Helbredelsesmodeller

Jonas Hughes Larsen
Souad Zerzouri El-khatib

Aalborg Universitet
3-6-2020

AALBORG UNIVERSITET
STUDENTERRAPPORT

Institut for Matematiske Fag
Matematik og Statistik
Skjernvej 4a
9220 Aalborg Øst
www.math.aau.dk

Titel:
Helbredelsesmodeller

Tema:
Overlevelsesanalyse

Projektperiode:
Speciale, matematik
Forårssemesteret 2020

Projektgruppe:
1.216b

Deltagere:
Jonas Hughes Larsen
Souad Zerzouri El-khatib

Vejleder:
Rasmus Plenge Waagepetersen

Bivejleder:
Lasse Hjort Jakobsen

Oplagsantal: 4

Sidetal: 98 (Inklusiv appendiks)

Afsluttet: 3. juni 2020

Synopsis:
I specialet beskrives helbredelsesmodeller, som er specielle typer af modeller inden for overlevelsesanalyse. Helbredelsesmodeller anvendes til at estimere andelen af helbredte patienter og overlevelsen for de ikke-helbredte patienter. Formålet med specialet er at belyse teorien bag helbredelsesmodeller og anvende denne til at analysere et coloncancer-datasæt. Først introduceres netto og relativ overlevelse, og det vises, at den relative overlevelse beskriver nettooverlevelse under visse antagelser. Herefter beskrives tre ikke-parametriske metoder til at estimere den relative overlevelse: Ederer I, Ederer II og Hakulinen. To typer af helbredelsesmodeller beskrives: Mixtur helbredelsesmodeller og ikke-mixtur helbredelsesmodeller. Disse modeller kan modelleres ved hjælp af parametriske fordelinger, såsom Weibull-, log-normal- og eksponential-fordelingen. Disse fordelinger kan dog være for simple til at opfange den underliggende tendens. Derfor introduceres ARS- og FMC-modellen, som er to fleksible parametriske helbredelsesmodeller. ARS-modellen er en speciel type ikke-mixtur helbredelsesmodel og modelleres ved hjælp af en baglæns-spline. FMC-modellen er derimod en mixtur helbredelsesmodel, som modelleres ved hjælp af en restringeret kubisk spline.

Indhold

Forord

Dette projekt er udarbejdet som speciale for kandidatuddannelsen på Aalborg Universitet. Det er udarbejdet på matematikstudiet under Instituttet for Matematiske Fag under Det Teknisk-Naturvidenskabelige Fakultet i perioden 1. februar 2020 til 3. juni 2020. Specialet er forfattet af Jonas Hughes Larsen og Souad Zerzouri El-khatib. Specialets overordnede tema er helbredelsesmodeller, og herunder behandles den teori, som ligger til baggrund for modellerne. Der udarbejdes desuden dataanalyse, som er lavet i softwareprogrammet R, [R Core Team, 2020], løbende igennem specialet med den relevante teori. Dataanalysen er hovedsageligt baseret på R-pakkerne cuRe, rstpm2 og relsurv, [Jakobsen, 2020], [Clements and Liu, 2019] og [Perme and Pavlic, 2018]. Alle figurerne i specialet er desuden lavet med R-pakken ggplot2, [Wickham, 2016].

Specialet er opbygget af nummererede kapitler med afsnit og underafsnit. Figurer, tabeller og så videre er nummererede efter kapitler og fortløbende og henvises til uden parenteser. Ligninger er nummererede på samme måde, og disse henvises til med parenteser. Kilder angives med firkantede parenteser samt med forfatter og årstal. Der henvises til kilderne, hvor de er brugt, og litteraturlisten kan findes til sidst i specialet. Desuden anvendes parenteserne (), [] og {} for at skelne mellem start- og slutparenteser. Der er altså ingen forskel på betydningen af (x), $[x]$ og $\{x\}$. Alle vektorer er desuden fremhævet med fed skrift.

Det antages, at læseren har viden svarende til en fuldført 3. semesters kandidatuddannelse på matematikstudiet på Aalborg Universitet. Helbredelsesmodeller er et emne inden for overlevelsesanalyse, og derfor kræves det yderligere, at læseren har kendskab hertil. Det anbefales, at læseren først læser Appendiks A, da denne indeholder teori inden for overlevelsesanalyse.

Mange tak til vores vejleder Rasmus Plenge Waagepetersen og bivejleder Lasse Hjort Jakobsen samt til Maja Sjørslev Petersen for at hjælpe til med grammatikken.

Aalborg Universitet 3. juni 2020.

Jonas Hughes Larsen

Souad Zerzouri El-khatib

Abstract

This master's thesis focuses on cure models, which are special types of models within survival analysis. Cure models can be used to estimate the cure proportion and the survival of the uncured patients, which is of great interest to patients and doctors. Cure models can be formulated in an overall or relative survival setting, but the thesis mainly focuses on the relative survival setting since it is more relevant when studying cancer survival. The relative survival function is the ratio between the patient and the general population survival function. The thesis describes three non-parametric estimators for the relative survival function, including the Ederer I, Ederer II and Hakulinen estimators.

Two types of cure models are included in the thesis; the mixture and the non-mixture cure models. The mixture cure models include the patient population as a mixture of cured and uncured patients, where the survival of the uncured is modelled parametric with various distributions. The non-mixture cure models have a more difficult interpretation compared to the mixture cure models, and they are therefore less preferable. However, both cure models allow for modelling of the cure proportion and the survival of the uncured.

The cure models can sometimes be too simple to capture the underlying distribution, especially if simple parametric models such as the Weibull distribution is used. Flexible parametric cure models are therefore introduced, which attempts to model the underlying distribution with splines. Two flexible cure models are included in the thesis; the ARS model and the FMC model. The ARS model is a flexible cure model, which uses a backwards spline. The model is shown to be a special case of a non-mixture cure model. The FMC model is a mixture cure model, which uses a restricted cubic spline to model the survival of the uncured.

We illustrate the usage of the different cure models using survival data on colon cancer. It is concluded in the analyses of the colon cancer data that older patients are weaker against colon cancer. The start of the follow-up is especially rough for the older patients. It is also concluded that health services has improved their treatment against colon cancer in the 1975-1994 period. The data analysis also showed that it can be hard to determine the best cure model from AIC alone. It is therefore recommended using AIC while comparing with non-parametric estimates for the relative survival function, such as Ederer I.

Kapitel 1 | Introduktion

Cancer er en større klasse af sygdomme, som dræber tusindvis af mennesker hvert år. I 2017 døde mere end 9.5 millioner mennesker af cancer på verdensplan, og det var den næststørste dødsårsag efter hjertesygdomme. I Danmark var cancer i 2017 den største dødsårsag med 17000 dødstilfælde, [Ritchie and Roser, 2020]. I 2018 blev over 43000 danskere diagnosticeret med cancer, hvoraf de hyppigste typer var bryst, prostata, lunge, hud og colon, [Cancerregisteret, 2019]. Det er interessant at undersøge overlevelsen for forskellige typer af cancer, da det gør det muligt at stille kræftpatienter en prognose. I forbindelse med kliniske forsøg undersøges overlevelsen også for at sammenligne behandlinger. Hvis den ene behandling resulterer i en bedre overlevelse end den anden behandling, anvendes denne fremover. Overlevelsen for kræftpatienter bestemmes ved tiden fra diagnosen til død af kræft og opsummeres ofte ved andelen af patienter i live til et givet tidspunkt, eksempelvis 5 år.

Hovedmålet for kræftpatienter er at blive helbredt, men det kan være svært at give en præcis definition af helbredelse for kræft. I praksis siges en patient at være i komplet remission, hvis der efter behandling ikke kan detekteres flere maligne celler i kroppen, og patienten derfor kan betragtes som helbredt. Ved kræftpatienter kan det dog være misvisende at definere helbredelse ved komplet remission. Dette skyldes, at der enten kan være dødelige bivirkninger eller fordi sygdommen genopstår. Da kræftpatienter også kan dø af andre årsager end kræft, er det i stedet blevet foreslået at definere helbredelse ved at sammenligne patientoverlevelsen med overlevelsen i en baggrundsbefolkning matchet på køn og alder, [Tralongo et al., 2017]. Hvis patienterne opnår den samme overlevelse som baggrundsbefolkningen, siges de overlevende patienter at være statistisk helbredte. Denne definition tager højde for de dødelige bivirkninger samt risikoen for, at sygdommen genopstår, [Surbone et al., 2013]. Tidspunktet, hvor statistisk helbredelse forekommer, kaldes for helbredelsestidspunktet. Andelen af statistisk helbredte patienter og overlevelsen for de ikke-helbredte patienter er af stor interesse for læger og patienter, da disse giver et direkte indblik i sygdommens alvorlighed. Andelen af statistisk helbredte patienter samt overlevelsen for de ikke-helbredte patienter kan estimeres ved hjælp af helbredelsesmodeller. Helbredelsesmodeller blev først anvendt i 1949, [Boag, 1949], men er ikke de mest anvendte overlevelsesmodeller.

1.1 Datasæt

I dette speciale laves der løbende analyser på et datasæt bestående af patienter diagnosticeret med coloncancer i perioden 1975-1994. Datasættet er en del af R-pakken cuRe, [Jakobsen, 2020]. Opfølgningstiden for patienterne er bestemt ved tiden fra diagnosedatoen til dødsdatoen eller censurering (31/12-1995). Vi har valgt at ekskludere patienter under 18 og over 90 år. Der er i alt 10741 patienter, der oplever begivenheden af interesse, som er dødsfald, og 4634 patienter, der censureres. Median opfølgningstiden er 8.9 år. Tabellen der følger giver et overblik over patienterne i datasættet.

Kalender periode	1975-1984	1985-1994	Total
Antal patienter (%)	6410(42)	8965(58)	15375(100)
Mand/kvinde forhold %	39/61	42/58	41/59
Aldersgruppe; n(%)			
18-44	353(48)	379(52)	732(100)
45-59	1029(43)	1339(57)	2368(100)
60-74	2894(44)	3699(56)	6593(100)
75-90	2134(38)	3548(62)	5682(100)
Median alder	70	71	71

Tabel 1.1: Demografisk tabel for coloncancer-datasættet.

1.2 Specialets formål

Formålet med dette speciale er at beskrive helbredelsesmodeller, som er en speciel type model inden for overlevelsesanalyse. Helbredelsesmodeller kan blandt andet anvendes til at estimere andelen af statistisk helbredte patienter for en given sygdom samt overlevelsen for de ikke-helbredte patienter. Specialet uddyber forskellige typer af helbredelsesmodeller, som anvendes til at analysere coloncancer-datasættet beskrevet i det ovenstående. I analyserne undersøges det, hvordan aldersgrupperne 18-44, 45-59, 60-74 og 75-90 klarer sig i forhold til coloncancer. Specialet vil blandt andet hjælpe med at besvare spørgsmål som; har alder en betydning for andelen af statistisk helbredte patienter for coloncancer? Er sundhedsvæsenet blevet bedre til at behandle coloncancer?

Kapitel 2 | Netto og relativ overlevelse

Nettooverlevelsen beskriver overlevelsen for patienter med en bestemt sygdom i det hypotetiske scenarie, hvor sygdommen er den eneste dødsårsag. Der haves $1, \ldots, n$ patienter $(X_i, \delta_i, \mathbf{z}_i)$, hvor $X_i = \min(T_i, C_i)$. I denne sammenhæng er T_i tiden til en givet begivenhed, og C_i er censureringstiden. Disse antages at være uafhængige givet kovariater $\mathbf{z}_i$. Der haves $\delta_i = \mathbb{1}[T_i \leq C_i]$, som beskriver om individ i oplever begivenheden af interesse eller censureres. Lad T_D være tid til død for en bestemt sygdom, og lad T_O være tid til død af andre årsager. Der observeres kun $T = \min(T_D, T_O)$, da T_D og T_O er konkurrerende begivenheder. Lad desuden $\mathbf{z}$ være en vektor af kovariater til patienterne, som eksempelvis kan bestå af alder, køn og kræftstadie.

Nettooverlevelsen for den specifikke sygdom er bestemt ved overlevelsesfunktionen $S_D(t \mid \mathbf{z}) = P(T_D > t \mid \mathbf{z})$. For at kunne udregne S_D, skal den specifikke dødsårsag være kendt. Dødsårsagen fremgår i en dødsattest, men det er ofte tilfældet, at den enten ikke er kendt eller er noteret forkert, [Begg and Schrag, 2002]. Det kan desuden være svært at komme til en entydig konklusion om, hvorvidt dødsårsagen skyldes den specifikke sygdom eller noget andet. Et alternativ til at bestemme nettooverlevelsen er derfor nødvendigt. Til dette formål defineres den relative overlevelse

$$R(t \mid \mathbf{z}) = \frac{S(t \mid \mathbf{z})}{S^*(t \mid \mathbf{z})}, \tag{2.1}$$

hvor $S(t \mid \mathbf{z})$ er overlevelsesfunktionen tilhørende den stokastiske variabel T, og $S^*(t \mid \mathbf{z})$ er overlevelsesfunktionen for en sammenlignelig gruppe fra baggrundsbefolkningen, som også kaldes den forventede overlevelse. Denne kan bestemmes ud fra levetidstabeller, hvilket uddybes i Afsnit 2.1. Ud fra Ligning (2.1) kan den totale overlevelse udtrykkes ved

$$S(t \mid \mathbf{z}) = S^*(t \mid \mathbf{z})R(t \mid \mathbf{z}). \tag{2.2}$$

Den relative overlevelse bruger ikke den specifikke dødsårsag og er derfor ikke afhængig af, om den er noteret korrekt. Derimod behøves information om overlevelsen

for en sammenlignelig gruppe fra baggrundsbefolkningen. Det kan være svært at anskaffe levetidstabeller for en baggrundsbefolkning uden at inkludere individer med sygdommen. Lad T^* være tid til død i baggrundsbefolkningen, og p være sandsynligheden for, at en person er patient med en bestemt sygdom. For hver person i baggrundsbefolkningen observeres

$$T^* = \begin{cases} T_O, & \text{hvis ikke patient} \\ \min(T_D, T_O), & \text{hvis patient.} \end{cases}$$

Dermed haves

$$\begin{aligned} S^*(t \mid \mathbf{z}) = P(T^* > t \mid \mathbf{z}) &= p \cdot P(\min(T_D, T_O) > t \mid \mathbf{z}) + (1-p) \cdot P(T_O > t \mid \mathbf{z}) \\ &\approx P(T_O > t \mid \mathbf{z}) = S_O(t \mid \mathbf{z}), \end{aligned}$$

hvis p er lille. Det virker derfor rimeligt at antage $S^*(t \mid \mathbf{z}) = S_O(t \mid \mathbf{z})$, hvis baggrundsbefolkningen eksempelvis er Danmarks befolkning. Hvis det derudover også antages, at T_D og T_O er betinget uafhængige givet $\mathbf{z}$, haves

$$\begin{aligned} R(t \mid \mathbf{z}) = \frac{S(t \mid \mathbf{z})}{S^*(t \mid \mathbf{z})} &= \frac{S(t \mid \mathbf{z})}{S_O(t \mid \mathbf{z})} = \frac{P(T_D > t, T_O > t \mid \mathbf{z})}{P(T_O > t \mid \mathbf{z})} \\ &= \frac{P(T_D > t \mid T_O > t, \mathbf{z}) P(T_O > t \mid \mathbf{z})}{P(T_O > t \mid \mathbf{z})} \\ &= P(T_D > t \mid T_O > t, \mathbf{z}) \\ &= P(T_D > t \mid \mathbf{z}) \\ &= S_D(t \mid \mathbf{z}). \end{aligned}$$

Det vil sige, at den relative overlevelse beskriver nettooverlevelsen under disse antagelser. Det kan dog ikke undersøges ud fra data, om T_D og T_O er betinget uafhængige, og det kan derfor være svært at retfærdiggøre denne antagelse.

Ved at anvende sammenhængen mellem overlevelsesfunktionen og hazard-funktionen, kan Ligning (2.1) også udtrykkes ved:

$$-\frac{d \ln[R(t \mid \mathbf{z})]}{dt} = h(t \mid \mathbf{z}) - h^*(t \mid \mathbf{z}),$$

hvor $h^*(t \mid \mathbf{z})$ er hazard-funktionen for baggrundsbefolkningen. En omskrivning af dette giver

$$h(t \mid \mathbf{z}) = h^*(t \mid \mathbf{z}) + \lambda(t \mid \mathbf{z}), \tag{2.3}$$

hvor $\lambda(t \mid \mathbf{z}) = -\frac{d \ln[R(t|\mathbf{z})]}{dt}$ er den forøgede hazard-funktion, der forekommer ved sygdommen. Det betyder også, at $R(t \mid \mathbf{z}) = \exp\left(-\int_0^t \lambda(u \mid \mathbf{z})du\right)$. Hazard-funktionen kan altså opdeles i to led. Det ene er den hazard, der forekommer ved sygdommen, mens det andet er den, der forekommer ved andre årsager, som antages at være den samme som baggrundsbefolkningens. Den forøgede hazard-funktion giver et direkte indblik i den forøgede dødelighed givet overlevelse til tiden t. Den forøgede hazard-funktion $\lambda(t \mid \mathbf{z})$ kan eksempelvis modelleres ved en proportional forøget hazard-model

$$\lambda(t \mid \mathbf{z}) = \lambda_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{z}), \tag{2.4}$$

hvor $\lambda_0(t)$ er den forøgede reference hazard-funktion, som ofte bestemmes parametrisk. Dette kan gøres ved at opdele tiden i K intervaller og antage en stykkevis konstant reference hazard-funktion $\lambda_0(t) = \sum_{j=1}^K \alpha_j \mathbb{1}[t \in I_j]$, hvor I_j er det j'te interval, og α_j er den forøgede hazard i det j'te interval for $\mathbf{z} = 0$. Ligning (2.4) estimeres ved at maksimere log-likelihoodfunktionen i Ligning (A.3), [Estève et al., 1990]. Ved integration af Ligning (2.3), kan den kumulerede hazard-funktion udtrykkes ved

$$H(t \mid \mathbf{z}) = H^*(t \mid \mathbf{z}) + \Lambda(t \mid \mathbf{z}),$$

hvor $\Lambda(t \mid \mathbf{z}) = \int_0^t \lambda(u \mid \mathbf{z})du = -\ln[R(t \mid \mathbf{z})]$ er den forøgede kumulerede hazard-funktion, der forekommer ved sygdommen, og $H^*(t \mid \mathbf{z})$ er baggrundsbefolkningens kumulerede hazard-funktion.

I følgende afsnit beskrives det, hvordan et ikke-parametrisk estimat bestemmes for den relative overlevelse.

2.1 Estimation af den relative overlevelse

I dette afsnit betragtes ikke-parametriske metoder til at estimere $S^*(t)$ og $R(t)$ i tilfældet, hvor kovariater ikke indgår. Metoderne eksemplificeres løbende med et konkret datasæt for Danmarks befolkning, samt coloncancer-datasættet, som blev introduceret i Afsnit 1.1. Den forventede overlevelse for Danmarks befolkning i form af 1-dags diskretiseret hazard stratificeret efter alder, køn og kalenderår kan findes på [The Human Mortality Database, 2002] og anvendes i dette speciale som baggrundsbefolkningen. Det kan eksempelvis aflæses, at den forventede daglige hazard for en 19-årig mand i henholdvis 2000 og 2001 er 0.000002410429 og 0.00000199941.

Ved at anvende $S(t) = \exp[-H(t)]$, kan hans 1 års forventede overlevelse fra den 31/10/2000 bestemmes ved

$$\exp(-62 \cdot 0.000002410429 - 303 \cdot 0.00000199941) = 0.999245.$$

I princippet skal hans præcise alder i dage anvendes for at få et præcist resultat. Det kunne eksempelvis være tilfældet, at han havde fødselsdag den 1/1. I så fald skulle den forventede daglige hazard for en 20-årig mand i 2001, som er 0.000002109009, anvendes for perioden efter den 1/1/2001, og udregningen ville i stedet blive

$$\exp(-62 \cdot 0.000002410429 - 303 \cdot 0.000002109009) = 0.9992118.$$

I data er det dog ofte kun alder i år, som er kendt, og ikke den præcise fødselsdag.

2.1.1 Estimation af den forventede overlevelse

Det ovenstående beskriver, hvordan den forventede overlevelse bestemmes for et hypotetisk individ i baggrundsbefolkningen. Det har ikke den store relevans at sammenligne en givet patientgruppe med den forventede overlevelse for hele baggrundsbefolkningen, hvis patientgruppen eksempelvis kun består af ældre patienter. I stedet kan den forventede overlevelse bestemmes for et udsnit af baggrundsbefolkningen, som matcher patientgruppen på de demografiske variable; alder, køn og kalenderår.

Det vil for eksempel sige, at hvis en 25-årig kvinde i patientgruppen er diagnosticeret med en givet sygdom i 1980, skal den matchende baggrundsbefolkning også inkludere en 25-årig kvinde i 1980. Hendes daglige hazard kan slås op i levetidstabellen, hvorefter den forventede overlevelse kan bestemmes ved tilsvarende udregninger som i det ovenstående. Lad $h_i^*(t)$ være hazard-funktionen for det hypotetiske individ fra den matchende baggrundsbefolkning med samme demografiske variable i diagnosetidspunktet som patient i. Den tilhørende forventede kumulerede hazard-funktion og overlevelsesfunktion er givet ved

$$H_i^*(t) = \int_0^t h_i^*(u)du,$$
$$S_i^*(t) = \exp[-H_i^*(t)].$$

For en patientgruppe med $i = 1, \ldots, n$ patienter, er den hyppigste måde at udregne den forventede kumulerede hazard-funktion og overlevelsesfunktion som følger,

[Therneau and Offord, 1999],

$$H^*(t) = \int_0^t \frac{\sum_{i=1}^n h_i^*(u) w_i(u)}{\sum_{i=1}^n w_i(u)} du, \tag{2.5}$$
$$S^*(t) = \exp[-H^*(t)],$$

hvor $w_i(t)$ er en vægt, som skal specificeres nærmere. Integranden i Ligning (2.5) er dermed et vægtet gennemsnit af $h_i^*(t)$. De mest anvendte metoder til at specificere vægtene er kendt under navnene Ederer I, Ederer II og Hakulinen, [Ederer et al., 1961], [Ederer and Heise, 1959] og [Hakulinen, 1982].

Ederer I anvender $w_i(t) = S_i^*(t)$, hvilket betyder, at vægten er sandsynligheden for, at patient i er i live til tiden t. Ved at anvende $\frac{d}{du}S(u) = -f(u) = -h(u)S(u)$ og kædereglen, haves

$$\begin{aligned}\frac{d}{du}\ln\left[\frac{1}{n}\sum_{i=1}^n S_i^*(u)\right] &= \frac{-\frac{1}{n}\sum_{i=1}^n h_i^*(u)S_i^*(u)}{\frac{1}{n}\sum_{i=1}^n S_i^*(u)} \\ &= -\frac{\sum_{i=1}^n h_i^*(u)S_i^*(u)}{\sum_{i=1}^n S_i^*(u)}.\end{aligned}$$

Ved at anvende dette resultat, haves

$$\begin{aligned}S^*(t) = \exp[-H^*(t)] &= \exp\left[-\int_0^t \frac{\sum_{i=1}^n h_i^*(u)S_i^*(u)}{\sum_{i=1}^n S_i^*(u)} du\right] \\ &= \exp\left[\int_0^t \frac{d}{du}\left\{\ln\left(\frac{1}{n}\sum_{i=1}^n S_i^*(u)\right)\right\} du\right] \\ &= \frac{1}{n}\sum_{i=1}^n S_i^*(t).\end{aligned}$$

Det vil sige, at Ederer I er et gennemsnit af de hypotetiske individers forventede overlevelse til tiden t. Et problem med Ederer I er, at levetidstabeller ikke altid er opdaterede til det nuværende kalenderår, hvilket betyder, at det kan være nødvendigt at bruge det sidste tilgængelige år i stedet.

Ederer II anvender vægten $w_i(t) = Q_i(t)$, hvor $Q_i(t) = \mathbb{1}[T_i \geq t, C_i \geq t]$. Der haves altså $Q_i(t) = 1$, hvis individ i er i risiko til tiden t og 0 ellers. Det i'te hypotetiske individ fra den matchende baggrundsbefolkning inkluderes altså kun i udregningen af $S^*(t)$, hvis patient i er i risiko til tiden t. Dette adskiller sig fra Ederer I, som

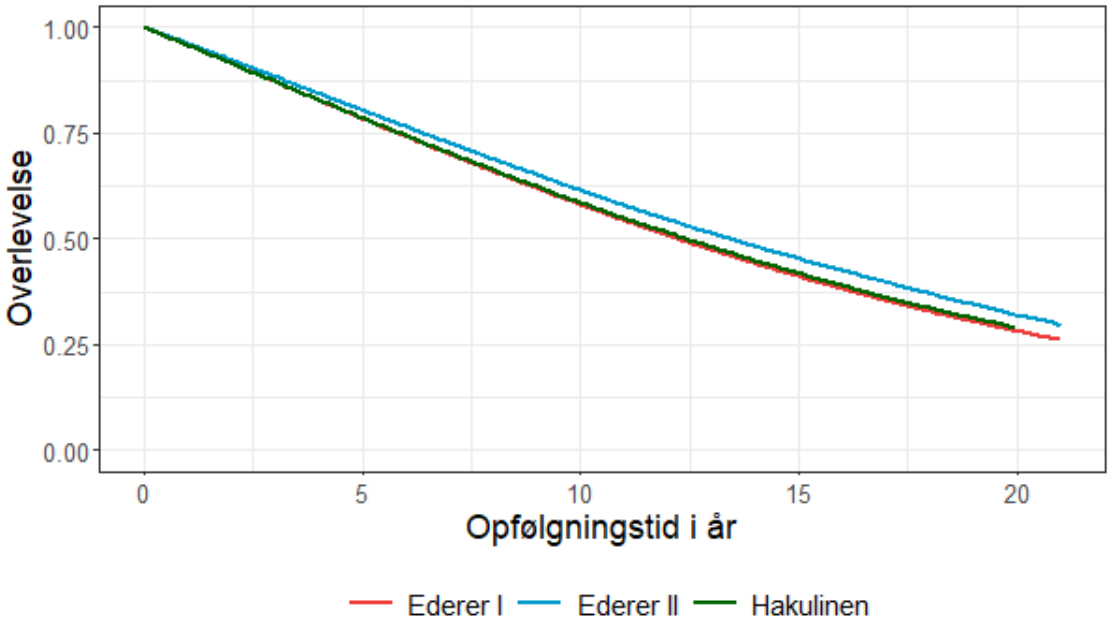
inkluderer alle de hypotetiske individer fra den matchende baggrundsbefolkning, uanset om patient i er i risiko. Dette er en fordel, når der er en interesse i betingede overlevelseskurver. Hvis det eksempelvis ønskes at bestemme $S^*(2)$ ved hjælp af Ederer II, betinges der med overlevelse i 2 år, da det kun er de patienter, som fortsat er i risiko efter 2 år, der inkluderes.

Det er tidligere blevet påvist, at den relative overlevelse bestemt ved Ederer I kan give biased resultater under informativ censurering, [Hakulinen, 1982]. Informativ censurering forekommer, hvis nogle af patienterne ikke længere kan deltage i opfølgningen af grunde relateret til studiet. Dette kunne eksempelvis være hvis behandlingen har gjort dem for syge. Hakulinen-metoden forsøger at reducere denne bias. Metoden antager, at der følges op på de hypotetiske individer på samme måde som for patienterne. Det antages, at studiets sluttidspunkt er kendt for alle patienterne, som eksempelvis kan være tidspunktet, hvor analyserne foretages. Lad henholdsvis C_i og $\widetilde{C}_i$ være censureringstiden og den potentielle opfølgningstid for patient i, hvor den potentielle opfølgningstid er tiden fra patient i's inklusion i studiet til studiets sluttidspunkt. Den potentielle opfølgningstid $\widetilde{C}_i$ er eksempelvis 5 år, hvis en patient dør 2 år inde i studiet, men ellers ville have været fulgt indtil 5 år. Hakulinen-metoden benytter vægten $w_i(t) = S_i^*(t)Q_i^*(t)$, hvor

$$Q_i^*(t) = \begin{cases} \mathbb{1}[C_i \geq t], & \text{hvis } \delta_i = 0 \\ \mathbb{1}[\widetilde{C}_i \geq t], & \text{hvis } \delta_i = 1. \end{cases}$$

Det vil sige, at det hypotetiske individ censureres, når patienten censureres. Hvis patienten derimod oplever begivenheden, er det hypotetiske individ i risiko indtil det potentielle opfølgningstidspunkt, hvorefter det hypotetiske individ censureres.

Figur 2.1 illustrerer den forventede overlevelse for coloncancer-datasættet bestemt ved Ederer I, Ederer II og Hakulinen.



Figur 2.1: Forventet overlevelse for coloncancer-datasættet.

Alle de tre ovenstående estimatorer for $S^*(t)$ kan anvendes til at bestemme et ikke-parametrisk estimat af $R(t) = \frac{S(t)}{S^*(t)}$. I specialet anvendes Kaplan-Meier-estimatet til at bestemme $S(t)$ i det ikke-parametriske estimat af $R(t)$, se Kaplan-Meier-estimatet i Ligning (A.1). Det er derfor underforstået, at Kaplan-Meier-estimatet også er anvendt, hvis det benævnes, at den relative overlevelse er bestemt ved Ederer I, Ederer II eller Hakulinen. Det er generelt forskelligt hvilket ikke-parametrisk estimat, der anvendes i litteraturen, og ingen skiller sig umiddelbart ud som værende den bedste.

2.1.2 Standardfejl og konfidensinterval

Standardfejlen af den ikke-parametriske relative overlevelsesfunktion estimeres ved at dividere standardfejlen af overlevelsesfunktionen for patientgruppen med den forventede overlevelsesfunktion, [Ederer et al., 1961]:

$$\text{se}[\widehat{R}(t)] = \frac{\text{se}[\widehat{S}(t)]}{\widehat{S}^*(t)} = \frac{\sqrt{\mathbb{V}\text{ar}[\widehat{S}(t)]}}{\widehat{S}^*(t)},$$

hvor $\widehat{S}^*(t)$ er bestemt ved Ederer I, Ederer II eller Hakulinen, og $\mathbb{V}\text{ar}[\widehat{S}(t)]$ kan bestemmes ved at anvende Ligning (A.2) i appendiks. Konfidensintervallet for $\widehat{R}(t)$ er dermed givet som

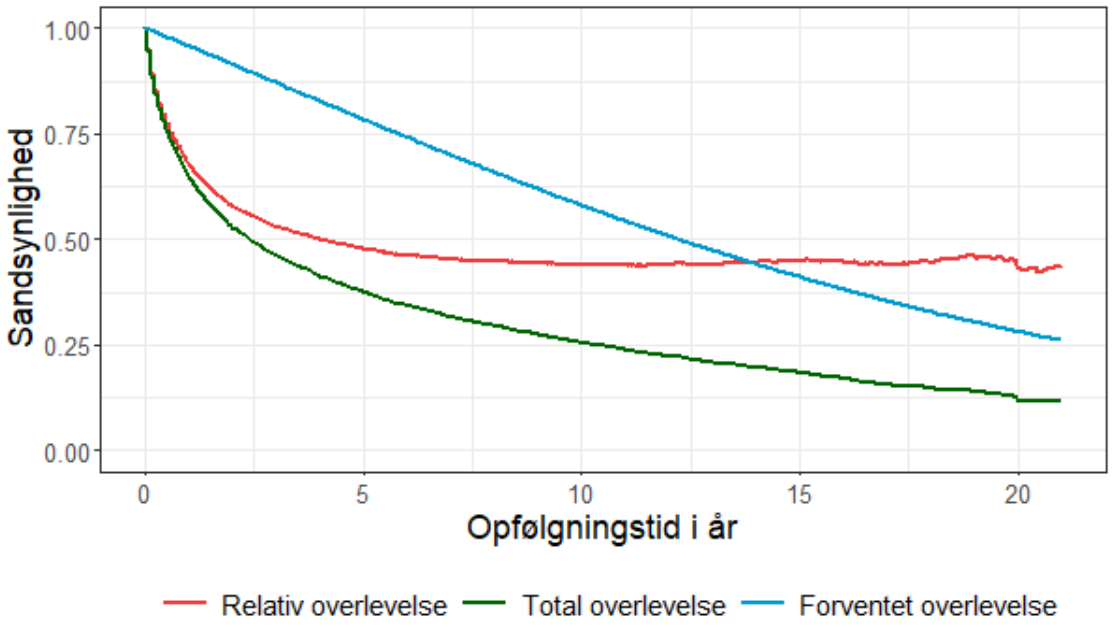
$$\widehat{R}(t) \pm 2 \cdot \text{se}[\widehat{R}(t)].$$

Konfidensintervallet kan også bestemmes ved $\log(\cdot)$ eller $\log[-\log(\cdot)]$ transformationer, hvorefter der transformeres tilbage ved $\exp(\cdot)$ eller $\exp[-\exp(\cdot)]$. I specialet anvendes $\log(\cdot)$ transformationen.

2.1.3 Statistisk helbredelse

Det er interessant at undersøge, hvornår overlevelsen for patienterne er den samme som ved den matchende baggrundsbefolkning. Dette forekommer, når $\lambda(t)$ i Ligning (2.3) er lig med 0, hvilket illustreres ved en udfladning i den relative overlevelsesfunktion, da $R(t)$ i så fald er konstant. Tidspunktet, hvor dette forekommer, er helbredelsestidspunktet, og de overlevende patienter efter helbredelsestidspunktet siges at være statistisk helbredte. Dette er en populationsbaseret definition af helbredelse og udelukker derfor ikke, at sygdommen kommer igen i løbet af den enkelte persons levetid. Det er ikke altid, at der forekommer en udfladning i den relative overlevelse. Et eksempel herpå er, hvis patienterne med sygdommen altid har en forøget hazard i forhold til den matchende baggrundsbefolkning. Desuden er den relative overlevelse heller ikke nødvendigvis en monotont aftagende funktion, da det er muligt, at patienterne med sygdommen har en bedre overlevelse end den matchende baggrundsbefolkning. Dette virker usandsynligt, men kan være en effekt af den ekstra kontakt, som patienterne har til sundhedsvæsenet. Det kan også være et resultat af patient-selektion. Et eksempel herpå er, hvis det kun er de patienter med en god overlevelse, der inkluderes i studiet.

Figur 2.2 illustrerer den totale overlevelse og den relative overlevelse for coloncancer-patienterne samt den forventede overlevelse. Den forventede overlevelse er bestemt ved Ederer I, mens den totale overlevelse er bestemt ved Kaplan-Meier. Den relative overlevelse er bestemt ved Kaplan-Meier og Ederer I.

