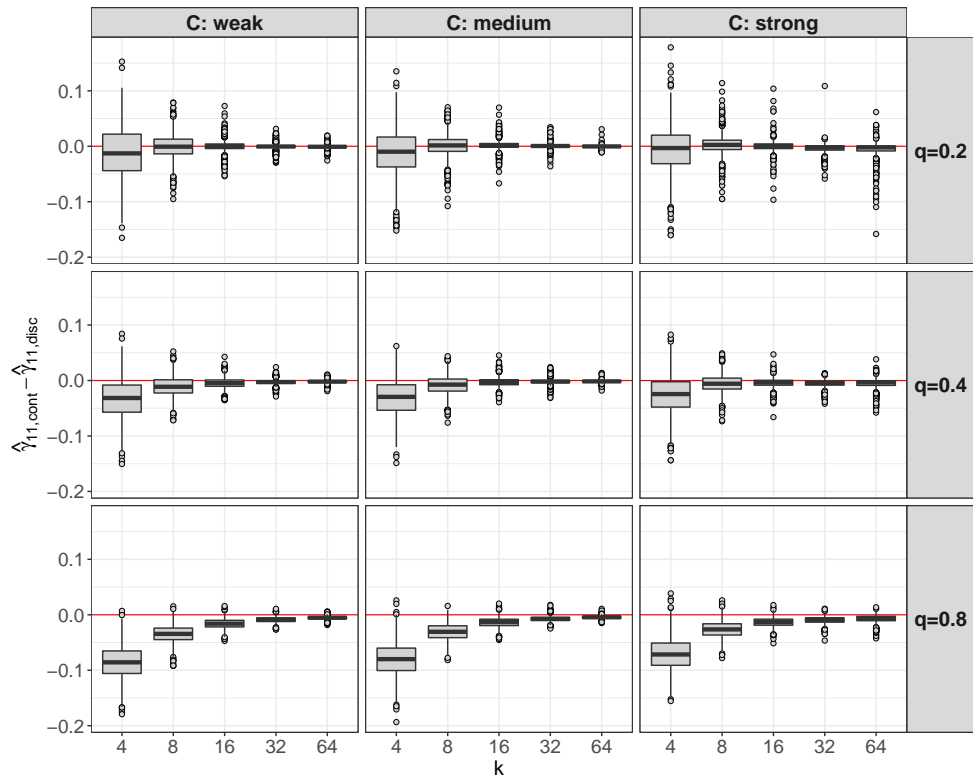


Fig. 5: Results of the simulation study. The boxplots visualize the differences between the discrete-time subdistribution hazard model and the continuous-time Fine & Gray model in the estimates of the coefficient $\gamma_{11} = 0.4$ ($n = 500$). Interval numbers on the x-axis are presented on the \log_2 scale.

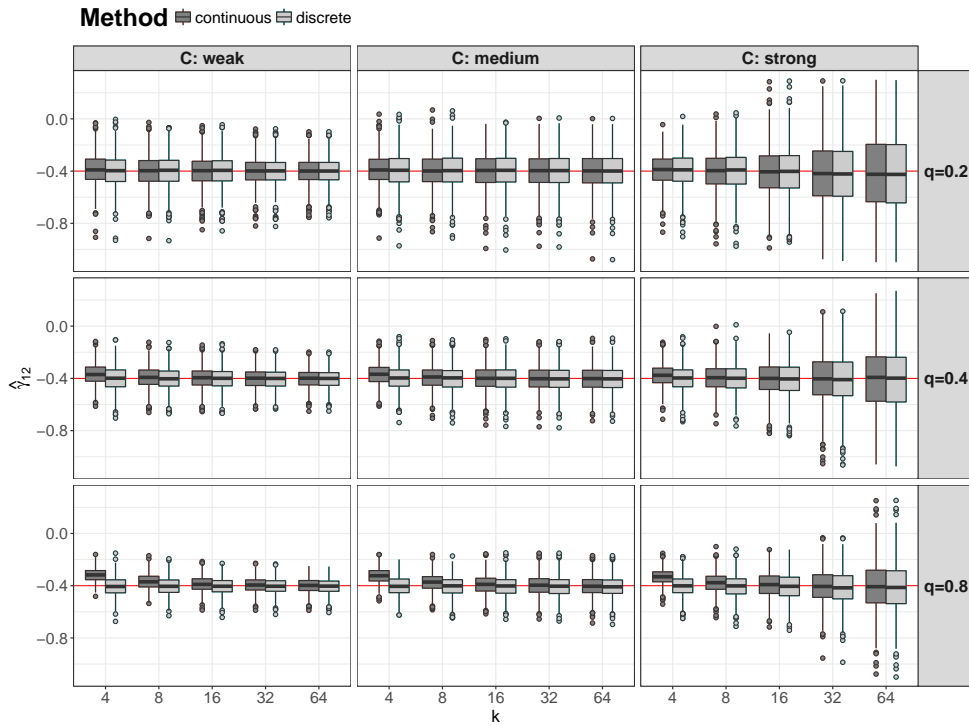


Fig. 6: Results of the simulation study. The boxplots visualize the estimates of the coefficient $\gamma_{12} = -0.4$, as obtained from the discrete-time subdistribution hazard model and the continuous-time Fine & Gray model ($n = 500$). The horizontal lines refer to the true values of γ_{12} . Interval numbers on the x-axes are presented on the \log_2 scale.

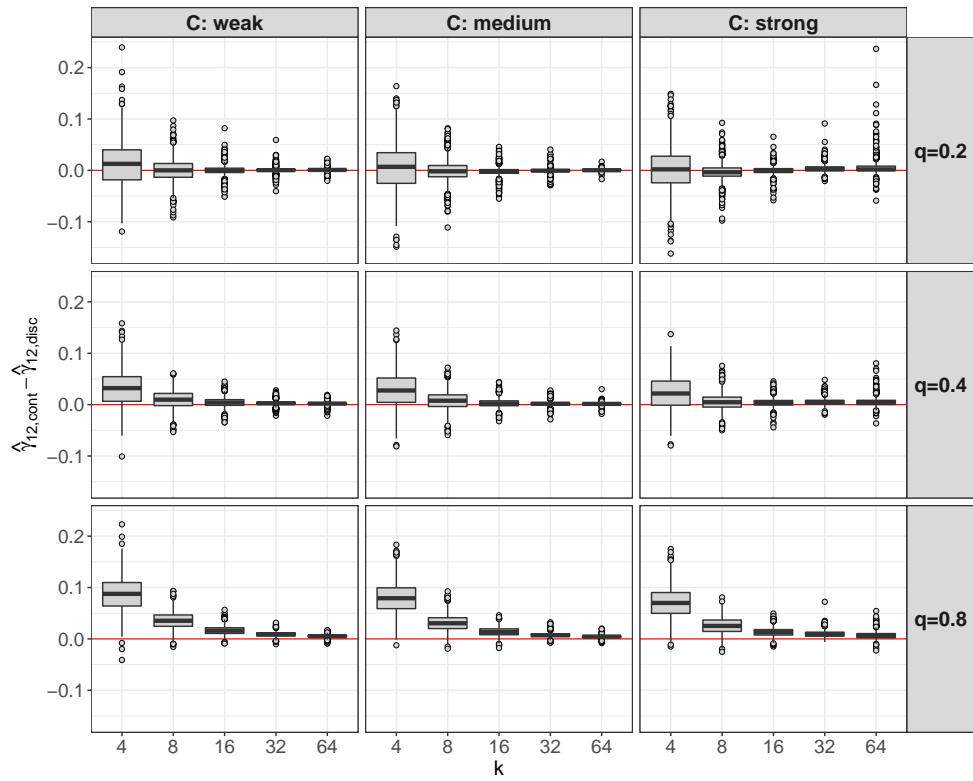


Fig. 7: Results of the simulation study. The boxplots visualize the differences between the discrete-time subdistribution hazard model and the continuous-time Fine & Gray model in the estimates of the coefficient $\gamma_{12} = -0.4$ ($n = 500$). Interval numbers on the x-axis are presented on the \log_2 scale.

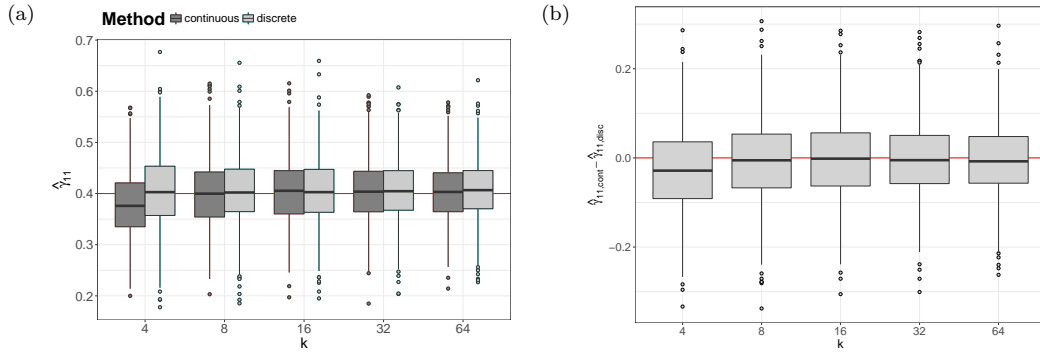


Fig. 8: Results of the simulation study. The boxplots in panel (a) visualize the the estimates of the coefficient $\gamma_{11} = 0.4$, as obtained from the discrete-time subdistribution hazard model and the continuous-time Fine & Gray model using the Efron method for tie handling ($n = 500$, $q = 0.8$, $b = 0.85$, horizontal line = true value of γ_{11}). The boxplots in panel (b) visualize the respective differences between the discrete-time subdistribution hazard model and the continuous-time Fine & Gray model in the estimates of γ_{11} . Interval numbers on the x-axis are presented on the \log_2 scale. The continuous-time Fine & Gray estimates were obtained using the SAS macro by Kohl *and others* (2015).

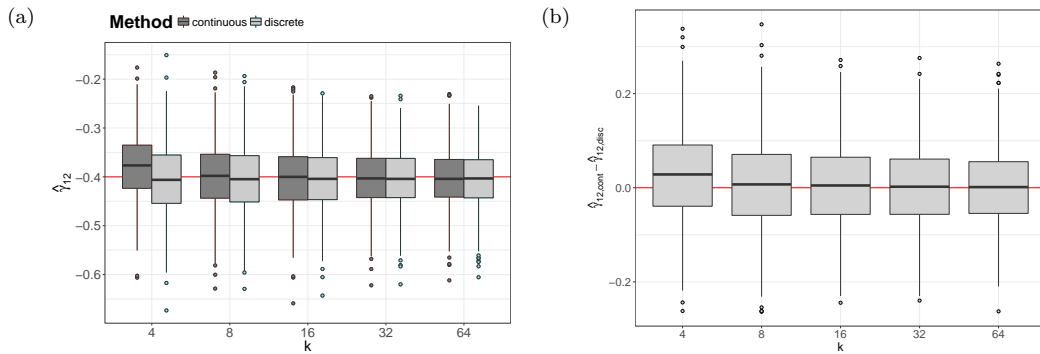


Fig. 9: Results of the simulation study. The boxplots in panel (a) visualize the the estimates of the coefficient $\gamma_{12} = -0.4$, as obtained from the discrete-time subdistribution hazard model and the continuous-time Fine & Gray model using the Efron method for tie handling ($n = 500$, $q = 0.8$, $b = 0.85$, horizontal line = true value of γ_{12}). The boxplots in panel (b) visualize the respective differences between the discrete-time subdistribution hazard model and the continuous-time Fine & Gray model in the estimates of γ_{12} . Interval numbers on the x-axis are presented on the \log_2 scale. The continuous-time Fine & Gray estimates were obtained using the SAS macro by Kohl *and others* (2015).

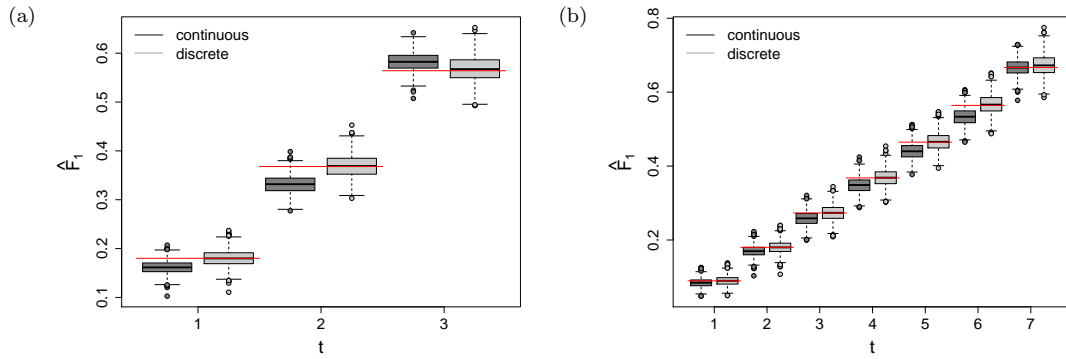


Fig. 10: Results of the simulation study. The boxplots visualize the estimates of the cumulative incidence function F_1 , as obtained from the discrete-time subdistribution hazard model and the continuous-time Fine & Gray model ($n = 500$, $q = 0.8$, $b = 0.85$). Panels (a) and (b) refer to the scenarios with $k = 4$ and $k = 8$, respectively. All estimates were averaged over the values of the covariates, i.e., each of the boxplots contains one estimate per simulation run. The continuous-time Fine & Gray estimates are based on the Breslow method for tie handling. The horizontal lines refer to the true values of $F_1(t|\cdot)$. The two panels demonstrate for almost all time points a downward bias of the Breslow approximation, which is in line with recent findings by Mehrotra and Zhang (2018), see their Figure 1. However, when t approaches the last time point, the bias vanishes (panel b) or turns into an upward bias (panel a). A possible explanation is that inference in the Fine & Gray model outside model conditions appears to capture the plateau of the cumulative incidence function, see Grambauer *and others* (2010) for a detailed investigation.

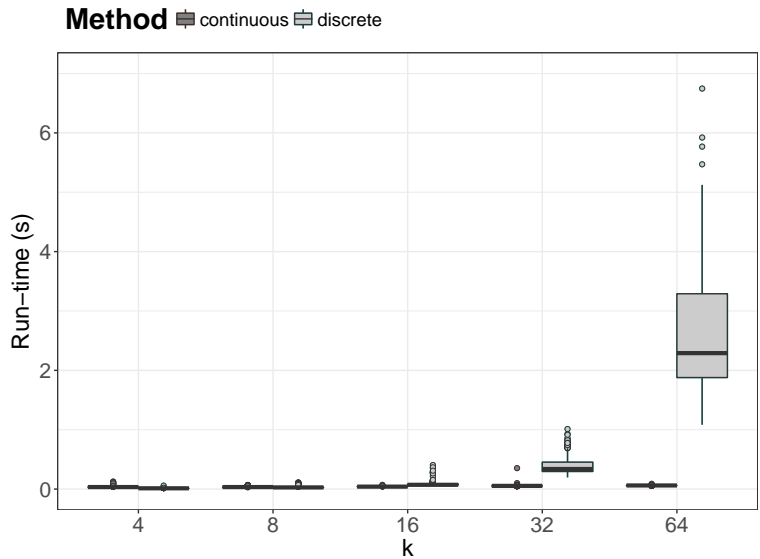


Fig. 11: Results of the simulation study. The boxplots illustrate the run-times that were obtained from fitting discrete-time subdistribution hazard models and continuous-time Fine & Gray models to the 1000 data sets ($n = 500$, $q = 0.8$, $b = 0.85$). Run-times refer to a computing system with a 2.66 GHz Intel Xeon X5650 CPU and 96 GB RAM. Interval numbers on the x-axis are presented on the \log_2 scale.

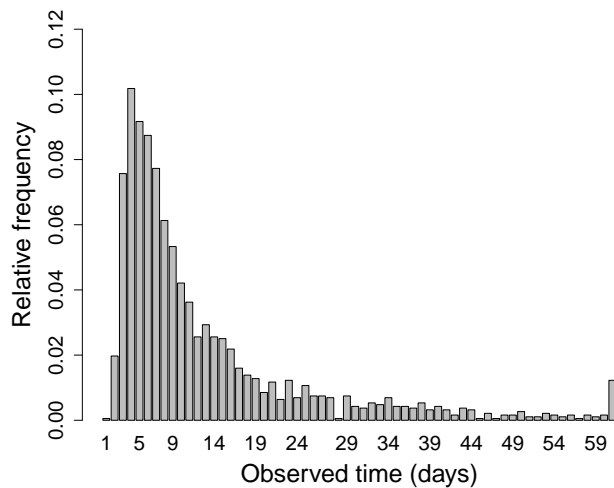


Fig. 12: Analysis of the NP infection data. The figure shows a bar plot of the observation times (measured in days) of the $n = 1,876$ patients included in the study. The median observation time was 8 days.

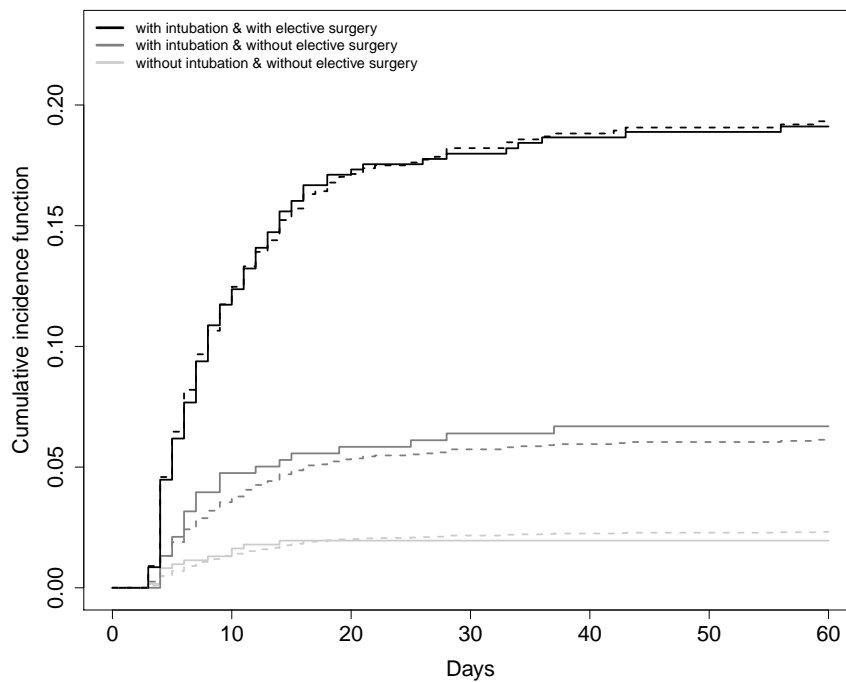


Fig. 13: Analysis of the NP infection data (model 1). The figure shows the estimated cumulative incidence functions for NP acquisition that were obtained from fitting covariate-free discrete subdistribution hazard models to subsets of the data. Solid lines refer to subgroups defined by intubation on admission and/or elective surgery at admission. Dashed lines refer to the respective average estimated cumulative incidence functions obtained from the complementary log-log model with all covariates. The figure does not suggest any major violations of the proportional subdistribution hazards assumption.

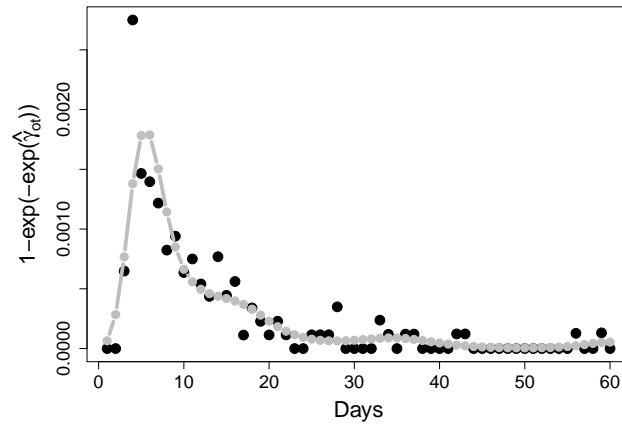


Fig. 14: Analysis of the NP infection data. The figure shows the estimated baseline coefficients of model 1 (black dots) and the estimated smooth baseline function of model 2 (gray dots). The baseline coefficients/function were transformed using the distribution function $1 - \exp(-\exp(\hat{\gamma}_{0t}))$ of the complementary log-log subdistribution hazard model. Both models indicate that the part of the subdistribution hazard that could not be explained by any of the predictor variables strongly increased up to day 4 and subsequently decreased until about day 25. It was close to zero beyond day 25.

REFERENCES

- CARROLL, R., RUPPERT, D., STEFANSKI, L. A. AND CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Boca Raton: Chapman and Hall/CRC.
- GRAMBAUER, N., SCHUMACHER, M. AND BEYERSMANN, J. (2010). Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified. *Statistics in Medicine* **29**, 875–884.
- KOHL, M., PLISCHKE, M., LEFFONDRE, K. AND HEINZE, G. (2015). PSHREG: A SAS macro for proportional and nonproportional subdistribution hazards regression. *Computer Methods and Programs in Biomedicine* **118**, 218–233.
- MEHROTRA, D. V. AND ZHANG, Y. (2018). Hazard ratio estimation and inference in clinical trials with many tied event times. *Statistics in Medicine*. doi:10.1002/sim.7843.

2.5 Berger et al., Canadian Journal of Statistics 50, 572-591

Bei der Entwicklung von neuen Vorhersagemodellen ist es entscheidend, dass die Güte der Modelle anhand von unabhängigen, externen Daten validiert wird (Moons et al., 2012a). Dabei unterscheidet man grundsätzlich zwischen Diskriminierung, d.h. wie gut sich Fälle von Kontrollen trennen lassen, und Kalibrierung, d.h. wie gut die vorhergesagten Werte und die beobachteten Werte übereinstimmen (Steyerberg et al., 2010). Während für die diskreten Hazard-Modelle, die in den Kapiteln 2.1 bis 2.3 behandelt wurden, bereits Methoden zur Beurteilung von Diskriminierung und Kalibrierung in der Literatur existieren, ist das für die neue Methode, die in Kapitel 2.4 eingeführt wurde, nicht der Fall.


