Interval-censored semi-competing risks data: a novel approach for modelling bladder cancer.

Núria Porta Bleda

Thesis directed by

Dr. Guadalupe Gómez Melis
Universitat Politècnica de Catalunya

Dr. M.Luz Calle Rosingana
Universitat de Vic

Universitat Politècnica de Catalunya

Barcelona, June 2010

Tesi presentada per obtenir el títol de Doctor per la Universitat Politècnica de Catalunya.

Amb el suport del Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya i del Fons Social Europeu.

Als meus pares i a l'Enric

Acknowledgements

Avui, abans d'escriure aquestes línees, m'he trobat la universitat plena d'alumnes d'institut que venien a fer la selectivitat. Avui, que he vingut a dipositar la tesi, he pensat que era bonic que tot just avui que ells comencen la seva aventura universitària, tot just acabi jo la meva.

A pesar de què el camí ha estat llarg i no sempre fàcil, tinc el privilegi d'haver fet el què m'agrada: descobrir, aprendre i aprofundir, i no necessàriament en aquest ordre. Si hi ha un lema en la vida de l'investigador, aquest hauria de ser 'No deixaré mai d'aprendre'.

Vull agrair enormement a les meves directores Lupe Gómez i Malu Calle l'intensa feina dels darrers anys. He tingut una sort immensa perquè m'heu recolzat i animat, i m'heu ajudat a adquirir un bagatge que va més enllà d'aquestes pàgines. Gràcies per les oportunitats però sobretot per la relació de sincer afecte que hem establert. I sé que la Lupe em perdonarà si menciono especialment a la Malu: em sento molt orgullosa de ser la teva primera doctoranda. Gràcies per tot.

Mil gràcies a la Dra. Núria Malats per permetre'm treballar amb les dades de l'Estudi Espanyol de Càncer de Bufeta (EPICURO), i per involucrar-me en aquest gran projecte: espero haver aportat el meu granet de sorra en el coneixement d'aquesta malaltia. Gràcies a la Cristiane Murtra i als companys amb qui he col·laborat en aquest projecte.

Agrair també la oportunitat de la Generalitat de Cataluya per la beca FI, sense la qual segurament aquest treball no hagués estat possible.

Al Dr. Josep Maria Arnau, responsable de la UCICEC-CAIBER de l'IDIBELL li agraeixo especialment la seva flexibilitat i recolzament durant els darrers mesos.

Thanks to the Department of Biostatistics in Harvard School of Public Health for the opportunity to visit them. I am specially grateful to Prof. Marvin Zelen and Prof. Steve Lagakos for their kind help. Thanks to Prof. Geert Molenberghs for my 3-month stage in Hasselt University, during which I started my research career.

Gràcies als companys del GRASS, que sempre m'han ofert el seu temps per escoltar-me i la seva valuosa opinió per millorar la meva feina. Gràcies a la Raquel i la Núria: sempre ens quedarà

Rotterdam i la final de Champions!

Gràcies als companys del departament EIO, la meva segona casa durant els darrers 4 anys, especialment al Toni per la seva paciència i als companys de doctorat amb qui he compartit penes, bons moments i temazos. Gràcies Laura i Mari Paz!

To the external referees, I greatly appreciate the time and effort you invested in the review of my work.

Sou molts els què heu estat al meu costat, els què us heu interessat per la meva feina, i els que heu fet aquesta tesi una mica vostra només per l'estima que em teniu. A tots, gràcies. A mis niñas de mates, i les meves amigues del cole, i molt especialment a la Bea: sin duda vamos a recordar este 2010, te quiero, rosío. A Alejandra y Fabi, por su apoyo incondicional. Gràcies Cris perquè sense tu tot se m'hagués fet molt més difícil. Gràcies per no deixar-me caure, i estimar-me molt.

Finalment, però no menys important, gràcies a la meva família: sense el seu recolzament i sacrifici no hagués pogut fer el què més m'agrada. Gracias Carlos, Marta, Gemma y Elena, por hacer siempre de hermanos mayores de esta peque que ha salido un poco empollona.

Gràcies Enric per compartir aquest camí i tots els què vindran, amb mi. T'estimo molt.

Gràcies als meus pares pel seu recolzament, per saber que si queia, sempre éreu allí per ajudar-me a aixecar. Gràcies pel vostre sacrifici, i perquè, malgrat jo poso l'esforç i les ganes d'aprendre, el què em fa evolucionar és la pasta de què estic feta i els valors que vosaltres m'heu transmès. Aquesta tesi és per vosaltres.