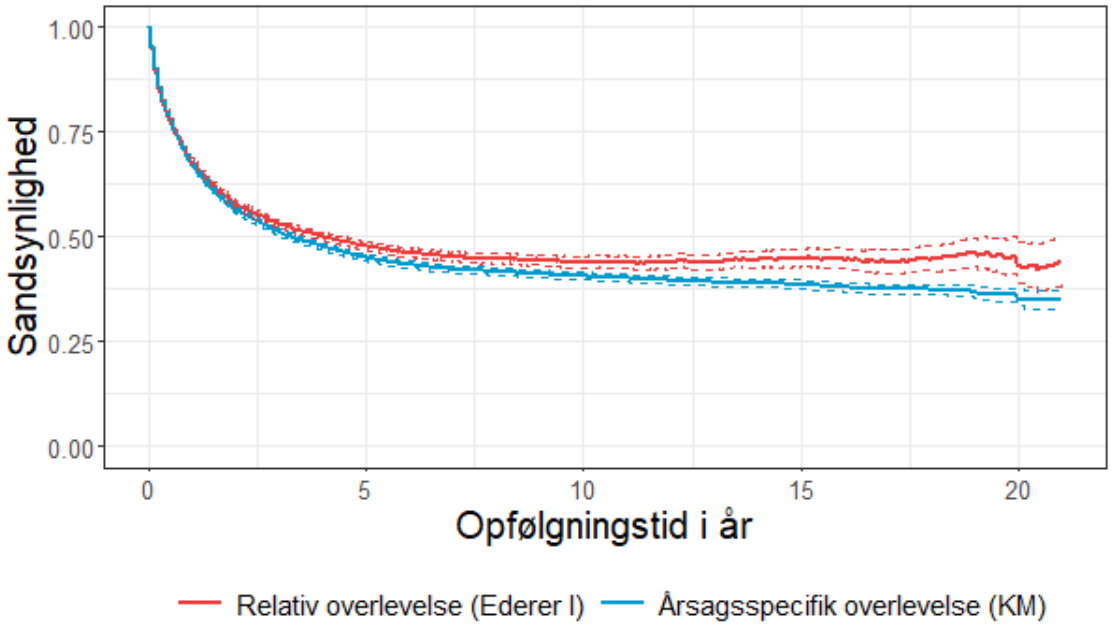


Figur 2.2: Den totale overlevelse og den relative overlevelse for coloncancer-patienter samt den forventede overlevelse.

Baseret på en grafisk inspektion af Figur 2.2 fremgår det, at der forekommer en udfladning i den relative overlevelse efter cirka 7 år. De overlevende patienter efter dette tidspunkt har den samme overlevelse som den matchende baggrundsbefolkning og betragtes som statistisk helbredte.

2.1.4 Årsagsspecifik overlevelse

Coloncancer-datasættet indeholder også den specifikke dødsårsag for patienterne. Hvis dødsårsagen er noteret korrekt, samt T_D og T_O er betinget uafhængige, kan nettooverlevelsen $S_D(t)$ estimeres. Dette kan gøres ved et Kaplan-Meier-estimat, hvor patienter, som er døde af andre årsager end coloncancer, censureres. Dette kaldes også for den årsagsspecifikke overlevelse. Det er dog som tidligere nævnt ikke altid tilfældet, at dødsårsagen er noteret korrekt. I coloncancer-datasættet er 8216 patienter noteret som død af cancer. På følgende figur sammenlignes den relative overlevelse bestemt ved Ederer I med den årsagsspecifikke overlevelse.



Figur 2.3: Den relative overlevelse bestemt ved Ederer I og den årsagsspecifikke overlevelse bestemt ved Kaplan-Meier for coloncancer-patienterne.

Det fremgår af Figur 2.3, at den årsagsspecifikke overlevelse og den relative overlevelse tilnærmelsesvist følger hinanden. Den årsagsspecifikke overlevelse ligger dog under den relative overlevelse i hele perioden. Det er muligt, at dødsårsagen ikke altid er noteret korrekt for coloncancer-datasættet, og derfor foretrækkes den relative overlevelse over den årsagsspecifikke overlevelse.

I det næste kapitel præsenteres helbredelsesmodeller. Disse kan beskrives i forhold til den totale overlevelse og den relative overlevelse.

Kapitel 3 | Helbredelsesmodeller

I analyser, der omhandler tid til en begivenhed, antages det typisk, at alle individer kan opleve begivenheden af interesse, og overlevelsesfunktionen, $S(t \mid \mathbf{z})$, vil gå mod 0 for $t \to \infty$. I nogle situationer kan det derimod antages, at en andel individer aldrig vil opleve begivenheden af interesse. Denne andel siges at være helbredte, og den observerede population kan derfor betragtes som en blanding af helbredte og ikke-helbredte individer.

Eksempelvis kan begivenheden af interesse være tilbagefald af en bestemt kræftsygdom efter behandling. I dette tilfælde vil nogle af patienterne aldrig opleve et tilbagefald, og de betragtes derfor som helbredte. Efter tilstrækkeligt lang tid vil overlevelsesfunktionen, $S(t \mid \mathbf{z}) = P(T > t \mid \mathbf{z})$, hvor T repræsenterer tiden fra diagnosen til et tilbagefald, flade ud. Dette skyldes, at alle patienterne, bortset fra de helbredte patienter, har oplevet et tilbagefald, når $t \to \infty$.

Et andet eksempel er tiden fra mødres første barnefødsel til anden barnefødsel, som her er begivenheden af interesse. Nogle mødre får kun et barn og anses som helbredte, mens mødrene, som får barn nummer to, anses som ikke-helbredte. I dette tilfælde vil overlevelsesfunktionen også flade ud efter tilstrækkeligt lang tid. I begge eksempler er essensen andelen af helbredte individer samt overlevelsesfunktionen for de ikke-helbredte individer. Klassen af modeller, som tager disse faktorer i betragtning, kaldes helbredelsesmodeller.

Lad T være en stokastisk variabel, der beskriver tiden til begivenheden af interesse. Individerne er i så fald helbredte, hvis begivenheden aldrig opstår, hvilket bekvemt kan skrives som $T = \infty$. Sandsynligheden for helbredelse er derfor $P(T = \infty \mid \mathbf{z})$, og sandsynligheden for ikke at være helbredt er $P(T < \infty \mid \mathbf{z})$. Fordelingsfunktionen for den stokastiske variabel T er givet ved $F(t \mid \mathbf{z}) = P(T \leq t \mid \mathbf{z})$ for $0 \leq t < \infty$, og dermed haves

$$\lim_{t\to\infty} S(t \mid \mathbf{z}) = \lim_{t\to\infty} 1 - F(t \mid \mathbf{z}) = 1 - P(T < \infty \mid \mathbf{z}).$$

Det vil sige, at sandsynligheden for helbredelse er lig med asymptoten for overlevelsesfunktionen $S(t \mid \mathbf{z})$.

I forhold til den specifikke sygdom siges en patient at være statistisk helbredt, hvis denne ikke dør af sygdommen, hvilket svarer til $T_D = \infty$. I tilfælde af kovariater $\mathbf{z}$ er sandsynligheden for dette givet ved $P(T_D = \infty \mid \mathbf{z})$, og $P(T_D < \infty \mid \mathbf{z})$

er sandsynligheden for ikke at være statistisk helbredt. Under antagelsen om, at $S_O(t \mid \mathbf{z}) = S^*(t \mid \mathbf{z})$ og betinget uafhængighed af T_D og T_O givet $\mathbf{z}$, haves

$$\begin{aligned}\lim_{t\to\infty} R(t \mid \mathbf{z}) &= \lim_{t\to\infty} \frac{S(t \mid \mathbf{z})}{S_O(t \mid \mathbf{z})} = \lim_{t\to\infty} \frac{P(T_D > t, T_O > t \mid \mathbf{z})}{P(T_O > t \mid \mathbf{z})} \\ &= \lim_{t\to\infty} \frac{P(T_D > t \mid T_O > t, \mathbf{z}) P(T_O > t \mid \mathbf{z})}{P(T_O > t \mid \mathbf{z})} \\ &= \lim_{t\to\infty} P(T_D > t \mid \mathbf{z}) = \lim_{t\to\infty} 1 - F_D(t \mid \mathbf{z}) \\ &= 1 - \lim_{t\to\infty} P(T_D \leq t \mid \mathbf{z}).\end{aligned}$$

Det vil sige, at asymptoten for $R(t \mid \mathbf{z})$ er lig med sandsynligheden for aldrig at dø af sygdommen og dermed også sandsynligheden for at være statistisk helbredt. Denne asymptote forekommer, når overlevelsen for patienterne når det samme niveau som overlevelsen i baggrundsbefolkningen. Alt dette er under antagelsen om betinget uafhængighed af T_D og T_O givet $\mathbf{z}$, som ikke kan undersøges ud fra data. Antagelsen kan virke utilstrækkelig i nogle situationer. Et eksempel herpå er, hvis en patient lider af flere komorbiditeter, kan denne være i risiko for at dø af flere årsager.

Alternativt kan helbredelsesmodeller anvendes. De to primære klasser af helbredelsesmodeller, der er beskrevet i litteraturen, er mixtur helbredelsesmodeller og ikke-mixtur helbredelsesmodeller, [Lambert et al., 2007b]. I følgende afsnit præsenteres mixtur helbredelsesmodeller.

3.1 Mixtur helbredelsesmodeller

I en mixtur helbredelsesmodel betragtes den observerede population som to grupper; de helbredte individer og de ikke-helbredte individer. Lad $Y_i = \mathbb{1}[T < \infty]$ være en stokastisk variabel, som beskriver helbredelsestilstanden for individ i. Det følger, at $Y_i = 1$ når $\delta_i = 1$, men censurering påvirker både de helbredte og ikke-helbredte individer, da begivenheden af interesse aldrig forekommer og $T = \infty$ ikke er en mulighed. Det vil sige, at Y_i er ukendt når $\delta_i = 0$, og derfor er $\mathbf{Y}$ kun delvist kendt. Overlevelsesfunktionen for populationen i denne model er givet ved

$$S(t \mid \mathbf{z}) = \pi(\mathbf{z}) + [1 - \pi(\mathbf{z})]\widetilde{S}_u(t \mid \mathbf{z}), \tag{3.1}$$

hvor $\pi(\mathbf{z}) = P(Y = 0 \mid \mathbf{z})$ er andelen af helbredte individer, og $1 - \pi(\mathbf{z})$ er andelen af ikke-helbredte individer, som på et tidspunkt vil opleve begivenheden af interesse. Overlevelsesfunktionen for de ikke-helbredte individer $\widetilde{S}_u(t \mid \mathbf{z})$ er en egentlig

overlevelsesfunktion med $\widetilde{S}_u(0 \mid \mathbf{z}) = 1$ og $\lim_{t\to\infty} \widetilde{S}_u(t \mid \mathbf{z}) = 0$. Dette betyder, at overlevelsesfunktionen for populationen, $S(t \mid \mathbf{z})$, er uegentlig med $\lim_{t\to\infty} S(t \mid \mathbf{z}) = \pi$. Overlevelsesfunktionen $\widetilde{S}_u(t \mid \mathbf{z})$ kan blandt andet modelleres parametrisk ved en Weibull-, log-normal- eller eksponential-fordeling, og $\pi(\mathbf{z})$ ved en logistisk regression, [Lambert et al., 2007b]. Kovariaterne $\mathbf{z}$ for $\pi(\mathbf{z})$ og $\widetilde{S}_u(t \mid \mathbf{z})$ kan være forskellige. Det kan for eksempel være, at $\widetilde{S}_u(t \mid \mathbf{z})$ kun er afhængig af en delmængde af kovariaterne for $\pi(\mathbf{z})$. For at holde notationen simpel, anvendes $\mathbf{z}$ for både $\pi(\mathbf{z})$ og $\widetilde{S}_u(t \mid \mathbf{z})$. Hazard-funktionen tilhørende overlevelsesfunktionen i Ligning (3.1) er

$$\begin{aligned} h(t \mid \mathbf{z}) &= -\frac{d\ln[S(t \mid \mathbf{z})]}{dt} = -\frac{1}{\pi(\mathbf{z}) + [1-\pi(\mathbf{z})]\widetilde{S}_u(t \mid \mathbf{z})}[1-\pi(\mathbf{z})]\frac{d\widetilde{S}_u(t \mid \mathbf{z})}{dt} \\ &= \frac{[1-\pi(\mathbf{z})]\widetilde{f}_u(t \mid \mathbf{z})}{\pi(\mathbf{z}) + [1-\pi(\mathbf{z})]\widetilde{S}_u(t \mid \mathbf{z})}, \end{aligned} \tag{3.2}$$

hvor $\widetilde{f}_u(t \mid \mathbf{z}) = -\frac{d\widetilde{S}_u(t|\mathbf{z})}{dt}$ er tæthedsfunktionen tilhørende overlevelsesfunktionen $\widetilde{S}_u(t \mid \mathbf{z})$. Log-likelihoodfunktionen for mixtur helbredelsesmodellen kan dermed bestemmes ved at substituere overlevelsesfunktionen fra Ligning (3.1) og hazard-funktionen fra Ligning (3.2) i Ligning (A.3):

$$\begin{aligned} l &= \sum_{i=1}^{n} \delta_i \ln[h(x_i \mid \mathbf{z}_i)] + \ln[S(x_i \mid \mathbf{z}_i)] \\ &= \sum_{i=1}^{n} \delta_i \ln\left\{\frac{[1-\pi(\mathbf{z}_i)]\widetilde{f}_u(x_i \mid \mathbf{z}_i)}{\pi(\mathbf{z}_i) + [1-\pi(\mathbf{z}_i)]\widetilde{S}_u(x_i \mid \mathbf{z}_i)}\right\} + \ln\left\{\pi(\mathbf{z}_i) + [1-\pi(\mathbf{z}_i)]\widetilde{S}_u(x_i \mid \mathbf{z}_i)\right\}. \end{aligned}$$

Helbredelsesmodellen tilpasses ved at maksimere denne log-likelihoodfunktion.

Helbredelsesmodeller kan anvendes til at analysere tid til død for kræftpatienter, hvor formålet er at estimere andelen af langtids-overlevende patienter. Det er ofte tilfældet, at den totale overlevelse anvendes til disse analyser, [Othus et al., 2012], men behandlinger og prognoser for mange sygdomme forbedres hele tiden, hvilket medvirker til længere opfølgningstider. Det er derfor almindeligt, at en stor andel af dødsfaldene er fra andre årsager end sygdommen. Særligt for ældre patienter er alderdom og andre sygdomme ofte en dødsårsag. Den totale overlevelse i Ligning (3.1) er derfor ikke specielt oplysende om overlevelsen for en specifik sygdom. I

stedet kan den totale overlevelse i Ligning (3.1) udvides til at inkludere den relative overlevelse ved, [Lambert et al., 2007b],

$$S(t \mid \mathbf{z}) = S^*(t \mid \mathbf{z}) \{\pi(\mathbf{z}) + [1 - \pi(\mathbf{z})]S_u(t \mid \mathbf{z})\} = S^*(t \mid \mathbf{z})R(t \mid \mathbf{z}). \tag{3.3}$$

I denne model antages det, at de helbredte individer har den samme overlevelse som baggrundsbefolkningen $S^*(t \mid \mathbf{z})$, mens de ikke-helbredte individer har en lavere ukendt overlevelse $S^*(t \mid \mathbf{z})S_u(t \mid \mathbf{z})$. Helbredelsesmodellen, der beskriver den relative overlevelse, er derfor en mixtur model givet ved

$$R(t \mid \mathbf{z}) = \pi(\mathbf{z}) + [1 - \pi(\mathbf{z})]S_u(t \mid \mathbf{z}). \tag{3.4}$$

I forbindelse med den relative overlevelse beskriver Y_i om, individ i er statistisk helbredt. Der haves $Y_i = 0$ for statistisk helbredte individer og $Y_i = 1$ for de ikke-statistisk helbredte individer. Det vil sige, at $\pi(\mathbf{z}) = P(Y = 0 \mid \mathbf{z})$ er andelen af statistisk helbredte individer, og $1 - \pi(\mathbf{z})$ er andelen af ikke-statistisk helbredte individer. Der skrives ofte blot helbredte individer og ikke-helbredte individer, selvom der er tale om statistisk helbredte individer og ikke-statistisk helbredte individer. Det fremgår af Ligning (3.4), at den relative overlevelse for de helbredte individer er konstant lig med 1, mens de ikke-helbredte individer har en lavere ukendt relativ overlevelse $S_u(t \mid \mathbf{z})$, som antages at gå mod 0 for $t \to \infty$. Det betyder også, at $R(t \mid \mathbf{z}) \to \pi(\mathbf{z})$ for $t \to \infty$. Overlevelsesfunktionen $S_u(t \mid \mathbf{z})$ kan blandt andet, ligesom $\widetilde{S}_u(t \mid \mathbf{z})$, modelleres parametrisk ved en Weibull-, log-normal- eller eksponential-fordeling, og $\pi(\mathbf{z})$ ved en logistisk regression, [Lambert et al., 2007b].

Fortolkningen af $\widetilde{S}_u(t \mid \mathbf{z})$ i Ligning (3.1) og $S_u(t \mid \mathbf{z})$ i Ligning (3.3) er forskellige. I Ligning (3.1) beskriver $\widetilde{S}_u(t \mid \mathbf{z})$ den totale overlevelse for de ikke-helbredte individer, hvorimod $S_u(t \mid \mathbf{z})$ i Ligning (3.3) beskriver den relative overlevelse for de ikke-helbredte individer. Denne adskillelse er vigtig og er årsagen til tilde-notationen i forbindelse med Ligning (3.1).

Overlevelsesfunktionen i Ligning (3.3) har den tilhørende hazard-funktion

$$\begin{aligned} h(t \mid \mathbf{z}) &= -\frac{d\ln[S(t \mid \mathbf{z})]}{dt} = -\frac{d\ln[S^*(t \mid \mathbf{z})]}{dt} - \frac{d\ln\{\pi(\mathbf{z}) + [1 - \pi(\mathbf{z})]S_u(t \mid \mathbf{z})\}}{dt} \\ &= h^*(t \mid \mathbf{z}) + \frac{[1 - \pi(\mathbf{z})]f_u(t \mid \mathbf{z})}{\pi(\mathbf{z}) + [1 - \pi(\mathbf{z})]S_u(t \mid \mathbf{z})}. \end{aligned} \tag{3.5}$$

Heraf er den forøgede hazard-funktion givet ved

$$\lambda(t \mid \mathbf{z}) = \frac{[1-\pi(\mathbf{z})]f_u(t \mid \mathbf{z})}{\pi(\mathbf{z}) + [1-\pi(\mathbf{z})]S_u(t \mid \mathbf{z})}.$$

Log-likelihoodfunktionen for denne mixtur helbredelsesmodel kan nu bestemmes ved at substituere overlevelsesfunktionen fra Ligning (3.3) og hazard-funktionen fra Ligning (3.5) i Ligning (A.3):

$$\begin{aligned} l &= \sum_{i=1}^{n} \delta_i \ln[h(x_i \mid \mathbf{z}_i)] + \ln[S(x_i \mid \mathbf{z}_i)] \\ &= \sum_{i=1}^{n} \delta_i \ln\left[h^*(x_i \mid \mathbf{z}_i) + \lambda(t_i \mid \mathbf{z}_i)\right] + \ln[S^*(x_i \mid \mathbf{z}_i)] + \ln\{\pi(\mathbf{z}_i) + [1-\pi(\mathbf{z}_i)]S_u(x_i \mid \mathbf{z}_i)\} \\ &= \sum_{i=1}^{n} \delta_i \ln\left[h^*(x_i \mid \mathbf{z}_i) + \lambda(x_i \mid \mathbf{z}_i)\right] + \ln[S^*(x_i \mid \mathbf{z}_i)] + \ln[R(x_i \mid \mathbf{z}_i)]. \end{aligned}$$

Dette kan reduceres til

$$l = \sum_{i=1}^{n} \delta_i \ln\left[h^*(x_i \mid \mathbf{z}_i) + \lambda(x_i \mid \mathbf{z}_i)\right] + \ln[R(x_i \mid \mathbf{z}_i)], \tag{3.6}$$

da $S^*(x_i \mid \mathbf{z}_i)$ er uafhængig af modellens ukendte parametre. Det er derfor unødvendigt at udregne $S^*(x_i \mid \mathbf{z}_i)$, hvilket kræver integration af den rapporterede hazard i levetidstabellerne. Den eneste ekstra nødvendige information er $h^*(t \mid \mathbf{z})$ evalueret i de observerede tidspunkter, som kan aflæses direkte ud fra levetidstabellerne.

Maksimering af Ligning (3.6) kan udføres ved hjælp af iterative metoder som for eksempel Newtons metode. Den inverse informationsmatrix giver desuden et asymptotisk estimat af kovariansmatricen for parameterestimaterne. En test kan udføres ved likelihood-ratio teststørrelsen for at undersøge, hvorvidt der er en signifikant forskel mellem to modeller, M_1 og M_2. Denne er givet ved

$$T = -2[\log(L_{M_1}) - \log(L_{M_2})] = 2[\log(L_{M_2}) - \log(L_{M_1})],$$

hvor M_1 er en reduceret model af M_2, mens L_{M_1} og L_{M_2} er den største værdi af likelihoodfunktionen for henholdsvis M_1 og M_2. Likelihood-ratio teststørrelsen følger en χ^2-fordeling med $M_2 - M_1$ frihedsgrader, [De Angelis et al., 1999].

3.1.1 Konfidensinterval

Variansen af $R(t), S_u(t)$ og π kan estimeres ved at anvende delta-metoden i Sætning A.2. Variansen er givet som

$$\mathbb{V}ar\left[R(t;\widehat{\boldsymbol{\beta}})\right] = \left(\frac{\partial R(t;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^{\top} \widehat{\Sigma} \left(\frac{\partial R(t;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right),$$

$$\mathbb{V}ar\left[S_u(t;\widehat{\boldsymbol{\beta}})\right] = \left(\frac{\partial S_u(t;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^{\top} \widehat{\Sigma} \left(\frac{\partial S_u(t;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right),$$

$$\mathbb{V}ar\left[\pi(\widehat{\boldsymbol{\beta}})\right] = \left(\frac{\partial \pi(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^{\top} \widehat{\Sigma} \left(\frac{\partial \pi(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right),$$

hvor $\widehat{\boldsymbol{\beta}}$ er maksimum likelihood estimatet, og $\widehat{\Sigma}$ er dets kovariansmatrix. Konfidensintervallerne for $R(t;\widehat{\boldsymbol{\beta}})$ og $S_u(t;\widehat{\boldsymbol{\beta}})$ bestemmes desuden punktvis. Det skal bemærkes, at $\widehat{\boldsymbol{\beta}}$ og $\widehat{\Sigma}$ er forskellige i $R(t;\widehat{\boldsymbol{\beta}})$, $S_u(t;\widehat{\boldsymbol{\beta}})$ og $\pi(\widehat{\boldsymbol{\beta}})$.

3.2 Dataanalyse for mixtur helbredelsesmodeller

I dette afsnit analyseres coloncancer-datasættet ved hjælp af mixtur helbredelsesmodellen for den relative overlevelse i Ligning (3.4). Der opstilles tre modeller, og i første omgang inkluderes der ikke kovariater i modellerne. Modellerne er givet ved

$$\begin{aligned} R(t) &= \pi + (1-\pi)S_u(t), \\ \pi &= \frac{\exp(\beta_0)}{1+\exp(\beta_0)}, \end{aligned} \tag{3.7}$$

og $S_u(t)$, som modelleres med henholdsvis en Weibull-, log-normal- og eksponentialfordeling, se fordelingerne i Afsnit A.2 i appendiks. Der henvises fremadrettet til modellerne ved Weibull-, log-normal- og eksponential-modellen.

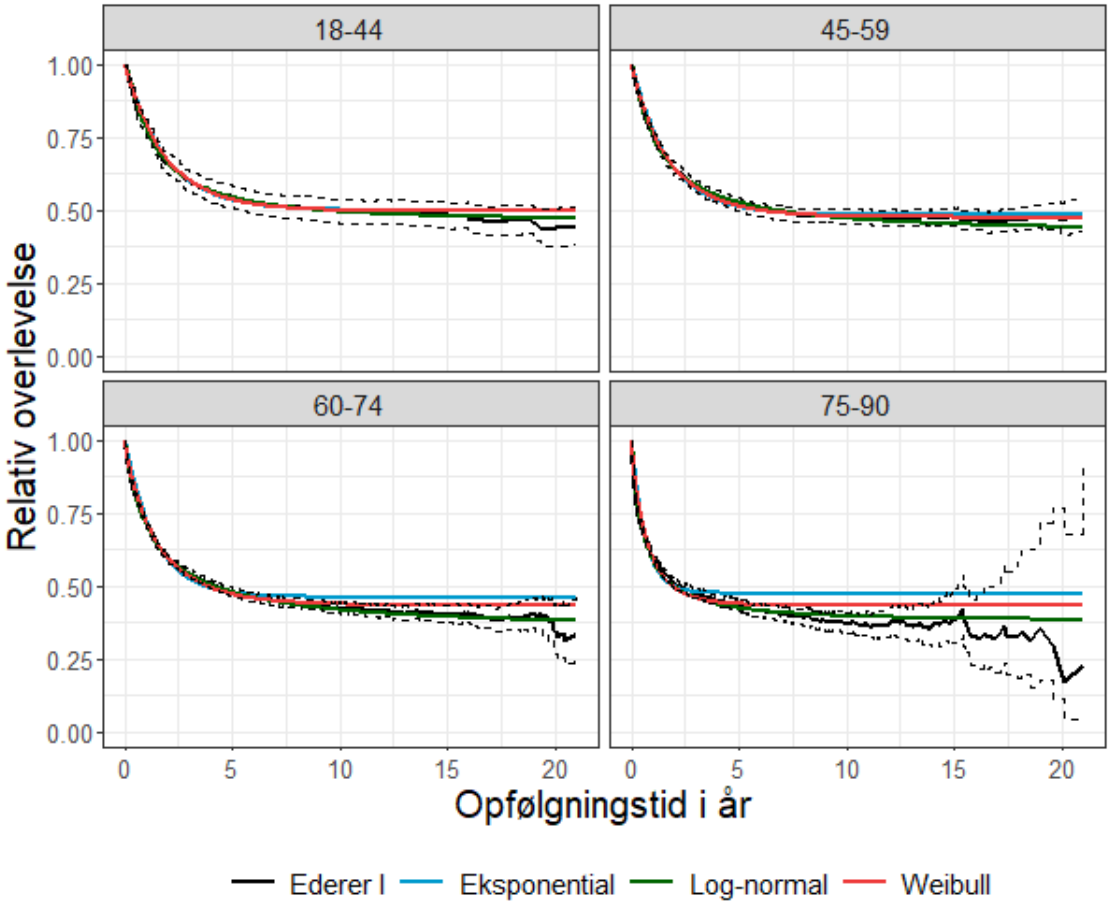
En oplagt mulighed er at sammenligne modellerne ved Akaikes informationskriterie (AIC), som udregnes ved

$$\text{AIC} = -2\log(L_{\max}) + 2P, \tag{3.8}$$

hvor $L_{\max}$ er den største værdi af likelihoodfunktionen, og P er antallet af parametre i modellen. Anvendelsen af AIC har dog tidligere givet problemer i forhold

til helbredelsesmodeller, da den ligger vægt på tilpasningen af $R(t)$ for de t, hvor begivenhederne indtræffer. Dette er ofte i starten af opfølgningen. Det betyder, at en model med en lavere AIC ikke nødvendigvis giver et bedre estimat af den relative overlevelse i slutningen af opfølgningen og derfor heller ikke i forhold til andelen af helbredte patienter, [Andersson, 2013] og [Lambert et al., 2007b].

Alternativt kan modellerne også sammenlignes med et ikke-parametrisk estimat for at få en idé om, hvorvidt modellerne beskriver den relative overlevelse godt. Dette er tidligere blevet gjort i litteraturen, [Andersson et al., 2010] og [Lambert et al., 2007a]. Vi har valgt at sammenligne modellerne ved hjælp af AIC, da der ikke findes noget bedre, så vidt vi ved. Det er dog vigtigt at tage problematikken med AIC i betragtning. Vi sammenligner derfor også modellerne med Ederer I for at sikre, at modellerne estimerer den relative overlevelse tilstrækkeligt i slutningen af opfølgningen. Det skal dog nævnes, at Ederer I ikke beskriver den sande relative overlevelse. Figur 3.1 illustrerer den estimerede relative overlevelse, $R(t)$, for de tre modeller sammen med et ikke-parametrisk estimat udregnet ved Ederer I. Der er stratificeret efter aldersgrupperne 18-44, 45-59, 60-74 og 75-90. Det vil sige, at modellerne tilpasses de fire aldersgrupper hver for sig. Dette er gjort for at få en idé om, hvorvidt modellerne er bedre for visse aldersgrupper.



Figur 3.1: Den relative overlevelse for coloncancer-datasættet udregnet ved Ederer I med tilhørende 95% konfidensinterval (stiplede linjer) samt Weibull-, log-normal- og eksponential-modellen for aldersgrupperne 18-44, 45-59, 60-74 og 75-90.

Følgende tabel opsummerer modellernes AIC for de fire aldersgrupper.

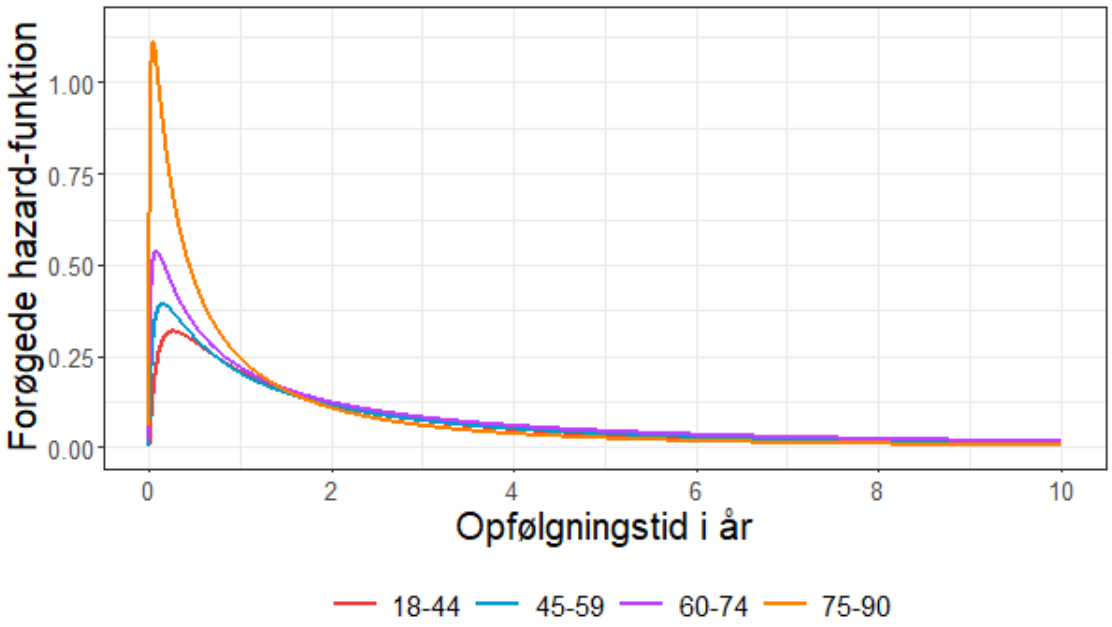
Model	18-44	45-59	60-74	75-90
Wei	2184.68	7069.72	20950.83	14757.63
Log	2157.42	7041.72	20792.64	14436.74
Exp	2184.94	7077.58	21028.97	14845.72

Tabel 3.1: AIC for Weibull-, log-normal- og eksponential-modellen for aldersgrupperne 18-44, 45-59, 60-74 og 75-90.

Det fremgår af Figur 3.1, at alle tre modeller følger Ederer I-estimatet ved aldersgrupperne 18-44 og 45-59. For aldersgruppen 60-74 ligger både Weibull- og

eksponential-modellen over Ederer I, mens log-normal-modellen stadig følger Ederer I-estimatet tilnærmelsesvist. For aldersgruppen 75-90 bemærkes det, at Ederer I-estimatet er utilregneligt efter 15 år. Dette skyldes en stor varians, som forekommer, da der ikke er mange patienter i risiko efter 15 år for denne aldersgruppe. I de første 15 år observerer vi, at Weibull- og eksponential-modellen ligger over Ederer I-estimatet, mens log-normal-modellen følger den tilnærmelsesvist. Vi observerer også, at modellerne flader ud ved omkring 0.5 for alle fire aldersgrupper. Dette sker en smule hurtigere for aldersgruppen 75-90, hvilket betyder, at deres relative overlevelse er dårligere i starten af opfølgningen end de resterende aldersgrupper. Dette giver god mening, da ældre patienter som regel er mindre robuste i forhold til sygdomme. Ud fra Tabel 3.1 observerer vi, at log-normal-modellen resulterer i den laveste AIC for alle fire aldersgrupper. Det konkluderes dermed, at log-normal-modellen er den bedste model, da den følger Ederer I bedst samt resulterer i den laveste AIC.

Følgende figur illustrerer den forøgede hazard-funktion for log-normal-modellen for de fire aldersgrupper.



Figur 3.2: Den forøgede hazard-funktion for log-normal-modellen for aldersgrupperne 18-44, 45-59, 60-74 og 75-90.

Det fremgår af Figur 3.2, at den forøgede hazard-funktion er størst for alle aldersgrupperne i starten af opfølgningen. Det virker urealistisk, at de forøgede hazard-

funktioner vokser kraftigt i starten af opfølgningen. Det havde været mere logisk, hvis de i stedet havde startet i deres højeste punkt og derefter været aftagende. Dette er en direkte konsekvens af anvendelsen af log-normal-modellen. De forøgede hazard-funktioner for Weibull- og eksponential-modellen ville have startet i deres højeste punkt og derefter været aftagende. Log-normal-modellen har dog vist at være den bedste af de tre modeller, hvilket umiddelbart skyldes, at de forøgede hazard-funktioner aftager bedre for log-normal-modellen. Det observeres også, at den forøgede hazard-funktion er markant højere for de ældre patienter i starten af opfølgningen. Efter den hårde startperiode begynder den forøgede hazard-funktion at aftage for alle aldersgrupperne, og efter cirka 1 år følger de forøgede hazard-funktioner tilnærmelsesvist hinanden. Det fremgår desuden, at patienterne næsten ikke oplever nogen forøget hazard efter 5 år.