In diesem Kapitel wird daher ein Instrumentarium zur Beurteilung der Kalibrierung von diskreten Subdistribution Hazard-Modellen (22) eingeführt. Dies beinhaltet (i) ein Diagramm zur grafischen Beurteilung der Kalibrierung, und (ii) ein Rekalibrierungsmodell zur formalen Beurteilung der Kalibrierung. Letzteres entspricht dem logistischen Regressionsmodell

$$\log \left(\frac{\lambda_1(t|X)}{1 - \lambda_1(t|X)} \right) = a + b \log \left(\frac{\hat{\lambda}_1(t|X)}{1 - \hat{\lambda}_1(t|X)} \right), \quad t = 1, \dots, k - 1. \quad (23)$$

Modellgleichung (23) enthält den Intercept a zur Messung der “calibration in the large”, der anzeigt, ob die vorhergesagten Wahrscheinlichkeiten systematisch zu niedrig oder systematisch zu hoch sind, und den Steigungsparameter b zur Messung des “refinement”, der anzeigt, ob die vorhergesagten Wahrscheinlichkeiten zu wenig oder zu viel Variation aufweisen. Beide Methoden bauen auf Ansätzen für binäre Regressionsmodelle auf (Miller et al., 1993; Hosmer et al., 2013).

In einer Simulationsstudie, deren Aufbau sich an den Simulationen in Kapitel 2.4 orientiert, wird gezeigt, dass die Methoden sehr gut funktionieren, um sowohl korrekt spezifizierte Modelle zu erkennen als auch Fehlspezifikationen aufzudecken. Insbesondere werden simulierte Daten betrachtet, für die die Annahme unabhängiger Zensierung (siehe auch Kapitel 1.1) verletzt ist. Im letzten Abschnitt wird die Kalibrierung des in Kapitel 2.4 entwickelten Vorhersagemodells für die Erkrankung an nosokomialer Lungenentzündung evaluiert. Sowohl die Kalibrierungsdiagramme, als auch die Rekalibrierungsmodelle, die durch wiederholtes Aufteilen des Datensatzes in Trainings- und Validierungsdaten erstellt wurden, weisen auf eine angemessene Kalibrierung des Modells hin.

Assessing the calibration of subdistribution hazard models in discrete time

Moritz BERGER^{1*}  and Matthias SCHMID¹

¹*Institute of Medical Biometry, Informatics and Epidemiology, Faculty of Medicine, University of Bonn, Bonn, Germany*

Key words and phrases: Calibration; competing risks; discrete time-to-event data; subdistribution hazard; validation

MSC 2020: Primary 62N01; secondary 62P10.

Abstract: The generalization performance of a risk prediction model can be evaluated by its calibration, which measures the agreement between predicted and observed outcomes on external validation data. Here, we propose methods for assessing the calibration of discrete time-to-event models in the presence of competing risks. Specifically, we consider the class of discrete subdistribution hazard models, which directly relate the cumulative incidence function of one event of interest to a set of covariates. We apply the methods to a prediction model for the development of nosocomial pneumonia. Simulation studies show that the methods are strong tools for calibration assessment even in scenarios with a high censoring rate and/or a large number of discrete time points. *The Canadian Journal of Statistics* 50: 572–591; 2022 © 2021 The Authors. The Canadian Journal of Statistics/La revue canadienne de statistique published by Wiley Periodicals LLC on behalf of Statistical Society of Canada.

Résumé: La performance de généralisation d'un modèle de prévision des risques peut être évaluée par sa calibration qui mesure la concordance entre les valeurs prédites et observées dans des données externes de validation. Les auteurs proposent des méthodes pour évaluer la calibration de modèles discrets de durée de vie en présence de risques concurrents. Plus précisément, ils considèrent la classe de modèles à sous-distribution discrète du risque qui relie directement la fonction d'incidence cumulative d'un événement à un ensemble de covariables. Les auteurs appliquent leurs méthodes à un modèle de prévision pour le développement de pneumonie nosocomiale. Ils présentent des études de simulation montrant que les méthodes sont d'excellents outils pour l'évaluation de la calibration, même dans les scénarios comportant un haut taux de censure et/ou un large nombre de points temporels discrets. *La revue canadienne de statistique* 50: 572–591; 2022 © 2021 Les auteurs. La revue canadienne de statistique/The Canadian Journal of Statistics, publiée par Wiley Periodicals LLC au nom de la Société statistique du Canada.

1. INTRODUCTION

Over the past decade, risk prediction models have become an indispensable tool for decision making in applied research. Popular examples include models for diagnosis and prognosis in the health sciences where risk prediction is used, such as screening and therapy decisions (Moons et al., 2012b; Liu et al., 2014; Steyerberg, 2019) and models for risk assessment in ecological research, which have become an established tool to quantify and forecast the ecological impact of technology and development (Gibbs, 2011).

Additional Supporting Information may be found in the online version of this article at the publisher's website.

* Corresponding author: moritz.berger@imbie.uni-bonn.de

© 2021 The Authors. The Canadian Journal of Statistics/La revue canadienne de statistique published by Wiley Periodicals LLC on behalf of Statistical Society of Canada.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

A key aspect in the development of risk prediction models is the validation of generalization performance. This task, which is usually performed by applying a previously derived candidate prediction model to one or more sets of independent external validation data, has been subject to extensive methodological research (Moons et al., 2012a; Steyerberg & Vergouwe, 2014; Harrell, 2015; Steyerberg & Harrell, 2016; Alba et al., 2017). As a result, strategies for investigating the discriminatory power (measuring how well a model separates cases from controls), calibration (measuring the agreement between predicted and observed outcomes) and prediction error (quantifying both discrimination and calibration aspects) of prediction models have been developed (Steyerberg et al., 2010). Alternative techniques that additionally involve decision analytic measures include, among others, net benefit analysis (Vickers, Van Calster & Steyerberg, 2016), decision curve analysis (Vickers & Elkin, 2006) and relative utility curve analysis (Baker et al., 2009; Kerr & Janes, 2017).

The aim of this article is to develop a set of methods for assessing the calibration of a prediction model with a time-to-event outcome. This class of models has been dealt with extensively during the past years, see, for example, Henderson & Keiding (2005), Witten & Tibshirani (2010), Soave & Strug (2018) and Braun et al. (2018). Here, we explicitly assume that event times are measured on a discrete time scale $t = 1, 2, \dots$ (Ding et al., 2012; Tutz & Schmid, 2016; Berger & Schmid, 2018), and that the event of interest may occur along with one or more “competing” events (Fahrmeir & Wagenpfeil, 1996; Fine & Gray, 1999; Lau, Cole & Gange, 2009; Beyersmann, Allignol & Schumacher, 2011; Austin, Lee & Fine, 2016; Lee, Feuer & Fine, 2018; Schmid & Berger, 2020). Scenarios of this type are frequently encountered in observational studies with a limited number of fixed follow-up measurements, for instance, in epidemiology (Andersen et al., 2012). Such study designs do not allow recording the exact (continuous) event times, so that it is only known whether or not an event of interest (or a competing event) occurred between two consecutive follow-up times a_{t-1} and a_t , implying that the discrete time scale $t = 1, 2, \dots$ refers to a special case of interval censoring with fixed intervals.

An important example, which will be considered in this article, is the duration to the development of nosocomial pneumonia (NP) in intensive care patients measured on a daily basis (Wolkewitz et al., 2008). As NP infections are associated with an increased length of hospital stay and have considerable impact on morbidity and mortality, it is highly relevant to build a statistical model that gives valid predictions for future patients. The case of observational hospital data is interesting in that early discrete-time competing risks analysis is found in the literature as early as the 1860s (Nightingale, 1863, Chapter IX). In Section 7 we validate a prediction model developed by Berger et al. (2020) for this type of data.

In recent years, several authors have developed measures and estimators for analyzing the generalization performance of discrete time-to-event models. For example, discrimination measures for discrete time-to-event models were proposed by Schmid, Tutz & Welchowski (2018). Measures of prediction error were considered in Tutz & Schmid (2016), Chapter 4. Graphical tools for assessing the calibration of discrete time-to-event predictions (not accounting for the occurrence of competing events) were explored in Berger & Schmid (2018). Methods for assessing the generalization performance of discrete cause-specific hazard models (a common approach for competing risks analysis) have been recently proposed by Heyard et al. (2020).

Here we propose to base the calibration assessments for discrete competing risks models on the *cumulative incidence function* $F_j(t|\mathbf{x}) := P(T \leq t, \epsilon = j|\mathbf{x})$, denoting by T the time to the first event, by \mathbf{x} a set of covariates, and by $\epsilon \in \{1, \dots, J\}$ a random variable that indicates the occurrence of one out of J competing events at T (Fine & Gray, 1999; Klein & Andersen, 2005). In the following, we will assume without loss of generality that the event of interest and its cumulative incidence function are defined by $\epsilon = 1$ and $F_1(t|\mathbf{x})$, respectively. A popular method to derive

predictions of $F_1(t|\mathbf{x})$ from a set of training data is to fit a proportional subdistribution hazard model (Fine & Gray, 1999). This approach is designed for the analysis of right-censored event times, and it will be considered here. This approach has been recommended to analysts “whenever the focus is on estimating incidence or predicting prognosis in the presence of competing risks” (Austin, Lee & Fine, 2016). While the original model proposed by Fine & Gray (1999) assumed the event times to be measured on a continuous time scale, the methods developed in this article are based on a recent extension of the subdistribution hazard modelling approach to discrete-time competing risks data (Berger et al., 2020). Specifically, the model proposed by Berger et al. (2020) is designed for estimating the *discrete subdistribution hazard* $\lambda_1(t|\mathbf{x}) := P(T = t, \epsilon = 1 | (T \geq t) \cup (T \leq t - 1, \epsilon \neq 1), \mathbf{x})$, which defines a one-to-one relationship with the discrete cumulative incidence function $F_1(t|\mathbf{x})$ (see Sections 2 and 3 for details). The calibration of a subdistribution hazard prediction model may thus be characterized by how well the subdistribution hazards *observed* in a validation sample can be approximated by the respective *predicted* subdistribution hazards that are obtained from applying the prediction model to the validation data.

The proposed methodology for assessing the calibration of a discrete-time subdistribution hazard model comprises two parts, both of which build on methods for binary regression: The first part (presented in Section 4) will be concerned with the derivation of an appropriate *calibration plot* that visualizes the agreement between the predicted and the observed subdistribution hazards. In the second part (Section 5), we will propose a *recalibration model* for discrete-time subdistribution hazard models that can be used to analyze calibration-in-the-large and refinement (i.e., the bias and the variation, respectively, of the predicted subdistribution hazards) along the lines of Cox (1958) and Miller et al. (1993). As will be shown in Sections 4 and 5, the weights used in the subdistribution hazard modelling approach proposed in Berger et al. (2020) allow defining appropriate versions of the observed and predicted hazards (to be depicted in the calibration plot) and for fitting a weighted logistic recalibration model (giving rise to point estimates and hypothesis tests on calibration-in-the-large and refinement).

The proposed calibration assessments will be illustrated by a simulation study (Section 6) and by the aforementioned prediction model for the duration to the development of NP (Section 7). Section 8 summarizes the main findings of the article.

2. DISCRETE SUBDISTRIBUTION HAZARD MODELS

Let T_i be the event time and C_i be the censoring time of an i.i.d. sample with n individuals $i = 1, \dots, n$. Both T_i and C_i are assumed to be independent random variables (random censoring) taking discrete values in $\{1, 2, \dots, k\}$, where k is a natural number. It is further assumed that the censoring mechanism is non-informative for T_i , in the sense that C_i does not depend on any parameters used to model the event time (Kalbfleisch & Prentice, 2002). For instance, in longitudinal studies with fixed follow-up visits, the discrete event times $1, \dots, k$ may refer to time intervals $[0, a_1), [a_1, a_2), \dots, [a_{k-1}, \infty)$, where $T_i = t$ means that the event has occurred in time interval $[a_{t-1}, a_t)$ with $a_k = \infty$. For right-censored data, the time period during which an individual is under observation is denoted by $\tilde{T}_i = \min(T_i, C_i)$, that is, \tilde{T}_i corresponds to the true event time if $T_i \leq C_i$ and to the censoring time otherwise. The random variable $\Delta_i := I(T_i \leq C_i)$ indicates whether \tilde{T}_i is right-censored ($\Delta_i = 0$) or not ($\Delta_i = 1$). Here, it is assumed that each individual can experience one out of J competing events and that the event type of the i th individual at T_i is denoted by $\epsilon_i \in \{1, \dots, J\}$. In accordance with Fine & Gray (1999), our interest is in modelling the cumulative incidence function $F_1(t) = P(T \leq t, \epsilon = 1)$ of a type 1 event conditional on covariates, taking into account that there are $J - 1$ competing events and also the censoring event ($\Delta_i = 0$).

For given values of a set of time-constant covariates $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$, the discrete subdistribution hazard function for a type 1 event is defined by

$$\lambda_1(t|\mathbf{x}_i) = P(T_i = t, \epsilon_i = 1 | (T_i \geq t) \cup (T_i \leq t - 1, \epsilon_i \neq 1), \mathbf{x}_i) \tag{1}$$

$$= P(\vartheta_i = t | \vartheta_i \geq t, \mathbf{x}_i), \tag{2}$$

where (2) is the discrete hazard function of the subdistribution time

$$\vartheta_i := \begin{cases} T_i, & \text{if } \epsilon_i = 1, \\ \infty, & \text{if } \epsilon_i \neq 1, \end{cases} \tag{3}$$

see Berger et al. (2020). The subdistribution time ϑ_i measures the time to the occurrence of a type 1 event first and is not finite if $\epsilon_i \neq 1$ (as a type 1 event will never be the first event as soon as a competing event has occurred). The discrete subdistribution hazard is linked to the cumulative incidence function by

$$F_1(t|\mathbf{x}_i) = 1 - \prod_{s=1}^t (1 - \lambda_1(s|\mathbf{x}_i)) = 1 - S_1(t|\mathbf{x}_i), \tag{4}$$

where $S_1(t|\mathbf{x}_i) = P(\vartheta_i > t|\mathbf{x}_i)$ is the discrete survival function for a type 1 event. Thus, a regression model for the discrete subdistribution hazard λ_1 has a direct interpretation in terms of the cumulative incidence function F_1 .

A class of regression models that relate the discrete subdistribution hazard function (2) to the covariates \mathbf{x}_i was proposed by Berger et al. (2020). It is defined by

$$\lambda_1(t|\mathbf{x}_i) = h(\eta_1(t, \mathbf{x}_i)), \tag{5}$$

where $h(\cdot)$ is a strictly monotone increasing distribution function. In line with classical hazard models for discrete event times (e.g., Tutz & Schmid 2016), it is assumed that the predictor function

$$\eta_1(t, \mathbf{x}_i) = \gamma_{0t} + \mathbf{x}_i^\top \boldsymbol{\gamma} \tag{6}$$

is composed of a set of time-varying intercepts $\gamma_{01}, \dots, \gamma_{0,k-1}$, referred to as *baseline coefficients*, and a linear function of the covariates with coefficients $\boldsymbol{\gamma} \in \mathbb{R}^p$ that do not depend on t . As in generalized additive models, it is also possible to extend $\eta_1(t, \mathbf{x}_i)$ by interactions and smooth (possibly nonlinear) functions. A popular choice of $h(\cdot)$ is the inverse complementary log–log function, which yields the *Gompertz model* $\lambda_1(t, \mathbf{x}_i) = 1 - \exp(-\exp(\eta_1(t, \mathbf{x}_i)))$, which is equivalent to the original subdistribution hazard model by Fine & Gray (1999) for continuous time-to-event data.

3. MODEL FITTING

In Berger et al. (2020), it was shown that consistent estimates of the model parameters in (6) can be derived using estimation techniques for weighted binary regression. This result is based on the observation that with i.i.d. data $(\tilde{T}_i, \Delta_i, \epsilon_i, \mathbf{x}_i)$, $i = 1, \dots, n$, the log likelihood of model (5) can be expressed as

$$\ell = \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} \{y_{it} \log(\lambda_1(t|\mathbf{x}_i)) + (1 - y_{it}) \log(1 - \lambda_1(t|\mathbf{x}_i))\} \tag{7}$$

with binary outcome values

$$(y_{i1}, \dots, y_{i, \tilde{T}_i}, \dots, y_{i, k-1}) = \begin{cases} (0, \dots, 0, 1, 0, \dots, 0), & \text{if } \Delta_i \varepsilon_i = 1, \\ (0, \dots, 0, 0, 0, \dots, 0), & \text{if } \Delta_i \varepsilon_i \neq 1. \end{cases} \quad (8)$$

For uncensored individuals that experience a type 1 event first ($\Delta_i \varepsilon_i = 1$) and for censored individuals ($\Delta_i \varepsilon_i = 0$) the weights w_{it} are defined as

$$w_{it} := \mathbf{I}(t \leq \tilde{T}_i), \quad (9)$$

whereas for uncensored individuals experiencing a competing event first ($\Delta_i \varepsilon_i > 1$) they are defined as

$$w_{it} := \begin{cases} 1 & \text{if } t \leq \tilde{T}_i, \\ \frac{\hat{V}(t-1)}{\hat{V}(\tilde{T}_i-1)} & \text{if } \tilde{T}_i < t \leq k-1. \end{cases} \quad (10)$$

The function $\hat{V}(t)$ in (10) is an estimate of the censoring survival function $V(t) = P(C_i > t)$, implying that the weights in (9) and (10) equal estimates of the individual-specific conditional probabilities of being (still) at risk for a type 1 event at time t . This is analogous to the respective weights for subdistribution hazard models in continuous time (Fine & Gray, 1999). As shown in Berger et al. (2020), maximization of the log likelihood (7) yields consistent and asymptotically normal estimators of the parameters γ_{0t} and γ . In Sections 4 and 5, we will show that the weights defined in (9) and (10) also play a key role in the calibration assessment of discrete-time subdistribution hazard models.

It should be noted that the inclusion of time-varying covariates in model (6) is not without problems. This is because the weighted log likelihood in (7) requires the time-dependent values of \mathbf{x}_i to be known up to time point $k-1$ and thus possibly beyond the observed event times \tilde{T}_i . When the covariates are not external, this is often unrealistic or even impossible. In particular, the cumulative incidence function (4) cannot be written as a function of the hazards when the model includes random (internal) time-varying covariates (Cortese & Andersen, 2010). Finding an adequate strategy for the analysis of such covariates in the subdistribution hazard modelling framework remains challenging (Poguntke et al., 2018; Schmid & Berger, 2020).

4. CALIBRATION PLOT

In the following we will assume that an i.i.d. training sample $(\tilde{T}_i, \Delta_i, \varepsilon_i, \mathbf{x}_i)$, $i = 1, \dots, n$ has been used to fit a statistical model that can be used to predict the individual-specific subdistribution hazards $\lambda_1(t|\mathbf{x})$ in some study population. We will further assume that the calibration of the fitted model is assessed by means of an independent i.i.d. validation sample with N individuals $(\tilde{T}_m, \Delta_m, \varepsilon_m, \mathbf{x}_m)$, $m = 1, \dots, N$. The starting point of our considerations is the calibration plot proposed in Berger & Schmid (2018), which applies to discrete hazard models with only a single type of event ($J = 1$). Note that both the specification of the subdistribution hazard model and the definition of its log-likelihood function remain valid in this case, as the scenario without competing events ($J = 1$) is a special case of Equations (5) and (7). The idea underlying the method by Berger & Schmid (2018) is to split the test data into G subsets D_g , $g = 1, \dots, G$, defined by the percentiles of the predicted hazards $\hat{\lambda}_1(t|\mathbf{x}_m) = \hat{P}(T_m = t | T_m \geq t, \mathbf{x}_m) = \hat{P}(y_{mt} = 1 | \mathbf{x}_m)$, $t = 1, \dots, \tilde{T}_m$, $m = 1, \dots, N$, which are obtained from the fitted binary model in (5). Following the approach by Hosmer, Lemeshow & Sturdivant (2013) for assessing the calibration of binary regression models, the average predicted hazards in the G groups are subsequently plotted against the empirical

hazards, which are given by the group-wise relative frequencies of outcome values with $y_{mt} = 1$. A well-calibrated model is indicated by a set of points that is close to the 45-degree line.

More formally, the predicted and empirical hazard estimates considered in Berger & Schmid (2018) can be written as

$$\begin{aligned} \bar{\hat{\lambda}}_{1g} &= \frac{1}{\sum_{m,t: \hat{\lambda}_1(t|\mathbf{x}_m) \in D_g} w_{mt}} \sum_{m,t: \hat{\lambda}_1(t|\mathbf{x}_m) \in D_g} \hat{\lambda}_1(t|\mathbf{x}_m) w_{mt}, \\ \text{and } \bar{y}_g &= \frac{1}{\sum_{m,t: \hat{\lambda}_1(t|\mathbf{x}_m) \in D_g} w_{mt}} \sum_{m,t: \hat{\lambda}_1(t|\mathbf{x}_m) \in D_g} y_{mt} w_{mt}, \\ t &= 1, \dots, k-1, \quad m = 1, \dots, N, \end{aligned} \tag{11}$$

respectively, where $w_{mt} := I(t \leq \tilde{T}_m) \in \{0, 1\}$ indicates whether individual m is at risk for a type 1 event at time point t , $t = 1, \dots, k-1$, or not. Note that the definition of w_{mt} in (11) is exactly the same as the definition of the weight w_{it} in (9). Also note that $\sum_{m,t: \hat{\lambda}_1(t|\mathbf{x}_m) \in D_g} w_{mt} = |D_g|$, as only the values $\hat{\lambda}_1(t|\mathbf{x}_m)$ with $w_{mt} = 1$ are used for defining the groups D_1, \dots, D_g . In a well-calibrated hazard model, the values $\bar{\hat{\lambda}}_{1g}$ should be close to their counterparts \bar{y}_g .

Now consider the scenario where, in addition to the type 1 event of interest, competing events of type 2, \dots, J may be observed. In this case, λ_1 becomes the subdistribution hazard of a type 1 event, as defined in (2). To obtain a calibration plot for a fitted subdistribution hazard model, we define the quantities $\hat{\lambda}_{1g}$ and \bar{y}_g analogous to the single-event scenario considered in (11). Unlike in the scenario with $J = 1$, however, the definition of the terms w_{mt} is not straightforward: the problem is that individuals experiencing a competing event first continue to be at risk beyond \tilde{T}_m until they experience the censoring event. Hence, as the censoring times C_m are unobserved if $C_m > \tilde{T}_m$, it usually cannot be determined whether these individuals would still be at risk at $t > \tilde{T}_m$. In accordance with Berger et al. (2020), we therefore propose to predict the probability of each individual $m = 1, \dots, N$, of being at risk for a type 1 event at time t and to set the terms w_{mt} equal to the predicted probabilities.