Abstract

This PhD thesis is concerned with survival analysis in the presence of multiple endpoints and complex censoring patterns. In this context, we propose a new methodology for interval-censored semi-competing risks data. This work is motivated by the Spanish Bladder Cancer/EPICURO Study, the most important study on bladder cancer ever done in Spain. Our contribution in this study was focused on modelling the course of the disease and on identifying prognostic factors of its evolution.

The course of complex diseases such as cancer or HIV-infection is characterized by the occurrence of multiple events on the same patient, for instance, the relapse of the disease, or death. These events can be terminating events, if the follow-up of the individual is stopped by their occurrence, or intermediate events, if the individual continues under observation after their occurrence. The presence of terminating events complicates the analysis since their occurrence prevents from observing other events later, inducing a possibly dependent censoring.

Appropriate methods are required in this context and, specifically, in this thesis we will focus on competing risks, multi-state models and semi-competing risks. These methodologies will be useful to describe important aspects of bladder cancer course. In particular, two novel contributions to the understanding of bladder cancer result from an appropriate use of competing risks and multi-state models: (1) the characterization of those patients with a high risk of progressing as a first event after diagnosis, and (2) the proposal of a dynamical prognostic model for progression.

Competing risks arises when we model the time until the first of K possible events, together with the indicator of the type of event observed. In the Spanish Bladder Cancer/EPICURO Study we are interested in the time until the first event is observed, distinguishing between recurrence, progression or death. The characterization of this first event is of paramount clinical importance in order to better target the adequate treatment for each patient.

Multi-state modelling is handled by describing all possible paths that the course of disease could follow, and establishing relationships between the events of interest. In the Spanish Bladder Cancer/EPICURO Study, for instance, a patient after is diagnosed could experience a recurrence, and

then die, or he/she could die in remission (before any disease-related event is observed). One interesting feature of multi-state models is the possibility to make updated predictions according to the occurrence of intermediate events along time. For the bladder cancer course, we will be able to assess the influence that the occurrence of a recurrence has on the posterior risk of progression.

A special kind of multi-state model is one with an intermediate event, $\mathcal{E}_1$, and a terminating event, $\mathcal{E}_2$. Denote by T_1 and T_2 their corresponding time-to-event endpoints. The study of the marginal law for T_1 is not addressed neither by the competing risks approach nor by the multi-state modelling. While the competing risks approach allows us to analyse the time T to the first between $\mathcal{E}_1$ and $\mathcal{E}_2$, that is, $T = \min(T_1, T_2)$, multi-state modelling focus on the conditional law of $T_2|T_1$, that is, in how the occurrence of $\mathcal{E}_1$ modifies the risk of $\mathcal{E}_2$. The distribution of T_1 is unidentifiable based only on observed data. The above situation is known as semi-competing risks data (Fine et al. 2001), where the occurrence of the terminating event prevents the observation of the intermediate event, and thus T_2 dependently censors T_1. The strategy of Fine and colleagues to solve this problem is to assume a joint model for (T_1, T_2), and then recover the distribution for T_1 derived from the assumed joint model.

Our contribution is focused towards the development of new methods in this area of survival analysis. Specifically, we propose a new methodology to deal with interval-censored semi-competing risks data, which arises when the time to the intermediate event, T_1, is interval-censored. In many longitudinal studies the occurrence of the intermediate event is evaluated at periodic visits, so T_1 is only known to lie between the times of two specific visits. Methods for right-censored semi-competing risks data are no longer valid in this scenario and a new approach is necessary. We extend the semi-parametric method proposed by Fine et al. (2001), which assumes a Clayton's copula model (1978) to describe the association between T_1 and T_2. Our methodology consists of an iterative estimation algorithm which jointly estimates the association structure of the model and the distribution of the intermediate event.

Resumen

La presente tesis trata sobre técnicas de análisis de supervivencia en situaciones con múltiples eventos y patrones complejos de censura. Proponemos una nueva metodología para tratar el problema de riesgos semi-competitivos cuando los datos están censurados en un intervalo. La motivación de este trabajo nace de nuestra colaboración con el estudio Español de Cáncer de Vejiga (SBC/EPICURO), el más grande estudio sobre cáncer de vejiga realizado en España hasta el momento. Nuestra participación en el mismo se centra en la modelización e identificación de factores pronósticos en el curso de la enfermedad.