I tabellen, der følger, er den 2 års relative overlevelse, den 5 års relative overlevelse, andelen af helbredte patienter og median relativ overlevelsestiden for de ikke-helbredte patienter angivet for Weibull-, log-normal- og eksponential-modellen stratificerede efter aldersgrupperne 18-44, 45-59, 60-74 og 75-90. Median relativ overlevelsestiden for de ikke-helbredte patienter er bestemt ved at løse

$$S_u(t) = 0.5 \iff t = S_u^{-1}(0.5).$$

Det vil sige tidspunktet, hvor den relative overlevelse for de ikke-helbredte patienter er 0.5. Denne beskriver, hvornår den relative overlevelse for de ikke-helbredte patienter er halvt så stor som den matchende baggrundsbefolkning. Den giver altså en idé om, hvor hurtigt $R(t)$ aftager. Der angives ikke konfidensintervaller for median relativ overlevelsestiden for de ikke-helbredte patienter, da estimatet ikke er implementeret i R-pakken cuRe. R-koden til vores implementation af median relativ overlevelsestiden for de ikke-helbredte patienter findes i Afsnit C.1 i appendiks.

	Model	18-44	45-59	60-74	75-90
2 års	Wei	0.67(0.64-0.7)	0.64(0.63-0.66)	0.6(0.59-0.61)	0.49(0.48-0.51)
relativ	Log	0.66(0.63-0.69)	0.64(0.62-0.66)	0.6(0.59-0.61)	0.5(0.48-0.51)
overlevelse	Exp	0.67(0.64-0.7)	0.64(0.63-0.66)	0.59(0.58-0.6)	0.5(0.48-0.51)
5 års	Wei	0.53(0.5-0.57)	0.51(0.49-0.53)	0.47(0.46-0.49)	0.44(0.42-0.46)
relativ	Log	0.54(0.51-0.58)	0.53(0.51-0.55)	0.48(0.47-0.49)	0.42(0.41-0.44)
overlevelse	Exp	0.53(0.5-0.57)	0.51(0.49-0.53)	0.48(0.46-0.49)	0.48(0.46-0.49)
Andel	Wei	0.5(0.45-0.54)	0.48(0.45-0.5)	0.43(0.42-0.45)	0.43(0.41-0.45)
helbredte	Log	0.46(0.42-0.51)	0.42(0.39-0.46)	0.35(0.33-0.38)	0.38(0.35-0.41)
patienter	Exp	0.5(0.46-0.54)	0.48(0.46-0.51)	0.46(0.45-0.48)	0.48(0.46-0.49)
Median	Wei	1.3	1.18	1.01	0.48
relativ over-	Log	1.28	1.25	1.19	0.5
levelsestid	Exp	1.3	1.19	0.97	0.44

Tabel 3.2: Den 2 års relative overlevelse, den 5 års relative overlevelse, andelen af helbredte patienter og median relativ overlevelsestiden i år for de ikke-helbredte patienter. De tilhørende 95% konfidensintervaller er angivet i parentes.

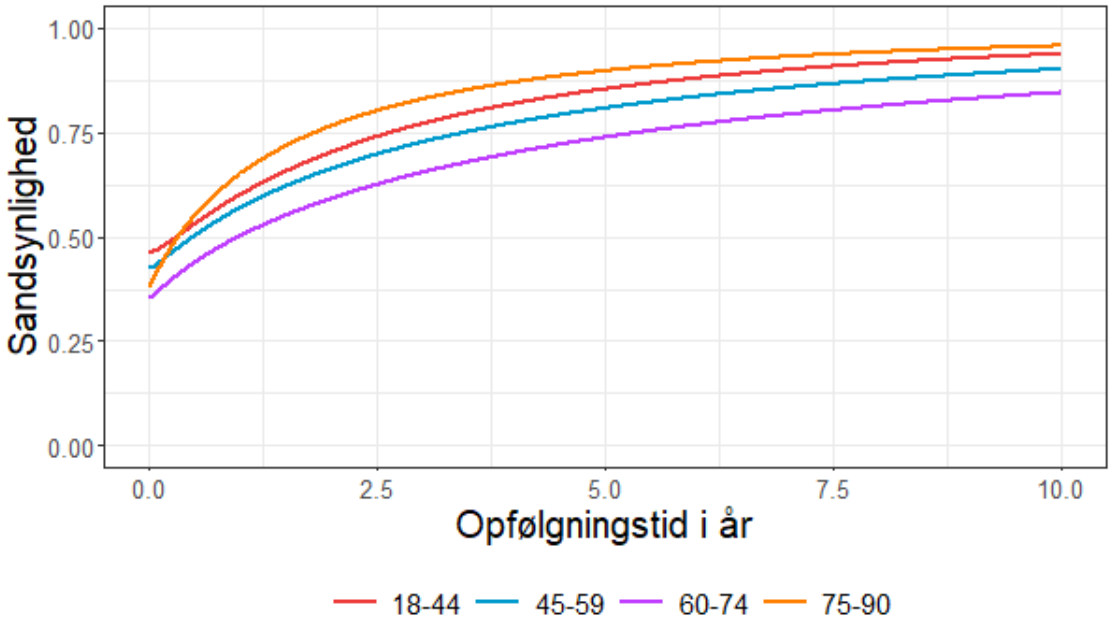
Det fremgår af Tabel 3.2, at estimaterne ikke varierer meget imellem modellerne. Den største variation forekommer ved andelen af helbredte patienter for de to ældste aldersgrupper. Det fremgår desuden, at den 2 års relative overlevelse falder med alderen. For aldersgruppen 18-44 ligger den 2 års relative overlevelse på cirka 0.67, mens den er cirka 0.5 for aldersgruppen 75-90. For den 5 års relative overlevelse er forskellen mellem aldersgrupperne mindre, men den er stadig lavere for aldersgruppen 75-90. Det observerede vi også i forbindelse med Figur 3.1. For andelen af helbredte patienter observeres der ikke den store forskel mellem 18-44 og 45-59 samt 60-74 og 75-90. De ældre patienter klarer sig dog dårligere end de yngre patienter. Aldersgruppen 75-90 har desuden en markant lavere median relativ overlevelsestid for de ikke-helbredte patienter i forhold til de resterende tre aldersgrupper. Det generelle billede fortæller os, at aldersgruppen 75-90 klarer sig dårligere i forhold til coloncancer end de resterende tre aldersgrupper. Dette gør sig særligt gældende i starten af opfølgningen, hvilket kan skyldes, at de er mindre robuste i forhold til sygdomme. Det fremgår tværtimod af den 5 års relative overlevelse og andelen af helbredte patienter, at de ældre patienter begynder at klare sig bedre, hvis de kommer igennem den hårde startperiode. Dette fremgik også af Figur 3.2, hvor vi observe-

rede en markant højere forøget hazard-funktion for de ældre patienter i starten af opfølgningen.

Andelen af helbredte patienter og median relativ overlevelsestiden for de ikke-helbredte patienter giver et indblik i sygdommens alvorlighed, men sandsynligheden for statistisk helbredelse som en funktion af tid kan også være nyttig. Denne er bestemt ved, [Sposto, 2002],

$$P(Y = 0 \mid T > t) = \frac{P(Y = 0, T > t)}{P(T > t)} = \frac{P(Y = 0)}{P(T > t)} = \frac{\pi}{\pi + (1 - \pi)S_u(t)}.$$

Følgende figur illustrerer sandsynligheden for helbredelse som en funktion af tiden for log-normal-modellen for de fire aldersgrupper.



Figur 3.3: Sandsynligheden for statistisk helbredelse for log-normal-modellen for aldersgrupperne 18-44, 45-59, 60-74 og 75-90.

Det observeres på Figur 3.3, at sandsynligheden for statistisk helbredelse er lavest for aldersgruppen 60-74 over hele opfølgningen, mens den er størst for aldersgruppen 75-90. Desuden fremgår det eksempelvis også, at sandsynligheden for statistisk helbredelse 5 år efter coloncancer-diagnosen er cirka 75% for en patient i aldersgruppen 60-74.

3.2.1 Kovariater

Udover at analysere patienterne i forhold til de fire aldersgrupper, er det også interessant at inddrage diagnoseperiode. Dette kan give et indblik i, om sundhedsvæsenet er blevet bedre til at behandle coloncancer i perioden 1975-1994. Der tilpasses en model, hvor $S_u(t \mid \mathbf{z})$ og $\pi(\mathbf{z})$ er afhængige af aldersgruppe og diagnoseperiode. Aldersgruppe inkluderes som en kategorisk variabel med aldersgrupperne 18-44, 45-59, 60-74 og 75-90, og diagnoseperiode inkluderes også som en kategorisk variabel med diagnoseperioderne 1975-1984 og 1985-1994. Der er desuden også inkluderet en vekselvirkning mellem aldersgruppe og diagnoseperiode. Modellen er givet ved

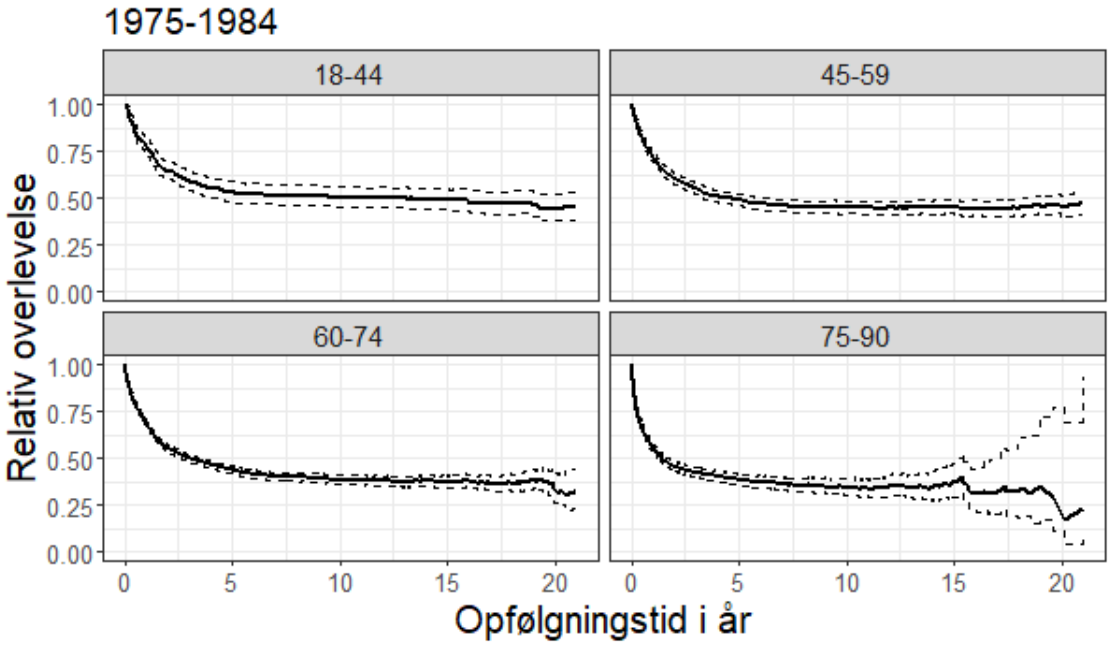
$$R(t \mid \mathbf{z}) = \pi(\mathbf{z}) + [1 - \pi(\mathbf{z})]S_u(t \mid \mathbf{z}),$$

$$\pi(\mathbf{z}) = \frac{\exp(\boldsymbol{\beta}^\top \mathbf{z})}{1 + \exp(\boldsymbol{\beta}^\top \mathbf{z})},$$

og $S_u(t \mid \mathbf{z})$, som modelleres med en Weibull-fordeling, hvor både γ_1 og γ_2 i fordelingen er afhængige af kovariaterne aldersgruppe, diagnoseperiode og en vekselvirkning mellem disse, se Weibull-fordelingen i Afsnit A.2 i appendiks.

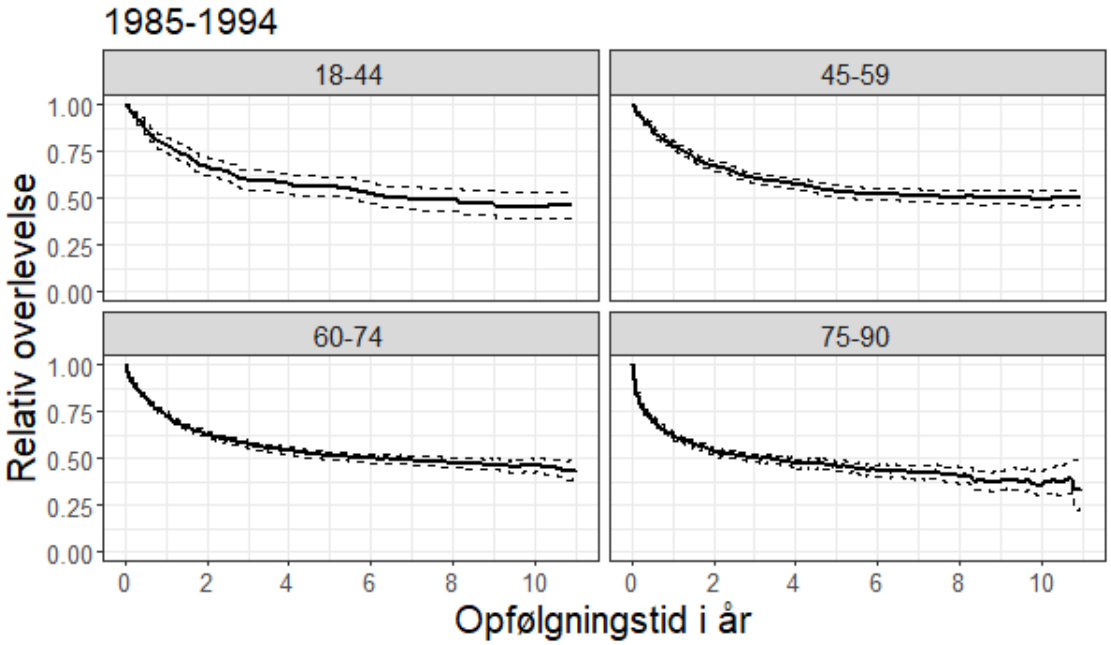
Vi har valgt at opstille en Weibull-model, da denne gav et bedre resultat end log-normal- og eksponential-modellen, når vi sammenlignede modellerne med modeller stratificerede efter aldersgruppe og diagnoseperiode. Dette betyder, at Weibull-modellen fanger effekten af kovariaterne bedst. Vi oplevede desuden også, at log-normal- og eksponential-modellen med kovariater resulterede i, at aldersgruppen 18-44 havde en dårligere relativ overlevelse end aldersgruppen 45-59, hvilket virker besynderligt. En likelihood-ratio test viser, at aldersgruppe, diagnoseperiode og vekselvirkningen mellem aldersgruppe og diagnoseperiode er signifikant ($P < 0.001$) for Weibull-modellen.

Det er vigtigt at bemærke, at opfølgningstiden for patienterne diagnosticeret i perioden 1985-1994 er kortere end for patienterne diagnosticeret i perioden 1975-1984. For aldersgruppen 18-44 er den maksimale opfølgningstid 10.89 år, mens den er 10.98 år for aldersgrupperne 45-59, 60-74 og 75-90. Vi ønsker derfor at undersøge, om det er tilstrækkeligt at antage statistisk helbredelse for alle grupperne. Følgende figur illustrerer Ederer I stratificeret efter aldersgruppe for diagnoseperioden 1975-1984.



Figur 3.4: Den relative overlevelse udregnet ved Ederer I med tilhørende 95% konfidensinterval (stiplede linjer) stratificeret efter aldersgruppe for diagnoseperioden 1975-1984.

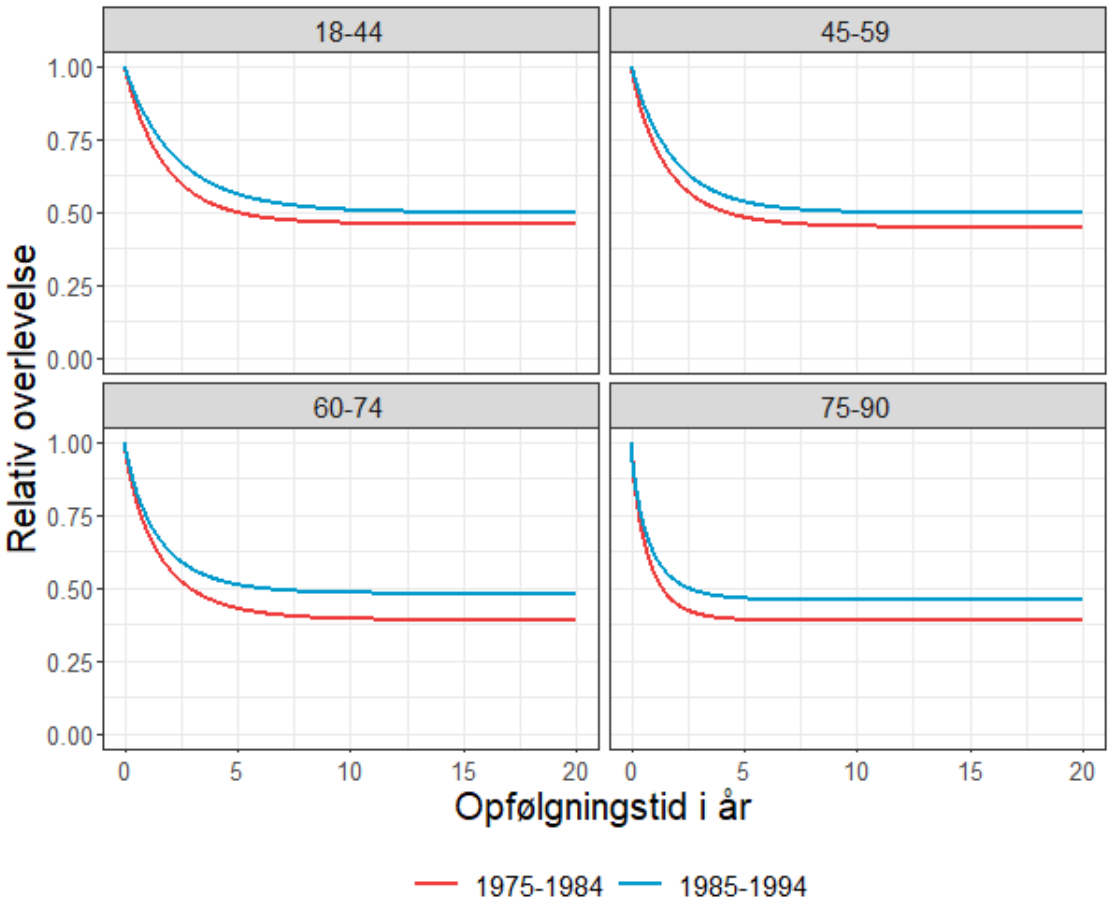
Det fremgår af Figur 3.4, at der forekommer en udfladning af Ederer I-estimatet i løbet af opfølgningen for aldersgrupperne 18-44, 45-59 og 60-74. For aldersgruppen 75-90 er variansen stor efter 15 år, fordi der ikke er mange patienter i risiko for denne gruppe derefter. Det tyder dog på, at der forekommer en udfladning i løbet af de første 15 år. På den næste figur undersøges diagnoseperioden 1985-1994.



Figur 3.5: Den relative overlevelse udregnet ved Ederer I med tilhørende 95% konfidensinterval (stiplede linjer) stratificeret efter aldersgruppe for diagnoseperioden 1985-1994.

Det observeres på Figur 3.5, at der også forekommer en udfladning af Ederer I i løbet af opfølgningen for aldersgrupperne 18-44 og 45-59, selvom opfølgningstiden er kortere. Det kan dog diskuteres, om det er tilfældet for de to ældste aldersgrupper. Det fremgår desuden af Figur B.1 og Figur B.2 i appendiks, at Weibull-modellen tilnærmelsesvist følger Ederer I for aldersgrupperne 45-59 og 60-74 for begge diagnoseperioder, men samtidig afviger for aldersgrupperne 18-44 og 75-90. Det kan derfor diskuteres, hvor godt Weibull-modellen med kovariater beskriver den relative overlevelse for aldersgrupperne 18-44 og 75-90 for diagnoseperioderne 1975-1984 og 1985-1994.

De to diagnoseperioder for Weibull-modellen sammenlignes på følgende figur.



Figur 3.6: Den relative overlevelse bestemt ved Weibull-modellen for aldersgrupperne 18-44, 45-59, 60-74 og 75-90, som er diagnosticerede med coloncancer i henholdsvis 1975-1984 og 1985-1994.

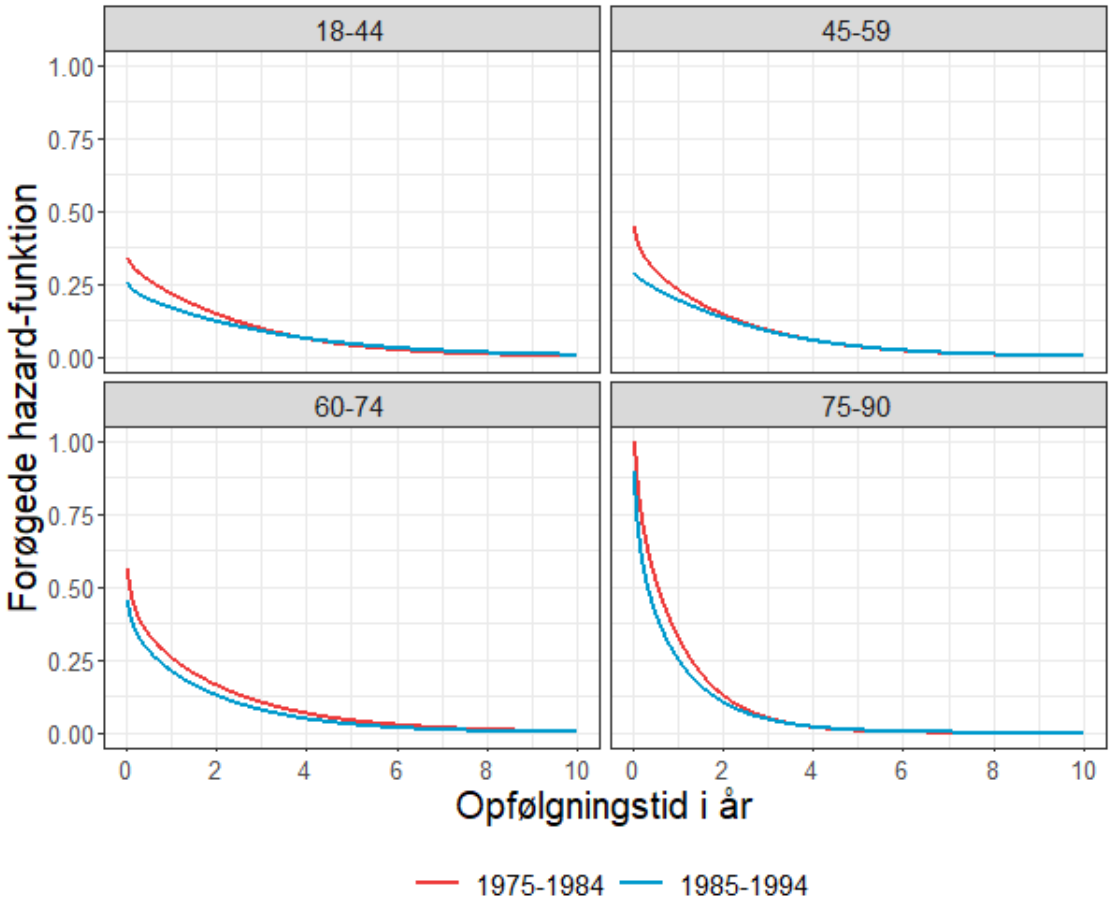
Det fremgår af Figur 3.6, at den relative overlevelse er højere i 1985-1994 end i 1975-1984 for alle aldersgrupperne. Det fremgår desuden, at forskellen mellem de to diagnoseperioder er større for aldersgrupperne 60-74 og 75-90 end for 18-44 og 45-59, hvilket er en direkte effekt af vekselvirkningen. Det tyder altså på, at sundhedsvæsenet er blevet bedre til at behandle coloncancer - særligt hos de ældre patienter. Følgende tabel opsummerer den 2 års relative overlevelse, den 5 års relative overlevelse, andelen af helbredte patienter og median relativ overlevelsestiden for de ikke-helbredte patienter.

	Periode	18-44	45-59	60-74	75-90
2 års relativ overlevelse	1975-1984	0.64(0.59-0.69)	0.61(0.58-0.64)	0.56(0.55-0.58)	0.45(0.42-0.47)
	1985-1994	0.71(0.67-0.75)	0.67(0.65-0.69)	0.62(0.61-0.64)	0.52(0.5-0.54)
5 års relativ overlevelse	1975-1984	0.5(0.45-0.55)	0.48(0.45-0.51)	0.43(0.41-0.45)	0.39(0.36-0.42)
	1985-1994	0.56(0.51-0.61)	0.53(0.51-0.56)	0.51(0.49-0.53)	0.46(0.44-0.49)
Andel helbredte patienter	1975-1984	0.46(0.4-0.51)	0.45(0.41-0.48)	0.39(0.37-0.41)	0.39(0.36-0.42)
	1985-1994	0.5(0.43-0.56)	0.5(0.46-0.53)	0.48(0.45-0.51)	0.46(0.43-0.49)
Median relativ overlevelsestid	1975-1984	1.22	1.04	0.99	0.47
	1985-1994	1.56	1.28	0.99	0.5

Tabel 3.3: Den 2 års relative overlevelse, den 5 års relative overlevelse, andelen af helbredte patienter og median relativ overlevelsestiden i år for de ikke-helbredte patienter. De tilhørende 95% konfidensintervaller er angivet i parentes.

Det fremgår af Tabel 3.3, at patienterne diagnosticerede i 1985-1994 klarer sig bedre end patienterne diagnosticerede i 1975-1984, som også fremgik af Figur 3.6. Derudover observerer vi, at aldersgruppen 75-90 klarer sig dårligst for begge diagnoseperioder. Dette er særligt gældende i starten af opfølgningen, hvilket fremgår af den markant lavere median relativ overlevelsestid for de ikke-helbredte patienter samt den 2 års relative overlevelse. De begynder derimod at klare sig bedre for begge diagnoseperioder, hvis de kommer igennem den hårde startperiode, hvilket fremgår af den 5 års relative overlevelse og andelen af helbredte patienter, som er sammenlignelig med de resterende tre aldersgrupper.

På den næste figur illustreres den forøgede hazard-funktion for de fire aldersgrupper i diagnoseperioderne 1975-1984 og 1985-1994.



Figur 3.7: Den forøgede hazard-funktion for aldersgrupperne 18-44, 45-59, 60-74 og 75-90 i diagnoseperioderne 1975-1984 og 1985-1994.

Det fremgår af Figur 3.7, at den forøgede hazard-funktion generelt er højere for perioden 1975-1984 end for 1985-1994 i starten af opfølgningen, men efter nogle år tilnærmer kurverne sig hinanden. Det tyder derfor på, at sundhedsvæsenet er blevet bedre til at håndtere coloncancer i starten af opfølgningen. Den forøgede hazard-funktion er desuden højest for de ældre patienter, hvilket igen tyder på, at de ældre patienter er mindre robuste i forhold til coloncancer. I forbindelse med Figur 3.2 kommenterede vi på, at det virkede urealistisk, at den forøgede hazard-funktion for log-normal-modellen voksede kraftigt i starten af opfølgningen. Vi nævnte også, at

det ikke havde været tilfældet for Weibull- eller eksponential-modellen. Det fremgår af Figur 3.7, at de forøgede hazard-funktioner for Weibull-modellen starter i deres højeste punkt.

I det næste afsnit præsenteres ikke-mixtur helbredelsesmodeller.

3.3 Ikke-mixtur helbredelsesmodeller

Ikke-mixtur helbredelsesmodellen definerer en asymptote for den kumulerede hazard-funktion, når andelen af helbredte individer eksisterer. Asymptoten for overlevelses-funktionen for hele populationen, når andelen af helbredte individer er større end nul, er $\lim_{t\to\infty} S(t) > 0$. Den kumulerede hazard-funktion defineres som $H(t) = \theta\overline{F}(t)$, hvor $\overline{F}$ er en egentlig fordelingsfunktion, altså $\lim_{t\to\infty} \overline{F}(t) = 1$, og $\theta > 0$. Det vil sige, at den kumulerede hazard-funktion er afgrænset, således at $\lim_{t\to\infty} H(t) = \theta < \infty$. Dermed kan overlevelsesfunktionen ved at inkludere kovariater skrives som, [Tsodikov, 2002],

$$S(t \mid \mathbf{z}) = \exp[-H(t \mid \mathbf{z})] = \exp[-\theta(\mathbf{z})\overline{F}(t \mid \mathbf{z})] = \pi(\mathbf{z})^{\overline{F}(t|\mathbf{z})}. \tag{3.9}$$

Modellen i Ligning (3.9) blev motiveret af en biologisk model til at analysere tiden til tilbagefald i kræftundersøgelser, [Yakovlev et al., 1993], hvilket introduceres i det næste underafsnit.

3.3.1 Biologisk motivation for en ikke-mixtur helbredelsesmodel

Lad $N \geq 0$ være antal kræftfremkaldende celler for en patient efter den første kræftbehandling. Det antages, at disse celler stadig kan være aktive, og at det tager en bestemt tid $\overline{T}_k$ for hver celle, $k = 1, \ldots, N$, til at blive en kræfttumor. For de ikke-helbredte individer, altså individer med $N \geq 1$, er overlevelsestiden defineret ved en stokastisk variabel T, således at $T = \min(\overline{T}_1, \ldots, \overline{T}_N)$. For de helbredte individer er der ingen kræftfremkaldende celler aktive. Dette vil sige, at $N = 0$, hvilket betyder, at $T = \infty$. Det antages, at N er Poisson-fordelt med middelværdi θ, og at $\overline{T}_k$'erne er uafhængige og identiske fordelte med henholdsvis fordelings- og overlevelsesfunktion $\overline{F}(t) = 1 - \overline{S}(t)$ og $\overline{S}(t)$, som er uafhængige af N. Overlevelsesfunktionen for T er

givet som

$$\begin{aligned}
S(t) &= P(N=0) + P(\overline{T}_1 > t, \ldots, \overline{T}_N > t, N \geq 1) \\
&= P(N=0) + \cup_{k=1}^{\infty} P(\overline{T}_1 > t, \ldots, \overline{T}_k > t, N=k) \\
&= P(N=0) + \sum_{k=1}^{\infty} P(\overline{T}_1 > t, \ldots, \overline{T}_k > t \mid N=k) P(N=k) \\
&= P(N=0) + \sum_{k=1}^{\infty} P(\overline{T}_1 > t, \ldots, \overline{T}_k > t) P(N=k) \\
&= \exp(-\theta) + \sum_{k=1}^{\infty} [\overline{S}(t)]^k \exp(-\theta) \frac{\theta^k}{k!} \\
&= \exp(-\theta) \sum_{k=0}^{\infty} [\overline{S}(t)]^k \frac{\theta^k}{k!} \\
&= \exp[-\theta + \theta \overline{S}(t)] \\
&= \exp[-\theta \overline{F}(t)] = \pi^{\overline{F}(t)}, \qquad (3.10)
\end{aligned}$$

hvor $\pi = \exp(-\theta)$ er sandsynligheden for helbredelse i modellen, da $\lim_{t \to \infty} S(t) = \pi$. Dette medfører, at $S(t)$ er en uegentlig overlevelsesfunktion, da $\pi = \exp(-\theta) > 0$. Det andetsidste lighedstegn i udledningen af Ligning (3.10) gælder på grund af eksponentialrækken, altså $\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$. Ved at inkludere kovariater til overlevelsesfunktionen i Ligning (3.10), haves

$$S(t \mid z) = \pi(\mathbf{z})^{\overline{F}(t|\mathbf{z})}, \qquad (3.11)$$

som svarer til overlevelsesfunktionen i Ligning (3.9).

For de ikke-helbredte individer betragtes overlevelsesfunktionen med mindst én kræftfremkaldende celle, som er bestemt ved

$$\begin{aligned}
P(T > t \mid N \geq 1) &= \frac{P(T > t, N \geq 1)}{P(N \geq 1)} \\
&= \frac{P(T > t) - P(N=0)}{1 - P(N=0)} \\
&= \frac{\exp[-\theta \overline{F}(t)] - \exp(-\theta)}{1 - \exp(-\theta)}.
\end{aligned}$$

Fordelings-, tætheds- og hazard-funktionen af T er henholdsvis givet som

$$\begin{aligned} F(t \mid \mathbf{z}) &= 1 - S(t \mid \mathbf{z}) = 1 - \pi(\mathbf{z})^{\overline{F}(t|\mathbf{z})}, \\ f(t \mid \mathbf{z}) &= \frac{\partial F(t \mid \mathbf{z})}{\partial t} = \frac{\partial}{\partial t}\left[1 - \pi(\mathbf{z})^{\overline{F}(t|\mathbf{z})}\right] \\ &= \frac{\partial}{\partial t}\left[-\exp\left\{\ln(\pi(\mathbf{z}))\overline{F}(t \mid \mathbf{z})\right\}\right] \\ &= -\ln[\pi(\mathbf{z})]\overline{f}(t \mid \mathbf{z})\exp\left\{\ln(\pi(\mathbf{z}))\overline{F}(t \mid \mathbf{z})\right\} \\ &= -\ln[\pi(\mathbf{z})]\overline{f}(t \mid \mathbf{z})\pi^{\overline{F}(t|\mathbf{z})}, \\ h(t \mid \mathbf{z}) &= \frac{f(t \mid \mathbf{z})}{S(t \mid \mathbf{z})} = -\ln[\pi(\mathbf{z})]\overline{f}(t \mid \mathbf{z}). \end{aligned} \tag{3.12}$$

Log-likelihoodfunktionen for ikke-mixtur helbredelsesmodellen kan bestemmes ved at substituere Ligning (3.11) og Ligning (3.12) ind i Ligning (A.3):

$$\begin{aligned} l &= \sum_{i=1}^{n} \delta_i \ln\left\{-\ln[\pi(\mathbf{z}_i)]\overline{f}(x_i \mid \mathbf{z}_i)\right\} + \ln\left[\pi(\mathbf{z}_i)^{\overline{F}(x_i|\mathbf{z}_i)}\right] \\ &= \sum_{i=1}^{n} \delta_i \ln\left\{-\ln[\pi(\mathbf{z}_i)]\right\} + \sum_{i=1}^{n} \delta_i \ln[\overline{f}(x_i \mid \mathbf{z}_i)] + \sum_{i=1}^{n} \ln[\pi(\mathbf{z}_i)]\overline{F}(x_i \mid \mathbf{z}_i). \end{aligned}$$

Hvis kovariater ikke optræder i $\overline{F}(t)$, skrives overlevelsesfunktionen i Ligning (3.11) som

$$S(t \mid \mathbf{z}) = \pi(\mathbf{z})^{\overline{F}(t)}. \tag{3.13}$$

Én af fordelene ved ikke-mixtur helbredelsesmodellen er, at den har en proportional hazard struktur, når $\overline{F}(t)$ er uafhængig af kovariater. Hazard-funktionen tilhørende overlevelsesfunktionen i Ligning (3.13) er givet som

$$h(t \mid \mathbf{z}) = -\ln[\pi(\mathbf{z})]\overline{f}(t),$$

og ved at betragte to individer med deres tilhørende vektorer af kovariater $\mathbf{z}_i \neq \mathbf{z}_j$, er ratioen mellem hazard-funktionerne givet ved

$$\frac{h(t \mid \mathbf{z}_i)}{h(t \mid \mathbf{z}_j)} = \frac{-\ln[\pi(\mathbf{z}_i)]\overline{f}(t)}{-\ln[\pi(\mathbf{z}_j)]\overline{f}(t)} = \frac{\ln[\pi(\mathbf{z}_i)]}{\ln[\pi(\mathbf{z}_j)]},$$

som er konstant over tid. Da modellen i Ligning (3.11) indeholder kovariater i $\overline{F}(t)$, opfylder den ikke proportional hazard-antagelsen.