More specifically, the proposed strategy comprises the following steps:

- (i) Sort the predicted subdistribution hazards $\hat{\lambda}(t|\mathbf{x}_m)$, $t = 1, \dots, k-1$, $m = 1, \dots, N$, obtained from the fitted subdistribution hazard model and form groups D_1, \dots, D_G defined by the percentiles of $\hat{\lambda}(t|\mathbf{x}_m)$.
- (ii) Compute the weights w_{mt} using the formulas in (9) and (10), where $\hat{V}(\cdot)$ is estimated from the training sample with individuals $i = 1, \dots, n$.
- (iii) Compute $\hat{\lambda}_{1g}$ and \bar{y}_g as in (11) using the weights obtained in step (ii). Note that by definition, $\sum_{m,t: \hat{\lambda}_1(t|\mathbf{x}_m) \in D_g} w_{mt} \leq |D_g|$.
- (iv) Plot $\hat{\lambda}_{1g}$ against \bar{y}_g (using proportional axes).
- (v) Assess the calibration of the fitted subdistribution hazard model by inspecting the plot generated in step (iv). A well-calibrated model is indicated by a set of points that is close to the 45-degree line.

For the choice of G , we propose to use the rule by Doane (1976), which was originally developed for univariate frequency classification. With this rule, the number of subsets is defined by

$$G = \left\lceil 1 + \log_2(\mathcal{N}) + \log_2(1 + |\kappa_{\hat{\lambda}_1}| / \sigma_{\kappa_{\hat{\lambda}_1}}) \right\rceil, \tag{12}$$

where $\kappa_{\hat{\lambda}_1}$ denotes the skewness of the predicted subdistribution hazards, $\mathcal{N} = N \cdot (k - 1)$, and $\sigma_{\kappa_{\hat{\lambda}_1}} = \sqrt{6(\mathcal{N} - 2)/(\mathcal{N} + 1)/(\mathcal{N} + 3)}$.

It should be emphasized that the calibration plot constitutes an exploratory approach, and that inspection of the plot in step (v) generally involves subjective impression. Formal tests on calibration are proposed in the next section.

5. RECALIBRATION MODEL

In addition to the graphical checks presented in Section 4, we propose a recalibration approach for discrete subdistribution hazard models originating from the method by Cox (1958). The idea of this method, which was originally developed for assessing the calibration of binary regression models, is to fit a logistic regression model to the test data in order to investigate the agreement between a set of predicted probabilities and the respective values of the binary outcome variable.

Based on the binary representation of the subdistribution hazard model in (5) and (7), we propose to adapt the recalibration framework by Cox (1958) as follows: assuming that calibration assessments are again based on a validation sample $(\tilde{T}_m, \Delta_m, \epsilon_m, \mathbf{x}_m)$, $m = 1, \dots, N$, we fit a logistic regression model of the form

$$\log \left(\frac{\lambda_1(t|\mathbf{x}_m)}{1 - \lambda_1(t|\mathbf{x}_m)} \right) = \eta_{\text{rc}}(t|\mathbf{x}_m) = a + b \log \left(\frac{\hat{\lambda}_1(t|\mathbf{x}_m)}{1 - \hat{\lambda}_1(t|\mathbf{x}_m)} \right),$$

$$t = 1, \dots, k - 1, m = 1, \dots, N, \quad (13)$$

where $\hat{\lambda}_1(t|\mathbf{x}_m)$ are the predicted hazards defined in Section 4.

In (13) a simple linear model is placed on the logits of the subdistribution hazards. Alternatively, one could also use other link functions, like the probit link or complementary log–log link. The intercept a in model (13) measures “calibration-in-the-large,” that is, it indicates whether the predicted hazards are systematically too low ($a > 0$) or too high ($a < 0$). Analogously, the slope b measures “refinement,” which indicates that the predicted hazards do not show enough variation ($b > 1$), show too much variation ($0 < b < 1$), or show the wrong general direction ($b < 0$, Miller et al., 1993).

To assess the fit of the predicted hazards, we propose to follow the suggestions by Miller et al. (1993) and to conduct recalibration tests on the following null hypotheses: (i) $H_0: a = 0, b = 1$, which refers to an overall test for calibration; (ii) $H_0: a = 0 \mid b = 1$, to test for calibration-in-the-large given appropriate refinement and (iii) $H_0: b = 1 \mid a$, to test refinement given corrected calibration-in-the-large.

Because the predicted hazards $\hat{\lambda}_1$ are derived from a subdistribution hazard model that was fitted using weighted maximum likelihood estimation, we fit the recalibration model in (13) by optimizing a weighted binary log likelihood of the form

$$\ell_{\text{rc}} = \sum_{m=1}^N \sum_{t=1}^{k-1} w_{mt} \{y_{mt} \log(\pi_1(t|\mathbf{x}_m)) + (1 - y_{mt}) \log(1 - \pi_1(t|\mathbf{x}_m))\}, \quad (14)$$

where the probabilities $\pi_1(t|\mathbf{x}_m)$ are given by $\pi_1(t|\mathbf{x}_m) = \exp(\eta_{\text{rc}}(t|\mathbf{x}_m)) / (1 + \exp(\eta_{\text{rc}}(t|\mathbf{x}_m)))$. The binary outcome values y_{mt} and the weights w_{mt} are defined in the same way as in Section 4. Note that $\hat{V}(\cdot)$ is again estimated from the training sample with individuals $i = 1, \dots, n$. In the case where $a = 0$ (referring to the tests in (i) and (ii) above), the log likelihood (14) can be written as

$$\ell_{\text{rc}} = b \sum_{m=1}^N \sum_{t=1}^{k-1} w_{mt} y_{mt} \log(\hat{\lambda}_1(t|\mathbf{x}_m))$$

$$\begin{aligned}
 &+ b \sum_{m=1}^N \sum_{t=1}^{k-1} w_{mt} (1 - y_{mt}) \log (1 - \hat{\lambda}_1(t|\mathbf{x}_m)) \\
 &- \sum_{m=1}^N \sum_{t=1}^{k-1} w_{mt} \log (\hat{\lambda}_1(t|\mathbf{x}_m)^b + (1 - \hat{\lambda}_1(t|\mathbf{x}_m))^b), \tag{15}
 \end{aligned}$$

which corresponds to the weighted log likelihood in equation (7) of Cox (1958). The derivation of (15) is given in Section 1 of the Supplementary Material. It follows that hypotheses (i)–(iii) can be examined using likelihood-ratio test statistics that asymptotically (as $N \rightarrow \infty$) follow χ^2 -distributions with one (hypotheses ii and iii) and two (hypothesis i) degrees of freedom, respectively.

6. NUMERICAL EXPERIMENTS

In this section we present the results of numerical experiments to evaluate the proposed calibration measures under known conditions. The main focus of the study was on measuring the performance in scenarios with different rates of type 1 events, different levels of censoring, a varying number of discrete time points, and various forms of misspecification.

6.1. Experimental Design

In order to generate data from a given subdistribution hazard model for type 1 events, we used a scheme adopted from Fine & Gray (1999). This procedure is also described in Beyersmann, Allignol & Schumacher (2011), where it was termed “indirect simulation.” In all simulation scenarios, we considered data with two competing events, $\epsilon_i \in \{1, 2\}$, that was generated under the assumption of proportional subdistribution hazards. More specifically, our discrete subdistribution hazard model was based on the discretization of the continuous model

$$F_1(t|\mathbf{x}_i) = P(T_{cont,i} \leq t, \epsilon_i = 1|\mathbf{x}_i) = 1 - (1 - q + q \exp(-t))^{\exp(\mathbf{x}_i^\top \boldsymbol{\gamma})}, \tag{16}$$

where $T_{cont,i} \in \mathbb{R}^+$ denotes the continuous time span of individual i , and $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)^\top$ is a set of regression coefficients. The parameter $q \in (0, 1)$ determines the probability of a type 1 event which, according to (16), was given by $\pi_{i1} := P(\epsilon_i = 1|\mathbf{x}_i) = 1 - (1 - q)^{\exp(\mathbf{x}_i^\top \boldsymbol{\gamma})}$. By definition, high values of q result in high probabilities of π_{i1} , and vice versa. The probability of a competing event was given by $\pi_{i2} := P(\epsilon_i = 2|\mathbf{x}_i) = 1 - \pi_{i1} = (1 - q)^{\exp(\mathbf{x}_i^\top \boldsymbol{\gamma})}$.

Continuous time spans for type 2 events were drawn from the exponential model

$$T_{cont,i} | \epsilon_i = 2, \mathbf{x}_i \sim \text{Exp}(\xi_2 = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})),$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ denotes a set of regression coefficients linking the rate parameter ξ_2 with the values of the covariates \mathbf{x} .

In order to obtain discrete event times $T_{disc,i}$, we generated data according to the indirect simulation scheme described above and grouped the resulting continuous event times into categories $k \in \{5, 10, 15\}$. The latter were defined by the quantiles of the continuous event times, which were pre-estimated from an independent sample with 1,000,000 observations. As a consequence, the same interval boundaries were used in each simulation run. Censoring times were generated from a discrete distribution with probability density function $P(C_{disc,i} = t) = u^{k-t+1} / \sum_{s=1}^k u^s$, $t = 1, \dots, k$, where the percentage of censored observations was controlled by the parameter $u \in \mathbb{R}^+$.

We considered two standard, normally distributed covariates $x_{i1}, x_{i2} \sim N(0, 1)$ and two binary covariates $x_{i3}, x_{i4} \sim \text{Bin}(1, 0.5)$. All covariates were independent, and the true regression coefficients were set to $\boldsymbol{\gamma} = (0.4, -0.4, 0.2, -0.2)^\top$ and $\boldsymbol{\beta} = (-0.4, 0.4, -0.2, 0.2)^\top$, see [Fine & Gray \(1999\)](#). We specified three different censoring rates, denoted by *weak*, *medium* and *strong*, where the degree of censoring was controlled by the parameter u of the censoring distribution. More specifically, we used the values $u = 0.85$ (weak), $u = 1$ (medium) and $u = 1.25$ (strong), resulting in the censoring rates shown in Figure S1 of the Supplementary Material. We also considered three different probabilities of a type 1 event, specifying $q \in \{0.2, 0.4, 0.8\}$. In total, this resulted in $3 \times 3 \times 3 = 27$ different scenarios. All scenarios were analyzed using 100 replications with 5000 independently drawn observations each, which were equally split into a training sample and a validation sample ($n = 2500$ and $N = 2500$), respectively.

Figure S1 of the Supplementary Material illustrates the relative frequencies of observed events for the nine scenarios with $k = 5$. It is seen that the rates of observed type 1 events increased with increasing values of q and that censoring rates increased with increasing values of u . For constant q and varying u , the ratio of observed type 1 and type 2 events remained approximately the same. For $q = 0.2$ and $q = 0.4$, we observed more events of type 2 than of type 1, and for $q = 0.8$ there were more events of type 1 than type 2. For the scenarios with $k = 10$ and $k = 15$, the observed relative frequencies were almost the same and are thus not shown.

In all scenarios, the following models were fitted to the training samples: (a) the discrete subdistribution hazard model (5) with the inverse complementary log–log function (Gompertz model), which defines the same values of $\boldsymbol{\gamma}$ as the Fine and Gray proportional subdistribution hazards model in continuous time; (b) model (5) with a logistic distribution function $h(\eta_1(t, \mathbf{x}_i)) = \exp(\eta_1(t, \mathbf{x}_i)) / (1 + \exp(\eta_1(t, \mathbf{x}_i)))$ and (c) a simple discrete hazard model (Gompertz model) for events of type 1, which does not account for the presence of competing type 2 events. In a fourth examination (d), we also fitted the discrete subdistribution hazard model as in (a), but slightly changed the data-generating process. The linear predictor $\mathbf{x}_i^\top \boldsymbol{\gamma}$ in (16) was replaced by $\sin(4x_1) + \sin(4x_2) + \gamma_3 x_3 - \gamma_4 x_4$, which defines nonlinear effects of the two standard normally distributed covariates. Finally, we considered a setting (e), where the independence assumption between T_i and C_i was violated in the training sample. For this setting, we generated the censoring times from the discrete distribution given above with parameters

$$u_i^{(e)} := \begin{cases} u + 0.25 & \text{if } T_{disc,i} < \text{median}(T_{disc}), \\ u - 0.25 & \text{if } T_{disc,i} \geq \text{median}(T_{disc}). \end{cases} \quad (17)$$

According to the data-generating process defined by (e), the probability of censoring was much higher for observations with relatively small event times than for observations with medium or large event times.

In (a) we fitted the true data-generating model, whereas in (b)–(e) the fitted models were misspecified.

6.2. Results under Correct Model Specification

The calibration plots for one randomly chosen replication of the nine simulation scenarios with $k = 5$ when fitting the true data-generating model (a) are presented in Figure 1. Using Equation (12), the appropriate number of subsets in the scenario with $q = 0.2$ and weak censoring for this example was $G = \lfloor 20.46 \rfloor$. Thus, we set $G = 20$ in all calibration plots of the simulation study. The plots show that the empirical hazards \bar{y}_g and the average predicted hazards $\bar{\hat{\lambda}}_{1g}$ coincided strongly, regardless of the degree of censoring and the rate of type 1 events. This result illustrates that the calibration plot defined in Section 4 is a strong tool for the graphical assessment of a correctly specified discrete subdistribution model. Figure 1 further shows that

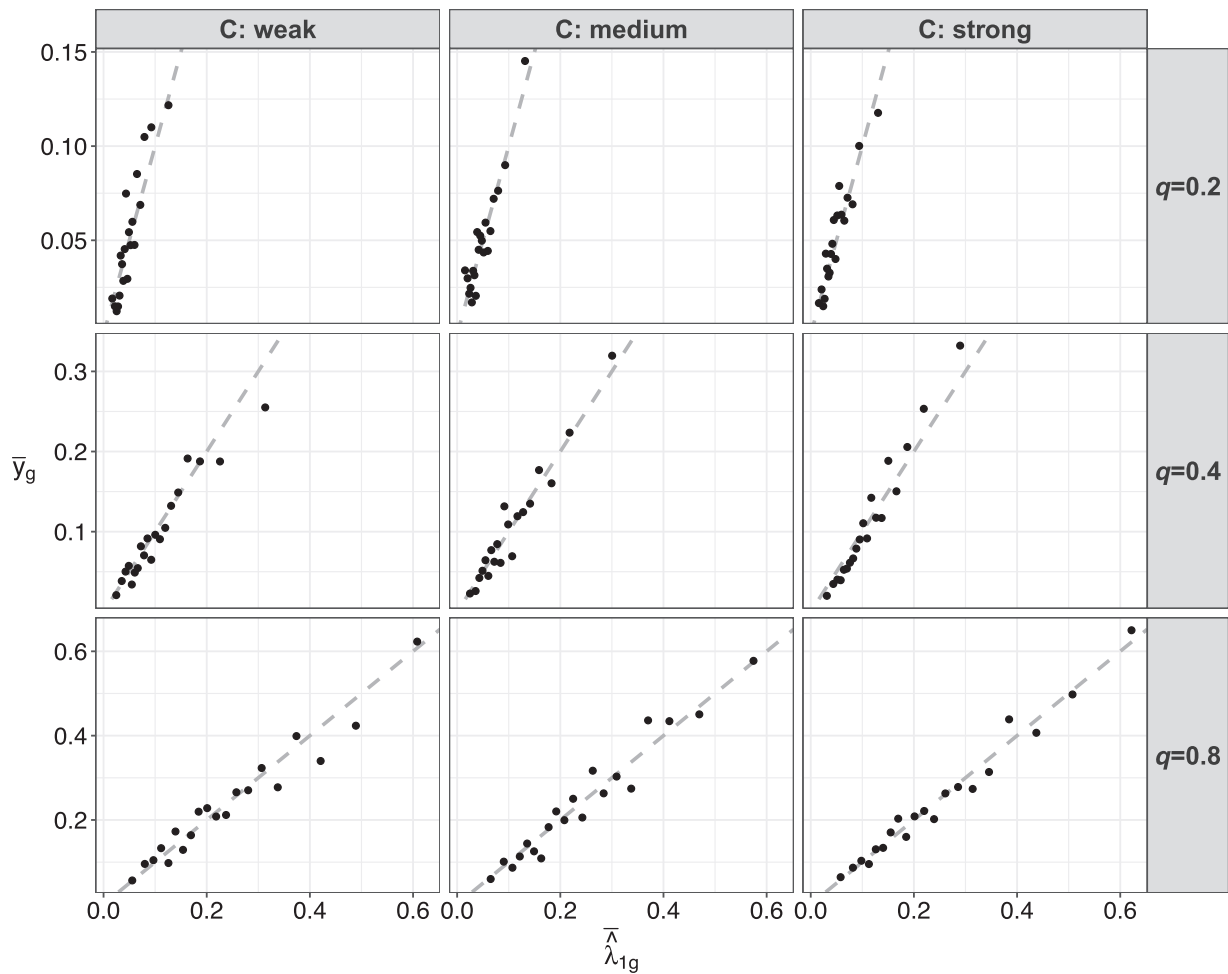


FIGURE 1: Results of the simulation study when fitting the true data-generating model. Calibration plots refer to one randomly chosen replication in each simulation scenario using $G = 20$ subsets ($k = 5$). Note that the y-axis limits differ across the rows, which is the reason why the points are not spread over the whole plots for $q = 0.2$ and $q = 0.4$. The 45-degree lines (dashed) indicate perfect calibration ($C =$ degree of censoring).

modelling the subdistribution hazard λ_1 also works well in the presence of only a relatively small number of type 1 events (for $q = 0.2$ only about 10% type 1 events were observed). When fitting the logistic recalibration model, for example to the dataset with strong censoring and $q = 0.2$ (upper right panel of Figure 1), we obtained the estimates $\hat{a} = -0.084$ and $\hat{b} = 0.957$, which are close to the values $a = 0$ and $b = 1$ of perfect calibration. Exemplary calibration plots for the scenarios with $k = 10$ and $k = 15$ are presented in Figures S2 and S3 of the Supplementary Material. Again, the plots suggest nearly perfect calibration, with the exception of the scenarios with $k = 15$ and strong censoring, where the variation and thus the deviation from the 45-degree lines is more apparent.

The estimates of the calibration parameters a and b for all scenarios with $k = 5$ are shown in Figure 2. It is seen from the boxplots that, on average, the estimates were very close to values $a = 0$ and $b = 1$. In particular, for $q = 0.8$ (lower panel) the results of the recalibration model correctly indicated nearly perfect calibration. It is also seen that the variance of the estimates of the intercept a increased with the decreasing rate of type 1 events. In contrast, the degree of censoring had only a small impact on the variance of the estimates. Figure 3 presents the corresponding P -values when conducting the recalibration tests (i)–(iii) specified in Section 5. Throughout all scenarios, the null hypotheses were kept in almost all replications (at the 5% type

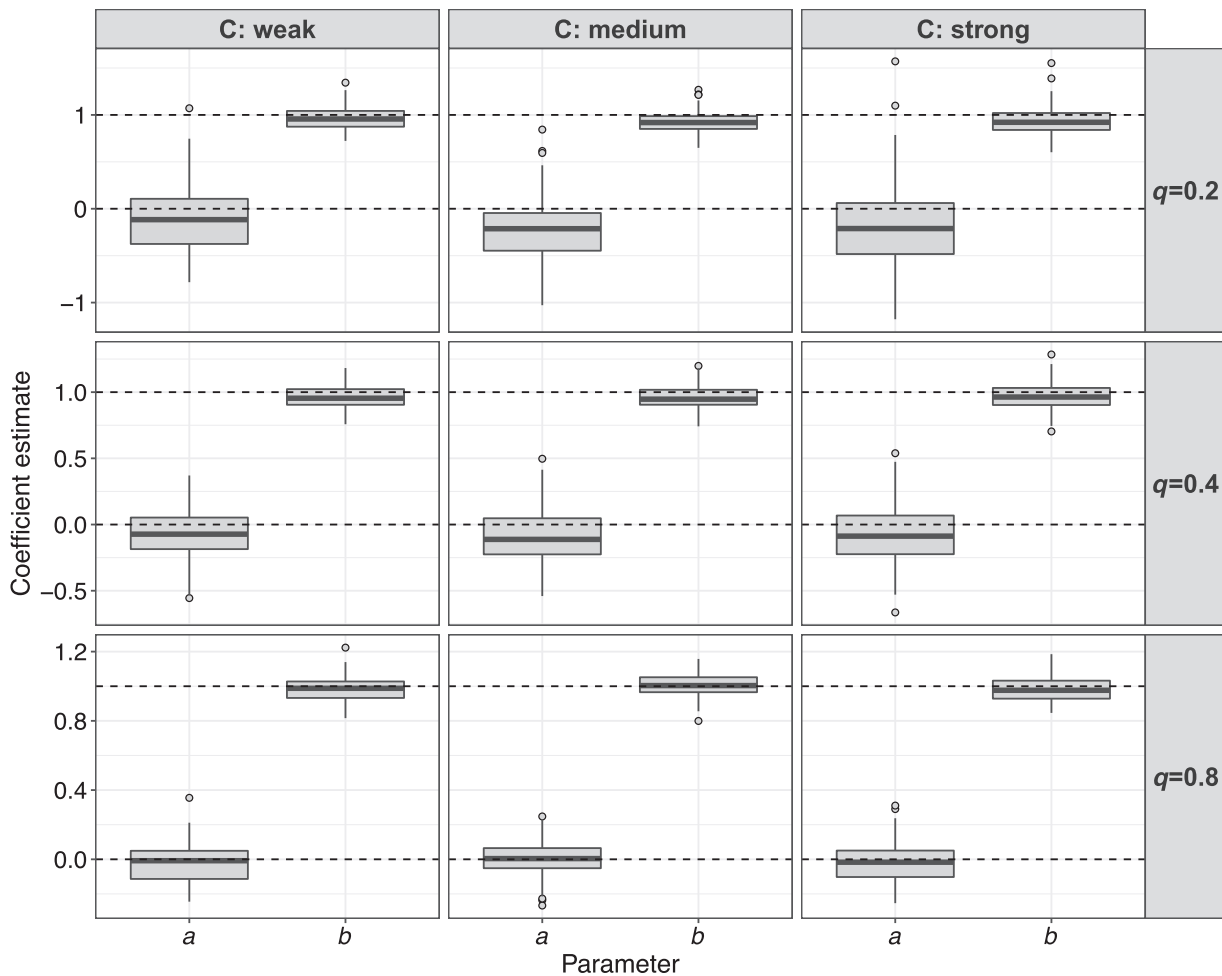


FIGURE 2: Results of the simulation study when fitting the true data-generating model. The boxplots visualize the estimates of the calibration intercepts a and calibration slopes b that were obtained from fitting the logistic recalibration model ($k = 5$).

1 error level). In particular, the tests for calibration-in-the-large given appropriate refinement (hypothesis ii) yielded very large P -values (corresponding to small negative log 10-transformed values). For example, in the scenario with $q = 0.2$ and strong censoring, this hypothesis was never rejected at the 5% type 1 error level. Overall, the results in Figures 2 and 3 illustrate that the proposed logistic recalibration model properly assessed the calibration of the fitted subdistribution hazard models, even in the case of strong censoring and a small rate of type 1 events.