El curso de enfermedades complejas tales como el cáncer o la infección por VIH, se caracteriza por la ocurrencia de múltiples eventos en el mismo paciente, como por ejemplo la recaída o la muerte. Estos eventos pueden ser finales, cuando el seguimiento del paciente termina con el evento, o bien intermedios, cuando el individuo sigue bajo observación. La presencia de eventos finales complica el análisis de los eventos intermedios, ya que impiden su completa observación, induciendo una posible censura dependiente.

En este contexto, se requieren metodologías apropiadas. Se utilizan los siguientes métodos: riesgos competitivos, modelos multi-estado y riesgos semi-competitivos. De la aplicación de métodos para riesgos competitivos y modelos multi-estado resultan dos aportaciones relevantes sobre el conocimiento de la enfermedad: (1) la caracterización de los pacientes con un alto riesgo de progresión como primer evento después del diagnóstico, y (2) la construcción de un modelo pronóstico y dinámico para el riesgo de progresión.

El problema de riesgos competitivos aparece cuando queremos describir el tiempo hasta el primero de K posibles eventos, junto con un indicador del tipo de evento observado. En el estudio SBC/EPICURO es relevante estudiar el tiempo hasta el primero entre recidiva, progresión o muerte. La caracterización de este primer evento permitiría seleccionar el tratamiento más adecuado de acuerdo con el perfil de riesgo basal del paciente.

Los modelos multi-estado describen las diferentes tipologías que el curso de la enfermedad puede seguir, estableciendo relaciones entre los eventos de interés. Por ejemplo, un paciente puede ex-

perimentar una recidiva y después morir, o bien puede morir sin haber tenido recaída alguna. El potencial interesante de los modelos multi-estado es que permiten realizar predicciones sobre el riesgo de futuros eventos dada la historia del paciente hasta ese momento. En el caso del cáncer de vejiga, podremos evaluar la influencia que tiene en el riesgo de progresar el haber tenido o no una recidiva previa.

Un caso especial de modelo multi-estado es el que contiene un evento intermedio $\mathcal{E}_1$ y uno final, $\mathcal{E}_2$. Sean T_1 y T_2 los tiempos hasta tales eventos, respectivamente. Ni el análisis de riesgos competitivos ni los modelos multi-estado permiten estudiar la distribución marginal de T_1. En efecto, el análisis de riesgos competitivos trata con la distribución del mínimo entre los dos tiempos, $T = \min(T_1, T_2)$, mientras que los modelos multi-estado se centran en la distribución condicional de T_2 dado T_1, $T_2|T_1$, en cómo la ocurrencia de E1 modifica el riesgo de E2. En ambos casos, la distribución de T1 no es identificable a partir de los datos observados. La situación anteriormente descrita donde un evento final impide la observación de un evento intermedio se conoce como riesgos semi-competitivos (Fine *et al.*, 2001). La estrategia de estos autores asume un modelo para la distribución conjunta (T_1, T_2) para así recuperar la distribución de T_1 derivada de ese modelo.

Proponemos una nueva metodología para tratar con riesgos semi-competitivos cuando el tiempo hasta el evento intermedio, T_1, esta censurado en un intervalo. En muchos estudios médicos longitudinales, la ocurrencia del evento de interés se evalúa en visitas periódicas al paciente, por lo que T_1 es desconocido, aunque se conoce que pertenece al intervalo comprendido entre los tiempos de dos visitas consecutivas. Los métodos para riesgos semi-competitivos en el contexto usual de censura por la derecha no son válidos en este caso y se requiere una nueva aproximación. En este trabajo ampliamos la metodología semi-paramétrica propuesta por Fine *et al.* (2001), que asume una cópula de Clayton (1978) (1978) para describir la dependencia entre T_1 y T_2. Bajo el mismo modelo de asociación, desarrollamos un algoritmo iterativo que estima conjuntamente el parámetro de asociación del modelo de cópula, así como la función de supervivencia del tiempo al evento intermedio T_1.

Resum

Aquesta tesi tracta sobre tècniques d'anàlisi de supervivència en situacions amb múltiples esdeveniments i patrons complexes de censura. Proposem una nova metodologia per tractar la situació de riscos semi-competitius quan les dades estan censurades en un interval. La motivació del treball neix de la nostra col·laboració amb l'Estudi Espanyol del Càncer de Bufeta (SBC/EPICURO), el més gran estudi sobre càncer de bufeta realitzat fins ara a l'Estat Espanyol. La nostra contribució en el projecte es centra en la modelització i identificació de factors pronòstics de l'evolució de la malaltia.