3.3.2 Relativ overlevelse

Ligesom mixtur helbredelsesmodellen kunne udvides til at inkludere den relative overlevelse, kan ikke-mixtur helbredelsesmodellen i Ligning (3.11) også udvides til at inkludere den relative overlevelse. Dette er givet som, [Lambert et al., 2007b],

$$S(t \mid \mathbf{z}) = S^*(t \mid \mathbf{z})\pi(\mathbf{z})^{\widetilde{F}(t|\mathbf{z})} = S^*(t \mid \mathbf{z}) \exp\left[\ln(\pi)\widetilde{F}(t \mid \mathbf{z})\right] = S^*(t \mid \mathbf{z})R(t \mid \mathbf{z}). \tag{3.14}$$

Helbredelsesmodellen, der beskriver den relative overlevelse, er således en ikke-mixtur model givet som

$$R(t \mid \mathbf{z}) = \pi(\mathbf{z})^{\widetilde{F}(t|\mathbf{z})}. \tag{3.15}$$

Overlevelsesfunktionen $\widetilde{S}(t \mid \mathbf{z})$, der tilhører $\widetilde{F}(t \mid \mathbf{z})$, har ikke en intuitiv beskrivelse ligesom $S_u(t \mid \mathbf{z})$. Fordelingsfunktionen $\widetilde{F}(t \mid \mathbf{z})$ kan eksempelvis modelleres med en Weibull-, log-normal- eller eksponential-fordeling, og $\pi(\mathbf{z})$ ved en logistisk regression.

Hazard-funktionen, der tilhører overlevelsesfunktionen i Ligning (3.14), udtrykkes som

$$\begin{aligned}
h(t \mid \mathbf{z}) &= -\frac{d \ln\left[S(t \mid \mathbf{z})\right]}{dt} \\
&= -\frac{d \ln\left[S^*(t \mid \mathbf{z})\pi(\mathbf{z})^{\widetilde{F}(t|\mathbf{z})}\right]}{dt} \\
&= -\frac{d \ln[S^*(t \mid \mathbf{z})]}{dt} - \frac{d \ln[\pi(\mathbf{z})^{\widetilde{F}(t|\mathbf{z})}]}{dt} \\
&= h^*(t \mid \mathbf{z}) - \frac{d\left[\widetilde{F}(t \mid \mathbf{z}) \ln\{\pi(\mathbf{z})\}\right]}{dt} \\
&= h^*(t \mid \mathbf{z}) - \ln[\pi(\mathbf{z})]\widetilde{f}(t \mid \mathbf{z}),
\end{aligned} \tag{3.16}$$

hvor den forøgede hazard-funktion er givet ved

$$\lambda(t \mid \mathbf{z}) = -\ln[\pi(\mathbf{z})]\widetilde{f}(t \mid \mathbf{z}).$$

Log-likelihoodfunktionen kan dermed bestemmes ved at substituere Ligning (3.14) og Ligning (3.16) ind i Ligning (A.3):

$$
\begin{aligned}
l &= \sum_{i=1}^{n} \delta_i \ln\left[h^*(x_i \mid \mathbf{z}_i) - \ln \pi(\mathbf{z}_i)\widetilde{f}(x_i \mid \mathbf{z}_i)\right] + \ln\left[S^*(x_i \mid \mathbf{z}_i)\pi(\mathbf{z}_i)^{\widetilde{F}(x_i|\mathbf{z}_i)}\right] \\
&= \sum_{i=1}^{n} \delta_i \ln\left[h^*(x_i \mid \mathbf{z}_i) + \lambda(x_i \mid \mathbf{z}_i)\right] + \ln\left[S^*(x_i \mid \mathbf{z}_i)\right] + \ln\left[S^*(x_i \mid \mathbf{z}_i)\right] + \ln\left[\pi(\mathbf{z}_i)^{\widetilde{F}(x_i|\mathbf{z}_i)}\right] \\
&= \sum_{i=1}^{n} \delta_i \ln\left[h^*(x_i \mid \mathbf{z}_i) + \lambda(x_i \mid \mathbf{z}_i)\right] + \ln\left[R(x_i \mid \mathbf{z}_i)\right],
\end{aligned}
$$

hvor $\ln\left[S^*(x_i \mid \mathbf{z}_i)\right]$ ignoreres i det sidste lighedstegn, da $S^*(x_i \mid \mathbf{z}_i)$ ikke afhænger af modellens ukendte parametre.

Denne ikke-mixtur helbredelsesmodel kan også skrives som en mixtur helbredelsesmodel. Ved at omskrive Ligning (3.15), haves

$$
\begin{aligned}
R(t \mid \mathbf{z}) &= \pi(\mathbf{z})^{\widetilde{F}(t|\mathbf{z})} \\
&= \pi(\mathbf{z}) + [1 - \pi(\mathbf{z})]\frac{\pi(\mathbf{z})^{\widetilde{F}(t|\mathbf{z})} - \pi(\mathbf{z})}{1 - \pi(\mathbf{z})},
\end{aligned} \tag{3.17}
$$

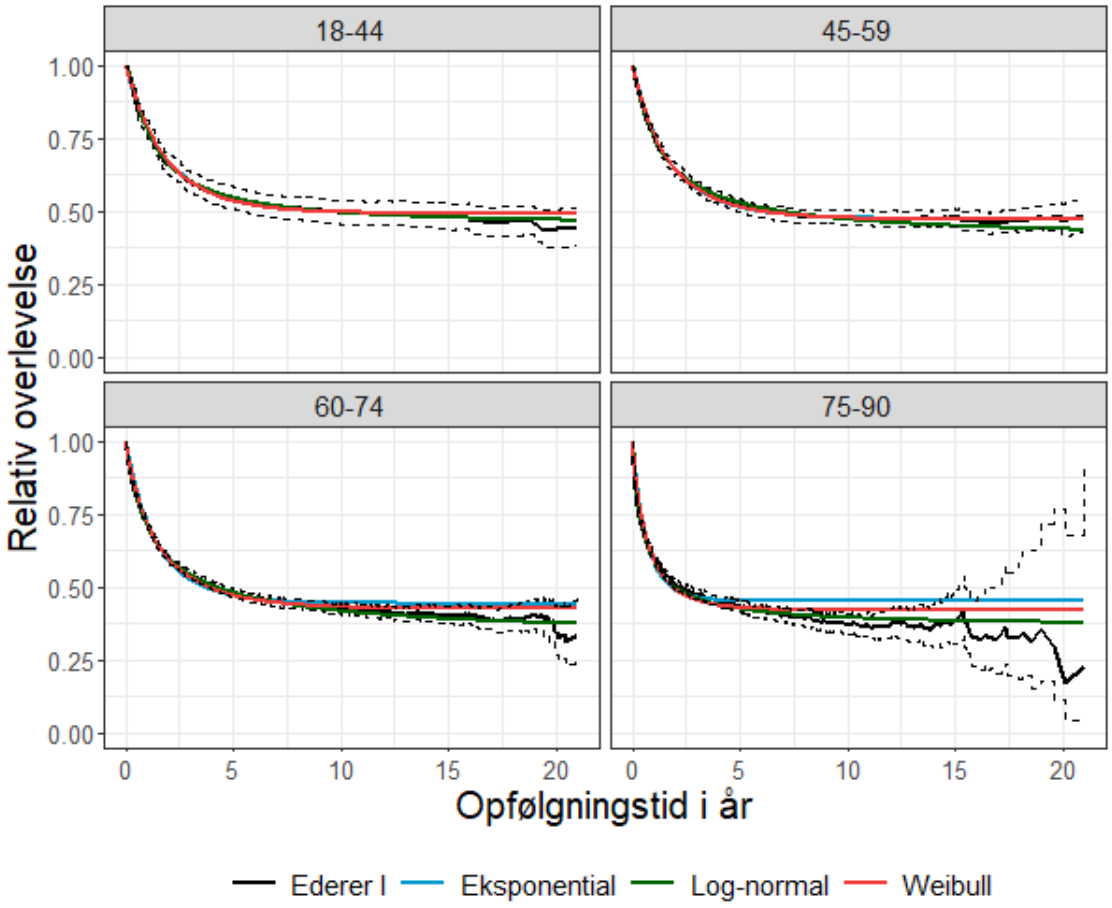
hvor den relative overlevelse er henholdsvis 1 og $S_u(t \mid \mathbf{z}) = \frac{\pi(\mathbf{z})^{\widetilde{F}(t|\mathbf{z})} - \pi(\mathbf{z})}{1-\pi(\mathbf{z})}$ for de helbredte og ikke-helbredte individer. Det fremgår altså, at $S_u(t \mid \mathbf{z})$ er afhængig af $\pi(\mathbf{z})$, hvilket gør ikke-mixtur helbredelsesmodellen sværere at fortolke.

3.4 Dataanalyse for ikke-mixtur helbredelsesmodeller

I dette afsnit analyseres coloncancer-datasættet ved hjælp af ikke-mixtur helbredelsesmodeller for den relative overlevelse i Ligning (3.15). Der opstilles tre ikke-mixtur helbredelsesmodeller uden kovariater. Modellerne er givet ved

$$
\begin{aligned}
R(t) &= \pi^{\widetilde{F}(t)}, \\
\pi &= \frac{\exp(\beta_0)}{1 + \exp(\beta_0)},
\end{aligned} \tag{3.18}
$$

og $\widetilde{F}(t)$, som modelleres med henholdsvis en Weibull-, log-normal- og eksponential-fordeling, se fordelingerne i Afsnit A.2 i appendiks. Modellerne sammenlignes i forhold til AIC, udregnet ved Ligning (3.8), og i forhold til Ederer I. Følgende figur illustrerer den estimerede relative overlevelse, $R(t)$, for de tre modeller sammen med Ederer I. Der er stratificeret efter aldersgrupperne 18-44, 45-59, 60-74, 75-90.



Figur 3.8: Den relative overlevelse for coloncancer-datasættet udregnet ved Ederer I med tilhørende 95% konfidensinterval (stiplede linjer), samt Weibull-, log-normal- og eksponential-modellen for aldergrupperne 18-44, 45-59, 60-74 og 75-90.

Følgende tabel opsummerer modellernes AIC for de fire aldersgrupper.

Model	18-44	45-59	60-74	75-90
Wei	2180.02	7061.5	20919.77	14706.45
Log	2157.41	7045.86	20798.78	14428.58
Exp	2180.42	7062.01	20952.63	14751.04

Tabel 3.4: AIC for Weibull-, log-normal- og eksponential-modellen for aldersgrupperne 18-44, 45-59, 60-74 og 75-90.

Det fremgår af Figur 3.8, at alle modellerne tilnærmelsesvist følger Ederer I-estimatet for aldersgrupperne 18-44 og 45-59. For aldersgrupperne 60-74 og 75-90 er log-normal-modellen tættere på Ederer I end Weibull- og eksponential-modellen. Ud fra Tabel 3.4 observerer vi desuden, at log-normal-modellen resulterer i den laveste AIC for alle fire aldersgrupper, og det kan dermed konkluderes, at log-normal-modellen er den bedste model. Vi konkluderede det samme i forbindelse med mixtur helbredelsesmodellerne i Afsnit 3.2. Det fremgår desuden af Tabel 3.1 og Tabel 3.4, at der ikke er den store forskel mellem mixtur og ikke-mixtur helbredelsesmodellernes AIC. Det har altså ikke den store betydning, om der anvendes en mixtur eller en ikke-mixtur helbredelsesmodel. Ikke-mixtur helbredelsesmodellen er dog sværere at fortolke, da $S_u(t)$ er afhængig af π, når $S_u(t)$ udledes, som i Ligning (3.17). Dette gør mixtur helbredelsesmodellen mere fordelagtig end ikke-mixtur helbredelsesmodellen.

I følgende tabel er den 2 års relative overlevelse, den 5 års relative overlevelse, andelen af helbredte patienter og median relativ overlevelsestiden for de ikke-helbredte patienter angivet for de tre ikke-mixtur helbredelsesmodeller stratificeret efter aldersgrupperne 18-44, 45-59, 60-74 og 75-90. For at kunne bestemme median relativ overlevelsestiden for de ikke-helbredte patienter, er ikke-mixtur helbredelsesmodellerne omskrevet til mixtur helbredelsesmodeller, som i Ligning (3.17). R-koden til vores implementation af median relativ overlevelsestiden for de ikke-helbredte patienter findes i Afsnit C.1 i appendiks.

	Model	18-44	45-59	60-74	75-90
2 års	Wei	0.67(0.64-0.7)	0.64(0.62-0.66)	0.6(0.58-0.61)	0.49(0.48-0.51)
relativ	Log	0.66(0.63-0.69)	0.64(0.62-0.66)	0.6(0.59-0.61)	0.5(0.48-0.51)
overlevelse	Exp	0.67(0.64-0.7)	0.64(0.62-0.66)	0.59(0.58-0.6)	0.49(0.47-0.5)
5 års	Wei	0.53(0.5-0.57)	0.51(0.49-0.53)	0.47(0.46-0.48)	0.43(0.41-0.45)
relativ	Log	0.54(0.51-0.58)	0.53(0.51-0.55)	0.48(0.47-0.49)	0.42(0.41-0.44)
overlevelse	Exp	0.54(0.5-0.57)	0.51(0.49-0.53)	0.47(0.46-0.48)	0.45(0.43-0.47)
Andel	Wei	0.49(0.45-0.53)	0.47(0.45-0.5)	0.43(0.41-0.44)	0.42(0.4-0.44)
helbredte	Log	0.45(0.4-0.5)	0.41(0.37-0.44)	0.33(0.3-0.36)	0.37(0.34-0.4)
patienter	Exp	0.49(0.45-0.53)	0.47(0.45-0.5)	0.44(0.43-0.46)	0.45(0.43-0.47)
Median	Wei	1.29	1.17	1.01	0.49
relativ over-	Log	1.3	1.31	1.28	0.51
levelsestid	Exp	1.28	1.17	0.97	0.45

Tabel 3.5: Den 2 års relative overlevelse, den 5 års relative overlevelse, andelen af helbredte patienter og median relativ overlevelsestiden i år for de ikke-helbredte patienter. De tilhørende 95% konfidensintervaller er angivet i parentes.

Det fremgår af Tabel 3.5, at estimaterne for den 2 og 5 års relative overlevelse for ikke-mixtur helbredelsesmodellerne næsten er identiske med mixtur helbredelsesmodellerne i Tabel 3.2. Desuden observeres der heller ikke den store forskel mellem andelen af helbredte patienter og median relativ overlevelsestiden for de ikke-helbredte patienter for ikke-mixtur og mixtur helbredelsesmodellerne. Det er muligt at udarbejde tilsvarende analyser som i Afsnit 3.2. Vi har imidlertid valgt ikke at inkludere disse, da resultaterne for ikke-mixtur helbredelsesmodellerne næsten er identiske med mixtur helbredelsesmodellerne.

I det næste afsnit introduceres identificerbarhed, som er en vigtig egenskab for helbredelsesmodellen.

3.5 Identificerbarhed

Identificerbarhed er en vigtig egenskab for en model. Det betyder, at der ikke eksisterer to forskellige parameterestimater, der genererer samme model. Identificerbarhed har konsekvenser for modellens statistiske inferens. Hvis en model ikke er identificerbar, er estimationen af dens parametre nemlig ustabil. Modellerne kræver generelt et stort datasæt med en lang opfølgningstid og mange censurerede observationer efter tidspunktet, hvor begivenhederne typisk forekommer. En opfølgningstid er tilstrækkeligt lang, hvis patienten oplever begivenheden af interesse inden censureringstidspunktet. Dette observeres, når der forekommer en udfladning i overlevelsesfunktionen, [Bremhorst and Lambert, 2016].

3.5.1 Identificerbarhed for mixtur helbredelsesmodeller

Der betragtes en mixtur helbredelsesmodel

$$S(t \mid \mathbf{z}) = \pi(\mathbf{z}) + [1 - \pi(\mathbf{z})]\widetilde{S}_u(t \mid \mathbf{z}),$$

som kan skrives i en ækvivalent form ved hjælp af den kumulerede fordelingsfunktion

$$\begin{aligned} S(t \mid \mathbf{z}) &= \pi(\mathbf{z}) + [1 - \pi(\mathbf{z})][1 - \widetilde{F}_u(t \mid \mathbf{z})] \\ &= 1 - [1 - \pi(\mathbf{z})]\widetilde{F}_u(t \mid \mathbf{z}). \end{aligned} \tag{3.19}$$

Ved at substituere $F(t \mid \mathbf{z}) = 1 - S(t \mid \mathbf{z})$ og $\bar{\pi}(\mathbf{z}) = 1 - \pi(\mathbf{z})$ ind i Ligning (3.19), haves

$$F(t \mid \mathbf{z}) = \bar{\pi}(\mathbf{z})\widetilde{F}_u(t \mid \mathbf{z}). \tag{3.20}$$

Inden begrebet identificerbarhed defineres, betragtes familien $\mathcal{P}$ og familien af kumulerede fordelingsfunktioner $\mathcal{F}$, hvorefter der foretages følgende antagelser, [Hanin and Huang, 2014]:

a) $0 < \bar{\pi}(\mathbf{z}) < 1 \quad \forall \mathbf{z} \quad \text{og} \quad \bar{\pi} \in \mathcal{P}$.

b) For ethvert $\mathbf{z}$ og $\widetilde{F}_u \in \mathcal{F}$ er funktionen $t \longmapsto \widetilde{F}_u(t \mid \mathbf{z})$ voksende, kontinuert fra højre i ethvert punkt $t \in [0; t_{\max}[$, således at $0 \leq \widetilde{F}_u < 1$, og opfylder betingelserne $\widetilde{F}_u(0 \mid \mathbf{z}) = 0$ og $\widetilde{F}_u(t \mid \mathbf{z}) > 0$ for $t > 0$. Hvis $\lim_{t \to t_{\max}} \widetilde{F}_u = 1$, siges det, at $\widetilde{F}_u$ er egentlig på $[0; t_{\max}[$.

Definition 3.1.
For $\mathcal{P}$ og $\mathcal{F}$ er modellen i Ligning (3.20) identificerbar, hvis

$$F_1(t \mid \mathbf{z}) = F_2(t \mid \mathbf{z}) \Rightarrow \bar{\pi}_1(\mathbf{z}) = \bar{\pi}_2(\mathbf{z}) \quad \text{og} \quad \widetilde{F}_{u1}(t \mid \mathbf{z}) = \widetilde{F}_{u2}(t \mid \mathbf{z})$$

for alle $\mathbf{z}$ og $0 \leq t < t_{\max}$, og for funktionerne $\bar{\pi}_1(\mathbf{z}), \bar{\pi}_2(\mathbf{z}), \widetilde{F}_{u1}$ og $\widetilde{F}_{u2}$, som er på formen af $\bar{\pi}(\mathbf{z})$ og $\widetilde{F}_u$ i Ligning (3.20).

Hvis modellen i Ligning (3.20) er identificerbar inden for familierne $\mathcal{P}$ og $\mathcal{F}$, er den også identificerbar inden for delfamilier af $\mathcal{P}$ og $\mathcal{F}$. Det ønskes at bestemme de største familier af $\mathcal{P}$ og $\mathcal{F}$, således at modellen i Ligning (3.20) stadig er identificerbar, [Hanin and Huang, 2014].

Proposition 3.2.
Antag, at $\mathcal{F}$ består af egentlige kumulerede fordelingsfunktioner. Det vil sige, at $\lim_{t \to t_{\max}} \widetilde{F}_u(t \mid \mathbf{z}) = 1 \quad \forall \mathbf{z}$ og $\widetilde{F}_u \in \mathcal{F}$. Modellen i Ligning (3.20) er dermed identificerbar.

Bevis
Antag, at

$$F_1(t \mid \mathbf{z}) = F_2(t \mid \mathbf{z}),$$

hvilket ifølge Definition 3.1 medfører, at

$$\bar{\pi}_1(\mathbf{z})\widetilde{F}_{u1}(t \mid \mathbf{z}) = \bar{\pi}_2(\mathbf{z})\widetilde{F}_{u2}(t \mid \mathbf{z}), \tag{3.21}$$

hvor $\bar{\pi}_1, \bar{\pi}_2 \in \mathcal{P}$ og $\widetilde{F}_{u1}, \widetilde{F}_{u2} \in \mathcal{F}$. Ved at lade $t \to t_{\max}$, er $\bar{\pi}_1(\mathbf{z}) = \bar{\pi}_2(\mathbf{z})$ for alle $\mathbf{z}$, hvilket medfører, at $\widetilde{F}_{u1}(t \mid \mathbf{z}) = \widetilde{F}_{u2}(t \mid \mathbf{z})$ for alle $\mathbf{z}$ og $0 \leq t < t_{\max}$. Dermed er modellen i Ligning (3.20) identificerbar. ■
Det følger af Proposition 3.2, at modellen i Ligning (3.20) er identificerbar, hvis $t_{\max} = \infty$.

I næste definition introduceres begrebet skalérbarhed.

Definition 3.3.

1. En familie $\mathcal{P}$ af funktioner $\bar{\pi}$ er skalérbar, hvis den udover $\bar{\pi}$ også indeholder skalar multiplum $c\bar{\pi}$ for $c > 0$, hvis $c\bar{\pi}$ opfylder antagelse a). Familien $\mathcal{P}$ siges at være svagt skalérbar, hvis $\mathcal{P}$ indeholder to forskellige funktioner $\bar{\pi}_1$ og $\bar{\pi}_2$, således at $\bar{\pi}_2 = c\bar{\pi}_1$ for en positiv konstant $c \neq 1$.

2. Tilsvarende er familien $\mathcal{F}$ af kumulerede fordelingsfunktioner $\widetilde{F}_u$, som opfylder antagelse b), skalerbar, hvis den udover $\widetilde{F}_u$ også indeholder alle $\widetilde{F}_u$'s skalar multiplum $c\widetilde{F}_u$, hvis $c\widetilde{F}_u(t \mid \mathbf{z}) \leq 1$ for alle $\mathbf{z}$. Familien $\mathcal{F}$ siges at være svagt skalerbar, hvis $\widetilde{F}_{u1} = c\widetilde{F}_{u2}$ for nogle funktioner $\widetilde{F}_{u1}, \widetilde{F}_{u2} \in \mathcal{F}$ og en positiv konstant $c \neq 1$.

Når der arbejdes med helbredelsesmodeller, kræves det normalt, at opfølgningstiden skal være tilstrækkeligt lang. Dette er imidlertid ikke altid tilfældet. Af den grund undersøges der for identificerbarhedsproblemer, når denne antagelse ikke er opfyldt. I følgende sætning fokuseres der på tilfældet, hvor $t_{\max} < \infty$, og det antages, at nogle kumulerede fordelingsfunktioner i $\mathcal{F}$ er uegentlige, [Hanin and Huang, 2014].

Sætning 3.4.
Lad $\|\bar{\pi}\| := \sup_{\mathbf{z}}\{\bar{\pi}(\mathbf{z})\},\ \bar{\pi} \in \mathcal{P}$, og $\|\widetilde{F}_u\| := \sup_{\mathbf{z}}\{\widetilde{F}_u(t \mid \mathbf{z})\},\ \widetilde{F}_u \in \mathcal{F}$.

1. Antag, at $\|\bar{\pi}\| = 1\ \forall \bar{\pi} \in \mathcal{P}$ eller $\|\widetilde{F}_u\| = 1\ \forall \widetilde{F}_u \in \mathcal{F}$. Modellen i Ligning (3.20) er dermed identificerbar.

2. Antag, at funktionerne i $\mathcal{P}$ og $\mathcal{F}$ ikke har fælles kovariater, og mindst én af de to familier ikke er svagt skalérbare. Modellen i Ligning (3.20) er dermed identificerbar.

3. Antag, at $\mathcal{P}$ og $\mathcal{F}$ er skalérbare. Hvis $\|\bar{\pi}\| < 1$ for et $\bar{\pi} \in \mathcal{P}$ eller $\|\widetilde{F}_u\| < 1$ for et $\widetilde{F}_u \in \mathcal{F}$, er modellen i Ligning (3.20) ikke identificerbar.

4. Hvis $\mathcal{P}$ består af konstante funktioner $\bar{\pi}$, hvor $0 < \bar{\pi} < 1$, og $\mathcal{F}$ er svagt skalérbar, er modellen i Ligning (3.20) ikke identificerbar.

Bevis

1. Antag, at én af de to antagelser a) og b) er opfyldt for $\mathcal{P}$ og $\mathcal{F}$, og at Ligning (3.21) er opfyldt. Lad

$$\tau_i(\mathbf{z}) := \sup\{t : 0 \leq t < t_{\max},\ F_i(t \mid \mathbf{z}) = 0\} \text{ for } i = 1, 2. \tag{3.22}$$

Ud fra Ligning (3.21) er $\tau_1(\mathbf{z}) = \tau_2(\mathbf{z})$, og der defineres $\tau(\mathbf{z}) := \tau_1(\mathbf{z}) = \tau_2(\mathbf{z})$, hvor $0 \leq \tau(\mathbf{z}) < t_{\max}$. Ligning (3.21) kan omskrives til

$$\frac{\bar{\pi}_1(\mathbf{z})}{\bar{\pi}_2(\mathbf{z})} = \frac{\widetilde{F}_{u2}(t \mid \mathbf{z})}{\widetilde{F}_{u1}(t \mid \mathbf{z})},\ \tau(\mathbf{z}) < t < t_{\max}.$$

Der eksisterer en funktion $c(\mathbf{z}) > 0$, således at

$$\bar{\pi}_1(\mathbf{z}) = c(\mathbf{z})[\bar{\pi}_2(\mathbf{z})], \tag{3.23}$$

$$\widetilde{F}_{u2}(t \mid \mathbf{z}) = c(\mathbf{z})\widetilde{F}_{u1}(t \mid \mathbf{z}),\ \tau(\mathbf{z}) < t < t_{\max}. \tag{3.24}$$

Ud fra Ligning (3.22) er Ligning (3.24) opfyldt for alle $0 \leq t < t_{\max}$. Ud fra Ligning (3.23) og Ligning (3.24) kan følgende udledes

$$\|\bar{\pi}_1\| = c(\mathbf{z})\|\bar{\pi}_2\|,$$
$$\|\widetilde{F}_{u2}\| = c(\mathbf{z})\|\widetilde{F}_{u1}\|.$$

Antag, at $\|\bar{\pi}\| = 1\ \forall \pi \in \mathcal{P}$ og $\forall \mathbf{z}$. Det vil sige, at

$$\|\bar{\pi}_1\| = \|\bar{\pi}_2\| \Rightarrow c(\mathbf{z}) = 1\ \forall \mathbf{z},$$

og dermed er $\bar{\pi}_1(\mathbf{z}) = \bar{\pi}_2(\mathbf{z})$ og $\widetilde{F}_{u2}(t \mid \mathbf{z}) = \widetilde{F}_{u1}(t \mid \mathbf{z})$. Det konkluderes ud fra Definition 3.1, at modellen i Ligning (3.20) er identificerbar.

2. Når $\bar{\pi}$ og $\widetilde{F}_u$ ikke har fælles kovariater, kan Ligning (3.21) skrives som

$$\bar{\pi}_1(\mathbf{y})\widetilde{F}_{u1}(t \mid \mathbf{z}) = \bar{\pi}_2(\mathbf{y})\widetilde{F}_{u2}(t \mid \mathbf{z}),$$

hvor $\mathbf{y}$ er en vektor af kovariater, som er disjunkt med $\mathbf{z}$. Ligesom det blev bevist tidligere for punkt 1, eksisterer der en konstant $c > 0$, således at

$$\bar{\pi}_1(\mathbf{y}) = c\bar{\pi}_2(\mathbf{y})\ \forall \mathbf{y},$$
$$\widetilde{F}_{u2}(t \mid \mathbf{z}) = c\widetilde{F}_{u1}(t \mid \mathbf{z})\ \forall \mathbf{z}.$$

Det antages, at mindst én af de to familier $\mathcal{P}$ og $\mathcal{F}$ ikke er svagt skalérbar. Det vil sige, at $c = 1$, og dermed er $\bar{\pi}_1 = \bar{\pi}_2$ og $\widetilde{F}_{u1} = \widetilde{F}_{u2}$. Det konkluderes derfor, at modellen i Ligning (3.20) er identificerbar.

3. Dette resultat vises ved modstrid. Antag, at $\mathcal{P}$ og $\mathcal{F}$ er skalérbare, og modellen i Ligning (3.20) er identificerbar. Antag yderligere, at $\|\overline{\pi}\| < 1$ for et $\overline{\pi} \in \mathcal{P}$. Der findes dermed en konstant $c > 1$, således at $c\overline{\pi} \in \mathcal{P}$. Vælg en arbitrær kumuleret fordelingsfunktion $\widetilde{F}_u \in \mathcal{F}$, som opfylder, at $\frac{\widetilde{F}_u}{c} \in \mathcal{F}$. Ved at sætte $\widetilde{\overline{\pi}} = c\overline{\pi}$ og $\widetilde{\overline{F}}_u = \frac{\widetilde{F}_u}{c}$, haves $\widetilde{\overline{\pi}}\widetilde{\overline{F}}_u = \overline{\pi}\widetilde{F}_u$, som er to forskellige faktoriseringer for samme kumuleret fordelingsfunktion. Dette medfører, at modellen i Ligning (3.20) ikke er identificerbar. Tilsvarende for et $\widetilde{F}_u \in \mathcal{F}$, hvis $\|\widetilde{F}_u\| < 1$, eksisterer der en konstant $c > 1$, således at $\widetilde{\overline{F}}_u := c\widetilde{F}_u \in \mathcal{F}$. Der vælges $\overline{\pi} \in \mathcal{P}$, som opfylder, at $\widetilde{\overline{\pi}} := \frac{\overline{\pi}}{c} \in \mathcal{P}$. Dermed er $\widetilde{\overline{\pi}}\widetilde{\overline{F}}_u = \overline{\pi}\widetilde{F}_u$. Dette betyder, at modellen i Ligning (3.20) ikke er identificerbar.

4. Da $\mathcal{F}$ er svagt skalérbar, eksisterer der funktioner $\widetilde{F}_u,\ \widetilde{\overline{F}}_u \in \mathcal{F}$ og en konstant $c > 1$, således at $\widetilde{\overline{F}}_u = c\widetilde{F}_u$. Ved at vælge en hvilken som helst $\overline{\pi} \in]0;1[$ og sætte $\widetilde{\overline{\pi}} := \frac{\overline{\pi}}{c}$, haves $\widetilde{\overline{\pi}}\widetilde{\overline{F}}_u = \overline{\pi}\widetilde{F}_u$, hvilket viser, at modellen i Ligning (3.20) ikke er identificerbar.

■

Bemærk, at betingelsen $\|\overline{\pi}\| < 1$ medfører, at andelen af helbredte individer er større end nul. Tilsvarende betyder betingelsen $\|\widetilde{F}_u\| < 1$, at opfølgningstiden $t_{\max}$ ikke er lang nok til, at begivenheden af interesse indtræffer for de ikke-helbredte individer.

3.5.2 Identificerbarhed for ikke-mixtur helbredelsesmodeller

Resultaterne i dette underafsnit er fra [Hanin and Huang, 2014] og bevises ikke.

Der betragtes en ikke-mixtur helbredelsesmodel

$$S(t \mid \mathbf{z}) = \pi(\mathbf{z})^{\overline{F}(t|\mathbf{z})} = \exp\left[-\theta(\mathbf{z})\overline{F}(t \mid \mathbf{z})\right], \quad 0 \le t < t_{\max}, \tag{3.25}$$

hvor $\theta(\mathbf{z})$ er en positiv funktion, der tilhører en familie Θ, og $\overline{F}(t \mid \mathbf{z}) \in \mathcal{F}$ er en fordelingsfunktion defineret i $[0; t_{\max}[$, som ikke nødvendigvis er egentlig. Identificerbarhed for ikke-mixtur helbredelsesmodellen i Ligning (3.25) introduceres i definitionen, der følger.

Definition 3.5.
Modellen i Ligning (3.25) er identificerbar inden for Θ og $\mathcal{F}$, hvis

$$S_1(t \mid \mathbf{z}) = S_2(t \mid \mathbf{z}),$$

hvor

$$S_i(t \mid \mathbf{z}) = \exp\left[-\theta_i(\mathbf{z})\overline{F}_i(t \mid \mathbf{z})\right], \quad i = 1, 2$$

medfører, at $\theta_1(\mathbf{z}) = \theta_2(\mathbf{z})$ $\forall \mathbf{z}$, og $\overline{F}_1(t \mid \mathbf{z}) = \overline{F}_2(t \mid \mathbf{z})$ $\forall \mathbf{z}$ og $0 \leq t < t_{\max}$. Funktionerne $\theta_1, \theta_2, \overline{F}_1$ og $\overline{F}_2$ er på formen af θ og $\overline{F}$ i Ligning (3.25).

Egentlige overlevelsesfunktioner er også vigtige for identificering af ikke-mixtur helbredelsesmodeller, hvilket illustreres i den næste proposition.

Proposition 3.6.
Antag, at familien $\mathcal{F}$ kun består af egentlige fordelingsfunktioner. Det vil sige, at $\lim_{t \to t_{\max}} \overline{F}(t \mid \mathbf{z}) = 1 \quad \forall \mathbf{z}$ og $\overline{F} \in \mathcal{F}$. Modellen i Ligning (3.25) er dermed identificerbar.