The parameter estimates \hat{a} and \hat{b} and the P -values for the scenarios with $k = 10$ and $k = 15$ are given in Figures S4–S7 of the Supplementary Material. These results largely confirmed the previous findings for $k = 5$. Although the estimated calibration parameters deviated more strongly from $a = 0$ and $b = 1$, the associated null hypotheses were still kept at the 5% type 1 error level. The only exception was the scenario with $k = 15$, strong censoring, and $q = 0.2$ (upper right panel of Figure S7 of the Supplementary Material), where about half of the null hypotheses (i) and (iii) were rejected. These results, which were clearly related to the number of time intervals, can be explained by the fact that very few type 1 events were observed at later points in time when k was increased. For example, with $k = 15$, strong censoring, and $q = 0.2$, fewer than four type 1 events occurred at time points $t > 10$ in most of the training and validation samples.

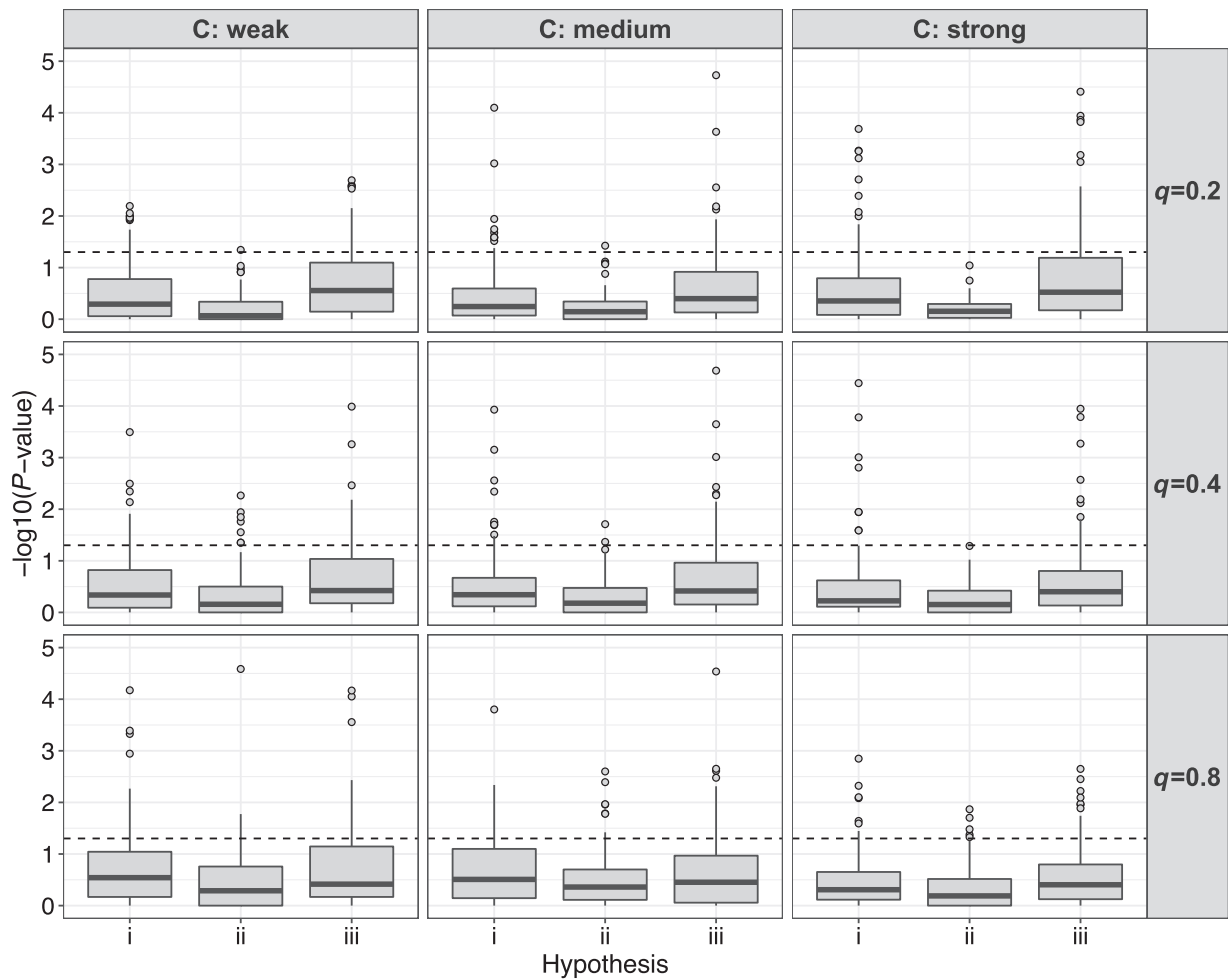


FIGURE 3: Results of the simulation study when fitting the true data-generating model. The boxplots visualize the negative log 10-transformed P -values obtained from the recalibration tests ($k = 5$). The dashed lines correspond to a P -value of 0.05. A value above the dashed line indicates a significant result at the 5% type 1 error level.

6.3. Results under Model Misspecification

In our second investigation (b), we fitted the discrete subdistribution hazard model with a logistic link function to the training samples. Regarding the calibration of the models, there was no noticeable difference to the correctly specified Gompertz model. Thus, the results were largely the same as those in Section 6.2 in all scenarios (not shown).

When fitting the simple discrete hazard model (c) to the training samples, the calibration of the models strongly deteriorated, which was clearly indicated by the proposed calibration measures. Exemplary calibration plots for the replication chosen in Figure 1 are shown in Figure 4. It is seen that, in particular, for the scenarios with weak and medium censoring, the set of points are mostly below the 45-degree line. Therefore, the predicted hazards were systematically too high. This result was also confirmed by the estimates of the recalibration intercepts a (Figure S8 of the Supplementary Material), which were all below zero. In the scenarios with a small number of type 1 events ($q = 0.2$), the mean of estimates \hat{a} were smaller than -1 . Accordingly, the associated recalibration tests of the null hypotheses (i) and (ii) were consistently rejected (Figure S11 of the Supplementary Material). Rejection of the conditional null hypothesis (ii) again confirmed a systematic shift of the predicted hazards to higher values. Only in the scenario with strong censoring and $q = 0.8$ (lower right panel of Figure S11 of the Supplementary

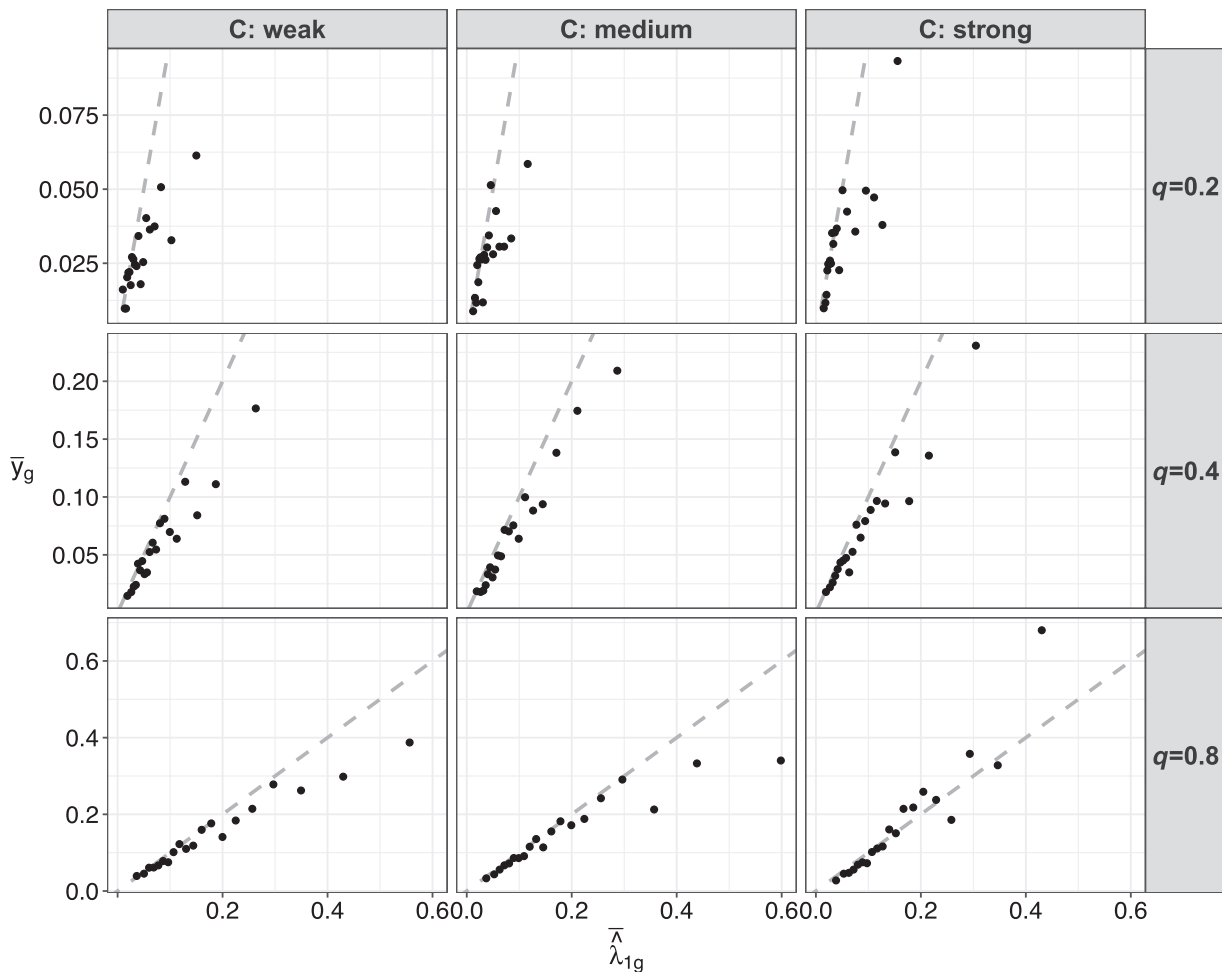


FIGURE 4: Results of the simulation study under the misspecified model (c). Calibration plots refer to one randomly chosen replication in each simulation scenario using $G = 20$ subsets ($k = 10$). The 45-degree lines (dashed) indicate perfect calibration ($C = \text{degree of censoring}$).

Material), the recalibration tests still indicated good calibration. This result is clearly related to the small number of type 2 events, which did not substantially affect the fit of the simple discrete hazard model in this scenario.

Figure 5 depicts exemplary calibration plots obtained from the model fits under the third source of misspecification (d), where the predictor of the model was falsely specified to be linear. In comparison to Figure S2 of the Supplementary Material, the points are spread considerably wider around the 45-degree line. This result was confirmed by the estimated recalibration slopes b (Figure S9 of the Supplementary Material), which were distinctly less than 1 (in particular, in the scenarios with $q = 0.2$). Remarkably, the test on null hypothesis (ii) was not affected by this form of misspecification throughout all scenarios (Figure S12 of the Supplementary Material). This demonstrates that calibration-in-the-large was still sufficient given appropriate refinement ($b = 1$). On the other hand, the two null hypotheses (i) and (iii) were more prone to the misspecification of the predictor function of the model, as they indicated poor calibration particularly in the scenarios with $q = 0.2$.

In the last setting (e) with violation of random censoring (calibration plots in Figure 6), the deviation from the 45-degree line is most evident in the scenarios with a high number of type 1 events ($q = 0.8$). This is also seen from the fits of the recalibration model (Figures S10 and S13 of the Supplementary Material), where the estimated coefficients were clearly too small and the three null hypotheses were rejected to a large extent. In contrast to the misspecification in (c),

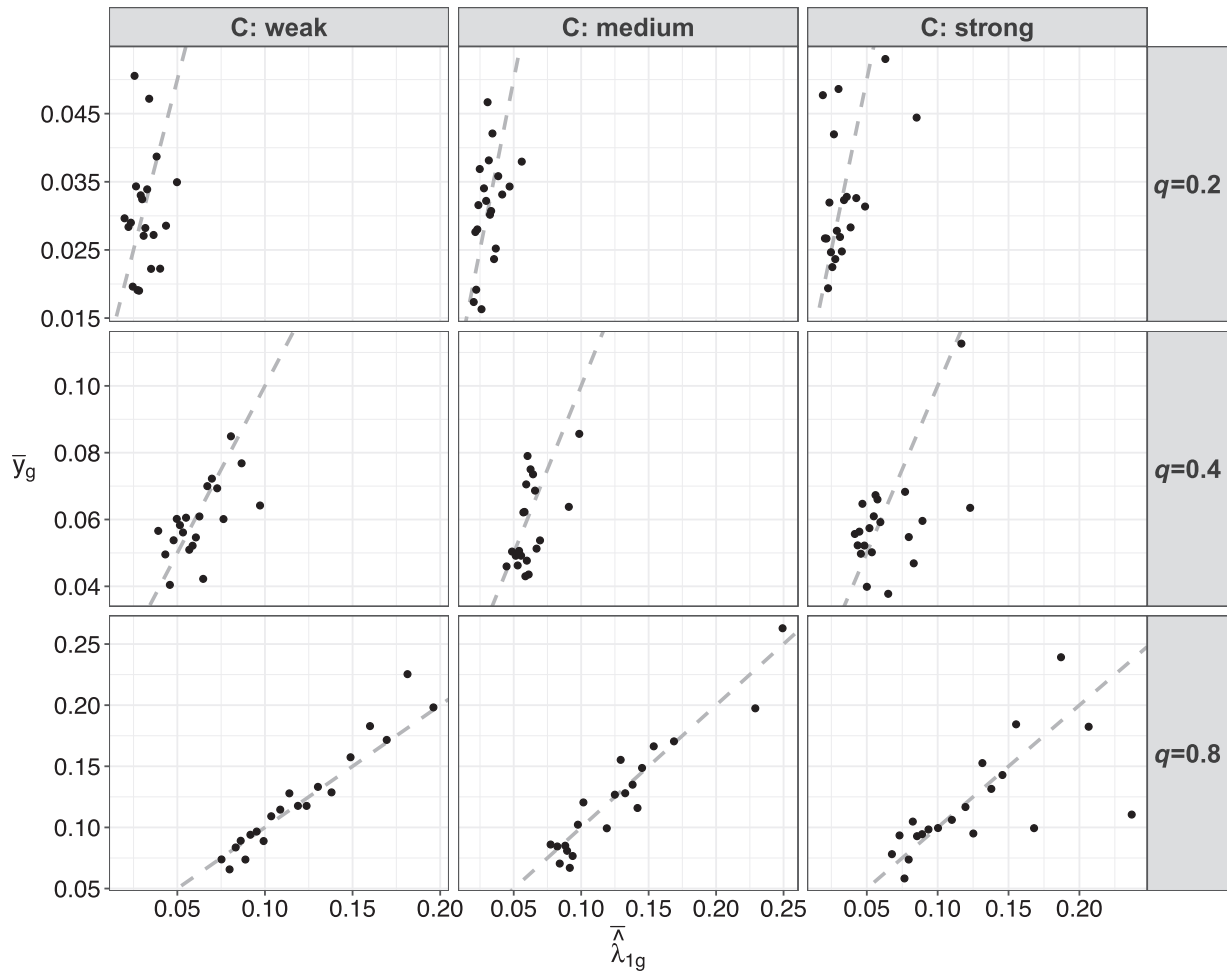


FIGURE 5: Results of the simulation study under the misspecified model (d). Calibration plots refer to one randomly chosen replication in each simulation scenario using $G = 20$ subsets ($k = 10$). The 45-degree lines (dashed) indicate perfect calibration ($C = \text{degree of censoring}$).

the scenario with strong censoring and $q = 0.8$ appeared to be most problematic. This is likely because the violation of random censoring mainly affects the estimation of \hat{V} , which is even more inaccurate in the case of a small number of type 2 events.

To sum up, the findings in cases (c)–(e) demonstrate that the proposed calibration measures are sensitive to the severity of misspecification of the fitted models. Here, calibration issues were most pronounced for models with an incorrect form of the predictor function (and weak censoring), and when the random censoring assumption was violated (and the number of type 1 events was small). Note that, because it did not gain any further insight, we reduced our considerations to the scenarios with $k = 10$ in this section.

7. NP INFECTION IN INTENSIVE CARE UNITS

To illustrate the use of the proposed calibration measures, we validated the prediction model by Berger et al. (2020) by analyzing a dataset on the development of pneumonia, which is a common nosocomial, that is, hospital-acquired infection in intensive care units (ICUs). This dataset was also considered earlier by Beyersmann et al. (2006), Wolkewitz et al. (2008), and other authors. As NP has a strong impact on the mortality of patients in ICUs, it is of high interest to determine the risk factors for the development of the disease.

The data were collected for a prospective cohort study at five ICUs in one university hospital, lasting 18 months (from February 2000 to July 2001) and comprising $n = 1876$ patients with a

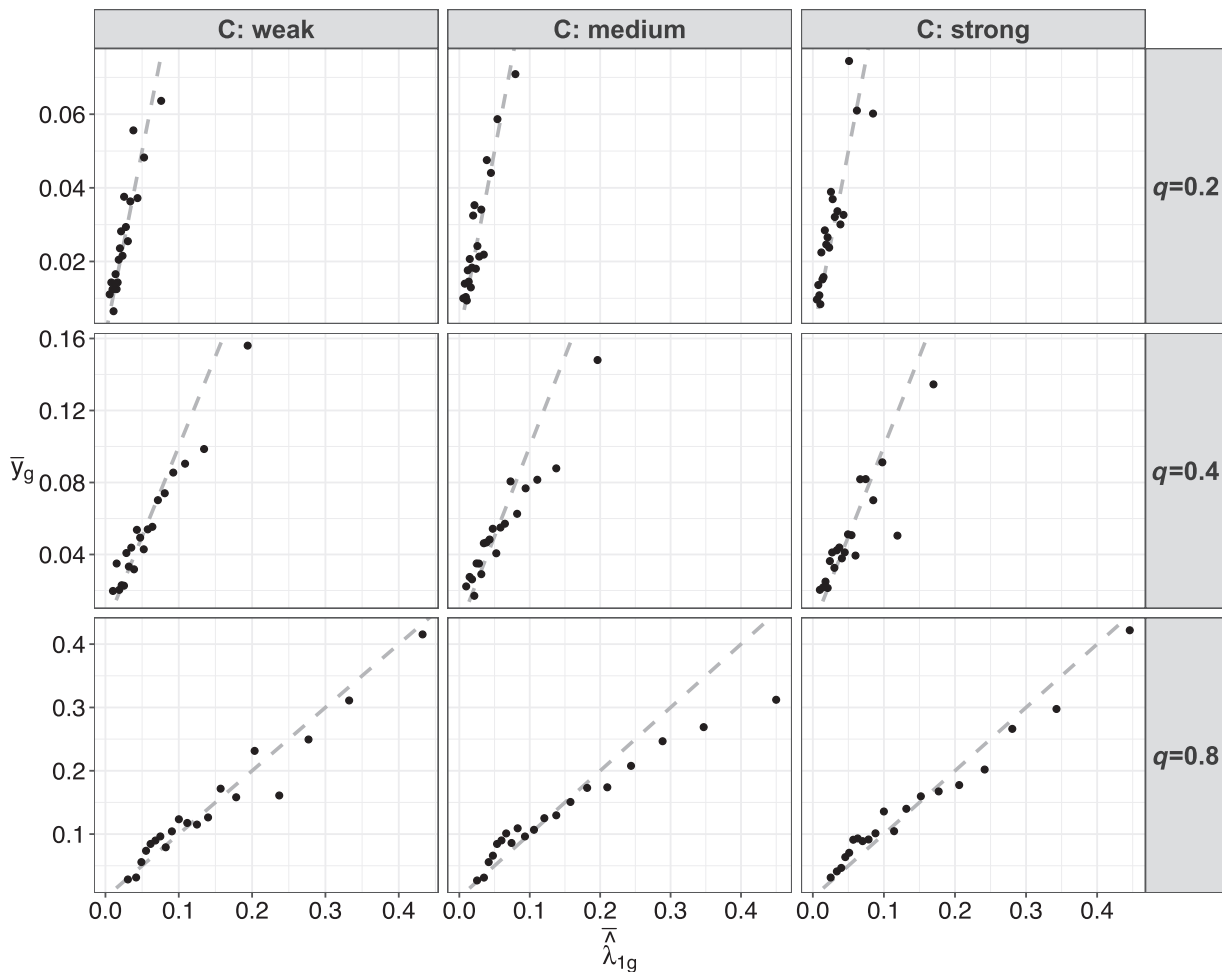


FIGURE 6: Results of the simulation study under the misspecified model (e). Calibration plots refer to one randomly chosen replication in each simulation scenario using $G = 20$ subsets ($k = 10$). The 45-degree lines (dashed) indicate perfect calibration ($C =$ degree of censoring).

duration of ICU stay of at least 2 days. The outcome of interest was the time to NP infection. Other possible events that were competing with the onset of NP (being the event of interest) were *death* and *discharge from hospital alive*. Owing to the study design, the observed event times were discrete, as they were measured on a daily basis. Berger et al. (2020) analyzed the data over a period of 60 days, resulting in 61 possible event times $t = 1, 2, \dots, 61$, where $t = k = 61$ referred to all individuals with event times ≥ 61 days. At the observed times, each patient acquired the NP infection ($n = 158$), died, was released from hospital ($n = 1695$), or was administratively censored ($n = 23$). Descriptive summary statistics of the baseline risk factors considered in the analysis were presented in Table 1 of Wolkewitz et al. (2008). In addition to the age of the patients (centred at 60 years), the gender of the patients, and the simplified acute physiology score (SAPS II), there were 11 binary risk factors characterizing the patients and their hospital stay. Note that SAPS II measures the severity of disease for patients admitted to ICUs aged 15 years or older. The score is calculated from 12 routine physiological measurements during the first 24 h, resulting in a range of $[0, 163]$ points (Le Gall, Lemeshow & Saulnier, 1993). The binary variables either referred to the time of ICU admission (*on admission*) or the time prior to ICU admission (*before admission*).