L'evolució de malalties complexes, com el càncer o la infecció VIH, es caracteritza per la ocurrència de múltiples esdeveniments en el mateix pacient: per exemple, la recaiguda de la malaltia o la mort. Aquests esdeveniments poden ser finals, quan el seguiment del pacient s'atura després de l'esdeveniment, o bé intermedis, quan l'individu continua sota observació. La presència d'esdeveniments finals complica l'anàlisi dels intermedis ja que n'impedeix la seva completa observació, induint una possible censura depenent.

En aquest context, es requereixen metodologies apropiades. En aquest treball els següents mètodes són emprats: riscos competitius, models multi-estat i riscos semi-competitius. A resultes de l'aplicació de mètodes per riscos competitius i models multi-estat, proposem dues noves aportacions rellevants al coneixement de la malaltia: (1) la caracterització dels pacients amb un alt risc de progressió com a primer esdeveniment després de la diagnosi, i (2) la construcció d'un model pronòstic dinàmic per al risc de progressió.

La situació de riscos competitius apareix quan es vol descriure el temps fins al primer esdeveniment d'entre K possibles, juntament amb un indicador del tipus d'esdeveniment observat. En l'estudi EPICURO, és rellevant estudiar el temps fins al primer esdeveniment, distingint entre una recidiva del tumor, progressió del mateix o mort. La caracterització d'aquest primer esdeveniment permetria seleccionar el millor tractament d'acord amb el perfil de risc basal del pacient.

Els models multi-estat descriuen les diferents evolucions que la malaltia pot seguir, establint relacions entre els esdeveniments d'interès: per exemple, un pacient pot experimentar una recidiva del

tumor primari, i després morir, o bé pot morir en remissió (morir abans de tenir una recaiguda relacionada amb la malaltia). Una característica interessant d'aquests models és que permeten fer prediccions del risc de futurs esdeveniments per a un pacient, d'acord amb la història que hagi pogut tenir fins aquell moment. En el cas de càncer de bufeta podrem avaluar la influència que té en el risc de progressar haver patit o no una recidiva prèvia.

Un cas especial de model multi-estat és aquell que conté un esdeveniment intermedi $\mathcal{E}_1$, i un esdeveniment final, $\mathcal{E}_2$. Siguin T_1 i T_2 els temps fins aquests esdeveniments, respectivament. Ni l'anàlisi de riscos competitius ni els models multi-estat permeten adreçar l'estudi de la distribució marginal de T_1. En efecte, l'anàlisi de riscos competitius tracta amb la distribució del mínim entre els dos temps, $T = \min(T_1, T_2)$, mentre que els models multi-estat es centren en la distribució condicional de $T_2|T_1$, és a dir, en com la ocurrència de $\mathcal{E}_1$ modifica el risc de $\mathcal{E}_2$. En aquest cas, la distribució de T_1 no és identificable a partir de les dades observades. La situació abans descrita, a on la ocurrència d'un esdeveniment final impedeix l'observació de l'esdeveniment intermedi, i per tant T_2 censura de forma possiblement depenent T_1, és coneguda com a riscos semi-competitius (Fine *et al.*, 2001). L'estratègia que seguiren passà per assumir un model per a la distribució conjunta (T_1, T_2), i aleshores recuperar la distribució marginal de T_1 derivada d'aquest model.

Proposem una nova metodologia per tractar amb riscos semi-competitius quan T_1, el temps fins a l'esdeveniment intermedi, està censurat en un interval. En molts estudis mèdics longitudinals, la ocurrència de l'esdeveniment d'interès s'avalua en visites periòdiques del pacient, i per tant, T_1 és desconegut, però es sap que pertany al interval comprès entre els temps de dues visites consecutives. Els mètodes per riscos semi-competitius en el context usual de censura per la dreta no són vàlids en aquest context i és necessària una nova aproximació. En aquest treball, ampliem la metodologia semi-paramètrica proposada per Fine *et al.* (2001), que assumeix un model de còpula de Clayton (1978) per a descriure la dependència entre T_1 i T_2. Prenent el mateix model per l'associació entre els temps d'interès, desenvolupem un algoritme iteratiu que estima conjuntament el paràmetre d'associació del model de còpula, així com la funció de supervivència del temps intermedi T_1.