I følgende sætning undersøges identificerbarhed for ikke-mixtur helbredelsesmodeller, når opfølgningstiden ikke er tilstrækkeligt lang. Det antages derfor, at $t_{\max} < \infty$, og at nogle kumulerede fordelingsfunktioner i $\mathcal{F}$ er uegentlige.

Sætning 3.7.
Lad $\|\overline{F}\| := \sup_{\mathbf{z}}\{\overline{F}(\mathbf{z})\}$.

1. Antag, at $\|\overline{F}\| = 1 \quad \forall \mathbf{z}$ og $\overline{F} \in \mathcal{F}$. Modellen i Ligning (3.25) er dermed identificerbar.

2. Antag, at funktionerne i Θ og $\mathcal{F}$ ikke har fælles kovariater, og mindst én af de to familier ikke er svagt skalérbar. Dermed er modellen i Ligning (3.25) ikke identificerbar.

3. Hvis en af de to familier Θ og $\mathcal{F}$ er skalérbar og den anden er svagt skalérbar, er modellen i Ligning (3.25) ikke identificerbar.

Resultaterne for identificerbarhed er kun beskrevet for total overlevelse, men resultaterne kan også generaliseres til relativ overlevelse.

3.5.3 Identificerbarhed for modellerne i dataanalyse

I dette underafsnit undersøger vi, hvorvidt modellerne, som vi har tilpasset coloncancer-datasættet, er identificerbare. Vi illustrerer det kun i tilfældet uden kovariater, da det kan vises tilsvarende for modeller med kovariater. Vi illustrerer identificerbarhed for modellerne, hvor S_u eller $\widetilde{F}$ er modelleret med en Weibull-fordeling. Samme fremgangsmåde kan anvendes for at undersøge andre modeller med andre fordelinger til at modellere S_u eller $\widetilde{F}$.

Vi skal illustrere, at modellen i Ligning (3.7) er identificerbar. Først omskrives modellen til formen i Ligning (3.20):

$$\begin{aligned} R(t) &= \pi + (1-\pi)S_u(t) \\ &= 1 - (1-\pi)F_u(t). \end{aligned}$$

Ved at substituere $\overline{R}(t) = 1 - R(t)$ og $\overline{\pi} = 1 - \pi$, haves

$$\overline{R}(t) = \overline{\pi} F_u(t).$$

Ved at modellere π med en logit link-funktion og F_u med en Weibull-fordeling, haves

$$\overline{R}(t) = \frac{1}{1+\exp(\beta_0)} \left[1 - \exp\left(-\gamma_1 t^{\gamma_2}\right)\right]. \tag{3.26}$$

Lad $\mathcal{P}$ være logit link familien og $\mathcal{F}$ være Weibull familien, hvor

$$\begin{aligned} \mathcal{P} &= \left\{ \overline{\pi} = \frac{1}{1+\exp(\beta_0)} \;\middle|\; \beta_0 \in \mathbb{R} \right\}, \\ \mathcal{F} &= \left\{ F_u = 1 - \exp\left(-\gamma_1 t^{\gamma_2}\right) \;\middle|\; \gamma_1, \gamma_2 > 0 \right\}. \end{aligned}$$

Det skal illustreres, at $\overline{R}(t)$ er identificerbar. Dette gøres i analog med Definition 3.1. Det antages, at

$$\begin{aligned} &\overline{R}_1(t) = \overline{R}_2(t) \\ \iff &\frac{1}{1+\exp(\beta_{01})} \left[1 - \exp\left(-\gamma_{11} t^{\gamma_{21}}\right)\right] = \frac{1}{1+\exp(\beta_{02})} \left[1 - \exp\left(-\gamma_{12} t^{\gamma_{22}}\right)\right] \end{aligned} \tag{3.27}$$

Det skal illustreres, at Ligning (3.27) medfører følgende ligninger

$$1 - \exp\left(-\gamma_{11} t^{\gamma_{21}}\right) = 1 - \exp\left(-\gamma_{12} t^{\gamma_{22}}\right), \tag{3.28}$$

$$\frac{1}{1+\exp(\beta_{01})} = \frac{1}{1+\exp(\beta_{02})}. \tag{3.29}$$

Dette illustreres for $t \to \infty$. Ligning (3.28) er opfyldt for $t \to \infty$, hvilket medfører, at Ligning (3.29) også er opfyldt for $t \to \infty$. Det konkluderes altså, at modellen i Ligning (3.26) er identificerbar. Det vil sige, at

$$\begin{aligned} \overline{R}_1(t) &= \overline{R}_2(t) \\ \Rightarrow 1 - \exp\left(-\gamma_{11} t^{\gamma_{21}}\right) &= 1 - \exp\left(-\gamma_{12} t^{\gamma_{22}}\right) \quad \text{og} \\ \frac{1}{1+\exp(\beta_{01})} &= \frac{1}{1+\exp(\beta_{02})}. \end{aligned}$$

Det skal også undersøges, om modellen for ikke-mixtur helbredelsesmodellen i Ligning (3.18) er identificerbar. Ved at skrive Ligning (3.25) i form af den relative overlevelse, haves

$$R(t) = \pi^{\widetilde{F}(t)} = \exp[-\theta \widetilde{F}(t)]. \tag{3.30}$$

For at gå frem i analog med Definition 3.5, skal θ isoleres. Først modelleres π med en logit link-funktion, og derefter isoleres θ

$$\begin{aligned} &\frac{\exp(\beta_0)}{1+\exp(\beta_0)} = \exp(-\theta) \\ &\quad \Longleftrightarrow \ \theta = -\log\left(\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right) = -\beta_0 + \log[1+\exp(\beta_0)]. \end{aligned}$$

Ved at modellere $\widetilde{F}$ med Weibull-fordelingen og anvende udtrykket for θ, kan det sidste lighedstegn i Ligning (3.30) skrives som

$$R(t) = \exp\left\{-(-\beta_0 + \log[1+\exp(\beta_0)])\right\}\left\{1 - \exp\left(-\gamma_1 t^{\gamma_2}\right)\right\}. \tag{3.31}$$

Nu skal det undersøges, om modellen i Ligning (3.31) er identificerbar for $t \to \infty$. Dette vises i analog med Definition 3.5. Først defineres to familier Θ og $\mathcal{F}$, hvor

$$\Theta = \{\theta = -\beta_0 + \log[1 + \exp(\beta_0)] \mid \beta_0 \in \mathbb{R}\},$$
$$\mathcal{F} = \left\{\widetilde{F}(t) = 1 - \exp\left(-\gamma_1 t^{\gamma_2}\right) \,\middle|\, \gamma_1, \gamma_2 > 0\right\}.$$

Det antages, at

$$R_1(t) = R_2(t)$$

$$\begin{aligned} &\iff \exp\left\{-\left(-\beta_{01} + \log\left[1 + \exp(\beta_{01})\right]\right)\right\}\left\{1 - \exp\left(-\gamma_{11} t^{\gamma_{21}}\right)\right\} \\ &= \exp\left\{-\left(-\beta_{02} + \log\left[1 + \exp(\beta_{02})\right]\right)\right\}\left\{1 - \exp\left(-\gamma_{12} t^{\gamma_{22}}\right)\right\}. \end{aligned} \tag{3.32}$$

Det skal illustreres, at Ligning (3.32) medfører følgende ligninger

$$1 - \exp\left(-\gamma_{11} t^{\gamma_{21}}\right) = 1 - \exp\left(-\gamma_{12} t^{\gamma_{22}}\right), \tag{3.33}$$
$$-\beta_{01} + \log\{1 + \exp(\beta_{01})\} = -\beta_{02} + \log\{1 + \exp(\beta_{02})\}. \tag{3.34}$$

Ligning (3.33) er opfyldt for $t \to \infty$, hvilket medfører, at Ligning (3.34) også er opfyldt for $t \to \infty$. Det konkluderes dermed, at modellen i Ligning (3.31) er identificerbar. Det vil sige, at

$$\begin{aligned} &R_1(t) = R_2(t) \\ \Rightarrow &-\beta_{01} + \log\{1 + \exp(\beta_{01})\} = -\beta_{02} + \log\{1 + \exp(\beta_{02})\} \quad \text{og} \\ &1 - \exp\left(-\gamma_{11} t^{\gamma_{21}}\right) = 1 - \exp\left(-\gamma_{12} t^{\gamma_{22}}\right). \end{aligned}$$

I det næste afsnit beskrives modelkontrol for mixtur helbredelsesmodeller.

3.6 Modelkontrol

I al almindelighed foretages en residualanalyse i forbindelse med modelkontrol. Dette indebærer forskellige plots af residualer, og inden for overlevelsesanalyse er det Cox-Snell-, martingale- og devians-residualerne, som er de mest udbredte. Cox-Snell-residualerne kan anvendes til at undersøge, hvorvidt den tilpassede model overordnet set beskriver datasættet godt. Martingale-residualer er baserede på Cox-Snell-residualerne og anvendes typisk til at undersøge antagelsen om den funktionale form af kovariaterne i modellen. En ulempe ved martingale-residualerne er, at de er definerede i intervallet $]-\infty;1]$. De er derfor typisk asymmetriske, hvilket betyder, at de ikke kan bruges til at bestemme outliers. I stedet kan devians-residualerne anvendes, som er en transformation af martingale-residualerne. Devians-residualerne er symmetriske omkring 0.

En residualanalyse inden for mixtur helbredelsesmodeller er mere kompliceret, da den skal inkludere en vurdering af flere dele. En residualanalyse for modellen i Ligning (3.1) skal eksempelvis inkludere en vurdering af henholdsvis overlevelsesfunktionen $\widetilde{S}_u(t \mid \mathbf{z})$ for de ikke-helbredte individer samt modellen for $\pi(\mathbf{z})$ og overlevelsesfunktionen $S(t \mid \mathbf{z})$ for alle individerne. Dette skyldes, at residualerne for $S(t \mid \mathbf{z})$ kun giver et indblik i, hvorvidt modellen er misspecificeret. Hvilken del, der er misspecificeret, er imidlertid uklart.

3.6.1 Cox-Snell-residualer for total overlevelse

I dette afsnit beskrives Cox-Snell-residualerne inden for mixtur helbredelsesmodellen for total overlevelse i Ligning (3.1). Cox-Snell-residualerne bruges til at undersøge overlevelsesfunktionen for de ikke-helbredte individer $\widetilde{S}_u(t \mid \mathbf{z})$ samt overlevelsesfunktionen $S(t \mid \mathbf{z})$ for alle individerne. Idéen bag Cox-Snell-residualerne er, at hvis en stokastisk variabel T har en egentlig overlevelsesfunktion $S(T > t)$, så følger Cox-Snell-residualerne $r^C = H(T) = -\ln[S(T)]$ en eksponential-fordeling med middelværdi 1, se Proposition A.1 i appendiks. Dette betyder, at tætheden for r^C er $f_{r^C}(t) = \exp(-t)$, og overlevelsesfunktionen er

$$S_{r^C}(t) = 1 - F_{r^C}(t) = 1 - \int_0^t \exp(-u)du = \exp(-t).$$

Dermed haves $H_{r^C}(t) = -\ln[S_{r^C}(t)] = t$, hvilket betyder, at et plot af $\widehat{H}(r^C)$ mod r^C forventes at give en ret linje igennem origo med hældningskoefficient 1.

For at undersøge hvor godt overlevelsesfunktionen $\widetilde{S}_u(t \mid \mathbf{z})$ for de ikke-helbredte individer passer til data, kan Cox-Snell-residualer benyttes. For en overlevelsestid x_i for et ikke-helbredt individ defineres Cox-Snell-residualet som

$$r_u^C(x_i) = -\ln[\widehat{\widetilde{S}}_u(x_i \mid \mathbf{z}_i)].$$

Under højre-censurering vides det ikke, om et individ er helbredt eller ikke-helbredt. Det vil sige, at Y_i er ukendt, hvis $\delta_i = 0$. Det er derfor blevet foreslået at erstatte Y_i med den forventede værdi, [Scolas et al., 2018],

$$\mathbb{E}(Y_i \mid \mathbf{z}_i, x_i, \delta_i) = P(Y_i = 1 \mid \mathbf{z}_i, x_i, \delta_i).$$

Der gælder, at $Y_i = 1$, når $\delta_i = 1$, og for $\delta_i = 0$ haves

$$\begin{aligned}
P(Y_i = 1 \mid \mathbf{z}_i, x_i, \delta_i = 0) &= P(X_i < \infty \mid \mathbf{z}_i, x_i, \delta_i = 0) \\
&= \frac{P(X_i < \infty, \mathbf{z}_i, x_i, \delta_i = 0)}{P(x_i, \mathbf{z}_i, \delta_i = 0)} \\
&= \frac{P(X_i < \infty, x_i, \delta_i = 0 \mid \mathbf{z}_i)}{P(x_i, \delta_i = 0 \mid \mathbf{z}_i)} \\
&= \frac{P(x_i < X_i < \infty \mid \mathbf{z}_i)}{P(X_i > x_i \mid \mathbf{z}_i)} \\
&= \frac{P(X_i < \infty \mid \mathbf{z}_i) - P(X_i \leq x_i \mid \mathbf{z}_i)}{P(X_i > x_i \mid \mathbf{z}_i)} \\
&= \frac{P(X_i < \infty \mid \mathbf{z}_i) - [1 - P(X_i > x_i \mid \mathbf{z}_i)]}{P(X_i > x_i \mid \mathbf{z}_i)} \\
&= \frac{P(X_i < \infty \mid \mathbf{z}_i) - [1 - S(x_i \mid \mathbf{z}_i)]}{S(x_i \mid \mathbf{z}_i)} \\
&= \frac{1 - \pi(\mathbf{z}_i) - [1 - \{\pi(\mathbf{z}_i) + (1 - \pi(\mathbf{z}_i))\widetilde{S}_u(x_i \mid \mathbf{z}_i)\}]}{\pi(\mathbf{z}_i) + [1 - \pi(\mathbf{z}_i)]\widetilde{S}_u(x_i \mid \mathbf{z}_i)} \\
&= \frac{[1 - \pi(\mathbf{z}_i)]\widetilde{S}_u(x_i \mid \mathbf{z}_i)}{\pi(\mathbf{z}_i) + [1 - \pi(\mathbf{z}_i)]\widetilde{S}_u(x_i \mid \mathbf{z}_i)}.
\end{aligned}$$

Den forventede værdi for de to tilfælde er dermed givet som

$$\mathbb{E}(Y_i \mid \mathbf{z}_i, x_i, \delta_i) = \delta_i + (1-\delta_i)\frac{[1-\pi(\mathbf{z}_i)]\widetilde{S}_u(x_i \mid \mathbf{z}_i)}{[1-\pi(\mathbf{z}_i)]\widetilde{S}_u(x_i \mid \mathbf{z}_i) + \pi(\mathbf{z}_i)}. \tag{3.35}$$

Denne forventede værdi giver kun mening i forhold til total overlevelse, da $Y_i = 1$ i relativ overlevelse ikke nødvendigvis betyder, at individ i har oplevet begivenheden, som er tilfældet for total overlevelse. Der er derfor behov for en anden løsning i forhold til relativ overlevelse.

Både $\pi(\mathbf{z}_i)$ og $\widetilde{S}_u(x_i \mid \mathbf{z}_i)$ er ukendte i Ligning (3.35), og derfor estimeres højresiden i ligningen ved

$$\widehat{\psi}_i = \delta_i + (1-\delta_i)\frac{[1-\widehat{\pi}(\mathbf{z}_i)]\widehat{\widetilde{S}}_u(x_i \mid \mathbf{z}_i)}{[1-\widehat{\pi}(\mathbf{z}_i)]\widehat{\widetilde{S}}_u(x_i \mid \mathbf{z}_i) + \widehat{\pi}(\mathbf{z}_i)}. \tag{3.36}$$

Estimatet $\widehat{\psi}_i$ kan antage værdier mellem 0 og 1. En grænse skal derfor vælges, før et individ kan klassificeres til at være ikke-helbredt eller helbredt. Et oplagt valg er at vælge grænsen til 0.5, således at individer med $\widehat{\psi}_i > 0.5$ klassificeres til at være ikke-helbredte, og individer med $\widehat{\psi}_i \leq 0.5$ til at være helbredte. For en overlevelsestid x_i med tilhørende estimat $\widehat{\psi}_i$ af Y_i defineres

$$r_u^C(x_i) = -\ln\left[\widehat{\widetilde{S}}_u(x_i \mid \mathbf{z}_i)\right] \text{ når } \widehat{\psi}_i > 0.5. \tag{3.37}$$

Hvis et plot af $\widehat{H}(r_u^C)$ mod r_u^C ikke giver en ret linje igennem origo med hældningskoefficient 1, betyder det, at overlevelsesfunktionen $\widetilde{S}_u(t \mid \mathbf{z})$ kan være utilstrækkelig. En anden fordeling til at modellere $\widetilde{S}_u(t \mid \mathbf{z})$ skal derfor overvejes. Det kan dog også skyldes, at modellen til at estimere $\pi(\mathbf{z})$ er dårlig. I forbindelse med Ligning (3.1) blev det beskrevet, at $\widetilde{S}_u(t \mid \mathbf{z})$ eksempelvis kan modelleres ved en Weibull-, log-normal- eller eksponential-fordeling. For at vælge den bedste model, kan residualplots for forskellige fordelinger være behjælpelige.

For overlevelsesfunktionen $S(t \mid \mathbf{z})$ i Ligning (3.1) er Cox-Snell-residualerne definerede ved

$$r^C(x_i) = -\ln\left[\widehat{S}(x_i \mid \mathbf{z}_i)\right] = -\ln\left\{\widehat{\pi}(\mathbf{z}_i) + [1-\widehat{\pi}(\mathbf{z}_i)]\widehat{\widetilde{S}}_u(x_i \mid \mathbf{z}_i)\right\}. \tag{3.38}$$

Antag, at T har værdier i $\mathbb{R}_+ \cup \{\infty\}$ og

$$S(t) = \begin{cases} \pi + (1-\pi)S_u(t), & \text{hvis } t < \infty \\ 0, & \text{hvis } t = \infty. \end{cases}$$

Fordelingsfunktionen er dermed givet ved

$$F(t) = \begin{cases} (1-\pi)F_u(t), & \text{hvis } t < \infty \\ 1, & \text{hvis } t = \infty, \end{cases}$$

hvor $F_u(T) = 1 - S_u(T)$ er en egentlig fordelingsfunktion. For $0 < q < 1 - \pi$, hvor $P(T < \infty) = 1 - \pi$, haves

$$\begin{aligned} P(F(T) \leq q) &= (1-\pi)P\left([1-\pi]F_u(T) \leq q \mid T < \infty\right) + \pi P(1 \leq q \mid T = \infty) \\ &= (1-\pi)P\left(F_u(T) \leq \frac{q}{1-\pi} \;\middle|\; T < \infty\right) = (1-\pi)\frac{q}{1-\pi} = q. \end{aligned}$$

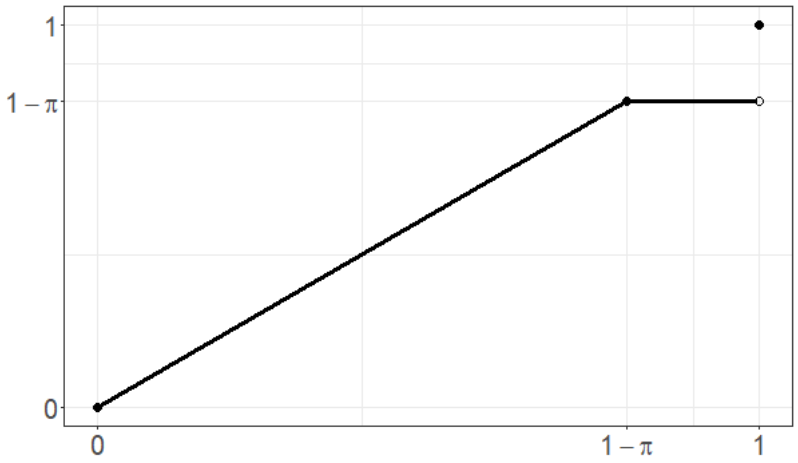
For $1 - \pi \leq q < 1$ haves

$$\begin{aligned} P(F(T) \leq q) &= (1-\pi)P\left([1-\pi]F_u(T) \leq q \mid T < \infty\right) + \pi P(1 \leq q \mid T = \infty) \\ &= (1-\pi)P\left(F_u(T) \leq \frac{q}{1-\pi} \;\middle|\; T < \infty\right) = 1 - \pi. \end{aligned}$$

For $q = 1$ haves

$$\begin{aligned} P(F(T) \leq 1) &= (1-\pi)P\left[(1-\pi)F_u(T) \leq q \mid T < \infty\right] + \pi P[1 \leq q \mid T = \infty] \\ &= (1-\pi)P\left[F_u(T) \leq \frac{q}{1-\pi} \;\middle|\; T < \infty\right] + \pi P[1 \leq q \mid T = \infty] \\ &= (1-\pi) + \pi = 1. \end{aligned}$$

Følgende figur illustrerer fordelingsfunktionen for T.



Figur 3.9: Fordelingsfunktionen for T.

Resultatet viser, at $F(T)$ opfører sig som en uniform stokastisk variabel i intervallet $[0; 1-\pi]$, hvilket betyder, at $S(T)$ opfører sig som en uniform stokastisk variabel i intervallet $[1; \pi]$. Det vil sige, at Cox-Snell-residualerne $r^C(x_i) = -\ln\left[\widehat{S}(x_i \mid \mathbf{z}_i)\right]$ følger en eksponential-fordeling med middelværdi 1 i intervallet $[0; -\ln(\pi)]$, jævnfør Proposition A.1. Det forventes derfor, at et plot af $\widehat{H}(r^C)$ mod r^C approksimerer en ret linje igennem origo med hældningskoefficient 1.

3.6.2 Modelkontrol for total overlevelse

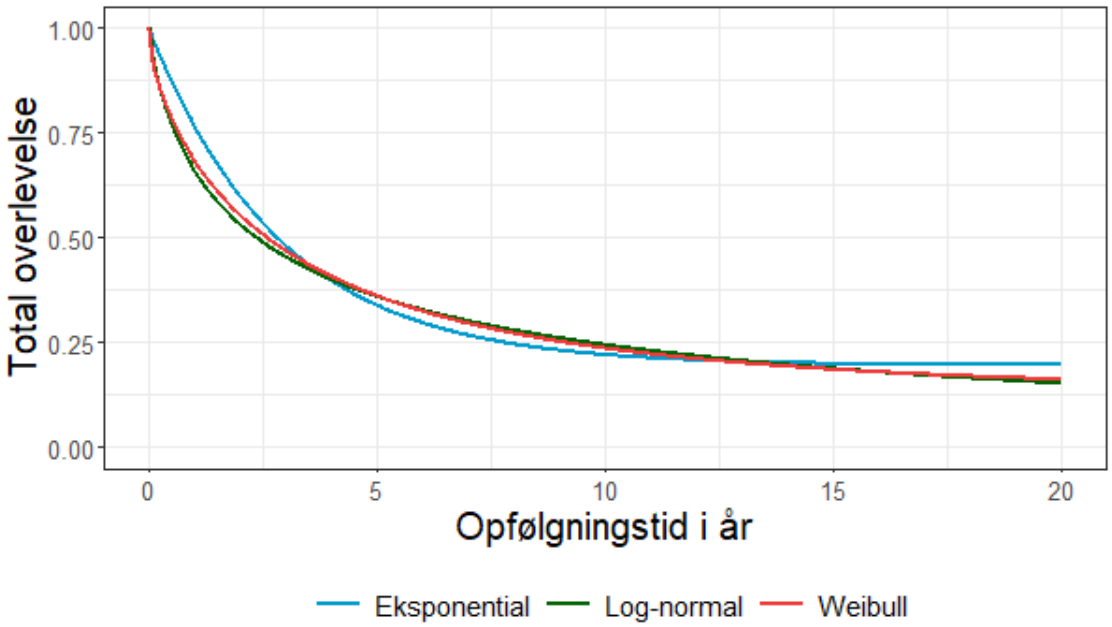
I dette underafsnit analyseres den totale overlevelse for coloncancer-patienterne. Der opstilles tre helbredelsesmodeller uden kovariater ved

$$S(t) = \pi + (1-\pi)\widetilde{S}_u(t),$$

$$\pi = \frac{\exp(\beta_0)}{1+\exp(\beta_0)},$$

og $\widetilde{S}_u(t)$, som modelleres med henholdsvis en Weibull-, log-normal- og eksponential-fordeling, se fordelingerne i Afsnit A.2 i appendiks.

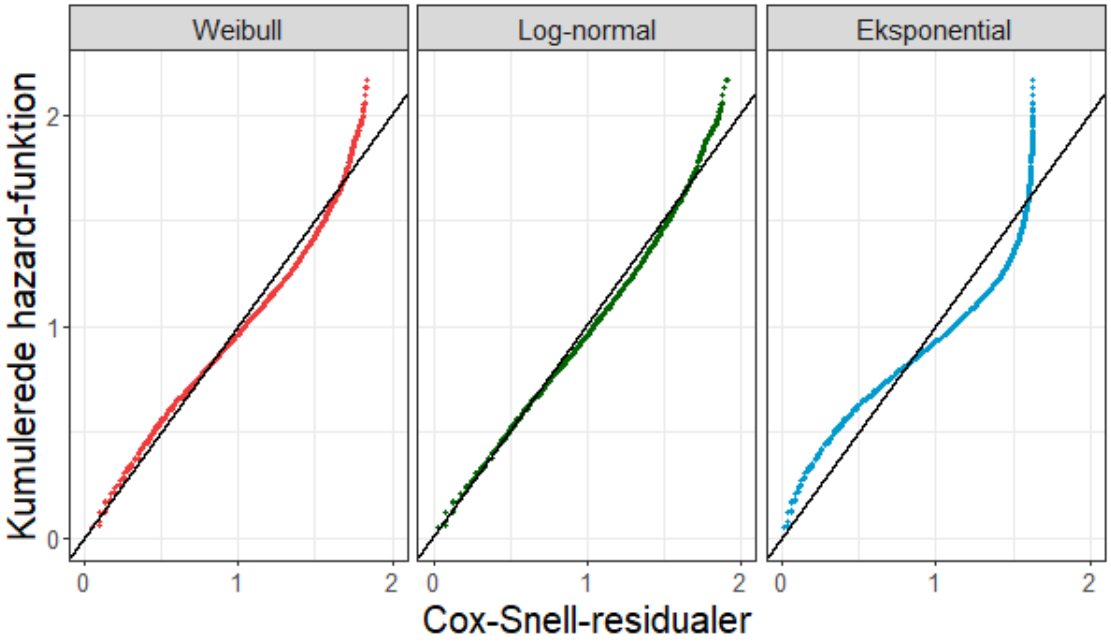
Følgende figur illustrerer den totale overlevelse for de tre modeller.



Figur 3.10: Den totale overlevelse for coloncancer-patienterne bestemt ved Weibull-, log-normal- og eksponential-modellen.

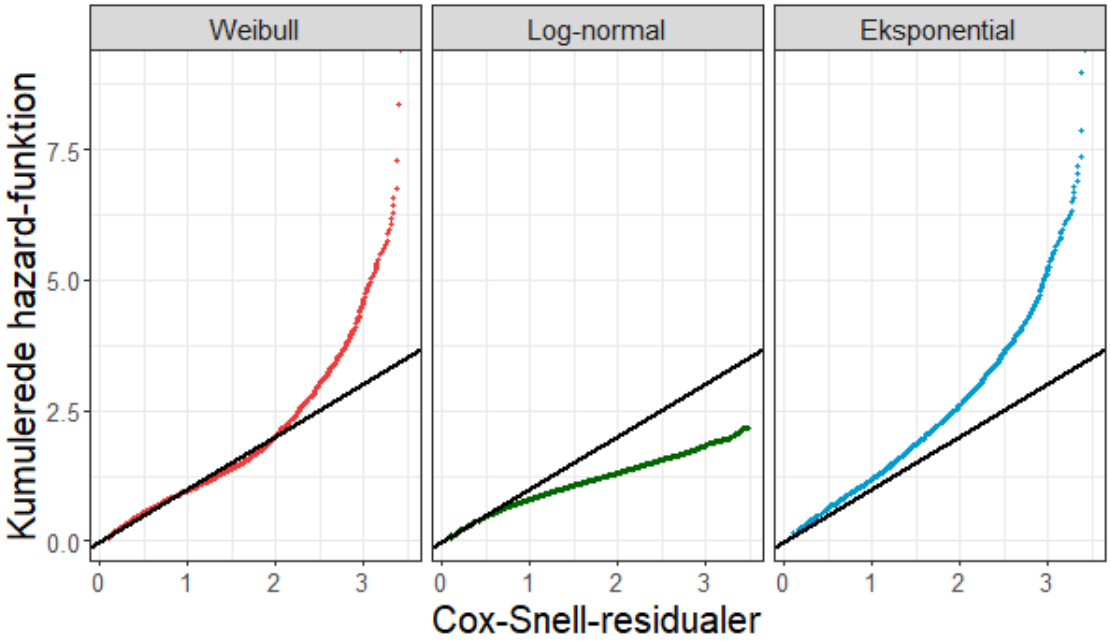
Det fremgår af Figur 3.10, at Weibull- og log-normal-modellen følger hinanden tilnærmelsesvist for hele opfølgningstiden, mens eksponential-modellen varierer mere. Eksponential-modellen estimerer en højere total overlevelse i starten og i slutningen af opfølgningen og en lavere total overlevelse omkring 5-10 år.

Vi ønsker nu at undersøge, om modellerne er tilstrækkelige. Dette kan gøres ved hjælp af Cox-Snell-residualerne for $S(t)$ i Ligning (3.38). Modellerne er tilstrækkelige, hvis et plot af $\widehat{H}(r^C)$ mod r^C approksimerer en ret linje igennem origo med hældningskoefficient 1. Vi bestemmer den kumulerede hazard-funktion af Cox-Snell-residualerne ved $\widehat{H}(r^C) = -\ln[\widehat{S}(r^C)]$, hvor $\widehat{S}(r^C)$ er et Kaplan-Meier-estimat af Cox-Snell-residualerne.



Figur 3.11: Cox-Snell-residual analyse for $S(t)$.

Det observeres på Figur 3.11, at Weibull- og log-normal-modellen følger den rette linje tilnærmelsesvist for t op til cirka 1.7 og dernæst begynder at afvige fra den rette linje. Log-normal-modellen afviger dog mindst fra den rette linje, og denne vurderes derfor til at være den bedste model. Eksponential-modellen afviger væsentligt mere fra den rette linje end Weibull- og log-normal-modellen, og derfor vurderes denne model til at være den dårligste af de tre modeller. En ulempe ved disse analyser er, at vi kun får et indblik i, om modellerne er tilstrækkelige, og ikke hvorfor de ikke skulle være det. Det eneste, der adskiller de tre modeller fra hinanden, er hvordan $\widetilde{S}_u(t)$ er modelleret. Det giver derfor god mening at undersøge $\widetilde{S}_u(t)$. Dette gør vi ved hjælp af Cox-Snell-residualerne i Ligning (3.37), hvor $\widehat{\boldsymbol{\psi}}$ er bestemt ved Ligning (3.36). I coloncancer-datasættet er der i alt 15375 patienter. For log-normal-modellen haves $\widehat{\boldsymbol{\psi}} > 0.5$ for alle patienterne, hvilket betyder, at alle patienterne er klassificerede til at være ikke-helbredte. For Weibull-modellen er $\widehat{\boldsymbol{\psi}} > 0.5$ for 13748 patienter, mens $\widehat{\boldsymbol{\psi}} > 0.5$ haves for 12409 patienter for eksponential-modellen. Den kumulerede hazard-funktion for Cox-Snell-residualerne er igen bestemt ved hjælp af Kaplan-Meier-estimatet.



Figur 3.12: Cox-Snell-residual analyse for $\widetilde{S}_u(t)$.

Det fremgår af Figur 3.12, at alle modellerne afviger i forhold til den rette linje. Eksponential-modellen afviger dog væsentligt mere end Weibull- og log-normal-modellen. Det tyder derfor på, at problemet med eksponential-modellen er, hvordan $\widetilde{S}_u(t)$ modelleres. Dette er som udgangspunkt ikke så overraskende, da det er $\widetilde{S}_u(t)$, der definerer modellen.

Kapitel 4 | Fleksible parametriske modeller

Helbredelsesmodellerne i Afsnit 3.1 og Afsnit 3.3 kræver en parametrisk fordeling til at beskrive $S_u(t \mid \mathbf{z})$ eller $\widetilde{F}(t \mid \mathbf{z})$, og det kan ofte være svært at tilpasse dem fleksibelt nok. Et eksempel herpå er, hvis der er en høj forøget hazard i starten af opfølgningsforløbet, hvilket ofte er tilfældet i kræftundersøgelser med ældre patienter. Dette kan medføre, at parametriske fordelinger ikke altid tilpasser data. Derfor er ældre patienter tidligere blevet ekskluderet i kræftundersøgelser i forbindelse med helbredelsesmodeller, [Lambert et al., 2007a]. For at undgå disse problemer, kan fleksible parametriske helbredelsesmodeller anvendes til at estimere andelen af helbredte patienter og overlevelsen for de ikke-helbredte patienter, [Andersson et al., 2011]. For at tilpasse den underliggende fordeling, anvender disse modeller en kubisk spline, som introduceres i det næste afsnit.

4.1 Kubisk spline

En kubisk spline $s(x \mid \boldsymbol{\eta})$ er en funktion, som er defineret stykkevist af tredjegradspolynomier, således at den er tvunget til at have kontinuerte første og anden ordens afledede i sammensætningspunkterne - også kaldet knuder. Et givet interval opdeles i $K-1$ delintervaller ved knuderne $k_1 < k_2 < \cdots < k_K$, hvor $k_2 < \cdots < k_{K-1}$ er indre knuder, mens k_1 og k_K er endeknuder. På hvert af de $K-1$ delintervaller tilpasses data ved et tredjegradspolynomium. En kubisk spline kan dermed beskrives ved

$$s(x \mid \boldsymbol{\eta}) = \begin{cases} s_1(x \mid \boldsymbol{\eta}), & \text{hvis } k_1 \leq x < k_2 \\ s_2(x \mid \boldsymbol{\eta}), & \text{hvis } k_2 \leq x < k_3 \\ \vdots & \\ s_{K-1}(x \mid \boldsymbol{\eta}), & \text{hvis } k_{K-1} \leq x \leq k_K, \end{cases}$$

hvor $s_j(x \mid \boldsymbol{\eta})$ er et tredjegradspolynomium for det j'te delinterval. Den kubiske spline kan også repræsenteres ved hjælp af en såkaldt plus-funktion, som er defineret

ved

$$u_+ = \begin{cases} u, & \text{hvis } u > 0 \\ 0, & \text{hvis } u \leq 0, \end{cases}$$

hvoraf den kubiske spline kan udtrykkes ved, [Andersson, 2013],

$$s(x \mid \boldsymbol{\eta}) = \eta_{00} + \eta_{01}x + \eta_{02}x^2 + \eta_{03}x^3 + \eta_1(x - k_1)_+^3 + \sum_{j=2}^{K-1} \eta_j(x - k_j)_+^3 + \eta_K(x - k_K)_+^3. \tag{4.1}$$

Den kubiske spline i Ligning (4.1) kan udvides til en restringeret kubisk spline, hvilket illustreres i det næste underafsnit.