We fitted the discrete subdistribution hazard model used in Berger et al. (2020) to the data described above. This model incorporates the baseline risk factors and a set of smooth baseline coefficients represented by cubic P-splines with a second-order difference penalty (fitted using

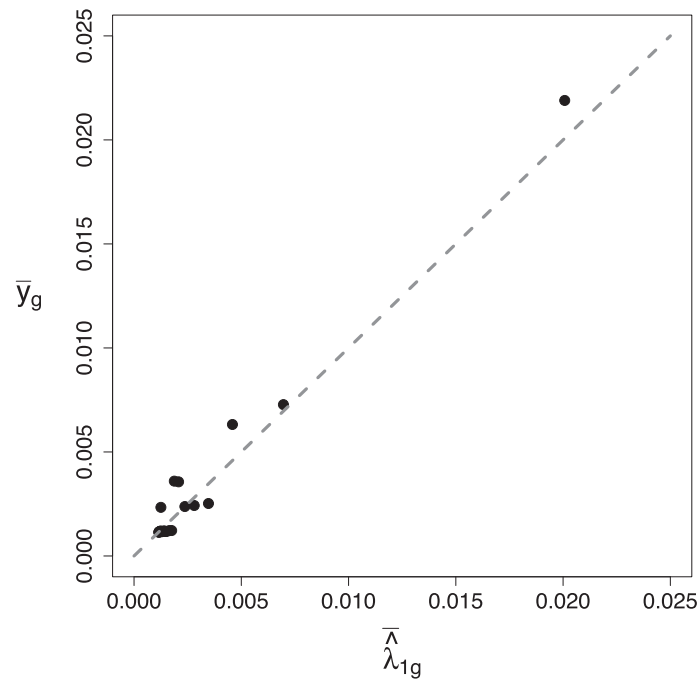


FIGURE 7: Analysis of the nosocomial pneumonia infection data. The calibration plot refers to a randomly chosen partition of the data into a training and a validation sample using $G = 24$ subsets. According to Equation (12), the appropriate number of subsets is $G = \lfloor 24.21 \rfloor$. The 45-degree line (dashed line) indicates perfect calibration.

the R package **mgcv**). This model was referred to as Model 2 in Berger et al. (2020). To assess the calibration of the model, we conducted a benchmark experiment which was based on 100 random partitions of the data. Each partition consisted of a training sample of size $n = 1500$ (80%) and a validation sample of size $N = 376$ (20%).

The main results in Berger et al. (2020) can be summarized as follows: risk factors significantly increasing the risk of NP acquisition at the 5% type I error level were (i) male gender, (ii) an intubation on admission, (iii) no pneumonia on admission, (iv) another infection on admission, (v) an elective or emergency surgery before admission and (vi) a cardio/pulmonary or neurological underlying disease.

Figure 7 presents the calibration plot of the model that was obtained for one randomly chosen partition of the data. It is seen that apart from the three subsets defined by the largest percentiles, the empirical hazards \bar{y}_g and the average predicted hazards $\bar{\lambda}_{1g}$ were very small (< 0.005). Furthermore, the plot showed strong agreement between \bar{y}_g and $\bar{\lambda}_{1g}$, indicating satisfactory calibration of the fitted model. The calibration plots obtained for 25 further partitions of the data are shown in Figure S14 of the Supplementary Material. Except single values, the plots do not reveal severe deviations from the 45-degree line. However, the bundle of small hazard values makes the evaluation of the plots rather difficult.

Boxplots of the estimated calibration parameters \hat{a} and \hat{b} and the P -values when performing the associated recalibration tests are shown in Figure 8. The estimates related to the calibration plot in Figure 7 were $\hat{a} = 0.039$ and $\hat{b} = 0.984$ with P -values 0.809 (hypothesis i), 0.518 (hypothesis ii), and 0.935 (hypothesis iii). The mean estimates of a and b (left panel of Figure 8) indicate that the predicted hazards tended to be too high ($a < 0$) and that they varied a little too much ($0 < b < 1$). Importantly, this trend was also seen in the simulations in Section 6.2 with weak censoring and $q = 0.2$ (see Figure 2 and Figures S4 and S5 of the Supplementary Material), which is the setting that is most comparable to the characteristics of the NP infection data. Also

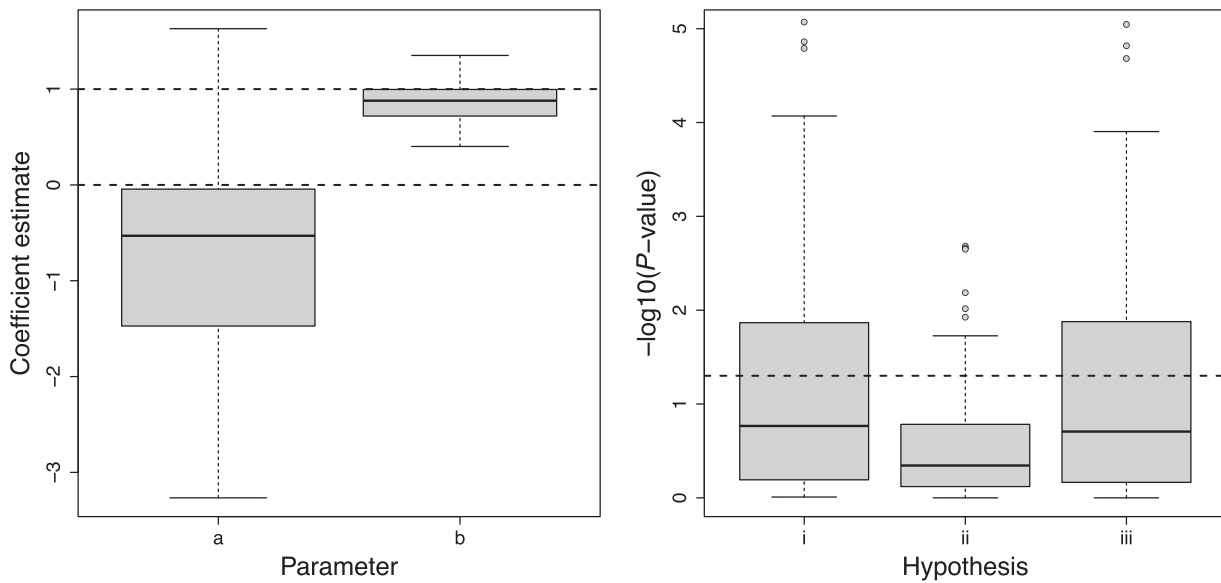


FIGURE 8: Analysis of the NP infection data. Estimates of the calibration intercepts a and calibration slopes b (left) and negative log 10-transformed P -values obtained from the recalibration tests (right) obtained for 100 partitions of the data into training and validation samples.

note that the number of observed type 1 events in the data is smaller than 3 at time points $t > 20$. According to the recalibration tests (right panel of Figure 8), the deviations of the calibration parameters from $a = 0$ and $b = 1$ were not substantial, as the majority of the null hypotheses on calibration-in-the-large and refinement were not rejected at the 5% type 1 error level. This result is again in line with the findings in the simulation study and also indicated a good calibration of the discrete subdistribution model derived in Berger et al. (2020).

8. CONCLUDING REMARKS

Discrete time-to-event models have gained widespread popularity in applied research in recent years (Tutz & Schmid, 2016; Lee, Feuer & Fine, 2018). Therefore, methodology for the proper validation of their generalization performance is increasingly necessary. In this regard, the methods presented here constitute a new set of tools to assess the calibration of discrete subdistribution hazard models for competing risks analysis. They consist of a calibration plot for graphical assessments as well as a recalibration model including tests on calibration-in-the-large and refinement. Both methods are well connected to analogous approaches for binary regression (Miller et al., 1993; Hosmer, Lemeshow & Sturdivant, 2013). In the single-event scenario, the graphical tool presented here naturally reduces to the calibration plot proposed in Berger & Schmid (2018).

Unlike Heyard et al. (2020), who proposed tools to assess the calibration of cause-specific hazard models, we considered the subdistribution framework originally proposed by Fine & Gray (1999) for competing risks data in continuous time. In contrast to cause-specific hazard modelling, this approach has the advantage that only one model needs to be considered if the interest is in the occurrence of one specific event. Subdistribution hazard modelling is of high practical importance, as it allows the interpretation of regression coefficients in terms of increasing/decreasing effects of the covariates on the incidence of the target event (Austin & Fine, 2017). Furthermore, Young et al. (2020) suggested using differences in the cumulative incidence functions as estimands in causal modelling. To evaluate the calibration of cumulative incidence functions, Lee (2017) generated an alternative kind of calibration plot that compared predictions of the cumulative incidence function to their respective nonparametric estimates.

The simulation study and the analysis of the NP infection data suggest that the methods work well under both correct and incorrect model specifications, even in “unfavourable” scenarios with a high censoring rate and few type 1 events. However, one should be careful in situations with a large number of time intervals when the observed number of type 1 events at later time points is rare.

All evaluations presented in this article were performed using the R add-on package **discSurv** (Welchowski & Schmid, 2019). It contains the function `dataLongSubDist()` to generate the binary outcome vectors (8) and the corresponding weights (9) and (10). Parameter estimates of the recalibration model were obtained by using the function `glm()` with the family function `binomial()` for logistic regression.

ACKNOWLEDGEMENTS

We thank Jan Beyersmann for fruitful discussions on subdistribution hazard modelling and for helpful suggestions on how to improve the manuscript. We thank the SIR-3 study investigators for providing us with the NP infection data. Support by the German Research Foundation (DFG), grant SCHM 2966/2-1, is gratefully acknowledged.

REFERENCES

- Alba, A. C., Agoritsas, T., Walsh, M., Hanna, S., Iorioand, A., Devereaux, P. J., McGinn, T., & Guyatt, G. (2017). Discrimination and calibration of clinical prediction models: Users’ guides to the medical literature. *Journal of the American Medical Association*, 318, 1377–1384.
- Andersen, P. K., Geskus, R. B., de Witte, T., & Putter, H. (2012). Competing risks in epidemiology: Possibilities and pitfalls. *International Journal of Epidemiology*, 41, 861–870.
- Austin, P. C., Lee, D. S., & Fine, J. P. (2016). Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133, 601–609.
- Austin, P. C. & Fine, J. P. (2017). Practical recommendations for reporting Fine–Gray model analyses for competing risk data. *Statistics in Medicine*, 36, 4391–4400.
- Baker, S. G., Cook, N. R., Vickers, A., & Kramer, B. S. (2009). Using relative utility curves to evaluate risk prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172, 729–748.
- Berger, M. & Schmid, M. (2018). Semiparametric regression for discrete time-to-event data. *Statistical Modelling*, 18, 322–345.
- Berger, M., Schmid, M., Welchowski, T., Schmitz-Valckenberg, S., & Beyersmann, J. (2020). Subdistribution hazard models for competing risks in discrete time. *Biostatistics*, 21, 449–466, <https://doi.org/10.1093/biostatistics/kxy069>.
- Beyersmann, J., Gastmeier, P., Grundmann, H., Bärwolff, S., Geffers, C., Behnke, M., Rüden, H., & Schumacher, M. (2006). Use of multistate models to assess prolongation of intensive care unit stay due to nosocomial infection. *Infection Control & Hospital Epidemiology*, 27, 493–499.
- Beyersmann, J., Allignol, A., & Schumacher, M. (2011). *Competing Risks and Multistate Models with R*. Springer, New York.
- Braun, D., Gorfine, M., Katki, H. A., Ziogas, A., & Parmigiani, G. (2018). Nonparametric adjustment for measurement error in time-to-event data: Application to risk prediction models. *Journal of the American Statistical Association*, 113, 14–25.
- Cortese, G. & Andersen, P. K. (2010). Competing risks and time-dependent covariates. *Biometrical Journal*, 52, 138–158.
- Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, 45, 562–565.
- Ding, A. A., Tian, S., Yu, Y., & Guo, H. (2012). A class of discrete transformation survival models with application to default probability prediction. *Journal of the American Statistical Association*, 107, 990–1003.
- Doane, D. P. (1976). Aesthetic frequency classifications. *The American Statistician*, 30, 181–183.
- Fahrmeir, L. & Wagenpfeil, S. (1996). Smoothing hazard functions and time-varying effects in discrete duration and competing risks models. *Journal of the American Statistical Association*, 91, 1584–1594.
- Fine, J. P. & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94, 496–509.

- Gibbs, M. (2011). Ecological risk assessment, prediction, and assessing risk predictions. *Risk Analysis: An International Journal*, 31, 1784–1788.
- Harrell, F. E. (2015). *Regression Modeling Strategies*. Springer, New York.
- Henderson, R. & Keiding, N. (2005). Individual survival time prediction using statistical models. *Journal of Medical Ethics*, 31, 703–706.
- Heyard, R., Timsit, J.-F., Held, L., & COMBACTE-MAGNET Consortium. (2020). Validation of discrete time-to-event prediction models in the presence of competing risks. *Biometrical Journal*, 62, 643–657.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. Hoboken, NJ, John Wiley & Sons.
- Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed., Wiley, Hoboken.
- Kerr, K. F. & Janes, H. (2017). First things first: Risk model performance metrics should reflect the clinical application. *Statistics in Medicine*, 36, 4503–4508.
- Klein, J. P. & Andersen, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*, 61, 223–229.
- Lau, B., Cole, S. R., & Gange, S. J. (2009). Competing risk regression models for epidemiologic data. *American Journal of Epidemiology*, 170, 244–256.
- Le Gall, J.-R., Lemeshow, S., & Saulnier, F. (1993). A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *Journal of the American Medical Association*, 270, 2957–2963.
- Lee, M. (2017). Inference for cumulative incidence on discrete failure times with competing risks. *Journal of Statistical Computation and Simulation*, 87, 1989–2001.
- Lee, M., Feuer, E. J., & Fine, J. P. (2018). On the analysis of discrete time competing risks data. *Biometrics*, 74, 1468–1481.
- Liu, D., Zheng, Y., Prentice, R. L., & Hsu, L. (2014). Estimating risk with time-to-event data: An application to the women’s health initiative. *Journal of the American Statistical Association*, 109, 514–524.
- Miller, M. E., Langefeld, C. D., Tierney, W. M., Hui, S. L., & McDonald, C. J. (1993). Validation of probabilistic predictions. *Medical Decision Making*, 13, 49–57.
- Moons, K. G. M., Kengne, A. P., Grobbee, D. E., Royston, P., Vergouwe, Y., Altman, D. G., & Woodward, M. (2012a). Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*, 98, 691–698.
- Moons, K. G. M., Kengne, A. P., Woodward, M., Royston, P., Vergouwe, Y., Altman, D. G., & Grobbee, D. E. (2012b). Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart*, 98, 683–690.
- Nightingale, F. (1863). *Notes on Hospitals*. Longman, Green, Longman, Roberts, and Green, London.
- Poguntke, I., Schumacher, M., Beyersmann, J., & Wolkewitz, M. (2018). Simulation shows undesirable results for competing risks analysis with time-dependent covariates for clinical outcomes. *BMC Medical Research Methodology*, 18, 79.
- Schmid, M. & Berger, M. (2020). Competing risks analysis for discrete time-to-event data. *Wiley Interdisciplinary Reviews: Computational Statistics*, e1529. <https://doi.org/10.1002/wics.1529>.
- Schmid, M., Tutz, G., & Welchowski, T. (2018). Discrimination measures for discrete time-to-event predictions. *Econometrics and Statistics*, 7, 153–164.
- Soave, D. M. & Strug, L. J. (2018). Testing calibration of Cox survival models at extremes of event risk. *Frontiers in Genetics*, 9, 177.
- Steyerberg, E. W. (2019). *Clinical Prediction Models*, 2nd ed., Springer, New York.
- Steyerberg, E. W. & Harrell, F. E. (2016). Prediction models need appropriate internal, internal–external, and external validation. *Journal of Clinical Epidemiology*, 69, 245–247.
- Steyerberg, E. W. & Vergouwe, Y. (2014). Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *European Heart Journal*, 35, 1925–1931.
- Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology*, 21, 128.
- Tutz, G. & Schmid, M. (2016). *Modeling Discrete Time-to-Event Data*. Springer, New York.
- Vickers, A. J. & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26, 565–574.

- Vickers, A. J., Van Calster, B., & Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*, 352, i6.
- Welchowski, T. & Schmid, M. (2019). *discSurv: Discrete Time Survival Analysis*. R package version 1.4.1. <http://cran.r-project.org/web/packages/discSurv>.
- Witten, D. M. & Tibshirani, R. (2010). Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research*, 19, 29–51.
- Wolkewitz, M., Vonberg, R. P., Grundmann, H., Beyersmann, J., Gastmeier, P., Bärwolff, S., Geffers, C., Behnke, M., Rüden, H., & Schumacher, M. (2008). Risk factors for the development of nosocomial pneumonia and mortality on intensive care units: Application of competing risks models. *Critical Care*, 12, R44.
- Young, J. G., Stensrud, M. J., Tchetgen Tchetgen, E. J., & Hernán, M. A. (2020). A causal framework for classical statistical estimands in failure-time settings with competing events. *Statistics in Medicine*, 39, 1199–1236.
-

Received 6 August 2020

Accepted 22 December 2020

Assessing the Calibration of Subdistribution Hazard Models in Discrete Time

– Supplementary Material –

Moritz Berger^{1*} and Matthias Schmid¹

¹*Institute of Medical Biometry, Informatics and Epidemiology, Faculty of Medicine, University of Bonn, Venusberg-Campus 1, D-53127 Bonn, Germany*

1. LOG-LIKELIHOOD OF THE RECALIBRATION MODEL (12)

To derive the log-likelihood of the logistic recalibration model for the discrete subdistribution hazard model, it is assumed that $a = 0$, hence the predictor reduces to

$$\eta_{rc}(t|\mathbf{x}_i) = b \log \left(\hat{\lambda}_1(t|\mathbf{x}_i) / (1 - \hat{\lambda}_1(t|\mathbf{x}_i)) \right) \quad (1)$$

and

$$\begin{aligned} \ell_{rc} &= \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} \{ y_{it} \log(\pi_1(t|\mathbf{x}_i)) + (1 - y_{it}) \log(1 - \pi_1(t|\mathbf{x}_i)) \} \\ &= \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} y_{it} \log \left(\frac{\exp \left(b \log \left(\frac{\hat{\lambda}_1(t|\mathbf{x}_i)}{1 - \hat{\lambda}_1(t|\mathbf{x}_i)} \right) \right)}{1 + \exp \left(b \log \left(\frac{\hat{\lambda}_1(t|\mathbf{x}_i)}{1 - \hat{\lambda}_1(t|\mathbf{x}_i)} \right) \right)} \right) \\ &\quad + \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} (1 - y_{it}) \log \left(1 - \frac{\exp \left(b \log \left(\frac{\hat{\lambda}_1(t|\mathbf{x}_i)}{1 - \hat{\lambda}_1(t|\mathbf{x}_i)} \right) \right)}{1 + \exp \left(b \log \left(\frac{\hat{\lambda}_1(t|\mathbf{x}_i)}{1 - \hat{\lambda}_1(t|\mathbf{x}_i)} \right) \right)} \right) \\ &= \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} y_{it} \log \left(\frac{\frac{\hat{\lambda}_1(t|\mathbf{x}_i)^b}{(1 - \hat{\lambda}_1(t|\mathbf{x}_i))^b}}{1 + \frac{\hat{\lambda}_1(t|\mathbf{x}_i)^b}{(1 - \hat{\lambda}_1(t|\mathbf{x}_i))^b}} \right) \\ &\quad + \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} (1 - y_{it}) \log \left(\frac{1}{1 + \frac{\hat{\lambda}_1(t|\mathbf{x}_i)^b}{(1 - \hat{\lambda}_1(t|\mathbf{x}_i))^b}} \right) \end{aligned}$$

* Author to whom correspondence may be addressed.
 E-mail: moritz.berger@imbie.uni-bonn.de

$$\begin{aligned}
&= \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} y_{it} \log \left(\frac{\hat{\lambda}_1(t|\mathbf{x}_i)^b}{(1 - \hat{\lambda}_1(t|\mathbf{x}_i))^b} \right) \\
&\quad + \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} \log \left(1 + \frac{\hat{\lambda}_1(t|\mathbf{x}_i)^b}{(1 - \hat{\lambda}_1(t|\mathbf{x}_i))^b} \right) \\
&= b \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} y_{it} \log \left(\hat{\lambda}_1(t|\mathbf{x}_i) \right) - b \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} y_{it} \log \left(1 - \hat{\lambda}_1(t|\mathbf{x}_i) \right) \\
&\quad - \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} \log \left(\hat{\lambda}_1(t|\mathbf{x}_i)^b + (1 - \hat{\lambda}_1(t|\mathbf{x}_i))^b \right) \\
&\quad + b \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} \log \left(1 - \hat{\lambda}_1(t|\mathbf{x}_i) \right) \\
&= b \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} y_{it} \log \left(\hat{\lambda}_1(t|\mathbf{x}_i) \right) + b \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} (1 - y_{it}) \log \left(1 - \hat{\lambda}_1(t|\mathbf{x}_i) \right) \\
&\quad - \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} \log \left(\hat{\lambda}_1(t|\mathbf{x}_i)^b + (1 - \hat{\lambda}_1(t|\mathbf{x}_i))^b \right). \tag{2}
\end{aligned}$$

2. FURTHER NUMERICAL RESULTS

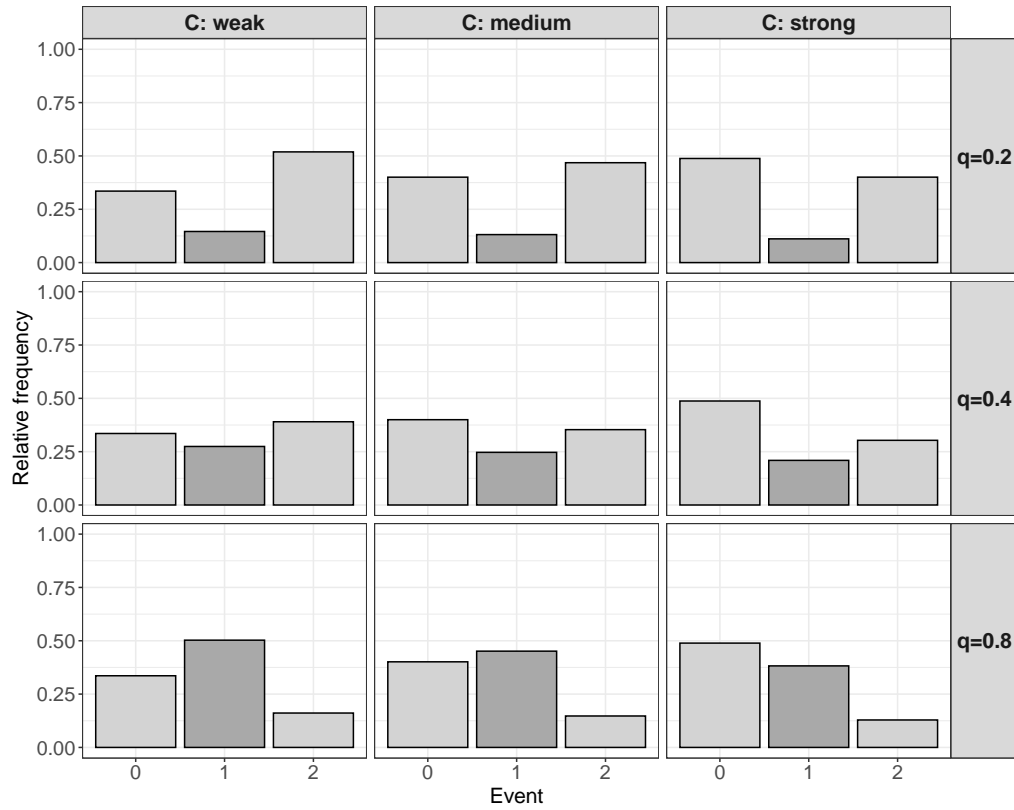


Figure S1: Illustration of the experimental design of the simulation study. The bars display the mean relative frequencies of observed events (0 = censoring event, 1 = event of interest, 2 = competing event) that were obtained from 100 simulated data sets ($k = 5$). The ratio of type 1 and type 2 events was approximately the same in each row (C = degree of censoring).

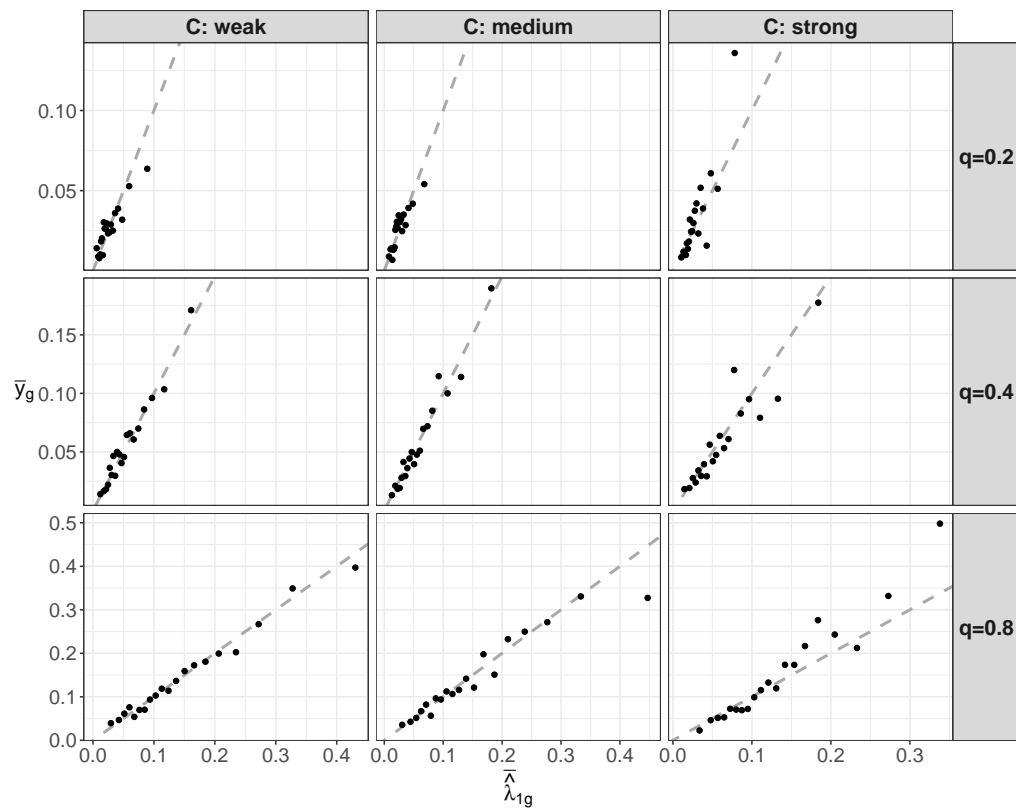


Figure S2: Results of the simulation study when fitting the true data-generating model. Calibration plots refer to one randomly chosen replication in each simulation scenario using $G = 20$ subsets ($k = 10$). The 45-degree lines (dashed) indicate perfect calibration ($C = \text{degree of censoring}$).

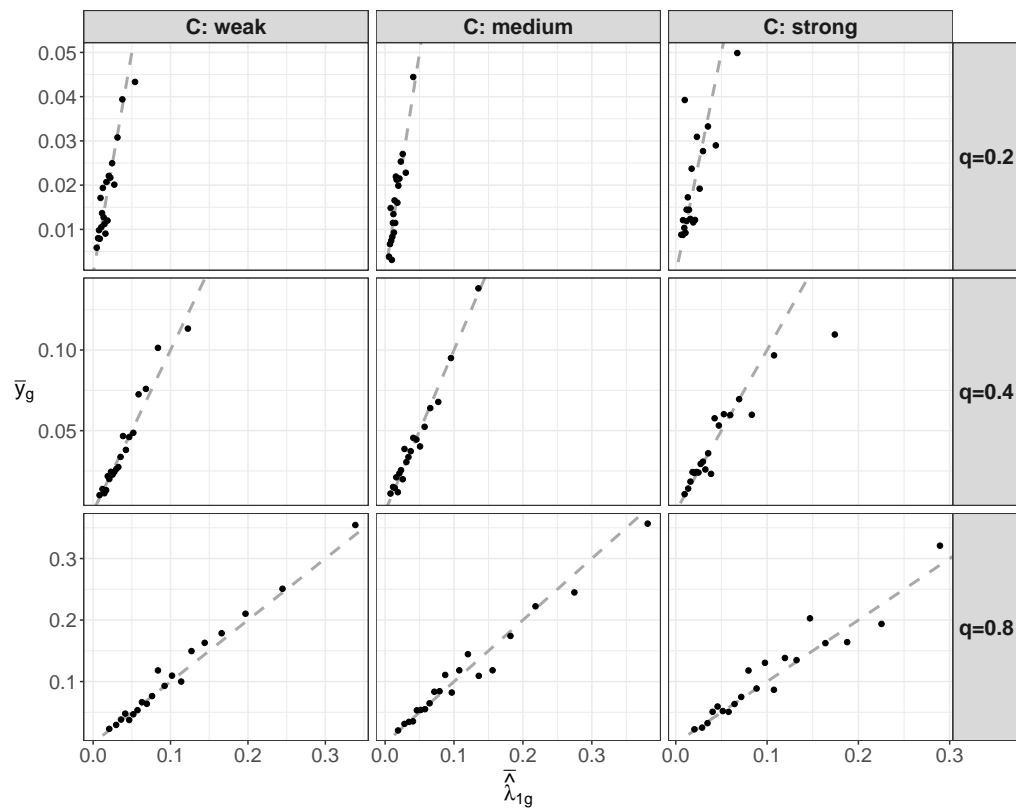


Figure S3: Results of the simulation study when fitting the true data-generating model. Calibration plots refer to one randomly chosen replication in each simulation scenario using $G = 20$ subsets ($k = 15$). The 45-degree lines (dashed) indicate perfect calibration ($C = \text{degree of censoring}$).

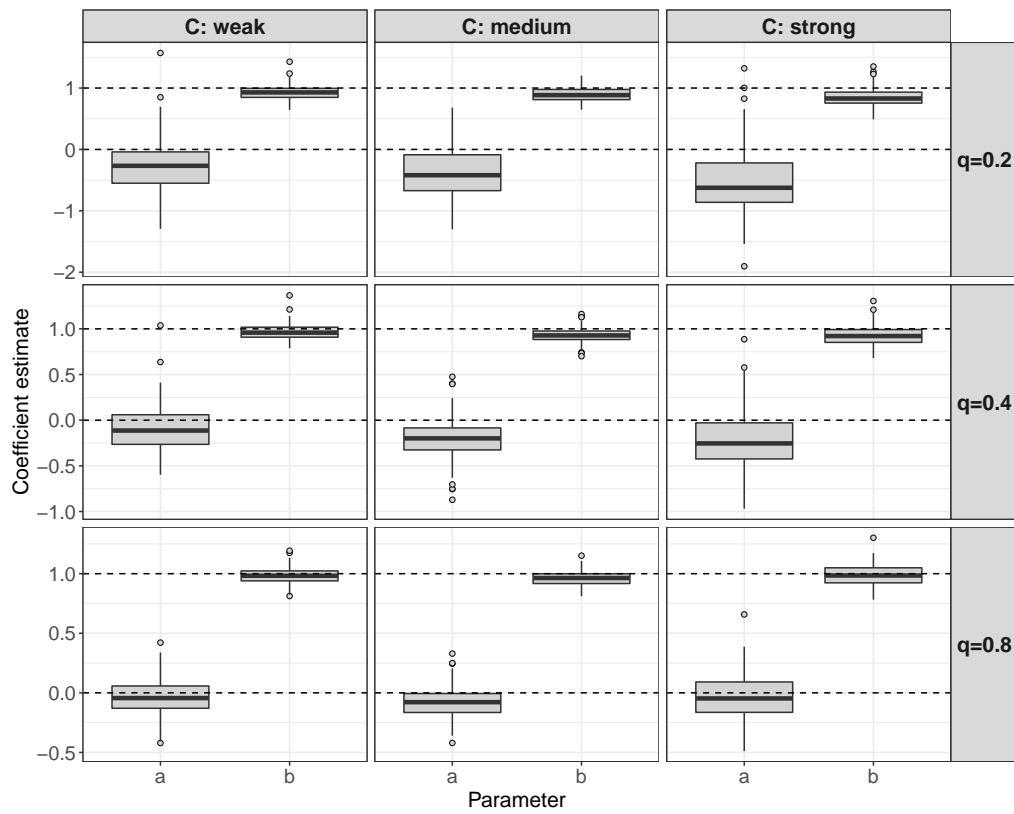


Figure S4: Results of the simulation study when fitting the true data-generating model. The boxplots visualize the estimates of the calibration intercepts a and calibration slopes b that were obtained from fitting the logistic recalibration model ($k = 10$).

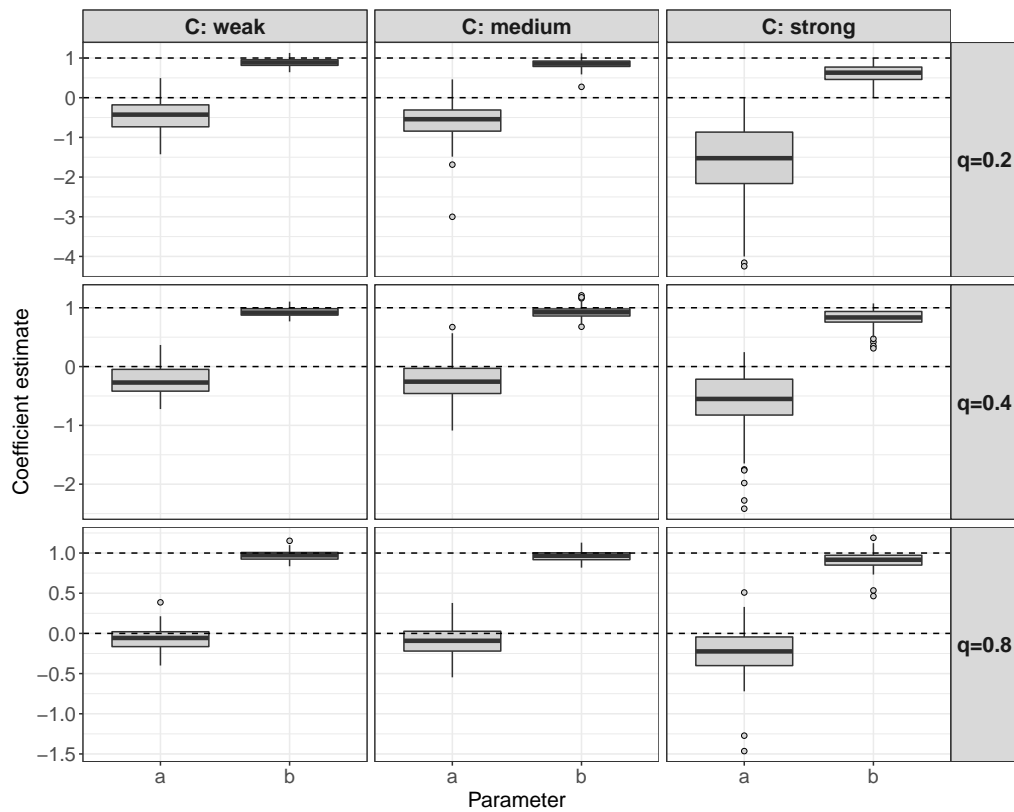


Figure S5: Results of the simulation study when fitting the true data-generating model. The boxplots visualize the estimates of the calibration intercepts a and calibration slopes b that were obtained from fitting the logistic recalibration model ($k = 15$).

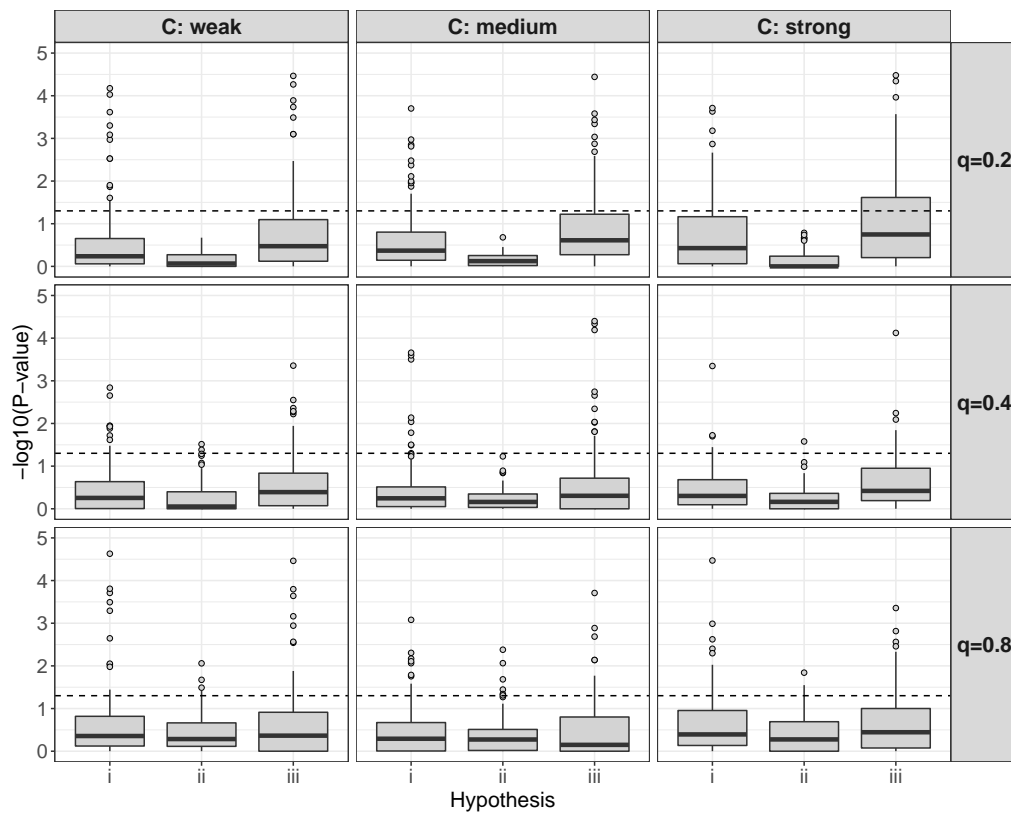


Figure S6: Results of the simulation study when fitting the true data-generating model. The boxplots visualize the negative log₁₀-transformed p -values obtained from the recalibration tests ($k = 10$). The dashed lines correspond to a p -value of 0.05. A value above the dashed line indicates a significant result at the 5% type 1 error level.

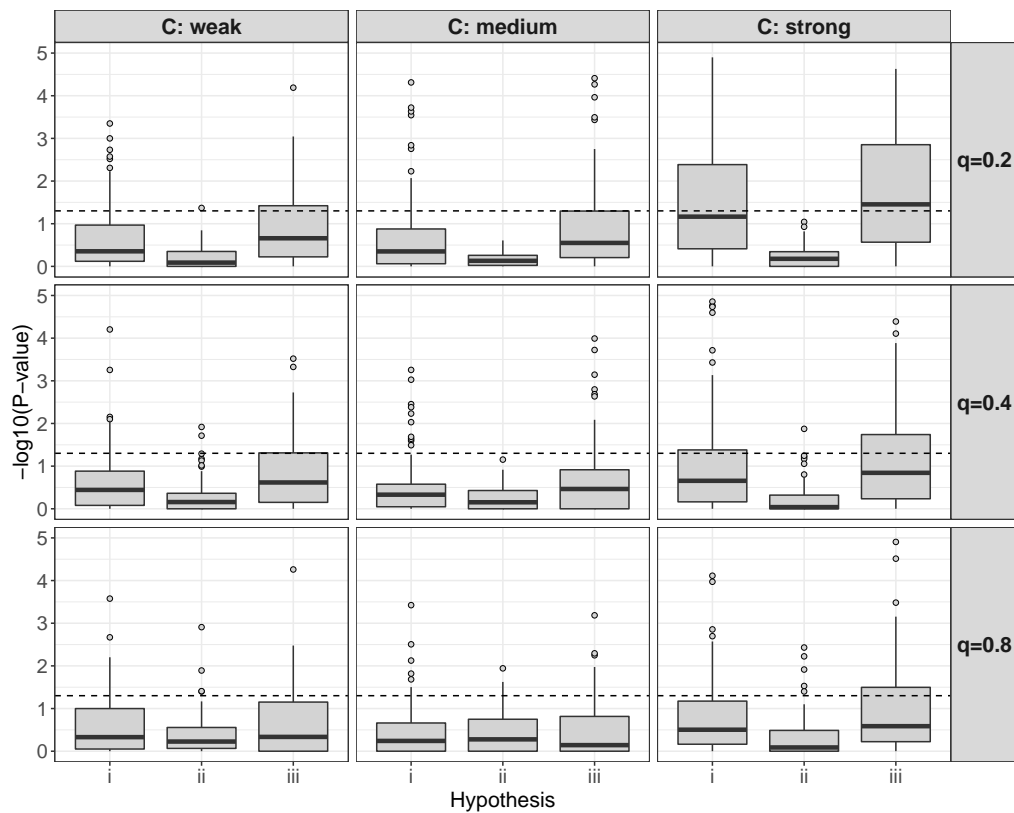


Figure S7: Results of the simulation study when fitting the true data-generating model. The boxplots visualize the negative log₁₀-transformed p -values obtained from the recalibration tests ($k = 15$). The dashed lines correspond to a p -value of 0.05. A value above the dashed line indicates a significant result at the 5% type 1 error level.

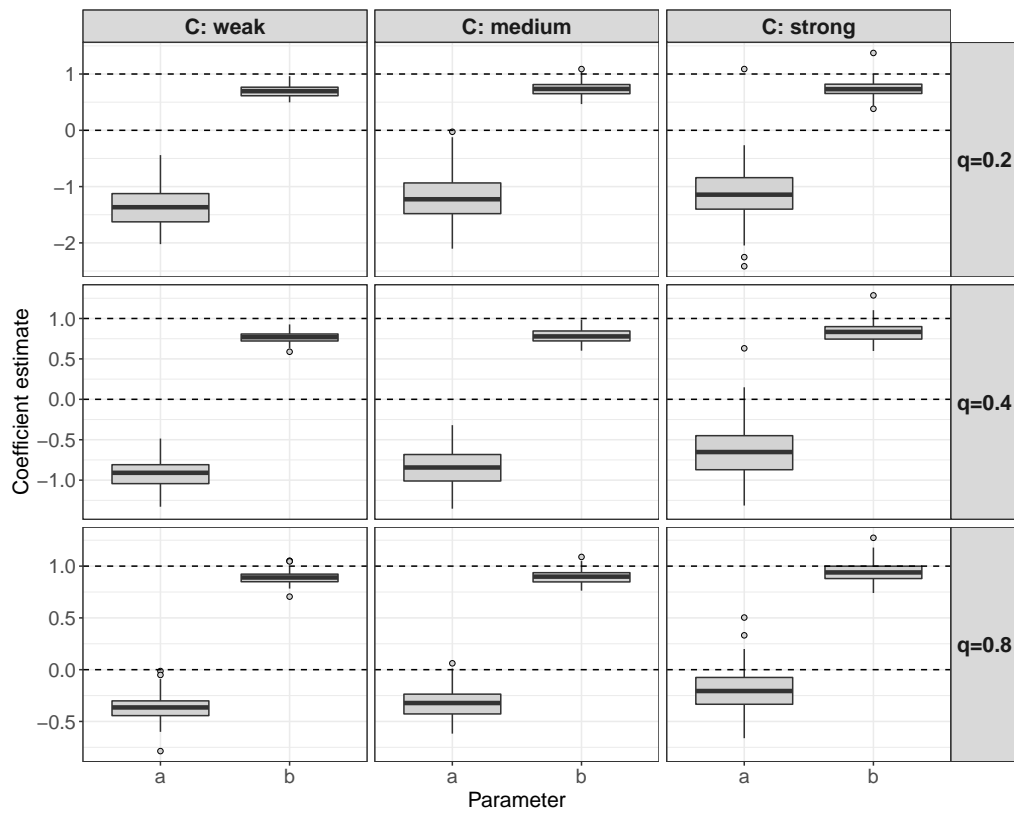


Figure S8: Results of the simulation study under the misspecified model (c). The boxplots visualize the estimates of the calibration intercepts a and calibration slopes b that were obtained from fitting the logistic recalibration model ($k = 10$).

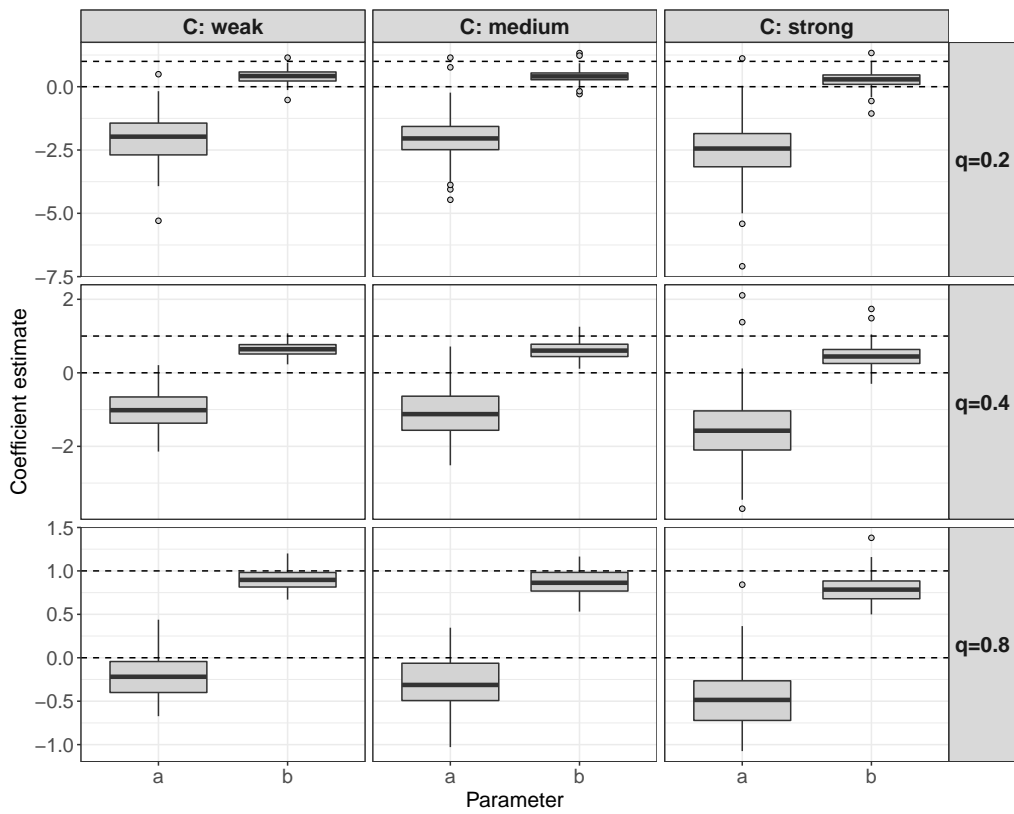


Figure S9: Results of the simulation study under the misspecified model (d). The boxplots visualize the estimates of the calibration intercepts a and calibration slopes b that were obtained from fitting the logistic recalibration model ($k = 10$).