4.1.1 Den restringerede kubiske spline

Den kubiske spline i Ligning (4.1) har $K + 4$ frihedsgrader. Derfor tilføjes der begrænsninger for at reducere antallet af frihedsgrader til K, således den restringerede kubiske spline (RKS) udledes. Der pålægges begrænsninger på koefficienterne, således at $s(x \mid \boldsymbol{\eta})$ er lineær for $x < k_1$ og $x > k_K$. Dette kan gøres ved at eliminere alle kvadratiske og kubiske led i Ligning (4.1). Linearitet for $x < k_1$ kan pålægges ved at sætte $\eta_{02} = \eta_{03} = 0$ i Ligning (4.1). For at pålægge linearitet for $x > k_K$, sættes de afledede af orden n for $n = 2, 3, \ldots$ lig med 0, da dette gælder for lineære funktioner. Dette er trivielt opfyldt for $n > 3$. For $x > k_K$ haves

$$s(x \mid \boldsymbol{\eta}) = \eta_{00} + \eta_{01}x + \eta_1(x - k_1)^3 + \sum_{j=2}^{K-1} \eta_j(x - k_j)^3 + \eta_K(x - k_K)^3, \tag{4.2}$$

$$s'(x \mid \boldsymbol{\eta}) = \eta_{01} + 3\eta_1(x - k_1)^2 + 3\sum_{j=2}^{K-1} \eta_j(x - k_j)^2 + 3\eta_K(x - k_K)^2,$$

$$s''(x \mid \boldsymbol{\eta}) = 6\eta_1(x - k_1) + 6\sum_{j=2}^{K-1} \eta_j(x - k_j) + 6\eta_K(x - k_K),$$

$$s'''(x \mid \boldsymbol{\eta}) = 6\eta_1 + 6\sum_{j=2}^{K-1} \eta_j + 6\eta_K.$$

De ukendte koefficienter η_1 og η_K elimineres ved at løse $s'''(x \mid \boldsymbol{\eta}) = 0$ i forhold til η_K og $s''(x \mid \boldsymbol{\eta}) = 0$ i forhold til η_1. Ved at løse $s'''(x \mid \boldsymbol{\eta}) = 0$ i forhold til η_K, haves

$$\eta_K = -\eta_1 - \sum_{j=2}^{K-1} \eta_j. \tag{4.3}$$

Nu udtrykkes η_1 i forhold til de resterende koefficienter. Dette gøres ved at løse $s''(x \mid \boldsymbol{\eta}) = 0$ i forhold til η_1 og indsætte $\eta_K = -\eta_1 - \sum_{j=2}^{K-1} \eta_j$. Der haves hermed

$$\begin{aligned}
\eta_1(x - k_1) &= -\sum_{j=2}^{K-1} \eta_j(x - k_j) - \eta_K(x - k_K) \\
&= -\sum_{j=2}^{K-1} \eta_j(x - k_j) + \left[\eta_1 + \sum_{j=2}^{K-1} \eta_j\right](x - k_K) \\
&= -\sum_{j=2}^{K-1} \eta_j(x - k_j) + \eta_1(x - k_K) + \sum_{j=2}^{K-1} \eta_j(x - k_K).
\end{aligned}$$

Ved at isolere η_1, haves

$$\begin{aligned}
\eta_1(x - k_1) &= -\sum_{j=2}^{K-1} \eta_j(x - k_j) + \eta_1(x - k_K) + \sum_{j=2}^{K-1} \eta_j(x - k_K) \iff \\
\eta_1(-k_1 + k_K) &= -\sum_{j=2}^{K-1} \eta_j x + \sum_{j=2}^{K-1} \eta_j k_j + \sum_{j=2}^{K-1} \eta_j x - \sum_{j=2}^{K-1} \eta_j k_K \iff \\
\eta_1 &= -\sum_{j=2}^{K-1} \eta_j \frac{k_K - k_j}{k_K - k_1} \\
&= -\sum_{j=2}^{K-1} \eta_j \phi_j,
\end{aligned} \tag{4.4}$$

hvor $\phi_j = \frac{k_K - k_j}{k_K - k_1}$. Ved at anvende Ligning (4.3) og Ligning (4.4), kan summen $\eta_1(x - k_1)^3 + \eta_K(x - k_K)^3$ i Ligning (4.2) skrives som

$$\begin{aligned}
&\eta_1(x - k_1)^3 + \eta_K(x - k_K)^3 \\
&= -\sum_{j=2}^{K-1} \eta_j \phi_j (x - k_1)^3 - \left(\eta_1 + \sum_{j=2}^{K-1} \eta_j\right)(x - k_K)^3
\end{aligned}$$

$$= -\sum_{j=2}^{K-1} \eta_j \phi_j (x-k_1)^3 - \left(-\sum_{j=2}^{K-1} \eta_j \phi_j + \sum_{j=2}^{K-1} \eta_j \right) (x-k_K)^3$$

$$= -\sum_{j=2}^{K-1} \eta_j \phi_j (x-k_1)^3 - \left(\sum_{j=2}^{K-1} \eta_j (1-\phi_j) \right) (x-k_K)^3$$

$$= \sum_{j=2}^{K-1} \eta_j \left[-\phi_j (x-k_1)^3 - (1-\phi_j)(x-k_K)^3 \right], \tag{4.5}$$

hvor to frihedsgrader η_1 og η_K er eliminerede. Ved at anvende Ligning (4.5), kan den RKS for alle værdier af x skrives som

$$s_R(x \mid \boldsymbol{\eta}) = \eta_{00} + \eta_{01} x + \eta_1 (x-k_1)_+^3 + \sum_{j=2}^{K-1} \eta_j (x-k_j)_+^3 + \eta_K (x-k_K)_+^3$$

$$= \eta_{00} + \eta_{01} x + \sum_{j=2}^{K-1} \eta_j \left[(x-k_j)_+^3 - \phi_j (x-k_1)_+^3 - (1-\phi_j)(x-k_K)_+^3 \right]$$

$$= \eta_{00} + \eta_{01} \nu_1(x) + \eta_2 \nu_2(x) + \cdots + \eta_{K-1} \nu_{K-1}(x), \tag{4.6}$$

hvor $\nu_j(x) = (x-k_j)_+^3 - \phi_j (x-k_1)_+^3 - (1-\phi_j)(x-k_K)_+^3$ for $j = 2, \ldots, K-1$ og $\nu_1(x) = x$. Det vil sige, at den RKS i Ligning (4.6) er opnået ved at eliminere fire koefficienter $\eta_{02}, \eta_{03}, \eta_1$ og η_K fra den kubiske spline i Ligning (4.1). Det er vigtigt at bemærke for den RKS, at alle spline-basisfunktionerne $\nu_j(x)$ bortset fra den lineære, $\nu_1(x) = x$, er 0 før den første knude. Denne egenskab anvendes i forbindelse med de fleksible parametriske helbredelsesmodeller, som introduceres i Afsnit 4.3.

I følgende afsnit introduceres, hvordan den RKS anvendes inden for overlevelsesmodeller.

4.2 Fleksible overlevelsesmodeller

Censurerede overlevelsesdata modelleres ofte ved hjælp af Cox proportional hazard-modellen (CPH-modellen). Denne model har den fordel at estimere kovariateffekter som log hazard-ratio, uden at der er behov for at estimere reference hazard-funktionen. Dette kan gøres ved at maksimere en såkaldt partiel likelihoodfunktion, [Klein and Moeschberger, 2003, Ligning (8.3.1)], som er uafhængig af reference hazard-funktionen. Et alternativ til CPH-modellen er parametriske modeller,

som specificerer reference hazard-funktionen parametrisk ved eksempelvis at anvende Weibull-fordelingen. Når hazard-funktionen imidlertid har en mere kompliceret form, for eksempel en U-form, er Weibull-fordelingen for simpel og kan derfor ikke opfange den underliggende tendens. En alternativ model til CPH-modellen og simple parametriske modeller, som Weibull-modellen, er Royston-Parmar-modellen, [Royston and Parmar, 2002]. Denne model er en mere fleksibel parametrisk overlevelsesmodel og karakteriseres ved at anvende en RKS, se Ligning (4.6).

Den kumulerede hazard-funktion kan modelleres ved Ligning (A.5). Ved at tage logaritmen af denne, haves

$$\ln\left[H(t \mid \mathbf{z})\right] = \ln\left[H_0(t)\right] + \boldsymbol{\beta}^\top \mathbf{z}.$$

Royston-Parmar-modellen anvender en RKS $s_R(x \mid \boldsymbol{\eta}_0)$ på log-tidsskalaen til at modellere $\ln[H_0(t)]$, og modellen er dermed givet som

$$\ln\left[H(t \mid \mathbf{z})\right] = s_R(x \mid \boldsymbol{\eta}_0) + \boldsymbol{\beta}^\top \mathbf{z}, \tag{4.7}$$

hvor $x = \ln(t)$ samt $\boldsymbol{\eta}_0$ og $\boldsymbol{\beta}$ er model-parametre. Dette er en Weibull-model givet ved

$$\ln[H(t \mid \mathbf{z})] = \ln(\gamma_1) + \gamma_2 \ln(t) + \boldsymbol{\beta}^\top \mathbf{z},$$

hvis $s_R(x \mid \boldsymbol{\eta}_0)$ erstattes med en lineær funktion af x. En vigtig egenskab ved at anvende den RKS i Ligning (4.7) er, at reference funktionen modelleres som en glat funktion i stedet for en trinfunktion. Overlevelsesfunktionen for Royston-Parmar-modellen er givet ved

$$S(t \mid \mathbf{z}) = \exp[-H(t \mid \mathbf{z})] = \exp\left[-\exp\left(s_R(x \mid \boldsymbol{\eta}_0) + \boldsymbol{\beta}^\top \mathbf{z}\right)\right]. \tag{4.8}$$

Den tilhørende hazard-funktion er bestemt ved

$$\begin{aligned} h(t \mid \mathbf{z}) = -\frac{d}{dx}\ln[S(t \mid \mathbf{z})] &= -\frac{d}{dx}\ln\left\{\exp\left[-\exp\left(s_R(x \mid \boldsymbol{\eta}_0) + \boldsymbol{\beta}^\top \mathbf{z}\right)\right]\right\} \\ &= \frac{d}{dx}\exp\left[s_R(x \mid \boldsymbol{\eta}_0) + \boldsymbol{\beta}^\top \mathbf{z}\right] \\ &= \exp\left[s_R(x \mid \boldsymbol{\eta}_0) + \boldsymbol{\beta}^\top \mathbf{z}\right]\frac{d}{dx}s_R(x \mid \boldsymbol{\eta}_0). \end{aligned} \tag{4.9}$$

Det vil sige, at Royston-Parmar-modellen er en proportional hazard-model i forhold til $\boldsymbol{\beta}$, når $\boldsymbol{\eta}_0$ er holdt fast, med

$$h_0(t) = \exp\left[s_R(x \mid \boldsymbol{\eta}_0)\right]\frac{d}{dx}s_R(x \mid \boldsymbol{\eta}_0).$$

Dette betyder, at $\boldsymbol{\beta}$ kan fortolkes ligesom i en CPH-model. Royston-Parmar-modellen tilpasses ved at maksimere log-likelihoodfunktionen, som kan bestemmes ved at substituere Ligning (4.8) og Ligning (4.9) ind i Ligning (A.3):

$$\begin{aligned} l &= \sum_{i=1}^{n} \delta_i \ln[h(x_i \mid \mathbf{z}_i)] + \ln[S(x_i \mid \mathbf{z}_i)] \\ &= \sum_{i=1}^{n} \delta_i \ln\left\{\exp\left[s_R(x_i \mid \boldsymbol{\eta}_0) + \boldsymbol{\beta}^\top \mathbf{z}_i\right] \frac{d}{dx_i} s_R(x_i \mid \boldsymbol{\eta}_0)\right\} + \ln\left\{\exp\left[-\exp\left(s_R(x_i \mid \boldsymbol{\eta}_0) + \boldsymbol{\beta}^\top \mathbf{z}_i\right)\right]\right\} \\ &= \sum_{i=1}^{n} \delta_i \left\{s_R(x_i \mid \boldsymbol{\eta}_0) + \boldsymbol{\beta}^\top \mathbf{z}_i + \ln\left[\frac{d}{dx_i} s_R(x_i \mid \boldsymbol{\eta}_0)\right]\right\} - \exp\left[s_R(x_i \mid \boldsymbol{\eta}_0) + \boldsymbol{\beta}^\top \mathbf{z}_i\right]. \end{aligned}$$

Fleksibiliteten af Royston-Parmar-modellen bestemmes ved antallet af knuder i den RKS. For at bestemme antallet af knuder, kan AIC eller lignende metoder anvendes, [Royston and Parmar, 2002].

Fremgangsmåden for Royston-Parmar-modellen kan også anvendes for relativ overlevelse ved at anvende en RKS til at modellere den log kumulerede forøgede hazard-funktion, [Nelson et al., 2007],

$$\ln\left[\Lambda(t \mid \mathbf{z})\right] = \ln\left\{-\ln[R(t \mid \mathbf{z})]\right\} = s_R(x \mid \boldsymbol{\eta}_0) + \boldsymbol{\beta}^\top \mathbf{z},$$

hvilket er en proportional forøget hazard-model. Ikke-proportionale forøgede hazard-modeller, som er modeller med tidsafhængige kovariateffekter, kan modelleres ved at inkludere interaktioner mellem kovariater og spline-funktionerne for tiden

$$\ln[\Lambda(t \mid \mathbf{z})] = \ln\left\{-\ln[R(t \mid \mathbf{z})]\right\} = s_R(x \mid \boldsymbol{\eta}_0) + \widetilde{\boldsymbol{\beta}}^\top \widetilde{\mathbf{z}} + \sum_{l=1}^{D} s_R(x \mid \boldsymbol{\eta}_l) z_l, \qquad (4.10)$$

hvor $s_R(x \mid \boldsymbol{\eta}_0)$ er spline-funktionen for den log kumulerede forøgede reference hazard-funktion, D er antallet af tidsafhængige kovariater, og $s_R(x \mid \boldsymbol{\eta}_l)$ er spline-funktionen for den l'te tidsafhængige effekt. Alle spline-funktionerne er lineære efter den sidste knude, da der anvendes en RKS. Antallet af knuder og deres placeringer behøver ikke være den samme for de forskellige spline-funktioner. Derudover er $\widetilde{\mathbf{z}}$ kovariatvektoren uden de D tidsafhængige kovariater med tilhørende kovariateffekter $\widetilde{\boldsymbol{\beta}}$. Ud fra Ligning (4.10) kan den relative overlevelse skrives som

$$R(t \mid \mathbf{z}) = \exp\left\{-\exp\left[s_R(x \mid \boldsymbol{\eta}_0) + \widetilde{\boldsymbol{\beta}}^\top \widetilde{\mathbf{z}} + \sum_{l=1}^{D} s_R(x \mid \boldsymbol{\eta}_l) z_l\right]\right\}, \qquad (4.11)$$

Ved at anvende $\lambda(t \mid \mathbf{z}) = -\frac{d \ln[R(t|\mathbf{z})]}{dx}$, haves

$$\lambda(t \mid z) = \left[\frac{1}{t}\frac{ds_R(x \mid \boldsymbol{\eta}_0)}{dx} + \sum_{l=1}^{D} z_l \frac{1}{t}\frac{ds_R(x \mid \boldsymbol{\eta}_l)}{dx}\right] \exp\left\{s_R(x \mid \boldsymbol{\eta}_0) + \widetilde{\boldsymbol{\beta}}^\top \widetilde{\mathbf{z}} + \sum_{l=1}^{D} s_R(x \mid \boldsymbol{\eta}_l) z_l\right\}, \tag{4.12}$$

hvor $\frac{ds_R(x|\boldsymbol{\eta}_0)}{dx} = \eta_{01} + \sum_{j=2}^{K-1} \eta_j \left[3(x-k_j)_+^2 - 3\phi_j(x-k_1)_+^2 - 3(1-\phi_j)(x-k_K)_+^2)\right]$.
Den totale overlevelse kan bestemmes ved at substituere Ligning (4.11) ind i Ligning (2.2), og den tilhørende hazard-funktion kan bestemmes ved at substituere Ligning (4.12) ind i Ligning (2.3), hvilket giver

$$S(t \mid \mathbf{z}) = S^*(t \mid \mathbf{z}) \exp\left\{-\exp\left[s_R(x \mid \boldsymbol{\eta}_0) + \widetilde{\boldsymbol{\beta}}^\top \widetilde{\mathbf{z}} + \sum_{l=1}^{D} s_R(x \mid \boldsymbol{\eta}_l) z_l\right]\right\}, \tag{4.13}$$

$$\begin{aligned} h(t \mid \mathbf{z}) = h^*(t \mid \mathbf{z}) + &\left[\frac{1}{t}\frac{ds_R(x \mid \boldsymbol{\eta}_0)}{dx} + \sum_{l=1}^{D} z_l \frac{1}{t}\frac{ds_R(x \mid \boldsymbol{\eta}_l)}{dx}\right] \\ &\exp\left\{s_R(x \mid \boldsymbol{\eta}_0) + \widetilde{\boldsymbol{\beta}}^\top \widetilde{\mathbf{z}} + \sum_{l=1}^{D} s_R(x \mid \boldsymbol{\eta}_l) z_l\right\}. \end{aligned} \tag{4.14}$$

Log-likelihoodfunktionen kan nu bestemmes ved at substituere Ligning (4.13) og Ligning (4.14) ind i Ligning (A.3):

$$\begin{aligned} l = \sum_{i=1}^{n} \delta_i \ln &\left\{h^*(x_i \mid \mathbf{z}_i) + \left[\frac{1}{t_i}\frac{ds_R(x_i \mid \boldsymbol{\eta}_0)}{dx_i} + \sum_{l=1}^{D} z_l \frac{1}{t_i}\frac{ds_R(x_i \mid \boldsymbol{\eta}_l)}{dx_i}\right]\right. \\ &\left.\exp\left[s_R(x_i \mid \boldsymbol{\eta}_0) + \widetilde{\boldsymbol{\beta}}^\top \widetilde{\mathbf{z}} + \sum_{l=1}^{D} s_R(x_i \mid \boldsymbol{\eta}_l) z_l\right]\right\} - \exp\left[s_R(x_i \mid \boldsymbol{\eta}_0) + \widetilde{\boldsymbol{\beta}}^\top \widetilde{\mathbf{z}} + \sum_{l=1}^{D} s_R(x_i \mid \boldsymbol{\eta}_l) z_l\right], \end{aligned}$$

hvor $S^*(x_i \mid \mathbf{z}_i)$ udelades, da den er uafhængig af modellens ukendte parametre.

Fleksible parametriske modeller kan udvides til helbredelsesmodeller, hvilket introduceres i det næste afsnit.

4.3 Fleksible parametriske helbredelsesmodeller

En mulighed for at introducere mere fleksibilitet i helbredelsesmodellerne er at inkorporere statistisk helbredelse i en fleksibel parametrisk overlevelsesmodel, således at den relative overlevelse bliver konstant efter den sidste knude, [Andersson et al., 2011]. Dette har vist sig at resultere i en ikke-mixtur helbredelsesmodel, og siden er det forsøgt at gøre modellerne mere fleksible ved at modellere $S_u(t \mid \mathbf{z})$ i mixtur helbredelsesmodellen ved hjælp af en RKS, [Jakobsen et al., 2019].

I den fleksible parametriske overlevelsesmodel i Ligning (4.11), kan statistisk helbredelse inkorporeres ved at tvinge modellen, som er lineær efter den sidste knude, til at have en hældningskoefficient lig med 0. Den sidste knude udgør dermed helbredelsestidspunktet, og andelen af helbredte individer kan således bestemmes i denne. Alle spline-basisfunktionerne, bortset fra $\nu_1(x) = x$, er lig med 0 før den første knude. Det vil sige, at hældningen før den første knude kan bestemmes ved at indføre begrænsninger på η_{01}. For at inkorporere statistisk helbredelse i den fleksible parametriske overlevelsesmodel, skal dette gøres efter den sidste knude i stedet. Dette kan gøres ved at betragte spline-basisfunktionerne i omvendt rækkefølge, hvor k_1 behandles som den sidste knude, og k_K behandles som den første knude. Dette bevirker, at alle spline-basisfunktionerne, bortset fra $v_1(x) = x$, er 0 efter den sidste knude i stedet for den første. Denne spline kaldes også for en baglæns-spline. Spline-basisfunktionerne $v_j(x)$ for denne baglæns-spline er defineret som, [Andersson et al., 2011],

$$v_j(x) = (k_{K-j+1} - x)_+^3 - \phi_j(k_K - x)_+^3 - (1 - \phi_j)(k_1 - x)_+^3, \tag{4.15}$$

for $j = 2, \ldots, K-1$ og $v_1(x) = x$, og hvor $\phi_j = \frac{k_{K-j+1} - k_1}{k_K - k_1}$. Statistisk helbredelse inkorporeres dermed ved at bruge denne baglæns-spline og begrænse η_{01} til at være lig med 0. Den relative overlevelse for denne fleksible parametriske helbredelsesmodel er givet som, [Andersson et al., 2011],

$$R(t) = \exp\left\{-\exp\left[\eta_{00} + \eta_{02}v_2(x) + \cdots + \eta_{0K-1}v_{K-1}(x)\right]\right\}. \tag{4.16}$$

Denne model kaldes også for ARS-modellen. Ved at sætte $\pi = \exp\left[-\exp(\eta_{00})\right]$, kan ARS-modellen skrives som

$$R(t) = \exp\left\{-\exp(\eta_{00})\exp\left[\eta_{02}v_2(x) + \cdots + \eta_{0K-1}v_{K-1}(x)\right]\right\}$$

$$= \exp\left[-\exp(\eta_{00})\right]^{\exp[\eta_{02}v_2(x)+\cdots+\eta_{0K-1}v_{0K-1}(x)]}$$
$$= \pi^{\exp[\eta_{02}v_2(x)+\cdots+\eta_{0K-1}v_{K-1}(x)]},$$

hvilket er et specialtilfælde af en ikke-mixtur helbredelsesmodel, hvor $\pi = \exp\left[-\exp(\eta_{00})\right]$ er andelen af helbredte individer, og fordelingsfunktionen er

$$\widetilde{F}(t) = \exp\left[\eta_{02}v_2(x) + \cdots + \eta_{0K-1}v_{K-1}(x)\right],$$

som er lig med 1 efter den sidste knude. Da ARS-modellen er konstant efter den sidste knude, kan andelen af helbredte individer bestemmes i den sidste knude, $\pi = R(K_1)$. Det betyder også, at placeringen af den sidste knude er essentiel for modellen. ARS-modellen har tidligere vist at give gode resultater, så længe knuderne placeres over hele opfølgningstiden og den sidste knude ved det sidste dødstidspunkt eller senere for at sikre, at der ikke pålægges et helbredelsespunkt for tidligt, [Andersson et al., 2011]. Ved at inkludere kovariater, kan den relative overlevelse i Ligning (4.16) skrives som

$$R(t \mid \mathbf{z}) = \exp\left\{-\exp(\eta_{00} + \widetilde{\boldsymbol{\beta}}^\top \widetilde{\mathbf{z}}) \exp\left[\eta_{02}v_2(x) + \cdots + \eta_{0K-1}v_{K-1}(x) + \sum_{l=1}^{D} s_B(x \mid \boldsymbol{\eta}_l) z_l\right]\right\},$$

hvor $s_B(x \mid \boldsymbol{\eta}_l)$ er den l'te baglæns-spline. Koefficienten η_{l1} tilhørende den lineære spline-basisfunktion skal begrænses til at være lig med 0 for alle $s_B(x \mid \boldsymbol{\eta}_l)$, som inkluderes i modellen. Det vil sige, at alle spline-basisfunktioner er 0 efter den sidste knude, hvilket betyder, at η_{00} er den log kumulerede forøgede hazard for referencegruppen efter den sidste knude. Parametrene η_{00} og $\widetilde{\boldsymbol{\beta}}$ anvendes derfor til at modellere andelen af helbredte individer, mens de tidsafhængige parametre anvendes til at modellere $\widetilde{F}(t)$.

ARS-modellens antagelse om statistisk helbredelse efter den sidste knude er stærk, og derfor introduceres en anden fleksibel parametrisk helbredelsesmodel, [Jakobsen et al., 2019]. Denne tager udgangspunkt i mixtur helbredelsesmodellen

$$R(t \mid \mathbf{z}) = \pi(\mathbf{z}) + [1 - \pi(\mathbf{z})]S_u(t \mid \mathbf{z}),$$

hvor $S_u(t \mid \mathbf{z})$ modelleres med RKS, i stedet for en simpel parametrisk fordeling, som

$$S_u(t \mid \mathbf{z}) = \exp\left[-\exp\left(s_R(x \mid \boldsymbol{\eta}_0) + \widetilde{\boldsymbol{\beta}}^\top \widetilde{\mathbf{z}} + \sum_{l=1}^{D} s_R(x \mid \boldsymbol{\eta}_l) z_l\right)\right].$$

Denne model kaldes også for FMC-modellen. Spline-funktionerne i FMC-modellen er lineære efter den sidste knude, men den relative overlevelse er nedadtil begrænset af $\pi(\mathbf{z})$. FMC-modellen gør det muligt at modellere den relative overlevelse fleksibelt uden at antage statistisk helbredelse efter den sidste knude. Samtidig inkluderes den mere intuitive fortolkning af en mixtur helbredelsesmodel. FMC-modellen kan tilpasses ved at maksimere log-likelihoodfunktionen for mixtur helbredelsesmodellen i Ligning (3.6).

I al almindelighed kan knudeplaceringerne bestemmes ved hjælp af AIC eller lignende metoder, men som beskrevet i Underafsnit 3.2, kan AIC give problemer i forbindelse med helbredelsesmodeller. Det anbefales derfor også at sammenligne de forskellige modeller med et ikke-parametrisk estimat for at sikre, at modellen med den laveste AIC også giver et tilstrækkeligt estimat af $\pi(\mathbf{z})$ og af $S_u(t \mid \mathbf{z})$. Det skal dog understreges, at et ikke-parametrisk estimat ikke er den sande relative overlevelse.

4.4 Dataanalyse for fleksible helbredelsesmodeller

I dette afsnit analyseres coloncancer-datasættet ved hjælp af fleksible parametriske helbredelsesmodeller. Til dette formål anvendes to forskellige helbredelsesmodeller: ARS- og FMC-modellen. Der inkluderes ikke kovariater i modellerne. ARS-modellen med K knuder er bestemt ved

$$R(t) = \exp\left\{-\exp\left[\eta_{00} + \eta_{02}v_2(x) + \cdots + \eta_{0K-1}v_{K-1}(x)\right]\right\}, \tag{4.17}$$

hvor $v_j(x)$ er givet i Ligning (4.15). FMC-modellen med K knuder er bestemt ved

$$\begin{aligned} R(t) &= \pi + (1-\pi)S_u(t), \\ \pi &= \frac{\exp(\beta_0)}{1+\exp(\beta_0)}, \\ S_u(t) &= \exp\left\{-\exp\left[s_R(x \mid \boldsymbol{\eta}_0)\right]\right\}, \end{aligned} \tag{4.18}$$

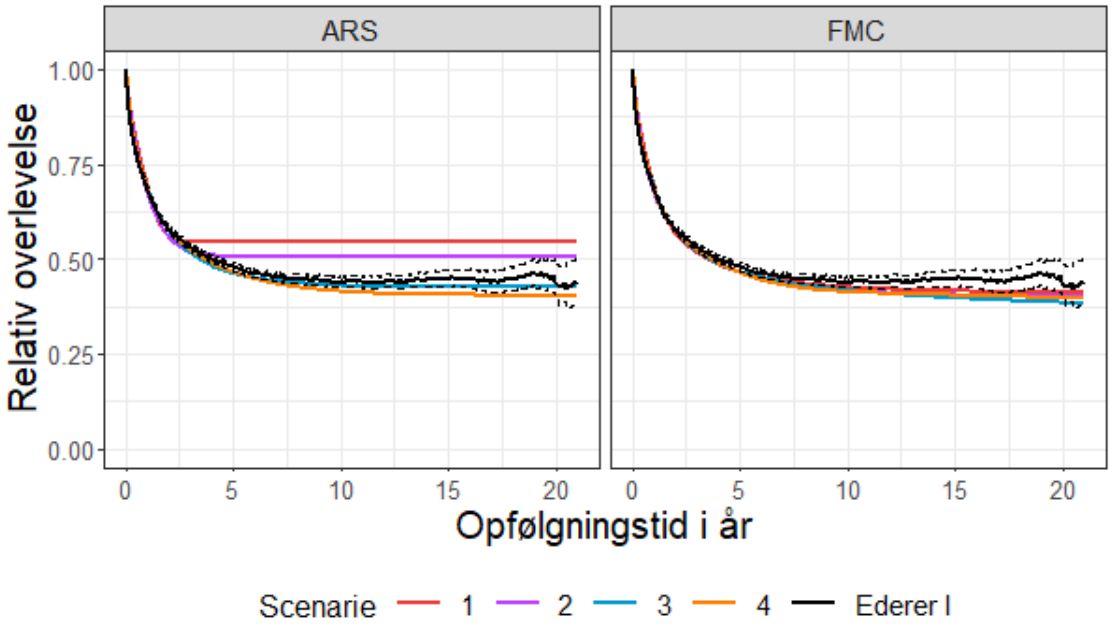
hvor $s_R(x \mid \boldsymbol{\eta}_0)$ er givet i Ligning (4.6). Det er nødvendigt for begge modeller at specificere antallet af knuder og placeringerne af disse. Placeringen af den sidste knude er særligt vigtig for ARS-modellen, da denne beskriver helbredelsestidspunktet for modellen.

Vi ønsker at undersøge, hvor sensitiv ARS- og FMC-modellen er i forhold til antallet af knuder og deres placeringer. Dette gør vi ved at sammeligne ARS- og FMC-modellen med Ederer I samt ved hjælp af AIC, som er bestemt ved Ligning (3.8). Tabel 4.1 angiver fire forskellige scenarier af knudeplaceringer, som ARS- og FMC-modellen opstilles for.

Scenarie	Antal knuder	Knudeplaceringer
1	3	1, 2 og 3 år
2	4	Første ikke-censurerede begivenhedstid, 1, 3 og 5 år.
3	6	0%, 20%, 40%, 60%, 80% og 100% kvartilerne af de ikke-censurerede begivenhedstider.
4	8	0%, 20%, 40%, 60%, 80% og 100% kvartilerne af de ikke-censurerede begivenhedstider. To ekstra knuder er placeret ved 8 og 15 år.

Tabel 4.1: Knudeplaceringer for ARS- og FMC-modellen. Knuderne for 0%, 20%, 40%, 60%, 80% og 100% kvartilerne af de ikke-censurerede begivenhedstider svarer til dag 15, 76, 229, 561, 1384 og 7470. Bemærk desuden, at den første ikke-censurerede begivenhedstid svarer til 0% kvartilen.

Følgende figur illustrerer den estimerede relative overlevelse for ARS- og FMC-modellen under scenarierne i Tabel 4.1 sammen med Ederer I-estimatet.



Figur 4.1: Den relative overlevelse for coloncancer-datasættet udregnet ved Ederer I med tilhørende 95% konfidensinterval (stiplede linjer), samt ARS- og FMC-modellen under forskellige knudeplacering-scenarier.

Følgende tabel opsummerer modellernes AIC under de forskellige scenarier.

Model	Scenarie 1	Scenarie 2	Scenarie 3	Scenarie 4
ARS	47492.06	46068.35	44926.47	44887.65
FMC	45324.86	45065.52	44879.1	44870.61

Tabel 4.2: AIC for ARS- og FMC-modellen under de forskellige scenarier.

Det fremgår af Figur 4.1, at ARS-modellen er meget sensitiv i forhold til knudeplacering. Dette var forventet, da den sidste knude beskriver helbredelsestidspunktet. Scenarie 1 og 2 for ARS-modellen ligger over Ederer I. Dette skyldes, at den sidste knude er placeret ved henholdsvis 3 og 5 år, og derfor flader den relative overlevelse ud herefter. Det er simpelthen for tidligt at antage statistisk helbredelse for coloncancer-patienterne efter 3 og 5 år. Scenarie 3 og 4 for ARS-modellen ligger til gengæld under Ederer I. Overordnet set er der ikke den store forskel mellem scenarie

3 og 4, men det fremgår af Tabel 4.2, at scenarie 4 resulterer i en lavere AIC, hvilket betyder, at scenarie 4 foretrækkes.

For FMC-modellen observerer vi, at der generelt ikke er den store forskel mellem scenarierne. De ligger alle under Ederer I, men kurverne ligger meget tæt på hinanden. Ud fra Tabel 4.2 fremgår det dog, at scenarie 4 resulterer i den laveste AIC, og derfor foretrækkes scenarie 4 også for FMC-modellen. Det generelle billede fortæller os, at ARS-modellen er mere sensitiv over for knudeplaceringerne end FMC-modellen. Under scenarie 4 er den 2 og 5 års relative overlevelse henholdsvis 0.58 (95% KI 0.57-0.59) og 0.46 (95% KI 0.45-0.47) for FMC-modellen samt 0.58 (95% KI 0.57-0.58) og 0.46 (95% KI 0.46-0.47) for ARS-modellen. Derudover er andelen af helbredte patienter 0.26 (95% KI 0.21-0.31) for FMC-modellen og 0.40 (95% KI 0.39-0.42) for ARS-modellen. Der er altså en forskel mellem andelen af helbredte patienter for de to modeller, men ikke for den 2 og 5 års relative overlevelse. Det er en smule overraskende, at estimatet af andelen af helbredte patienter er så lavt for FMC-modellen. Dette stemmer ikke overens med Ederer I, og som det fremgår af Figur 4.1, forekommer estimatet efter 21 år. Det tager altså lang tid, inden der opnås en komplet udfladning af den relative overlevelse for FMC-modellen under scenarie 4. Det tyder derfor på, at FMC-modellen ikke giver et tilstrækkeligt estimat af andelen af helbredte patienter, hvis scenarie 4 anvendes. Dette er også tilfældet for scenarie 3. Det kan muligvis skyldes, at FMC-modellen er for fleksibel under scenarie 3 og 4. På Figur 4.1 observerede vi dog, at alle scenarierne fulgte hinanden tilnærmelsesvist. Hvis scenarie 2 i stedet anvendes for FMC-modellen, får vi henholdsvis 0.58 (95% KI 0.57-0.59), 0.46 (95% KI 0.46-0.47) og 0.40 (95% KI 0.38-0.42) for den 2 års relative overlevelse, den 5 års relative overlevelse og andelen af helbredte patienter. Dette er et bedre resultat, når der sammenlignes med Ederer I. Det konkluderes derfor, at scenarie 2 er bedre end scenarie 3 og 4 for FMC-modellen på trods af, at AIC-værdien er mindre for scenarie 3 og 4.