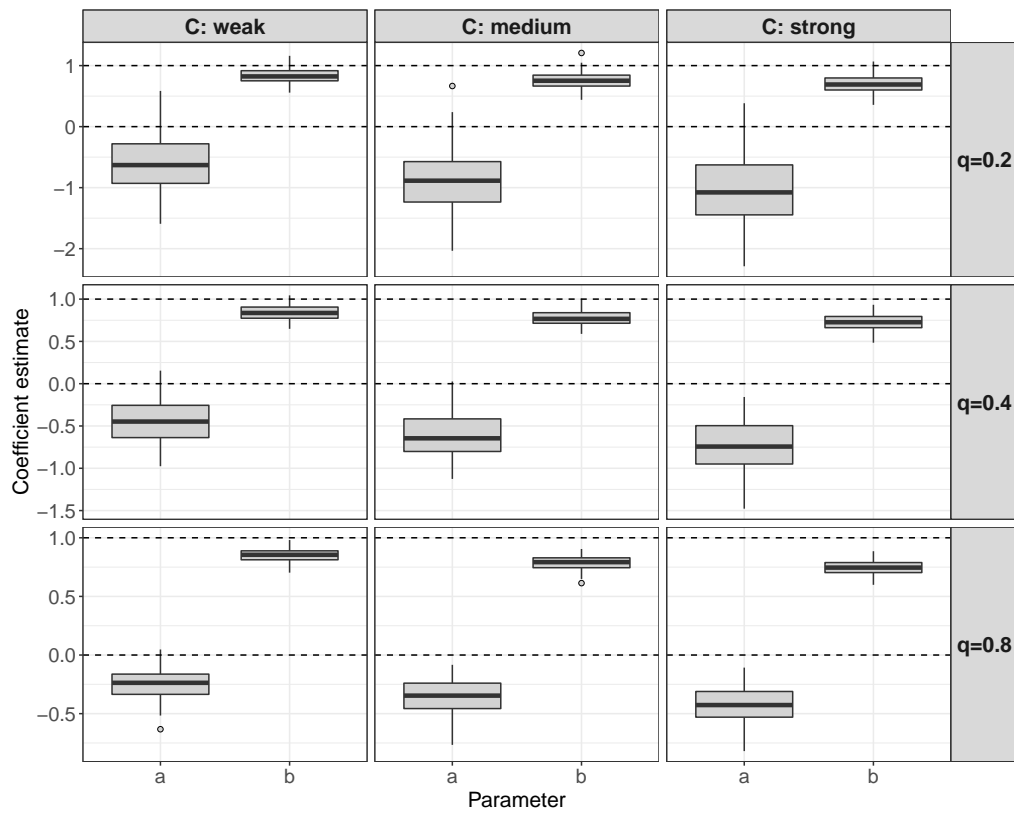


Figure S10: Results of the simulation study under the misspecified model (e). The boxplots visualize the estimates of the calibration intercepts a and calibration slopes b that were obtained from fitting the logistic recalibration model ($k = 10$).

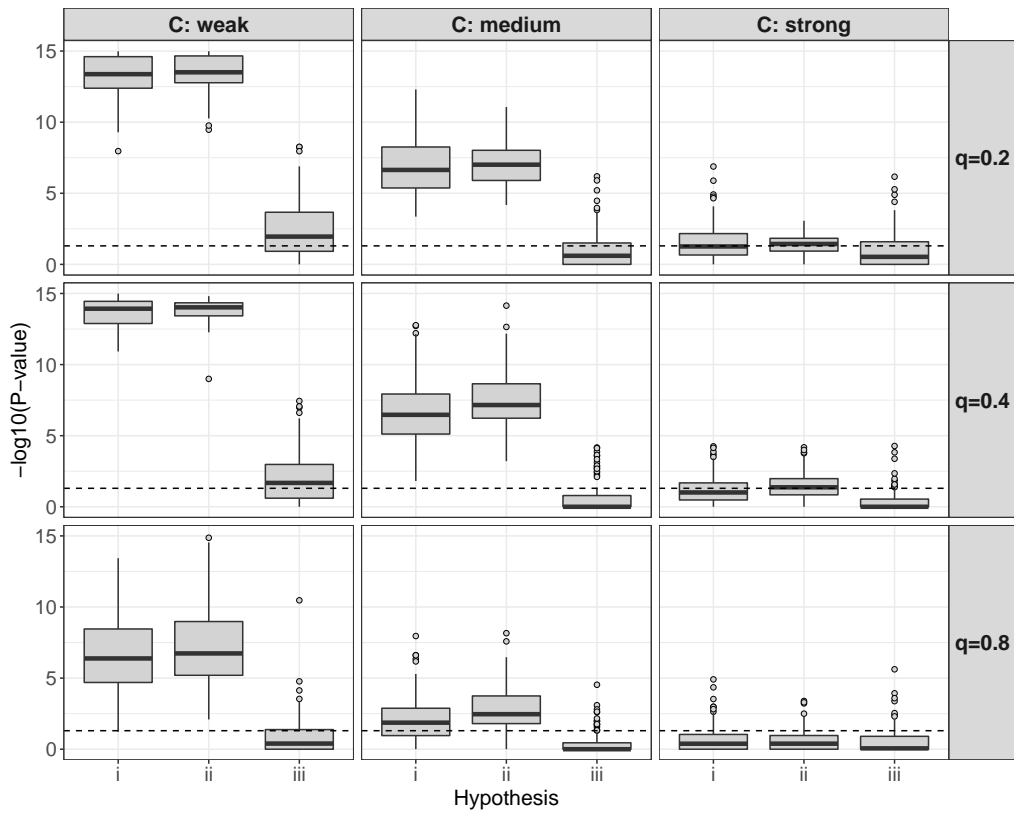


Figure S11: Results of the simulation study under the misspecified model (c). The boxplots visualize the negative log₁₀-transformed *p*-values obtained from the recalibration tests (*k* = 10). The dashed lines correspond to a *p*-value of 0.05. A value above the dashed line indicates a significant result at the 5% type 1 error level.

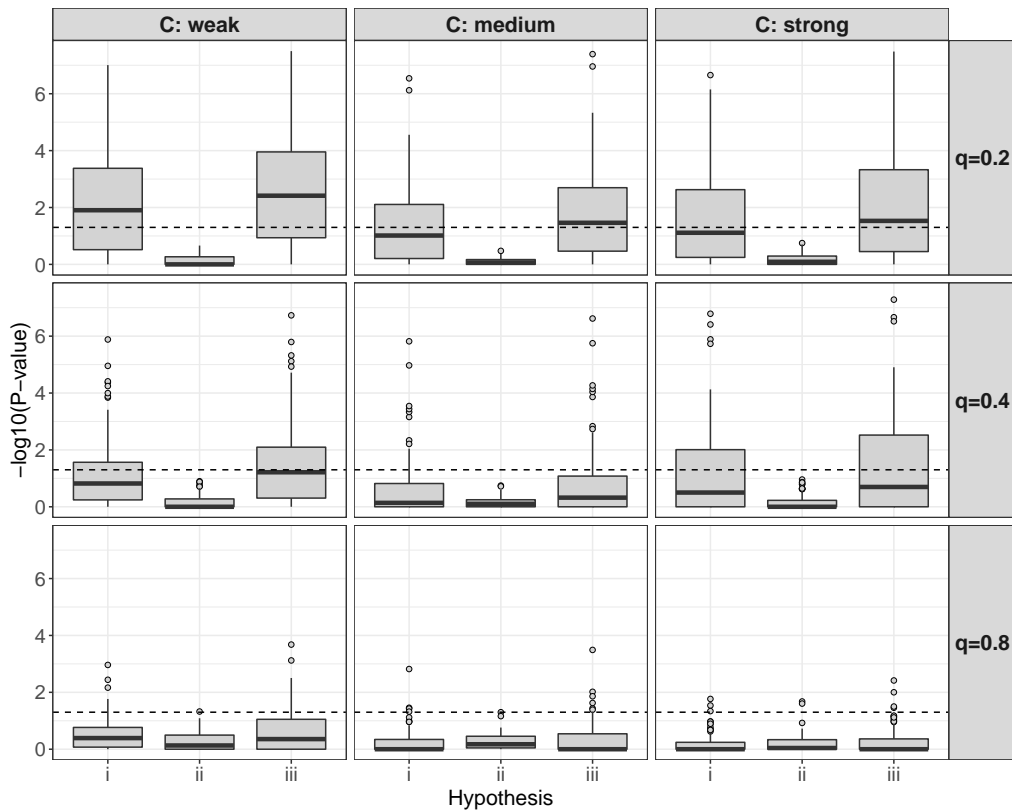


Figure S12: Results of the simulation study under the misspecified model (d). The boxplots visualize the negative log₁₀-transformed p -values obtained from the recalibration tests ($k = 10$). The dashed lines correspond to a p -value of 0.05. A value above the dashed line indicates a significant result at the 5% type 1 error level.

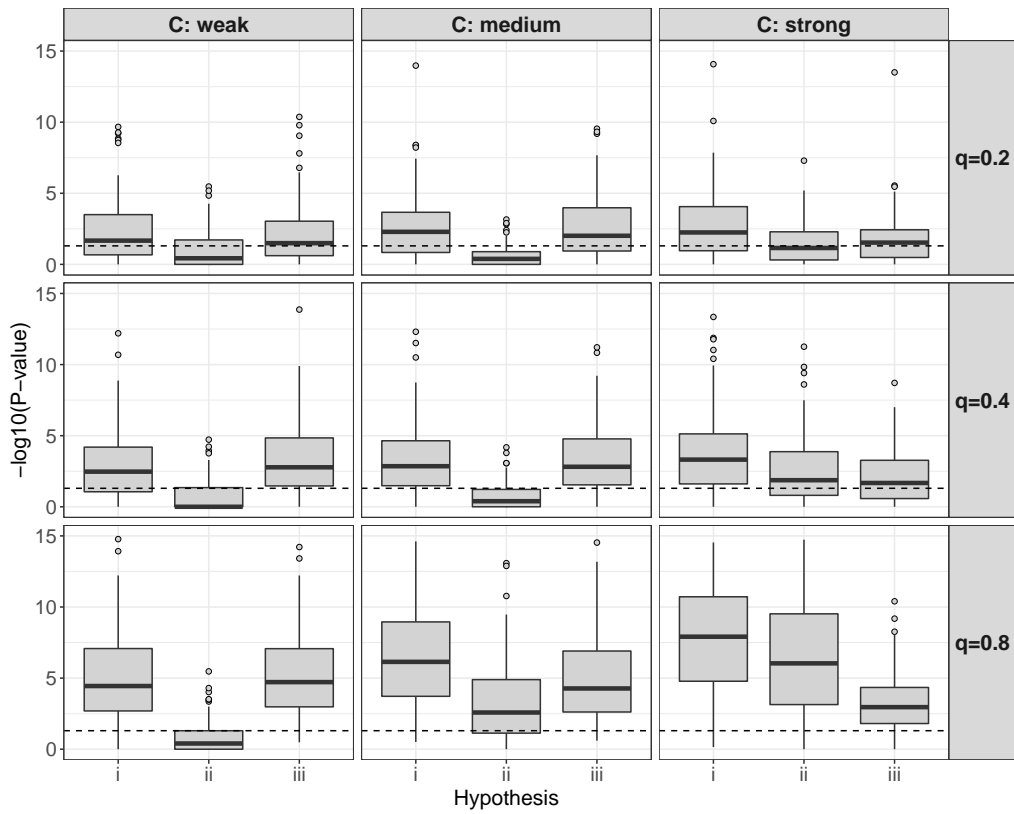


Figure S13: Results of the simulation study under the misspecified model (e). The boxplots visualize the negative log₁₀-transformed *p*-values obtained from the recalibration tests (*k* = 10). The dashed lines correspond to a *p*-value of 0.05. A value above the dashed line indicates a significant result at the 5% type 1 error level.

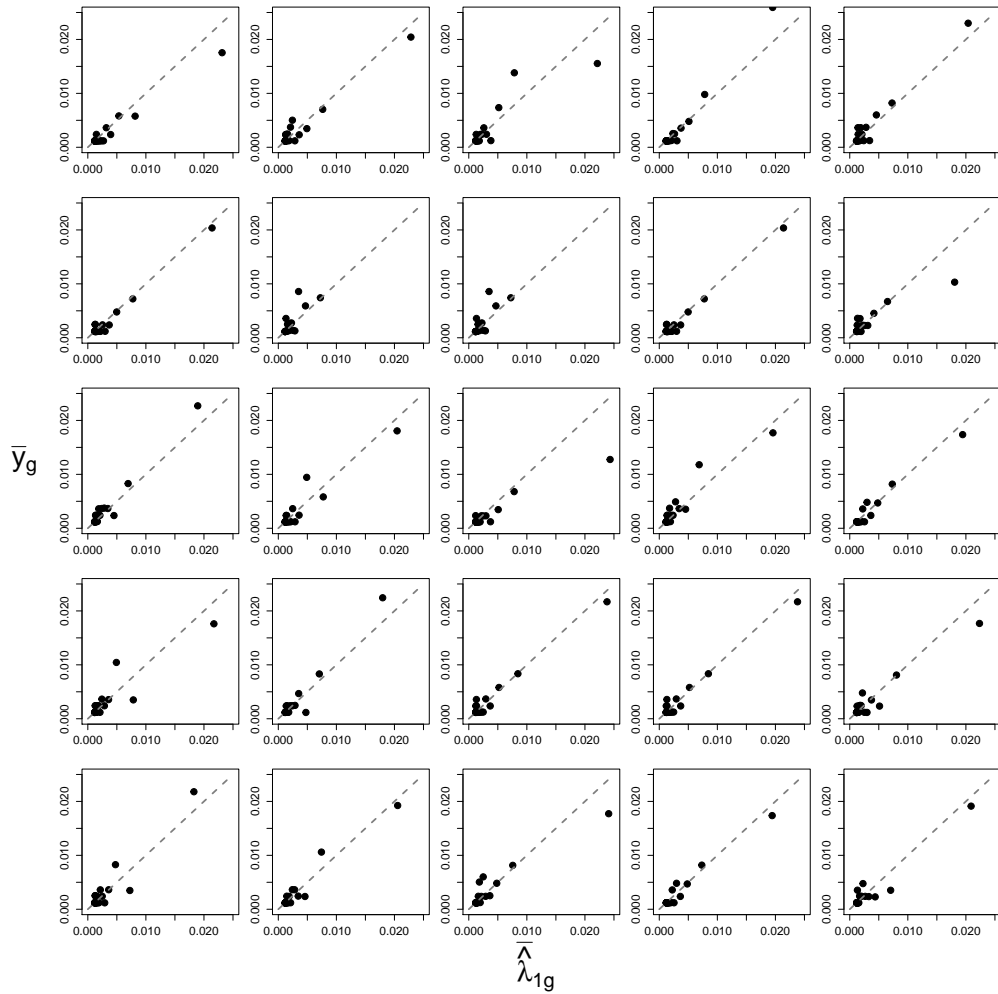


Figure S14: Analysis of the NP infection data. Calibration plots refer to 25 randomly chosen partitions of the data into a training and a validation sample using $G = 24$ subsets. The 45-degree line (dashed line) indicates perfect calibration.

3 Diskussion

Im Rahmen dieser Arbeit wurden eine Vielzahl von Methoden für die Analyse von diskreten Ereigniszeitdaten entwickelt, die in umfangreichen Simulationsstudien und anhand von verschiedenen Anwendungsbeispielen evaluiert wurden.

Alle Methoden können ganz allgemein nicht nur in klinischen und epidemiologischen Studien, sondern auch in allen Bereichen der angewandten Forschung, wie beispielsweise in den Sozial- und Wirtschaftswissenschaften, eingesetzt werden. Dabei ist nicht relevant, ob es sich bei den Daten um gruppierte stetige Ereigniszeiten oder um immanent diskrete Ereigniszeiten handelt (vgl. Kapitel 1.2). Es ist zu beachten, dass es sich bei gruppierten Daten auch um einen Spezialfall von intervallzensierten Daten mit festen Intervallgrenzen handelt (Lindsey und Ryan, 1998; Sun, 2007). Methoden für intervallzensierte Ereigniszeitdaten stellen ein eigenes Gebiet aktueller Forschung dar, das über diese Arbeit hinausgeht, siehe z.B. Gómez et al. (2009), Bogaerts et al. (2017), Fu und Simonoff (2017) und Yao et al. (2021).

Ein großer Vorteil der in den Kapiteln 2.1 bis 2.4 vorgestellten Hazard-Modelle ist, dass sie sich auf die Struktur von klassischen generalisierten Regressionsmodellen für binäre und kategoriale Zielvariablen zurückführen lassen. Dies ermöglicht die Verwendung von etablierten Softwareprogrammen zur Schätzung der Regressionskoeffizienten und gegebenenfalls zur Berechnung von zugehörigen Varianzschätzern. Wichtige R Funktionen, die in dieser Arbeit herangezogen wurden, sind `glm()` aus dem Basispaket und `gam()` aus dem Zusatzpaket **mgcv** sowie `rpart()` und `TSVC()` aus den jeweils gleichnamigen Zusatzpaketen. Um die Likelihood-Funktionen der Modelle zu bilden und die Schätzungen durchzuführen, müssen die originalen Daten im Vorfeld jeweils in eine erweiterte Datenmatrix (engl., “augmented data matrix”) umgewandelt werden, die sich wiederum aus kleineren Datenmatrizen für jede/n Patienten/Patientin zusammensetzt. Die konkrete Form der erweiterten Datenmatrix unterscheidet sich je nach Modell und wurde in den Ergebnisteilen im Detail beschrieben. Zur Aufbereitung der Daten und zur Erstellung der erweiterten Datenmatrix wurden jeweils die Funktionen des Zusatzpaketes **discSurv** in R verwendet.

Klassische Regressionsmodelle für stetige Ereigniszeiten setzen die Proportionalität der Hazardfunktion über die Zeit voraus (Cox, 1972; Fine und Gray, 1999). Diskrete Hazard-Modelle hingegen sind flexibler, da sie auch die Modellierung

von nicht proportionalen Hazardfunktionen erlauben. Die Wahl der Antwortfunktion $h(\cdot)$ entscheidet dabei jeweils, welche Proportionalitätseigenschaft durch das Modell abgebildet wird (Tutz und Schmid, 2016). Das logistische Hazard-Modell, das in Kapitel 2.1 und 2.2 betrachtet wurde, wird im Englischen auch entsprechend als “proportional continuation ratio model” bezeichnet. Denn es kann abgeleitet werden, dass für zwei Patienten/Patientinnen mit erklärenden Variablen X und \tilde{X} das Verhältnis der sogenannten Fortsetzungsraten

$$\frac{P(T = t|X)}{P(T > t|X)} / \frac{P(T = t|\tilde{X})}{P(T > t|\tilde{X})}, \quad t = 1, \dots, k, \quad (24)$$

nämlich der Verhältnisse der Wahrscheinlichkeit, dass das Ereignis zum Zeitpunkt t eintritt zur Wahrscheinlichkeit, dass das Ereignis zu einem späteren Zeitpunkt eintritt, über die Zeit als proportional angenommen wird. Für das Gompertz-Modell mit inverser komplementärer log-log-Funktion kann gezeigt werden, dass die Proportionalität über die Zeit für das Verhältnis der Logarithmen der Überlebensfunktionen

$$\log(S(t|X)) / \log(S(t|\tilde{X})), \quad t = 1, \dots, k, \quad (25)$$

gilt. Für weitere Alternativen sei auf Tutz und Schmid (2016) verwiesen. Eine wichtige Eigenschaft des Gompertz-Modells, die in Kapitel 2.4 und 2.5 genutzt wird, ist dessen Verknüpfung zu stetigen Ereigniszeiten. Tutz und Schmid (2016) zeigen, dass das Gompertz-Modell gilt, falls den diskreten Ereigniszeiten gruppierte, stetige Daten zugrunde liegen, die die Proportionalität der Hazardfunktionen über die Zeit erfüllen. Für das diskrete Subdistribution Hazard-Modell bedeutet dies, dass bei Verwendung der inversen komplementären log-log-Funktion die Parameter γ denen des stetigen Modells nach Fine und Gray (1999) entsprechen.

Die Einbettung der diskreten Hazard-Modelle in die Klasse der binären und kategorialen Regressionsmodelle bietet, wie in den Ergebnisteilen beschrieben, die einfache Möglichkeit zahlreicher Erweiterungen der klassischen Parametrisierung. Kapitel 2.2 befasst sich dabei explizit mit der Modellierung von zeit-variierenden Koeffizienten. Diese können als glatte Funktionen, z.B. über P-Splines, oder als stückweise konstante Funktionen über die Zeit mithilfe des vorgeschlagenen Baum-basierten Algorithmus spezifiziert werden. Für diskrete Ereigniszeiten ist zweitens eine Betrachtungsweise attraktiv, da es sehr plausibel ist, dass sich Effek-

te über bestimmte Zeitintervalle hinweg sprunghaft ändern. Auch die Baum-basierten Modelle in Kapitel 2.1 und 2.3 ermöglichen es, zeit-variierende Effekte abzubilden, da sowohl die erklärenden Variablen X als auch die Zeit t für die Baumkonstruktion herangezogen werden und somit jeder Endknoten einem Unterraum der erklärenden Variablen und einem Zeitintervall entspricht. Ein Vergleich gängiger Methoden für die Modellierung von zeit-variierenden Effekten für stetige Ereigniszeiten findet sich in Bansal und Heagerty (2019).