4.4.1 Stratificeret efter aldersgruppe

I dette underafsnit tilpasses ARS- og FMC-modellen i Ligning (4.17) og Ligning (4.18) til aldersgrupperne 18-44, 45-59, 60-74 og 75-90. Figur B.3-B.6 i appendiks illustrerer den estimerede relative overlevelse for modellerne under de forskellige aldersgrupper. Tabel B.1-B.4 opsummerer desuden modellernes AIC under de forskellige scenarier. Det er vigtigt at bemærke, at kvartilerne af de ikke-censurerede begivenhedstider er forskellige for aldersgrupperne. Knudeplaceringerne for kvartilerne er angivet i dage under deres repræsentative figur i Afsnit B.2 i appendiks. Det fremgår eksempelvis, at knudeplaceringerne i scenarie 3 og 4 er gode for alle aldersgrupperne for ARS-modellen, når der sammenlignes med Ederer I. Vi observerede et lignende resultat i forbindelse med Figur 4.1. Det fremgår også, at AIC-værdierne for scenarie 3 og 4 stort set er identiske for aldersgrupperne 18-44 og 45-59. For aldersgrupperne 60-74 og 75-90 er AIC-værdien derimod en smule mindre for scenarie 4 end scenarie 3. Vi konkluderer derfor, at knudeplaceringerne i scenarie 4 er de bedste for ARS-modellen.

For FMC-modellen er der stort set ingen forskel mellem scenarierne for aldersgrupperne 18-44, 45-59 og 60-74, når der sammenlignes med Ederer I. For aldersgruppen 75-90 er der derimod en væsentlig forskel, og det observeres, at scenarie 2 og 3 er de bedste, når der sammenlignes med Ederer I. Det fremgår dog af AIC-værdierne, at scenarie 4 er bedst for aldersgrupperne 18-44 og 60-74, mens scenarie 3 er bedst for aldersgruppen 70-95. For aldersgruppen 45-59 er AIC-værdien for scenarie 3 og 4 stort set identiske. Vi foretrækker dog alligevel knudeplaceringerne i scenarie 2, da scenarie 3 og 4 igen viser sig at resultere i utilstrækkelige estimater for andelen af helbredte patienter.

Følgende tabel opsummerer den 2 års relative overlevelse, den 5 års relative overlevelse, andelen af helbredte patienter og median relativ overlevelsestiden for de ikke-helbredte patienter for aldersgrupperne 18-44, 45-59, 60-74 og 75-90 under ARS- og FMC-modellen. For ARS-modellen er scenarie 4 anvendt, mens scenarie 2 er anvendt for FMC-modellen. R-koden til vores implementation af median relativ overlevelsestiden for de ikke-helbredte patienter findes i Afsnit C.1 i appendiks.

	Model	18-44	45-59	60-74	75-90
2 års relativ overlevelse	ARS	0.66(0.63-0.7)	0.64(0.63-0.66)	0.59(0.58-0.61)	0.51(0.49-0.52)
	FMC	0.66(0.63-0.69)	0.64(0.63-0.66)	0.6(0.59-0.61)	0.51(0.5-0.53)
5 års relativ overlevelse	ARS	0.54(0.51-0.58)	0.52(0.5-0.54)	0.48(0.47-0.49)	0.41(0.39-0.43)
	FMC	0.55(0.51-0.58)	0.52(0.49-0.54)	0.48(0.46-0.49)	0.41(0.39-0.43)
Andel helbredte patienter	ARS	0.48(0.44-0.53)	0.47(0.44-0.5)	0.4(0.38-0.43)	0.31(0.26-0.36)
	FMC	0.46(0.39-0.53)	0.47(0.44-0.5)	0.38(0.34-0.42)	0.34(0.29-0.39)
Median relativ overlevelsestid	ARS	1.25	1.13	1.07	0.65
	FMC	1.26	1.11	1.07	0.55

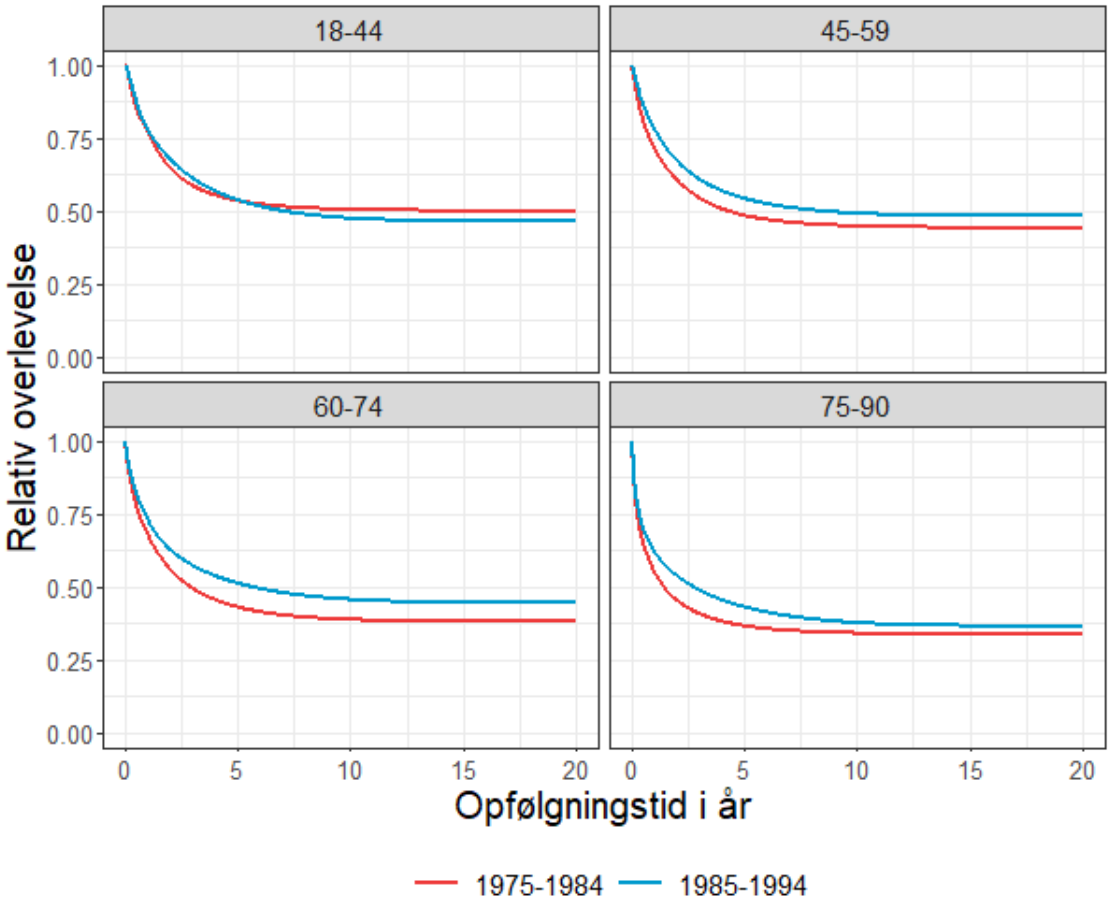
Tabel 4.3: Den 2 års relative overlevelse, den 5 års relative overlevelse, andelen af helbredte patienter og median relativ overlevelsestiden i år for de ikke-helbredte patienter. De tilhørende 95% konfidensintervaller er angivet i parentes.

Det fremgår af Tabel 4.3, at resultaterne for ARS- og FMC-modellen stort set er identiske for alle aldersgrupperne. For aldersgruppen 75-90 er median relativ overlevelsestiden for de ikke-helbredte patienter dog en smule højere for ARS-modellen. Det observeres desuden, at aldersgruppen 75-90 klarer sig dårligst, hvilket vi også observerede i forbindelse med de simple parametriske modeller i Tabel 3.2 og Tabel 3.5. Det fremgår også, at der overordnet set ikke er den store forskel mellem estimaterne for de fleksible helbredelsesmodeller i Tabel 4.3 og de simple parametriske helbredelsesmodeller i Tabel 3.2 og Tabel 3.5.

4.4.2 Stratificeret efter aldersgruppe og diagnoseperiode

I dette underafsnit ønsker vi at tilpasse ARS-modellen i Ligning (4.17) til de otte undergrupper: 18-44, 45-59, 60-74 og 75-90, som er diagnosticerede med coloncancer i henholdsvis 1975-1984 og 1985-1994. Vi har valgt at fokusere på ARS-modellen, da FMC-modellen resulterede i ustabile estimater. Det fremgik af Figur 3.4 og Figur 3.5, at der forekom en udfladning af Ederer I for alle aldersgrupperne for diagnoseperioden 1975-1984. For diagnoseperioden 1985-1994 kan det dog diskuteres, om det også var tilfældet for de to ældste aldersgrupper.

Nu tilpasses ARS-modellen til de otte undergrupper. Vi har valgt at placere knuderne ved 0%, 20%, 40%, 60%, 80% og 100% kvartilerne af de ikke-censurerede begivenhedstider, men for diagnoseperioden 1985-1994 placeres der en ekstra knude ved 20 år. Vi har valgt at tilføje denne ekstra knude for diagnoseperioden 1985-1994, da det ikke er klart, hvorvidt statistisk helbredelse forekommer i løbet af opfølgningen for de to ældste aldersgrupper. Knudeplaceringerne for kvartilerne til de forskellige undergrupper er angivet i dage i Tabel B.5 i appendiks. Det fremgår desuden af Figur B.7 og Figur B.8 i appendiks, at ARS-modellen tilnærmelsesvist følger Ederer I for alle otte undergrupper. De to diagnoseperioder sammenlignes på figuren, der følger.



Figur 4.2: Den relative overlevelse bestemt ved ARS-modellen for aldersgrupperne 18-44, 45-59, 60-74 og 75-90, som er diagnosticerede med coloncancer i 1975-1984 og 1985-1994.

Det fremgår af Figur 4.2, at aldersgrupperne 45-59, 60-74 og 75-90 har en højere relativ overlevelse i 1985-1994 end i 1975-1984 for hele opfølgningen. For aldersgruppen 18-44 observerer vi dog efter cirka 5 år, at den relative overlevelse er højere i 1975-1984 end i 1985-1994. Følgende tabel opsummerer den 2 års relative overlevelse, den 5 års relative overlevelse, andelen af helbredte patienter og median relativ overlevelsestiden for de ikke-helbredte patienter. R-koden til vores implementation af median relativ overlevelsestiden for de ikke-helbredte patienter findes i Afsnit C.1 i appendiks.

	Periode	18-44	45-59	60-74	75-90
2 års relativ overlevelse	1975-1984	0.65(0.6-0.7)	0.61(0.58-0.64)	0.56(0.54-0.58)	0.46(0.43-0.48)
	1985-1994	0.68(0.63-0.73)	0.67(0.65-0.7)	0.63(0.61-0.64)	0.54(0.52-0.56)
5 års relativ overlevelse	1975-1984	0.53(0.49-0.59)	0.48(0.45-0.52)	0.43(0.41-0.45)	0.37(0.34-0.39)
	1985-1994	0.54(0.49-0.59)	0.54(0.51-0.57)	0.51(0.49-0.53)	0.43(0.41-0.45)
Andel helbredte patienter	1975-1984	0.5(0.45-0.56)	0.44(0.41-0.48)	0.38(0.36-0.4)	0.34(0.31-0.37)
	1985-1994	0.46(0.4-0.53)	0.48(0.45-0.52)	0.45(0.42-0.48)	0.37(0.33-0.41)
Median relativ overlevelsestid	1975-1984	1.15	0.97	0.95	0.46
	1985-1994	1.41	1.32	1.07	0.61

Tabel 4.4: Den 2 års relative overlevelse, den 5 års relative overlevelse, andelen af helbredte patienter og median relativ overlevelsestiden i år for de ikke-helbredte patienter. De tilhørende 95% konfidensintervaller er angivet i parentes.

Det fremgår af Tabel 4.4, at patienterne diagnosticerede i 1985-1994 generelt klarer sig bedre end patienterne diagnosticerede i 1975-1984. Det fremgår desuden, at de ældre patienter klarer sig dårligere end de yngre patienter, hvilket særligt er gældende i starten af opfølgningen. Vi konkluderede det samme i forbindelse med Weibull-modellen med kovariater i Underafsnit 3.2.1.

Kapitel 5 | Diskussion

Én af problematikkerne med helbredelsesmodeller er, at litteraturen er sparsom i forhold til modelkontrol. I al almindelighed foretages en residualanalyse i forbindelse med modelkontrol, og inden for overlevelsesanalyse kan Cox-Snell-residualerne blandt andre anvendes. Vi har beskrevet, hvordan Cox-Snell-residualerne kan bestemmes for mixtur helbredelsesmodellen i Afsnit 3.6. Disse Cox-Snell-residualer tager udgangspunkt i den totale overlevelse, og de kan derfor ikke anvendes i forbindelse med den relative overlevelse, som har været det primære fokus i dette speciale. Vi mener dog, at det må være muligt at udvide Cox-Snell-residualerne til den relative overlevelse, men til vores kendskab er dette ikke blevet gjort før. Det kunne være interessant at bruge mere tid på dette område.

Helbredelsesmodeller giver altid et estimat af andelen af helbredte patienter - også i situationer, hvor statistisk helbredelse ikke giver mening. Det vil sige, når kurven for den relative overlevelse ikke flader ud. Det er derfor vigtigt at undersøge, om statistisk helbredelse giver mening for den givne situation, når helbredelsesmodeller anvendes. Til dette formål kan et ikke-parametrisk estimat af den relative overlevelse være behjælpelig. I vores analyser har vi sammenlignet helbredelsesmodellerne med Ederer I for at sikre, at de tilpassede helbredelsesmodeller stemmer nogenlunde overens med data. Vi kunne også have sammenlignet med Ederer II eller Hakulinen til dette formål. Disse sammenligninger fungerer som en form for modelkontrol. Det er dog vigtigt at bide mærke i, at et ikke-parametrisk estimat af den relative overlevelse ikke beskriver den sande relative overlevelse. Til vores kendskab findes der imidlertid ikke et bedre alternativ. Dette understreger relevansen for udviklingen af modelkontrol inden for relativ overlevelse.

En andet problematik inden for helbredelsesmodeller er anvendelsen af AIC, når forskellige modeller sammenlignes. Generalt foretrækkes modellen med den laveste AIC, men denne giver ikke nødvendigvis et bedre estimat af andelen af helbredte patienter. Dette skyldes, at modellerne ligger vægt på tilpasningen af $R(t)$ for de t, hvor begivenhederne indtræffer. Vi oplevede blandt andet i Afsnit 4.4, at AIC-værdien for FMC-modellen var lavest under knudeplaceringerne i scenarie 4, se Tabel 4.2. Denne model viste sig at resultere i et for lavt estimat af andelen af helbredte patienter, som forekom efter 21 år, når vi sammenlignede med Ederer I. Dette understreger, hvor vigtigt det er at sammenligne resultaterne for helbredelsesmodellerne med grafiske tjek. Det var desuden særligt et problem, at FMC-modellen gav ustabile estimater,

når vi stratificerede efter aldersgruppe og diagnoseperiode, hvilket er årsagen til, at vi fokuserede på ARS-modellen i disse analyser.

Et alternativ til de stratificerede analyser er at inkludere kovariater, ligesom der blev gjort for Weibull-modellen i Underafsnit 3.2.1. Dette viste sig at give problemer i forhold til ARS- og FMC-modellen. ARS-modellen estimerede en relativ overlevelse, som afveg markant i forhold til Ederer I for nogle af grupperne, og FMC-modellen havde konvergeringsproblemer. Disse problemer kan skyldes, at modellerne ikke er identificerbare, og vi valgte derfor at fokusere på de stratificerede analyser for ARS-modellen, som viste sig at give gode resultater, når vi sammenlignede med Ederer I, se Figur B.7 og Figur B.8 i appendiks. Vi foreslår derfor også at sammenligne med stratificerede analyser, når kovariater inkluderes, da dette giver en indikation af, hvorvidt modellerne fanger effekten af kovariaterne tilstrækkeligt. Dette gjorde vi også i Underafsnit 3.2.1, hvor vi observerede, at log-normal- og eksponential-modellen gav utilstrækkelige resultater, når vi sammenlignede med de tilhørende stratificerede analyser. Dette skyldes muligvis også identificerbarhedsproblemer. En ulempe ved de stratificerede analyser er, at de ikke giver anledning til en p-værdi, som analyser med kovariater gør. Det kan derfor være svært at konkludere, om aldersgruppe og diagnoseperiode er signifikante på samme måde. Det generelle billede i Tabel 4.4 fortæller os dog, at aldersgruppe og diagnoseperiode har en betydning for den relative overlevelse for coloncancer-patienterne.

ARS-modellen er defineret ved hjælp af en baglæns-spline, men vi mener, at det også må være muligt at definere ARS-modellen ved hjælp af en RKS. En RKS er både lineær før den første knude og efter den sidste. Det må derfor være muligt at begrænse den RKS på en tilsvarende måde, således at den RKS er konstant efter den sidste knude. Det er dog tydeligere, hvilken koefficient der skal begrænses for at inkorporere statistisk helbredelse, når en baglæns-spline anvendes. Dette kan være årsagen til, at modellen er defineret ved en baglæns-spline fremfor en RKS.

ARS-modellen antager statistisk helbredelse efter den sidste knude, hvilket betyder, at den sidste knude er essentiel for modellens udfald. Det er derfor vigtigt, at den sidste knude er placeret, hvor det er acceptabelt at antage statistisk helbredelse. ARS-modellen er i princippet en reduceret model af en fleksibel parametrisk overlevelsesmodel defineret ved en baglæns-spline. Det er derfor muligt at foretage en likelihood-ratio test for at sammenligne de to modeller og dermed undersøge, om hældningskoefficienten η_{01} tilhørende den lineære spline-basisfunktion, $v_1(x) = x$, er signifikant. Hvis η_{01} er signifikant, er statistisk helbredelse ikke opnået. Det er

dog tidligere set, at η_{01} har været signifikant, selvom der forekom en udfladning af den relative overlevelse for et ikke-parametrisk estimat, [Andersson et al., 2011]. Det kan derfor være svært at stole udelukkende på en likelihood-ratio test i denne sammenhæng. Helbredelsesmodeller giver desuden nogle oplysende estimater, som gør ARS-modellen mere fordelagtig.

I Afsnit 4.4 undersøgte vi, hvor sensitiv ARS- og FMC-modellen er i forhold til antallet af knuder og deres placeringer. I denne sammenhæng sammenlignede vi hovedsageligt med Ederer I, men som tidligere nævnt, er Ederer I ikke den sande relative overlevelse. Vi kunne i stedet have undersøgt betydningen af antallet af knuder og deres placeringer ved hjælp af simulationer. Vi kunne eksempelvis have genereret data ud fra en Weibull-model, og dermed ville den sande relative overlevelse være kendt.

Formålet med de fleksible parametriske helbredelsesmodeller er at inkludere mere fleksibilitet i modellerne, end der tillades i de simple parametriske helbredelsesmodeller. Dette kan alternativt gøres ved at anvende to fordelinger i stedet for én til at modellere $S_u(t)$ eller $\widetilde{F}(t)$ i de simple parametriske helbredelsesmodeller. Den såkaldte Weibull-Weibull-model anvender to Weibull-fordelinger, mens Weibull-eksponential-modellen anvender én Weibull- og én eksponential-fordeling, [Lambert et al., 2010]. Mixtur helbredelsesmodellen for den relative overlevelse beskrevet ved to Weibull-fordelinger er eksempelvis givet ved

$$R(t) = \pi + (1-\pi)\left[p\exp\left(-\gamma_1 t^{\gamma_2}\right) + (1-p)\exp\left(-\gamma_3 t^{\gamma_4}\right)\right].$$

Model-selektionen i denne model kan være kompleks, da parametrene p, γ_1, γ_2, γ_3 og γ_4 kan være afhængige af kovariater. Vi oplevede desuden, at modellen havde konvergeringsproblemer, når p, γ_1, γ_2, γ_3, γ_4 og π er afhængige af kovariaterne aldersgruppe og diagnoseperiode.

I forbindelse med blandt andre Tabel 3.2, Tabel 3.3 og Tabel 3.5 udregnede vi median relativ overlevelsestiden for de ikke-helbredte patienter. Dette estimat er ikke implementeret i R-pakkerne cuRe og rstpm2, som dataanalysen er udarbejdet med. Det er også årsagen til, at vi ikke har angivet et konfidensinterval for estimatet. Det er dog muligt at bestemme et konfidensinterval på tilsvarende vis som for $R(t)$, π og $S_u(t)$ i Afsnit 3.1.1. Median relativ overlevelsestiden for de ikke-helbredte patienter er bestemt ved $t = S_u^{-1}(0.5)$, og den tilhørende varians bestemt ved delta-metoden

i Sætning A.2 er

$$\mathbb{V}\mathrm{ar}\left[S_u^{-1}(0.5;\widehat{\boldsymbol{\beta}})\right] = \left(\frac{\partial S_u^{-1}(0.5;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^{\top} \widehat{\Sigma} \left(\frac{\partial S_u^{-1}(0.5;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right).$$

På trods af nogle af de ovenstående problematikker for helbredelsesmodeller er vores vurdering, at modellerne giver et godt grundlag for at analysere overlevelsen for kræftpatienter.

Kapitel 6 | Konklusion

Formålet med specialet er at beskrive helbredelsesmodeller og anvende disse til at analysere et coloncancer-datasæt. Til dette formål gennemgik vi teorien bag helbredelsesmodeller. Vi introducerede netto og relativ overlevelse. Nettooverlevelsen beskriver overlevelsen for patienter med en bestemt sygdom i det hypotetiske scenarie, hvor sygdommen er den eneste dødsårsag. Den relative overlevelse er defineret som forholdet mellem overlevelsesfunktionen for patientgruppen og den forventede overlevelse, hvor den sidstnævnte er overlevelsen for en sammenlignelig gruppe fra baggrundsbefolkninge. Denne bestemmes ud fra levetidstabeller. Vi har vist, at den relative overlevelse beskriver nettooverlevelsen under visse antagelser. Vi beskrev tre ikke-parametriske metoder til at estimere den relative overlevelse: Ederer I, Ederer II og Hakulinen.

To typer helbredelsesmodeller blev beskrevet: Mixtur helbredelsesmodeller og ikke-mixtur helbredelsesmodeller. I mixtur helbredelsesmodellerne betragtes den observerede population som to grupper; de helbredte individer og de ikke-helbredte individer. Andelen af helbredte individer kan modelleres ved en logistisk regression, og overlevelsen for de ikke-helbredte individer kan blandt andet modelleres parametrisk ved en Weibull-fordeling. Ikke-mixtur helbredelsesmodeller er sværere at fortolke end mixtur helbredelsesmodeller. Det er dog også muligt at modellere andelen af helbredte individer og overlevelsen for de ikke-helbredte individer for ikke-mixtur helbredelsesmodellerne. Dernæst introducerede vi identificerbarhed, som betyder, at der ikke findes to parameterestimater, der genererer samme model, og vi gennemgik, hvornår en helbredelsesmodel er identificerbar. Herefter blev modelkontrol introduceret for mixtur helbredelsesmodeller. Cox-Snell-residualer for total overlevelse blev udledt for $\widetilde{S}_u(t \mid \mathbf{z})$ og $S(t \mid \mathbf{z})$.

De parametriske fordelinger, der anvendes til at modellere helbredelsesmodellerne, kan nogle gange være for simple til at opfange den underliggende tendens. Derfor introducerede vi fleksible parametriske modeller, som i stedet anvender splines. Først blev en fleksibel parametrisk overlevelsesmodel præsenteret. Vi udvidede herefter denne til en fleksibel parametrisk helbredelsesmodel ved at inkorporere statistisk helbredelse ved hjælp af en baglæns-spline. Denne model kaldes også for en ARS-model og er et specialtilfælde af en ikke-mixtur helbredelsesmodel. I ARS-modellen bestemmes andelen af helbredte individer i den sidste knude, og derfor er placeringen af denne essentiel for modellen. Herefter beskrev vi FMC-modellen, som

er en mixtur helbredelsesmodel, hvor $S_u(t \mid \mathbf{z})$ modelleres med RKS.

I specialet blev der foretaget løbende dataanalyse. I Afsnit 3.2 opstillede vi tre mixtur helbredelsesmodeller for den relative overlevelse uden kovariater. Overlevelsesfunktionen $S_u(t)$ blev modelleret med henholdsvis en Weibull-, log-normal- og eksponential-fordeling, og π med en logit link-funktion. Modellerne blev sammenlignet med AIC og Ederer I. Vi stratificerede også modellerne efter aldersgrupperne 18-44, 45-59, 60-74 og 75-90. Det blev konkluderet, at log-normal-modellen var den bedste model, da den fulgte Ederer I bedst og havde den laveste AIC. Dernæst opstillede vi en Weibull-model med kovariaterne aldersgruppe og diagnoseperiode samt en vekselvirkning mellem disse. Her konkluderede vi, at ældre patienter er mindre robuste i forhold til coloncancer, og at sundhedsvæsenet er blevet bedre til at behandle coloncancer i perioden 1975-1994. Dette gør sig specielt gældende for de ældre patienter. I Afsnit 3.4 foretog vi en lignende analyse for ikke-mixtur helbredelsesmodellerne, og resultaterne for disse viste sig at være næsten identiske med mixtur helbredelsesmodellerne.

I Afsnit 4.4 tilpassede vi ARS- og FMC-modellen uden kovariater for fire forskellige knudeplacering-scenarier. Dette blev gjort for at undersøge, hvor sensitiv modellerne er i forhold til antallet af knuder og deres placeringer. Det blev konkluderet, at ARS-modellen er meget sensitiv i forhold til knudeplacering, hvilket hovedsageligt skyldes, at den sidste knude beskriver helbredelsestidspunktet. Til sidst tilpassede vi ARS-modellen, hvor vi stratificerede efter aldersgruppe og diagnoseperiode. Dette blev kun gjort for ARS-modellen, da FMC-modellen resulterede i ustabile estimater. I denne stratificerede analyse konkluderede vi også, at de ældre patienter klarer sig dårligere i forhold til coloncancer, og at sundhedsvæsenet er blevet bedre til at behandle coloncancer i perioden 1975-1994.

Appendiks A | Overlevelsesanalyse teori

I dette kapitel introduceres de grundlæggende begreber inden for overlevelsesanalyse baseret på [Klein and Moeschberger, 2003].

Antag, at der haves n observationer $(X_i, \delta_i, \mathbf{z}_i)$ for $i = 1, \ldots, n$, hvor $X_i = \min(T_i, C_i)$. I denne sammenhæng er T_i tiden til en givet begivenhed, og C_i er censureringstiden. Disse antages at være uafhængige givet kovariater $\mathbf{z}_i$. Der haves $\delta_i = \mathbb{1}[T_i \leq C_i]$, som beskriver om, hvorvidt individ i oplever begivenheden af interesse eller censureres. Overlevelsesfunktionen er defineret ved

$$S(t \mid \mathbf{z}) = P(T > t \mid \mathbf{z}) = 1 - F(t \mid \mathbf{z}) = \int_t^\infty f(u \mid \mathbf{z}) du,$$

hvilket betyder, at tæthedsfunktionen er bestemt ved $f(t \mid \mathbf{z}) = -\frac{dS(t|\mathbf{z})}{dt}$. For en egentlig overlevelsesfunktion haves $S(0 \mid \mathbf{z}) = 1$ og $\lim_{t\to\infty} S(t \mid \mathbf{z}) = 0$. Et ikke-parametrisk estimat af $S(t)$ kan eksempelvis bestemmes ved Kaplan-Meier-estimatet, som er givet ved

$$\widehat{S}(t) = \widehat{P}(T > t) = \prod_{t^* \in D, t^* \leq t} \left[1 - \frac{d(t^*)}{r(t^*)}\right], \tag{A.1}$$

hvor D er mængden af forskellige begivenheder, $d(t^*)$ er antallet af individer, som har oplevet begivenheden af interesse til tidspunktet t^*, og $r(t^*)$ er antallet af individer, som er i risiko til tidspunktet t^*. Variansen af $\widehat{S}(t)$, som er estimeret ved hjælp af delta-metoden i Sætning A.2, er givet som

$$\widehat{\mathbb{V}ar}\{\widehat{S}(t)\} = \widehat{S}(t)^2 \widehat{\mathbb{V}ar}\{\log[\widehat{S}(t)]\} = \widehat{S}(t)^2 \sum_{t^* \leq t} \frac{d(t^*)}{[r(t^*) - d(t^*)]r(t^*)}. \tag{A.2}$$

Denne formel kaldes også for Greenwoods formel.

Hazard-funktionen er defineret ved

$$h(t \mid \mathbf{z}) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t, \mathbf{z})}{\Delta t},$$

og sammenhængen mellem tætheds-, overlevelses- og hazard-funktionen er dermed

$$h(t \mid \mathbf{z}) = \frac{f(t \mid \mathbf{z})}{S(t \mid \mathbf{z})} = -\frac{d \ln [S(t \mid \mathbf{z})]}{dt}.$$

Den kumulerede hazard-funktion er givet ved

$$H(t \mid \mathbf{z}) = \int_0^t h(u \mid \mathbf{z}) du = \int_0^t -\frac{d}{dt} \ln [S(u \mid \mathbf{z})] \, du = -\ln [S(t \mid \mathbf{z})] .$$

Der haves dermed også

$$S(t \mid \mathbf{z}) = \exp[-H(t \mid \mathbf{z})] = \exp \left[- \int_0^t h(u \mid \mathbf{z}) du \right].$$

Likelihoodfunktionen for højre-censurerede data er givet ved

$$L = \prod_{i=1}^{n} [f(x_i \mid \mathbf{z}_i)]^{\delta_i} [S(x_i \mid \mathbf{z}_i)]^{1-\delta_i} = \prod_{i=1}^{n} [h(x_i \mid \mathbf{z}_i)]^{\delta_i} S(x_i \mid \mathbf{z}_i),$$

som giver log-likelihoodfunktionen

$$l = \ln(L) = \sum_{i=1}^{n} \delta_i \ln[h(x_i \mid \mathbf{z}_i)] + \ln[S(x_i \mid \mathbf{z}_i)], \tag{A.3}$$

En vigtig model inden for overlevelsesanalyse til at beskrive effekten af kovariater er Cox proportional hazard-modellen (CPH-modellen),

$$h(t \mid \mathbf{z}) = h_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{z}), \tag{A.4}$$

hvor $h_0(t)$ er en vilkårlig positiv reference hazard-funktion. Den tilhørende kumulerede hazard-funktion til hazard-funktionen i Ligning (A.4) er givet ved

$$H(t \mid \mathbf{z}) = H_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{z}), \tag{A.5}$$

hvor $H_0(t) = \int_0^\infty h_0(u \mid \mathbf{z}) du$ er den kumulerede reference hazard-funktion. I CPH-modellen kan kovariaterne fortolkes ved at se på en enkelt kovariat af gangen og holde de resterende kovariater konstant. Lad z_1 være en kategorisk variabel for køn, med $z_1 = 0$ hvis kvinde, og $z_1 = 1$ hvis mand. Lad de resterende kovariater være konstant. Hazard-ratioen mellem mænd og kvinder er dermed givet ved

$$\frac{h(t \mid z_1 = 1)}{h(t \mid z_1 = 0)} = \frac{h_0(t) \exp(\beta_1)}{h_0(t) \exp(0)} = \exp(\beta_1).$$

Det vil sige, at hazard-funktionen vokser eller aftager med β_1 for mænd.

A.1 Teoretiske resultater

Proposition A.1.
Lad T være en kontinuert stokastisk variabel med overlevelsesfunktion $S(T)$ og kumuleret hazard-funktion $H(T)$. Dermed gælder, at

$$S(T) \sim \text{Unif}(]0;1[), \quad H(T) \sim \text{Exp}(1).$$

Bevis
Denne proposition kan bevises i to dele:

1) Hvis T er en stokastisk variabel med fordelingsfunktion $F(t)$, er $U = F(T) \sim \text{Unif}([0;1])$. Hvis $U = F(T) \sim \text{Unif}([0;1])$, er $S(T) = 1 - F(T)$ også uniform fordelt.

2) Hvis $U \sim \text{Unif}(]0;1])$, er $Y = -\log(U) \sim \text{Exp}(1)$.

ad 1): Det skal vises, at U er uniform fordelt

$$\begin{aligned} F_U(u) &= P(U \leq u) \\ &= P(F(T) \leq u) \\ &= P(T \leq F^{-1}(u)) \\ &= F(F^{-1}(u)) = u. \end{aligned}$$

ad 2): Nu skal det vises, at $Y \sim \text{Exp}(1)$

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(-\log[U] \leq y) \\ &= P(U \geq \exp[-y]) \\ &= 1 - F_U(\exp[-y]) \\ &= 1 - \exp(-y), \end{aligned}$$

hvor det sidste lighedstegn gælder, da U er uniform fordelt. Nu differentieres $F_Y(y) = 1 - \exp(-y)$ med hensyn til y på begge sider af lighedstegnet, og der haves

$$f_Y(y) = \exp(-y),$$

som er tæthedsfunktionen for Y. Det vil sige, at $Y = -\log(U) \sim \text{Exp}(1)$, og da $H(T) = -\log[S(T)]$, haves $H(T) \sim \text{Exp}(1)$. ■

I den næste sætning introduceres delta-metoden, [Kulperger, 2017].

Sætning A.2 (Delta-metoden)**.**
Antag, at Y har en asymptotisk normal-fordeling, således at

$$\sqrt{n}(Y - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Antag, at g er en kontinuert funktion, og at $g'(\mu) \neq 0$. Det gælder dermed, at

$$\sqrt{n}(g(Y) - g(\mu)) \xrightarrow{D} N(0, g'(\mu)^2\sigma^2).$$

A.2 Fordelinger

Overlevelsesfunktionen, hazard-funktionen og parametrene for Weibull-, log-normal- og eksponential-fordelingen er givet i tabellen, der følger.

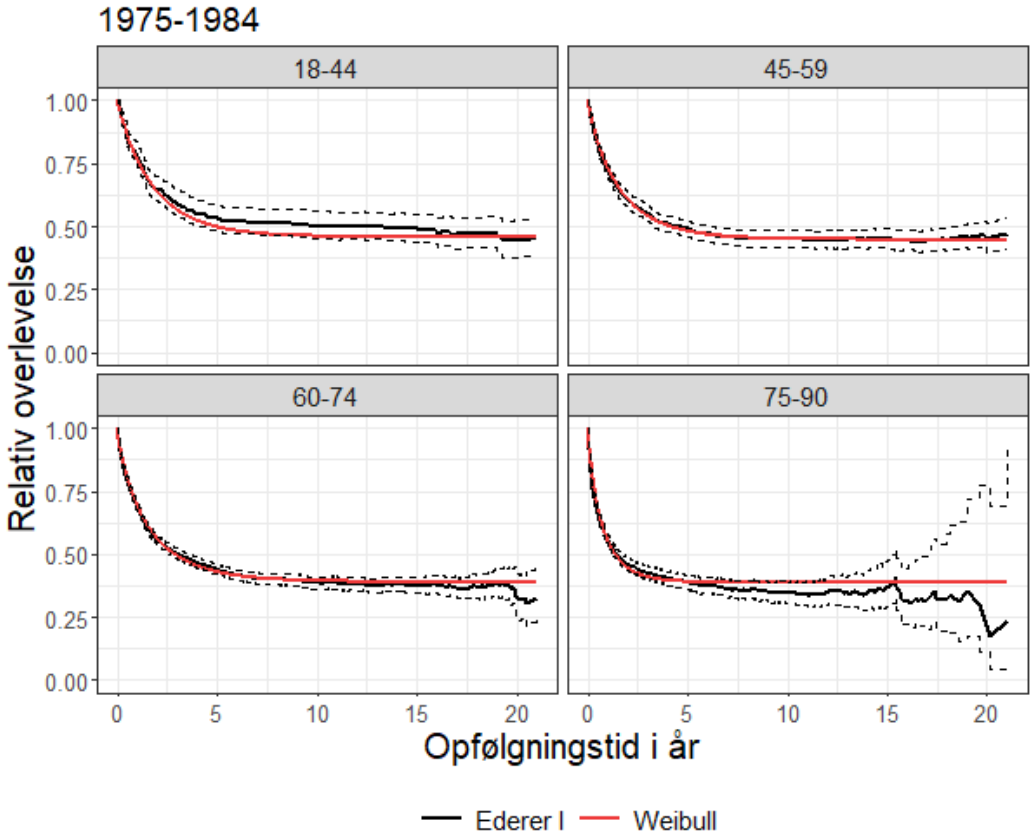
Fordeling	Overlevelsesfunktion	Hazard-funktion	Parametre
Weibull	$S(t) = \exp(-\gamma_1 t^{\gamma_2})$	$h(t) = \gamma_1\gamma_2 t^{\gamma_2-1}$	$\gamma_1, \gamma_2 > 0$
Log-normal	$S(t) = 1 - \Phi\left(\frac{\ln(t)-\gamma_1}{\gamma_2}\right)$	$h(t) = \frac{\exp\left(-\frac{1}{2}\left(\frac{\ln(t)-\gamma_1}{\gamma_2}\right)^2\right)}{t(2\pi)^{\frac{1}{2}}\gamma_2\left(1-\Phi\left(\frac{\ln(t)-\gamma_1}{\gamma_2}\right)\right)}$	$\gamma_1, \gamma_2 > 0$
Eksponential	$S(t) = \exp(-\gamma_1 t)$	$h(t) = \gamma_1$	$\gamma_1 > 0$

Tabel A.1: Overlevelsesfunktionen, hazard-funktionen og parametrene for Weibull-, log-normal- og eksponential-fordelingen. Notationen $\Phi(\cdot)$ i log-normal-fordelingen beskriver en standard normal-fordeling.

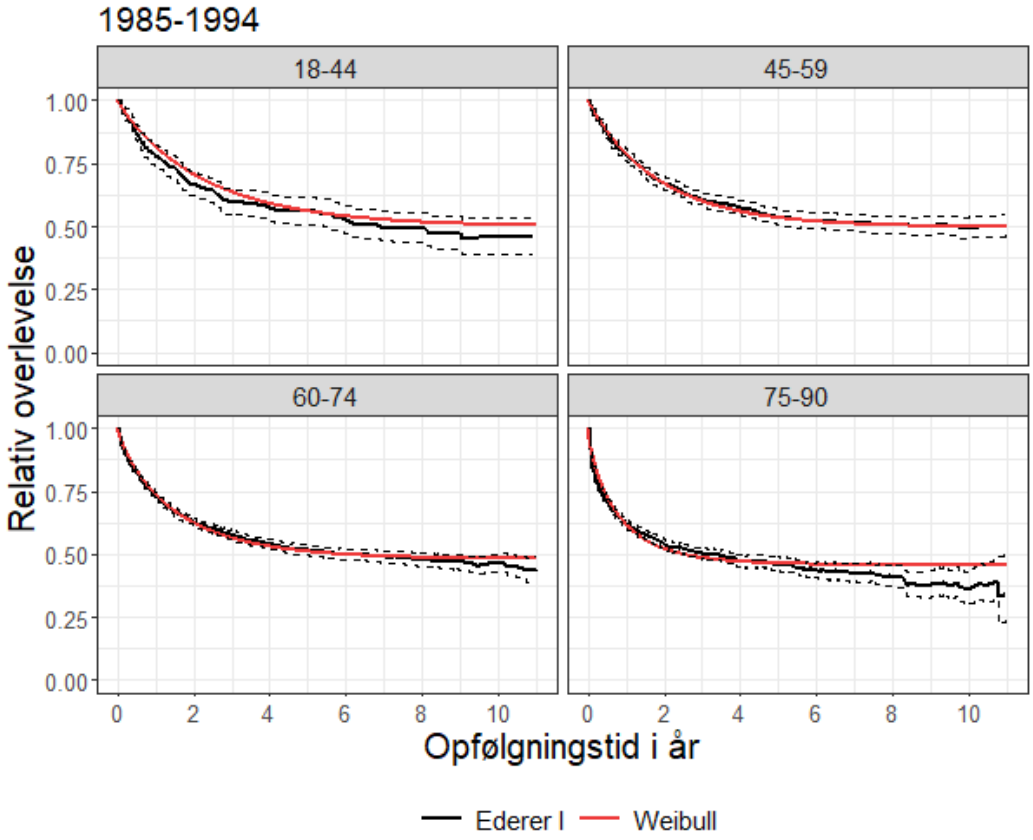
Når $\gamma_2 > 0$, er hazard-funktionen monotont voksende, men når $\gamma_2 < 1$, er hazard-funktionen monotont aftagende. For $\gamma_2 = 1$ reduceres Weibull-fordelingen til eksponential-fordelingen med en konstant hazard-funktion.

Appendiks B | Tabeller og figurer

B.1 Weibull-model med kovariater



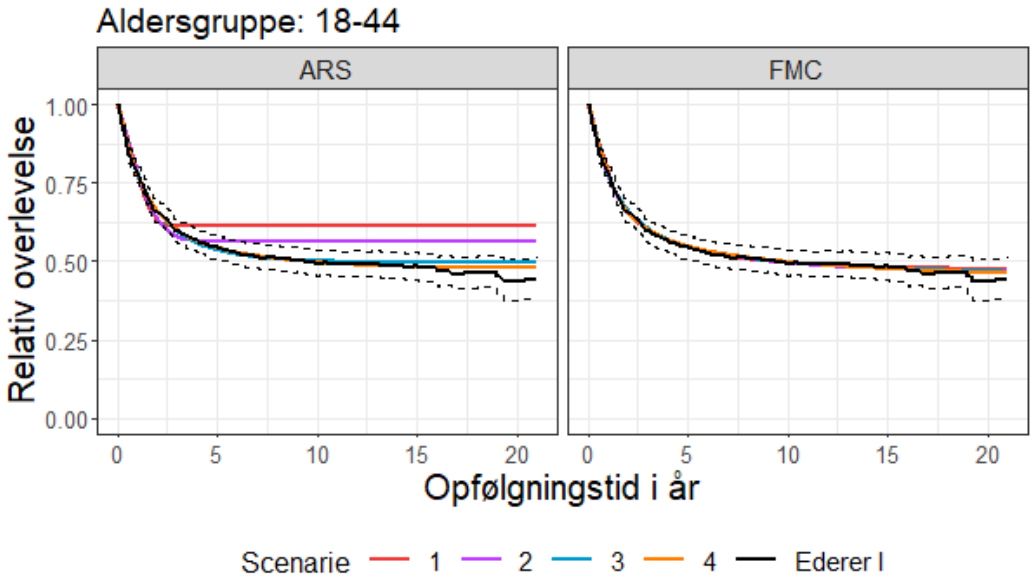
Figur B.1: Den relative overlevelse udregnet ved Ederer I med tilhørende 95% konfidensinterval (stiplede linjer) samt Weibull-modellen for diagnoseperioden 1975-1984.



Figur B.2: Den relative overlevelse udregnet ved Ederer I med tilhørende 95% konfidensinterval (stiplede linjer) samt Weibull-modellen for diagnoseperioden 1985-1994.

B.2 Aldersgruppe analyse for ARS- og FMC-modellen

I dette afsnit illustreres knudeplacering-scenarierne i Tabel 4.1 for aldersgrupperne 18-44, 45-59, 60-74 og 75-90.

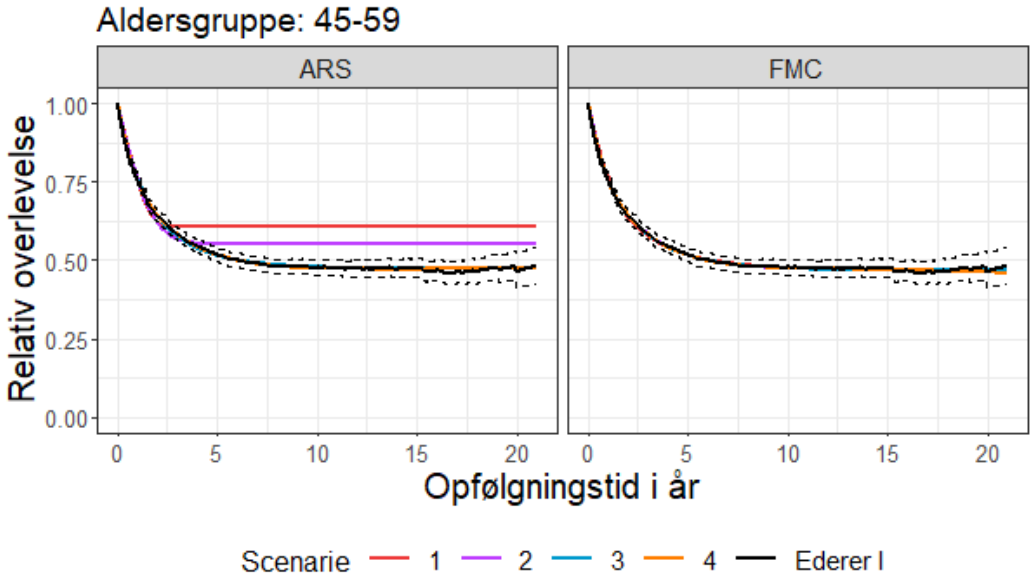


Figur B.3: Den relative overlevelse for coloncancer-datasættet udregnet ved Ederer I med tilhørende 95% konfidensinterval (stiplede linjer), samt ARS- og FMC-modellen under forskellige knudeplacering-scenarier for aldersgruppen 18-44.

Knuderne for 0%, 20%, 40%, 60%, 80% og 100% kvartilerne af de ikke-censurerede begivenhedstider svarer til dag 15, 137, 301.8, 564.6, 1080.8 og 7045.

Model	Scenarie 1	Scenarie 2	Scenarie 3	Scenarie 4
ARS	2505.13	2295.75	2171.32	2171.41
FMC	2173.22	2161.89	2163.9	2156.41

Tabel B.1: AIC for ARS- og FMC-modellen under de forskellige scenarier for aldersgruppen 18-44.

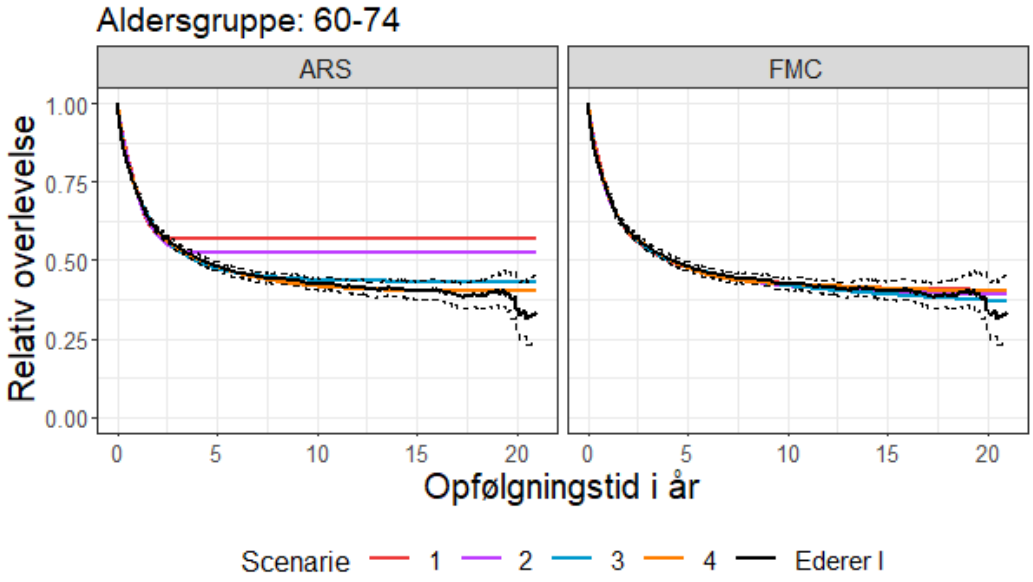


Figur B.4: Den relative overlevelse for coloncancer-datasættet udregnet ved Ederer I med tilhørende 95% konfidensinterval (stiplede linjer), samt ARS- og FMC-modellen under forskellige knudeplacering-scenarier for aldersgruppen 45-59.

Knuderne for 0%, 20%, 40%, 60%, 80% og 100% kvartilerne af de ikke-censurerede begivenhedstider svarer til dag 15, 107, 289, 564, 1231 og 7289.

Model	Scenarie 1	Scenarie 2	Scenarie 3	Scenarie 4
ARS	7721.49	7318.7	7049.8	7050.11
FMC	7064.66	7041.83	7036.81	7036.67

Tabel B.2: AIC for ARS- og FMC-modellen under de forskellige scenarier for aldersgruppen 45-59.

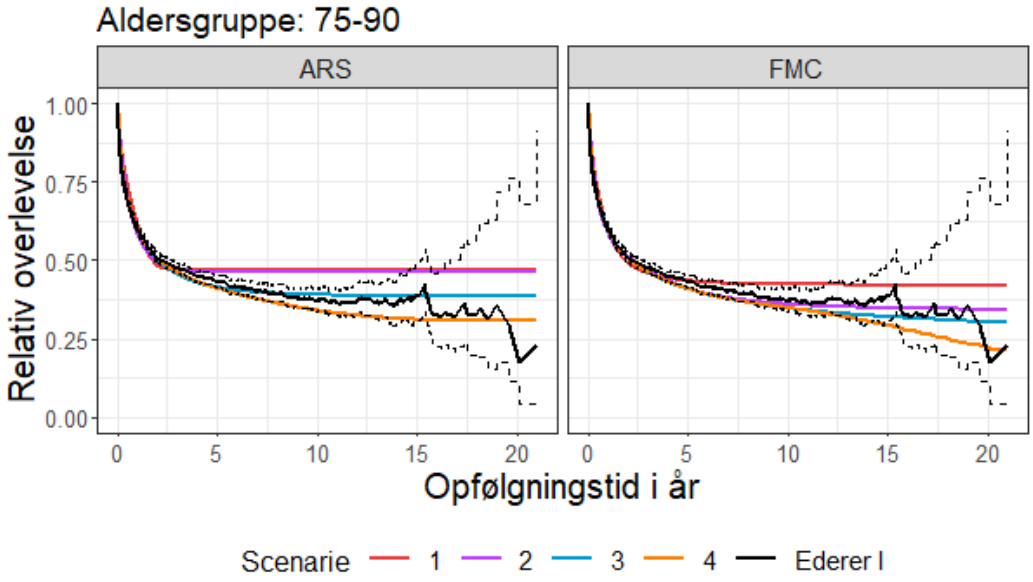


Figur B.5: Den relative overlevelse for coloncancer-datasættet udregnet ved Ederer I med tilhørende 95% konfidensinterval (stiplede linjer), samt ARS- og FMC-modellen under forskellige knudeplacering-scenarier for aldersgruppen 60-74.

Knuderne for 0%, 20%, 40%, 60%, 80% og 100% kvartilerne af de ikke-censurerede begivenhedstider svarer til dag 15, 106, 318, 626, 1568 og 7470.

Model	Scenarie 1	Scenarie 2	Scenarie 3	Scenarie 4
ARS	21800.3	21264.44	20839.74	20816.84
FMC	20918.76	20853.36	20805.13	20801.62

Tabel B.3: AIC for ARS- og FMC-modellen under de forskellige scenarier for aldersgruppen 60-74.



Figur B.6: Den relative overlevelse for coloncancer-datasættet udregnet ved Ederer I med tilhørende 95% konfidensinterval (stiplede linjer), samt ARS- og FMC-modellen under forskellige knudeplacering-scenarier for aldersgruppen 75-90.

Knuderne for 0%, 20%, 40%, 60%, 80% og 100% kvartilerne af de ikke-censurerede begivenhedstider svarer til dag 15, 46, 168, 442, 1262.8 og 7351.

Model	Scenarie 1	Scenarie 2	Scenarie 3	Scenarie 4
ARS	15116.22	14720.03	14425.49	14408.34
FMC	14752.12	14534.36	14393.81	14397.08

Tabel B.4: AIC for ARS- og FMC-modellen under de forskellige scenarier for aldersgruppen 75-90.

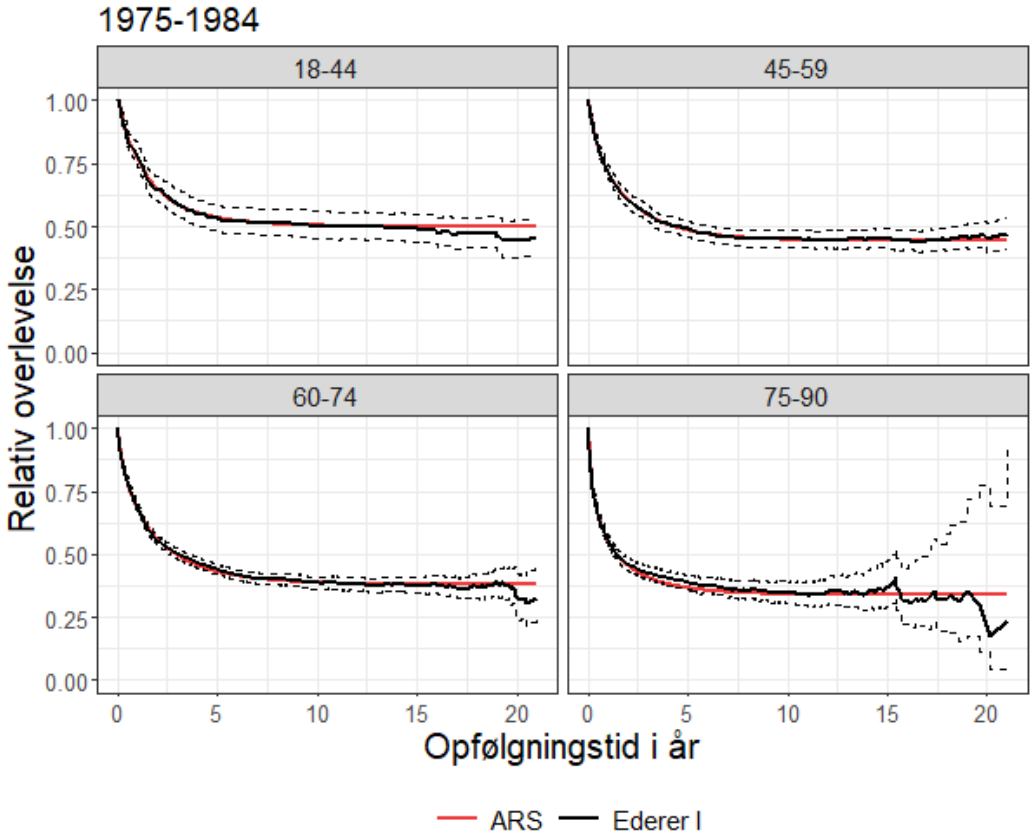
B.3 Analyse af aldersgruppe og diagnoseperiode for ARS-modellen

Knuderne for 0%, 20%, 40%, 60%, 80% og 100% kvartilerne af de ikke-censurerede begivenhedstider for de forskellige undergrupper i Underafsnit 4.4.2 er angivet i dage i tabellen, der følger.

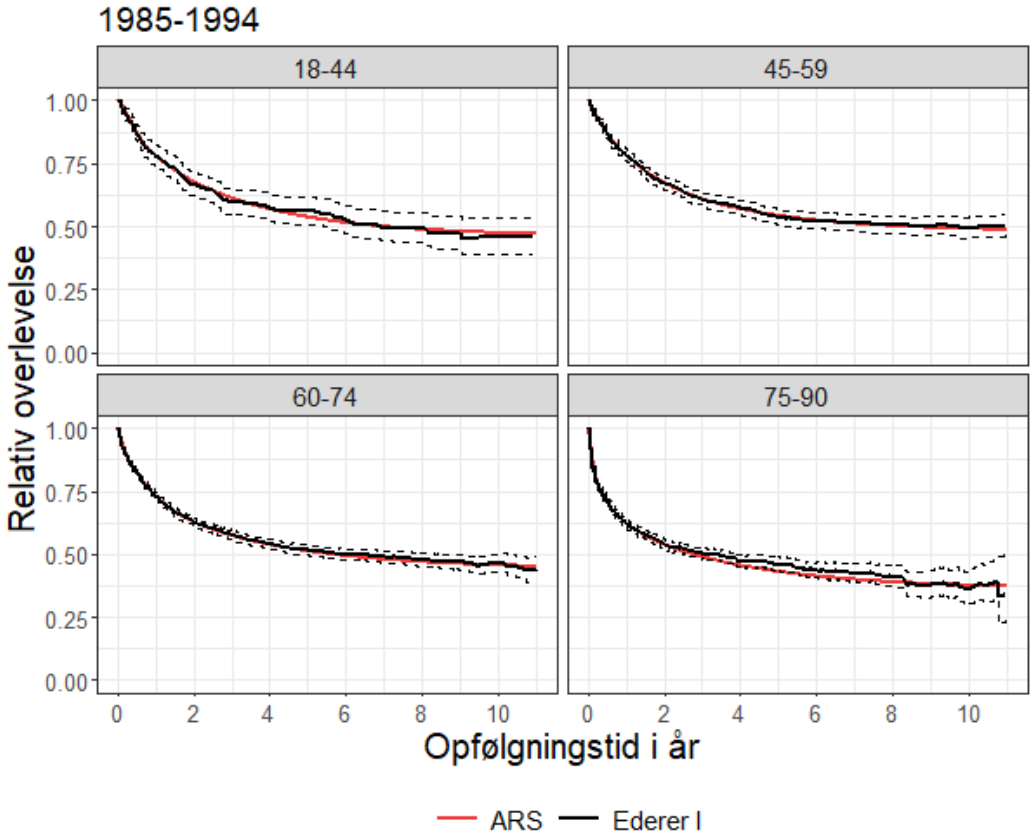
	0%	20%	40%	60%	80%	100%
1975-1984						
18-44	15	137.2	349	594.6	1263.6	7045
45-59	15	107	290	672.2	1568	7289
60-74	15	107	351	867	2321.4	7470
75-90	15	74	198	564	1841	7351
1985-1994						
18-44	15	137	276.8	563.4	988.2	3302
45-59	15	135	288	532	958	3575
60-74	15	77	257	473	1021	3909
75-90	15	46	138	380	929	3942

Tabel B.5: Knuderne for 0%, 20%, 40%, 60%, 80% og 100% kvartilerne af de ikke-censurerede begivenhedstider for de forskellige stratum i dage.

Figurerne, der følger, illustrerer ARS-modellen sammen med Ederer I for aldersgrupperne 18-44, 45-59, 60-74 og 75-90 for diagnoseperioderne 1975-1984 og 1985-1994.



Figur B.7: Den relative overlevelse udregnet ved Ederer I med tilhørende 95% konfidensinterval (stiplede linjer), samt ARS-modellen stratificeret efter aldersgruppe for diagnoseperioden 1975-1984.



Figur B.8: Den relative overlevelse udregnet ved Ederer I med tilhørende 95% konfidensinterval (stiplede linjer), samt ARS-modellen stratificeret efter aldersgruppe for diagnoseperioden 1985-1994.

Appendiks C | R-kode

C.1 Median relativ overlevelsestid for de ikke-helbredte patienter

Mixtur helbredelsesmodel

```
survivaluncured = function(fit, uncured=0.5){
  f = function(time) predict(fit, time=time,
  type="survuncured")[[1]]$Estimate-uncured
  Estimate = round(uniroot(f, lower=0.001, upper=20)$root,2)

  paste0(Estimate)
}
```

Ikke-mixtur helbredelsesmodel

```
SuNonMixture = function(fit, uncured=0.5){
  cure.pred_E = predict(fit, type="curerate")[[1]]$Estimate
  f = function(time) (predict(fit, time=time)[[1]]$Estimate -
      cure.pred_E) / (1 - cure.pred_E)-uncured
  Estimate = round(uniroot(f, lower=0.001, upper=20)$root,2)

  paste0(Estimate)
}
```

ARS-model

```
SU_ARS = function(fit, uncured=0.5){
  cure.pred = predict(fit, newdata = data.frame(FUyear = 25))
  f = function(time) (predict(fit, newdata=data.frame(FUyear
     = time)) - cure.pred) / (1 - cure.pred)-uncured
  Estimate = round(uniroot(f, lower=0.001, upper=20)$root,2)

  paste0(Estimate)
}
```

FMC-model

```
SU_FMC = function(fit, uncured=0.5){
  f = function(time) predict(fit, time=time, type="
     survuncured")[[1]]$Estimate-uncured
  Estimate = round(uniroot(f, lower=0.001, upper=20)$root,2)

  paste0(Estimate)
}
```

Litteratur

Therese M.-L. Andersson. Quantifying cancer patient survival: extensions and applications of cure models and life expectancy estimation, 2013. URL https://www.semanticscholar.org/paper/Quantifying-cancer-patient-survival-%3A-extensions-of-Andersson/e1270219a5d1e09a045fac9f00d8df9c0c994cdb.

Therese M.-L. Andersson, Paul C. Lambert, Åsa Rangert Derolf, Sigurdur Yngvi Kristinsson, Sandra Eloranta, Ola Landgren, Magnus Björkholm, and Paul W. Dickman. Temporal trends in the proportion cured among adults diagnosed with acute myeloid leukaemia in Sweden 1973–2001, a population-based study. *British Journal of Haematology*, 148(6):918–924, 2010. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2141.2009.08026.x.

Therese M.-L. Andersson, Paul C. Lambert, Paul W. Dickman, and Sandra Eloranta. Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. *BMC Medical Research Methodology*, 11(96), 2011. URL https://doi.org/10.1186/1471-2288-11-96.

Colin B. Begg and Deborah Schrag. Attribution of Deaths Following Cancer Treatment. *JNCI: Journal of the National Cancer Institute*, 94(14):1044–1045, 07 2002. URL https://doi.org/10.1093/jnci/94.14.1044.

John W. Boag. Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy. *Journal of the Royal Statistical Society: Series B (Methodological)*, 11(1):15–53, 1949. URL www.jstor.org/stable/2983694.

Vincent Bremhorst and Philippe Lambert. Flexible estimation in cure survival models using Bayesian P-splines. *Computational Statistics & Data Analysis*, 93: 270 – 284, 2016. ISSN 0167-9473. URL http://www.sciencedirect.com/science/article/pii/S0167947314001492.

Cancerregisteret. Cancerregisterets årsrapporter (2018), 2019. URL https://sundhedsdatastyrelsen.dk/da/tal-og-analyser/

analyser-og-rapporter/sygdomme/kraeft-_-cancerregisteret. Accessed on 06-04-2020.

Mark Clements and Xing-Rong Liu. *rstpm2: Smooth Survival Models, Including Generalized Survival Models*, 2019. URL https://CRAN.R-project.org/package=rstpm2. R package version 1.5.1.

R. De Angelis, R. Capocaccia, T. Hakulinen, B. Soderman, and A. Verdecchia. Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in Medicine*, 18(4):441–454, 1999.

F. Ederer and H. Heise. Instructions to IBM 650 programmers in processing survival computations. *National Cancer Institute*, 1959. Methodological note No. 10, End Results Evaluation Section.

F. Ederer, L. M. Axtell, and S. J. Cutler. The relative survival rate: a statistical methodology. *National Cancer Institute*, 6:101—121, September 1961. ISSN 0083-1921. URL http://europepmc.org/abstract/MED/13889176.

J. Estève, E. Benhamou, M. Croasdale, and L. Raymond. Relative survival and the estimation of net survival: Elements for further discussion. *Statistics in Medicine*, 9(5):529–538, 1990. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780090506.

Timo Hakulinen. Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics*, 38(4):933–42, 1982.

Leonid Hanin and Li-Shan Huang. Identifiability of cure models revisited. *Journal of Multivariate Analysis*, 130:261 – 274, 2014. ISSN 0047-259X. URL http://www.sciencedirect.com/science/article/pii/S0047259X14001328.

Lasse H Jakobsen, Therese M-L Andersson, Jorne L Biccler, Tarec C El-Galaly, and Martin Bøgsted. Estimating the loss of lifetime function using flexible parametric relative survival models. *B M C Medical Research Methodology*, 19 (1):1–13, January 2019. ISSN 1471-2288.

Lasse Hjort Jakobsen. *cuRe: Parametric Cure Model Estimation*, 2020. URL http://github.com/LasseHjort/cuRe. R package version 1.0.1.

John P. Klein and Melvin L. Moeschberger. *Survival Analysis Techniques for Censored and Truncated Data*. Springer, second edition, 2003.

Reg Kulperger. Delta Method, 2017. URL `http://fisher.stats.uwo.ca/faculty/kulperger/SS3858/Handouts/DeltaMethod.pdf`. Accessed on 01-06-2020.

P. C. Lambert, P. W. Dickman, C. L. Weston, and J. R. Thompson. Estimating the cure fraction in population-based cancer studies by using finite mixture models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59 (1):35–55, 2010. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2009.00677.x`.

Paul C. Lambert, Paul W. Dickman, Pia Österlund, Therese Andersson, Risto Sankila, and Bengt Glimelius. Temporal trends in the proportion cured for cancer of the colon and rectum: A population-based study using data from the Finnish Cancer Registry. *International Journal of Cancer*, 121(9):2052–2059, 2007a. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.22948`.

Paul C. Lambert, John R. Thompson, Claire L. Weston, and Paul W. Dickman. Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, 8(3):576–594, 10 2007b. ISSN 1465-4644. URL `https://doi.org/10.1093/biostatistics/kxl030`.

Christopher P. Nelson, Paul C. Lambert, Iain B. Squire, and David R. Jones. Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine*, 26(30):5486–5498, 2007. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3064`.

Megan Othus, Bart Barlogie, Michael L. LeBlanc, and John J. Crowley. Cure Models as a Useful Statistical Tool for Analyzing Survival. *Clinical Cancer Research*, 18(14):3731–3736, 2012.

Maja Perme and Klemen Pavlic. Nonparametric Relative Survival Analysis with the R Package relsurv. *Journal of Statistical Software, Articles*, 87(8):1–27, 2018. ISSN 1548-7660. URL `https://www.jstatsoft.org/v087/i08`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org.

Hannah Ritchie and Max Roser. Causes of Death. *Our World in Data*, 2020. URL https://ourworldindata.org/causes-of-death.

Patrick Royston and Mahesh K. B. Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in medicine*, 21:2175–2197, 2002.

Sylvie Scolas, Catherine Legrand, Abderrahim Oulhaj, and Anouar El Ghouch. Diagnostic checks in mixture cure models with interval-censoring. *Statistical Methods in Medical Research*, 27(7):2114–2131, 2018. URL https://doi.org/10.1177/0962280216676502. PMID: 27815495.

Richard Sposto. Cure model analysis in cancer: an application to data from the Children's Cancer Group. *Statistics in Medicine*, 21(2):293–312, 2002. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.987.

A. Surbone, M. A. Annunziata, A. Santoro, U. Tirelli, and P. Tralongo. Cancer patients and survivors: Changing words or changing culture? *Annals of Oncology*, 24(10):2468–2471, 2013. ISSN 0923-7534.

The Human Mortality Database. Department of Demography at the University of California in Berkeley (USA), and Max Planck Institute for Demographic Research in Rostock (Germany), 2002. URL https://www.mortality.org/.

Terry M. Therneau and Jan Offord. Expected Survival Based on Hazard Rates (Update). *Technical report 63*, 1999. Mayo Clinic Department of Health Science Research.

Paolo Tralongo, Mary S. McCabe, and Antonella Surbone. Challenge For Cancer Survivorship: Improving Care Through Categorization by Risk. *Journal of Clinical Oncology*, 35(30):3516–3517, 2017. URL https://doi.org/10.1200/JCO.2017.74.3450. PMID: 28834437.

A. Tsodikov. Semi-parametric models of long- and short-term survival: an application to the analysis of breast cancer survival in Utah by age and stage. *Statistics in Medicine*, 21(6):895–920, 2002. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1054`.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

A.Y. Yakovlev, Bernard Asselain, V.J. Bardou, A. Fourquet, Thu Hoang, A. Rochefediere, and A.D. Tsodikov. A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. *Biometrie et Analyse de Donnees Spatio-Temporelles*, 12:66–82, 01 1993.