In den Darstellungen dieser Arbeit wird weitgehend davon ausgegangen, dass die Werte der erklärenden Variablen X über die Zeit konstant sind. Diese Einschränkung kann prinzipiell aufgehoben werden, indem zeit-variierende Werte der erklärenden Variablen $X_t = (X_{1t}, \dots, X_{pt})^\top$ in die erweiterte Datenmatrix, die zur Schätzung der Modelle gebildet wird, eingefügt werden. Im logistischen Hazard-Modell (17), im multinomialen logistischen Hazard-Modell (19), sowie im Baum-basierten Modell (20) ist dies ohne Weiteres möglich. Auch eine Kombination von zeit-variierenden Koeffizienten und zeit-variierenden Variablen in Modell (18) ist denkbar. Im Subdistribution Hazard-Modell (22) sind zeit-variierende Variablen dagegen problematisch. Der Grund dafür ist, dass die gewichtete Likelihood-Funktion, die für die Schätzung herangezogen wird, die (zeit-variierenden) Werte der erklärenden Variablen bis zum Zeitpunkt $k - 1$ benötigt, d.h. über die beobachtete Ereigniszeit \tilde{T} hinaus. Sind die erklärenden Variablen nicht extern (d.h. an den untersuchten Patienten/Patientinnen gemessen), sind diese Werte in der Regel jedoch nicht bekannt (Kalbfleisch und Prentice, 2002). Der Umgang mit zeit-variierenden erklärenden Variablen mit einem Fokus auf das stetige Cox-Modell und das Konzept von internen und externen Variablen wird unter anderem in Fisher und Lin (1999) diskutiert.

Parametrische und semi-parameterische diskrete Hazard-Modelle erlauben über die Ansätze in dieser Arbeit hinaus die Verwendung von regularisierten Schätzverfahren wie Lasso (Tibshirani, 1996) oder Boosting (Bühlmann und Hothorn, 2007) zur Selektion von einflussreichen Variablen. Dies ist in hochdimensionalen Situationen von Nutzen, insbesondere wenn die Anzahl an Parametern die Anzahl an Beobachtungen in den Daten übersteigt. Eine Lasso-Schätzung des logistischen Hazard-Modells (17) kann in R beispielsweise mit dem Zusatzpaket **penalized** (Goeman, 2018) durchgeführt werden. Eine penalisierte Schätzmethode, die speziell auf die Struktur des multinomialen logistischen Hazard-Modells (19) zugeschnitten ist, wurde von Möst et al. (2016) vorgeschlagen (siehe auch Ka-

pitel 2.3). Bei entsprechender Spezifikation der Strafterme bewirkt die Methode nach Möst et al. (2016), dass alle Koeffizienten, die mit einem der Ereignisse vom Typ j assoziiert sind, gemeinsam in Richtung Null geschrumpft werden.

Wichtige Alternativen zu den genannten regularisierten Schätzverfahren, die ebenfalls eine datengesteuerte Selektion von einflussreichen Variablen ermöglichen, sind nicht-parametrische, Baum-basierte Methoden. Dazu zählen in erster Linie das Verfahren von Schmid et al. (2016), das in Kapitel 2.1 rekapituliert wurde, und dessen Erweiterung für konkurrierende Ereignisse, das in Kapitel 2.3 eingeführt wurde. Weitere Baum-basierte Verfahren zur Modellierung diskreter Ereigniszeiten (inkl. Random-Forest-Algorithmen) wurden unter anderem von Bou-Hamad et al. (2009), Bou-Hamad et al. (2011), Janitza und Tutz (2015), Tiendrébéogo et al. (2019), Kretowska (2019), Schmid et al. (2020) und Moradian et al. (2021) vorgeschlagen.

Neben dem Vergleich der neu entwickelten Methoden mit existierenden Verfahren lag ein Fokus der Simulationsstudien in dieser Arbeit darauf, die Güte der neuen Methoden in Situationen mit (i) unterschiedlich hohem Anteil zensierter Beobachtungen, und (ii) unterschiedlicher Anzahl an diskreten Zeitpunkten k zu untersuchen. In allen Fällen zeigte sich, dass die vorgeschlagenen Methoden auch bei verhältnismäßig hoher Zensierung (von bis zu 75%) sehr gut funktionieren. Des Weiteren wurde offensichtlich, dass die diskreten Modelle den Methoden für stetige Ereigniszeiten überlegen sind, wenn die Anzahl an diskreten Zeitpunkten klein (z.B. $k = 5$), und damit die Anzahl an Bindungen groß ist. Der systematische Vergleich anhand des Subdistribution Hazard-Modells in Kapitel 2.4 zeigte auf, dass die Schätzungen des stetigen Modells nach Fine und Gray (1999) sowohl mit der Korrekturmethode nach Breslow als auch nach Efron systematisch nach unten verzerrt waren.

Bei der Analyse von Ereigniszeiten mit konkurrierenden Ereignissen kann man prinzipiell entweder jedes Ereignis über die ereignis-spezifischen Hazardfunktionen modellieren (wie in Kapitel 2.3) oder nur ein einzelnes Ereignis über die Subdistribution Hazardfunktion (wie in Kapitel 2.4). Während die Schätzung der ereignis-spezifischen Hazardfunktionen einfacher umsetzbar ist, ist zu beachten, dass die Herleitung und Interpretation der kumulativen Inzidenzfunktionen dabei erschwert sind (Beyersmann et al., 2011). Der Grund dafür ist, dass die kumulative Inzidenzfunktion für ein Ereignis vom Typ j von den ereignis-spezifischen Hazardfunktionen aller möglichen Ereignisse $\lambda_1, \dots, \lambda_J$ abhängt, siehe auch Gleichung

chung (9). Die Subdistribution Hazardfunktion für ein Ereignis vom Typ j hat hingegen eine direkte Verknüpfung zur zugehörigen kumulativen Inzidenzfunktion (vgl. Kapitel 2.4). Eine Übersicht aktueller Methoden für diskrete Ereigniszeitdaten mit konkurrierenden Ereignissen inklusive einer Schritt-für-Schritt-Analyse findet sich auch in Schmid und Berger (2020).

Neben der Entwicklung neuartiger Methoden zur Erstellung von Vorhersagemodellen widmete sich diese Arbeit auch der externen Validierung neuer Modelle. Zur Beurteilung der Kalibrierung stehen für die Anwender insgesamt die Methoden aus Kapitel 2.1 für diskrete Hazard-Modelle, von Heyard et al. (2019) für diskrete ereignis-spezifische Hazard-Modelle und aus Kapitel 2.5 für diskrete Subdistribution Hazard-Modelle zur Verfügung.

4 Zusammenfassung

Die Ereigniszeitanalyse stellt ein weitverbreitetes, wichtiges Instrument für die angewandte Forschung dar. Wie der Name schon sagt, wird dabei die Zeit bis zum Eintreten eines oder mehrerer interessierender Ereignisse in Abhängigkeit von Risikofaktoren analysiert. Dies kann vor allem im Rahmen der Präzisionsmedizin einen entscheidenden Beitrag zur Individualisierung von Medikamenten und zur genauen Abstimmung von Therapien leisten. Da in der Praxis die Ereigniszeiten generell auf einer diskreten Skala gemessen werden, bedarf es dafür geeigneter Methoden. In dieser Arbeit wurden zahlreiche parametrische, semi-parametrische und nicht-parametrische Verfahren vorgestellt, die auf diskrete Daten zugeschnitten sind. Diese sind den jeweiligen Methoden für stetige Ereigniszeiten vor allem dann vorzuziehen, wenn die Anzahl diskreter Zeitpunkte verhältnismäßig klein ist.

Ein wichtiger Vorteil von diskreten Hazardfunktionen im Vergleich zu stetigen Hazardfunktionen ist, dass sie bedingte Wahrscheinlichkeiten beschreiben und damit für Anwender deutlich einfacher zu interpretieren sind als Hazardraten aus stetigen Modellen. Daraus ergibt sich auch die Analogie von diskreten Hazard-Modellen zur Klasse der sequentiellen Regressionsmodelle für ordinale Zielvariablen, die in ihrer Form übereinstimmen (Tutz, 2012). Sequentielle Modelle wurden jüngst über den Kontext der diskreten Ereigniszeitanalyse hinaus verwendet, um die Verteilung von Einkommenskategorien (Tutz und Berger, 2020) und Zähldaten (Berger und Tutz, 2021) zu modellieren.

Die verschiedenen Anwendungsbeispiele in dieser Arbeit illustrieren den Nutzen und den Mehrwert der neuen, diskreten Methoden. Die Analyse der Studie unter Patienten/Patientinnen mit odontogenen Infektionen (siehe Kapitel 2.2) deckte beispielsweise auf, dass die Erkrankung an Diabetes Typ 2 einen wichtigen Risikofaktor darstellt, der die Liegedauer dieser Patienten/Patientinnen deutlich verlängert. Durch die Auswertungen der Daten der pairfam-Studie wurde der zeit-variierende Effekt des Ausbildungsniveaus und der Lebenszufriedenheit der Frauen auf die Zeit bis zur Geburt des ersten Kindes deutlich. Die Analyse der MODIAMD-Studie in Kapitel 2.3 unterstrich außerdem den relevanten Effekt von refraktilen Drusen und des Alters der Patienten/Patientinnen auf die Entwicklung von altersbedingter Makuladegeneration im Spätstadium.

Aufbauend auf der Arbeit von Tutz und Schmid (2016) bilden die hier entwickelten Methoden eine umfangreiche Auswahl an Alternativen für die Analyse diskreter Ereigniszeiten. Fragestellungen, die über den Rahmen dieser Arbeit hinausgehen, sind unter anderem (i) die Analyse von Daten mit Messwiederholungen (siehe Tutz und Schmid, 2016, Kapitel 9), und (ii) die Analyse von sogenannten Multi-Spell-Daten, in denen ein oder mehrere mögliche Ereignisse wiederholt auftreten können (siehe Tutz und Schmid, 2016, Kapitel 10).

Literatur

- Andersen, P. K., Geskus, R. B., de Witte, T., und Putter, H. (2012). Competing risks in epidemiology: Possibilities and pitfalls. *International Journal of Epidemiology* 41, 861–870.
- Austin, P. C., Lee, D. S., und Fine, J. P. (2016). Introduction to the analysis of survival data in the presence of competing risks. *Circulation* 133, 601–609.
- Bansal, A. und Heagerty, P. (2019). A comparison of landmark methods and time-dependent roc methods to evaluate the time-varying performance of prognostic markers for survival outcomes. *Diagnostic and Prognostic Research* 3, 1–13.
- Berger, M. (2020). *TSVC: Tree-Structured Modelling of Varying Coefficients*. R package version 1.2.1.
- Berger, M. und Tutz, G. (2021). Transition models for count data: a flexible

- alternative to fixed distribution models. *Statistical Methods & Applications, online first* .
- Beyersmann, J., Allignol, A., und Schumacher, M. (2011). *Competing Risks and Multistate Models with R*. Springer, New York.
- Beyersmann, J., Gastmeier, P., Grundmann, H., Bärwolff, S., Geffers, C., Behnke, M., Rüden, H., und Schumacher, M. (2006). Use of multistate models to assess prolongation of intensive care unit stay due to nosocomial infection. *Infection Control & Hospital Epidemiology* 27, 493–499.
- Bogaerts, K., Komarek, A., und Lesaffre, E. (2017). *Survival Analysis with Interval-Censored Data: A Practical Approach with Examples in R, SAS, and BUGS*. Chapman & Hall / CRC, New York.
- Bou-Hamad, I., Larocque, D., und Ben-Ameur, H. (2011). Discrete-time survival trees and forests with time-varying covariates: application to bankruptcy data. *Statistical Modelling* 11, 429–446.
- Bou-Hamad, I., Larocque, D., Ben-Ameur, H., Mâsse, L. C., Vitaro, F., und Tremblay, R. E. (2009). Discrete-time survival trees. *Canadian Journal of Statistics* 37, 17–32.
- Breiman, L. (1996). Technical note: Some properties of splitting criteria. *Machine Learning* 24, 41–47.
- Breiman, L., Friedman, J. H., Olshen, R. A., und Stone, J. C. (1984). *Classification and Regression Trees*. Wadsworth, Monterey, CA.
- Bühlmann, P. und Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting (with discussion). *Statistical Science* 22, 477–505.
- Cameron, A. C. und Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. University Press, Cambridge.
- Cieslak, D. A., Hoens, T. R., Chawla, N. V., und Kegelmeyer, W. P. (2012). Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery* 24, 136–158.
- Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* 34, 187–220.

- Croissant, Y. (2016). *Ecdat: Data Sets for Econometrics*. R package version 0.3-1.
- De Boor, C. (1978). *A practical guide to splines*. Springer, New York.
- Eilers, P. H. und Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science* 11, 89–102.
- Fahrmeir, L., Kneib, T., Lang, S., und Marx, B. D. (2013). *Regression*. Springer, New York.
- Fine, J. P. und Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* 94, 496–509.
- Fisher, L. und Lin, D. (1999). Time-dependent covariates in the cox proportional-hazards regression model. *Annual Review of Public Health* 20, 145–157.
- Fu, W. und Simonoff, J. (2017). Survival trees for interval-censored survival data. *Statistics in Medicine* 36, 4831–4842.
- Goeman, J. (2018). *Penalized R package*. R package version 0.9-51.
- Gómez, G., Calle, M., Oller, R., und Langohr, K. (2009). Tutorial on methods for interval-censored data and their implementation in r. *Statistical Modelling* 9, 259–297.
- Gray, B. (2020). *cmprsk: Subdistribution Analysis of Competing Risks*. R package version 2.2-10.
- Hastie, T., Tibshirani, R., und Friedman, J. (2009). *The Elements of Statistical Learning (2nd ed.)*. Springer, New York.
- Heim, N., Berger, M., Wiedemeyer, V., Reich, R. H., und Martini, M. (2019). A mathematical approach improves the predictability of length of hospitalization due to acute odontogenic infection. A retrospective investigation of 303 patients. *Journal of Cranio-Maxillofacial Surgery* 47, 334–340.
- Heyard, R., Timsit, J.-F., Held, L., und COMBACTE-MAGNET consortium (2019). Validation of discrete time-to-event prediction models in the presence of competing risks. *Biometrical Journal* 62, 643–657.

- Hosmer, D., Lemeshow, S., und Sturdivant, R. (2013). *Applied Logistic Regression*. John Wiley & Sons, Hoboken: New Jersey.
- Hothorn, T., Hornik, K., und Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15, 651–674.
- Huinink, J., Brüderl, J., Nauck, B., Walper, S., Castiglioni, L., und Feldhaus, M. (2011). Panel analysis of intimate relationships and family dynamics (pairfam): Conceptual framework and design. *Journal of Family Research* 23, 77–101.
- Janitza, S. und Tutz, G. (2015). Prediction models for time discrete competing risks. *Ludwig-Maximilians-Universität München, Department of Statistics* Technical Report 177.
- Kalbfleisch, J. und Lawless, J. (1991). Regression models for right truncated data with applications to aids incubation times and reporting lags. *Statistica Sinica* 1, 19–32.
- Kalbfleisch, J. und Prentice, R. (2002). *The Survival Analysis of Failure Time Data (2nd ed.)*. Wiley, New Jersey.
- Klein, J. und Moeschberger, M. (2003). *Survival Analysis: Statistical Methods for Censored and Truncated Data*. Springer, New York.
- Klein, J. P. und Andersen, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics* 61, 223–229.
- Kleinbaum, D. G. und Klein, M. (2012). *Survival Analysis (3rd ed.)*. Springer, New York.
- Kretowska, M. (2019). Oblique survival trees in discrete event time analysis. *IEEE Journal of Biomedical and Health Informatics* 24, 247–258.
- Lawless, J. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley, New Jersey.
- Lee, M. (2017). Inference for cumulative incidence on discrete failure times with competing risks. *Journal of Statistical Computation and Simulation* 87, 1989–2001.

- Lindsey, J. und Ryan, L. (1998). Methods for interval-censored data. *Statistics in Medicine* 17, 219–238.
- McCullagh, P. und Nelder, J. A. (2019). *Generalized Linear Models*. Routledge, Boca Raton.
- Miller, M., Langefeld, C., Tierney, W., Hui, S., und McDonald, C. (1993). Validation of probabilistic predictions. *Medical Decision Making* 13(1), 49–57.
- Moons, K., Kengne, A., Grobbee, D., Royston, P., Vergouwe, Y., Altman, D., und Woodward, M. (2012a). Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 98, 691–698.
- Moons, K., Kengne, A., Woodward, M., Royston, P., Vergouwe, Y., Altman, D., und Grobbee, D. (2012b). Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart* 98, 683–690.
- Moradian, H., Yao, W., Larocque, D., Simonoff, J. S., und Frydman, H. (2021). Dynamic estimation with random forests for discrete-time survival data. *Canadian Journal of Statistics*, online first .
- Möst, S., Pößnecker, W., und Tutz, G. (2016). Variable selection for discrete competing risks models. *Quality & Quantity* 50, 1589–1610.
- Petrie, A. und Sabin, C. (2019). *Medical Statistics at a Glance*. Wiley, New Jersey.
- Pößnecker, W. (2014). *MRSP: Multinomial response models with structured penalties*. R package version 0.4.3.
- Prentice, R. L., Kalbfleisch, J. D., Peterson, J., Flournoy, N., Farewell, V. T., und Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* 34, 541–554.
- Putter, H., Fiocco, M., und Geskus, R. B. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine* 26, 2389–2430.
- Quinlan, J. R. (1993). *Programs for Machine Learning*. Morgan Kaufmann PublisherInc., San Francisco.
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

- Scheike, T. H. und Keiding, N. (2006). Design and analysis of time-to-pregnancy. *Statistical Methods in Medical Research* 15, 127–140.
- Schmid, M. und Berger, M. (2020). Competing risks analysis for discrete time-to-event data. *Wiley Interdisciplinary Reviews: Computational Statistics* e1529.
- Schmid, M., Küchenhoff, H., Hörauf, A., und Tutz, G. (2016). A survival tree method for the analysis of discrete event times in clinical and epidemiological studies. *Statistics in Medicine* 35, 734–751.
- Schmid, M., Welchowski, T., Wright, M., und Berger, M. (2020). Discrete-time survival forests with hellinger distance decision trees. *Data Mining & Knowledge Discovery* 34, 812–832.
- Steinberg, J. S., Göbel, A. P., Thiele, S., Fleckenstein, M., Holz, F. G., und Schmitz-Valckenberg, S. (2016). Development of intraretinal cystoid lesions in eyes with intermediate age-related macular degeneration. *Retina* 36, 1548–1556.
- Steyerberg, E. (2019). *Clinical Prediction Models*. Springer, New York, 2nd edition.
- Steyerberg, E., Vickers, A., Cook, N., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M., und Kattan, M. (2010). Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology* 21, 128.
- Sun, J. (2007). *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer, New York.
- Therneau, T. und Atkinson, B. (2019). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58, 267–288.
- Tiendrébéogo, S., Some, B., Kouanda, S., und Dossou-Gbété, S. (2019). Survival analysis of data of hiv infected persons receiving antiretroviral therapy using a model-based binary tree approach. *Journal of Mathematics and Statistics* 15, 354–365.
- Tutz, G. (2012). *Regression for Categorical Data*. University Press, Cambridge.

- Tutz, G. und Berger, M. (2020). The effect of explanatory variables on income: A tool that allows a closer look at the differences in income. *Econometrics and Statistics* 16, 28–41.
- Tutz, G. und Schmid, M. (2016). *Modeling Discrete Time-to-Event Data*. Springer, New York.
- Welchowski, T. und Schmid, M. (2019). *discSurv: Discrete Time Survival Analysis*. R package version 1.4.1.
- Wolkewitz, M., Vonberg, R. P., Grundmann, H., Beyersmann, J., Gastmeier, P., Bärwolff, S., Geffers, C., Behnke, M., Rüden, H., und Schumacher, M. (2008). Risk factors for the development of nosocomial pneumonia and mortality on intensive care units: Application of competing risks models. *Critical Care* 12, R44.
- Wood, S. (2017). *Generalized Additive Models: An Introduction with R (2nd ed.)*. Chapman & Hall, Boca Raton.
- Wood, S. (2021). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*. R package version 1.8-36.
- Yao, W., Frydman, H., und Simonoff, J. (2021). An ensemble method for interval-censored time-to-event data. *Biostatistics* 22, 198–213.
- Yee, T. W. (2021). *VGAM: Vector generalized linear and additive models*. R package version 1.1-5.

Danksagung

Diese Arbeit entstand im Rahmen meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Medizinische Biometrie, Informatik und Epidemiologie der Universität Bonn. Ich danke allen, die zur Entstehung dieser Arbeit beigetragen haben. Ein besonderer Dank geht an ...

... Prof. Dr. Matthias Schmid für die enge, inhaltliche Zusammenarbeit und die stetige Förderung meiner wissenschaftlichen und angewandten Projekte.

... Prof. Dr. Nadja Klein und Prof. Dr. Jörg Rahnenführer, die sich freundlicherweise bereit erklärt haben die Aufgabe der Gutachter zu übernehmen.

... Prof. Dr. Andreas Mayr für den kollegialen Austausch, viele lehrreiche Diskussion im Journal Club und ganz viel Schokolade.

... die aktuellen und ehemaligen Kolleginnen und Kollegen unserer Arbeitsgruppe für die tolle Zusammenarbeit und die angenehme Arbeitsatmosphäre. Im Speziellen danke ich Leonie und Marie für die gemeinsamen Projekte zu beschränkten Zielgrößen und diskreten Ereigniszeiten. Danke, dass ihr auch immer ein offenes Ohr für mich habt.

... alle Kolleginnen und Kollegen am IMBIE für zahlreiche Mittagspausen und gemeinsame Aktivitäten neben der Arbeit.

... meine Freunde und meine Familie für deren uneingeschränkte Unterstützung. In erster Linie danke ich meinen Eltern und meiner Schwester, die mir stets zur Seite stehen.

Inhaltliche Überlappung mit anderen kumulativen Habilitationsschriften

Eine inhaltliche Überlappung durch gemeinsame Autorenschaften mit anderen kumulativen Habilitationsschriften ist nicht anzunehmen.

Kenntnis der Richtlinien der guten wissenschaftlichen Praxis der Universität Bonn (§3 der Habilitationsordnung)

Hiermit bestätige ich, dass ich die Richtlinien zur guten wissenschaftlichen Praxis der Universität Bonn, laut Habilitationsordnung, zur Kenntnis genommen habe und ich versichere, dass ich sie beim Verfassen der Habilitationsschrift beachtet habe. Insbesondere versichere ich, dass ich alle in der Habilitationsschrift benutzten Quellen und Hilfsmittel angegeben habe.