Outline

Introduction **1**

I Modelling the evolution of bladder cancer **7**

1 The Spanish Bladder Cancer/EPICURO Study **9**

2 Competing risks analysis of the SBC/EPICURO Study **19**

3 Multi-state models: a dynamical model for progression **49**

II Interval-censored semi-competing risks data **67**

4 Methods for semi-competing risks data **69**

5 Methods for interval-censored data **85**

6 Interval-Censored Semi-Competing Risks Data **97**

7 Asymptotic theory **119**

8 ICSCR analysis of the SBC/EPICURO Study **127**

9 Simulation Study **139**

10 Software contributions **157**

11 Discussion and future research **167**

Bibliography **173**

Appendix **183**

A The Spanish Bladder Cancer Study **185**

B Theoretical Aspects **193**

C Tables of simulation results **211**

D R Programmes **223**

Contents

Introduction **1**

I Modelling the evolution of bladder cancer **7**

1 The Spanish Bladder Cancer/EPICURO Study **9**

1.1 Bladder cancer and the SBC/EPICURO Study 9

1.1.1 Bladder Cancer 9

1.1.2 The Spanish Bladder Cancer/EPICURO Study 11

1.2 Description of the data base 12

1.2.1 Prognostic factors 12

1.2.2 Descriptive characteristics 12

1.3 Lifetime endpoints of interest 14

1.3.1 Events involved in the evolution of bladder cancer 14

1.3.2 Lifetime endpoints considered 15

1.4 Follow-up and censoring patterns 16

1.5 Modelling the course of bladder cancer 17

2 Competing risks analysis of the SBC/EPICURO Study **19**

2.1 Methods for competing risks 20

2.1.1 Model specification 20

2.1.2 Likelihood function 22

2.1.3 Nonparametric estimation . . . 23
2.1.4 Regression modelling . . . 24
2.1.4.1 Cox proportional hazards model for the cause-specific hazards $\lambda_j(t)$: 24
2.1.4.2 Fine and Gray's model for the cumulative incidence functions $F_j(t)$: 25
2.1.4.3 Other regression models . . . 26
2.1.5 Predictions . . . 26
2.1.5.1 Nomograms for competing risks . . . 27
2.1.5.2 Calibration curves for competing risks . . . 29
2.1.6 Existing software for competing risks . . . 29
2.1.7 Final comments on competing risks . . . 30
2.2 Analysis of Event Free Survival in the SBC/EPICURO Study . . . 31
2.2.1 Competing risks for Event Free Survival . . . 31
2.2.2 Prognostic Factors for Event Free Survival. Univariate nonparametric analysis. 32
2.2.3 Multivariate regression model for Event Free Survival. . . . 35
2.3 Analysis for Progression Free Survival in the SBC/EPICURO Study . . . 37
2.3.1 Competing risks for Progression Free Survival . . . 37
2.3.2 Prognostic factors for Progression Free Survival. Nonparametric analysis. . . 38
2.3.3 Multivariate regression model for Progression Free Survival . . . 40
2.3.4 Prediction of the probability of progression . . . 42
2.4 Characterization of the first relapse . . . 44
2.4.1 Motivation . . . 44
2.4.2 Prognostic factors for the first event . . . 45

3 Multi-state models: a dynamical model for progression 49
3.1 Review of multi-state models . . . 50
3.1.1 Model specification . . . 51
3.1.2 Regression modelling . . . 51
3.1.3 Predictive Process . . . 53
3.1.4 Existing software for for multi-state models . . . 56
3.2 A dynamical model for the SBC/EPICURO Study . . . 56
3.2.1 The complete picture of the disease . . . 56
3.2.2 Model fitting . . . 57
3.3 Predictive process of the risk of progression . . . 60
3.3.1 Residual cumulative incidence of progression . . . 62

3.3.2 Conditional risk of progression . . . 63
3.3.3 Updated classification in risk groups . . . 64

II Interval-censored semi-competing risks data 67

4 Methods for semi-competing risks data 69

4.1 Concepts of bivariate survival data . . . 71
4.1.1 Notation . . . 71
4.1.2 Measures of dependence . . . 72
4.1.3 Copula models for bivariate survival data . . . 73
4.2 Semi-competing risks . . . 75
4.2.1 Semi-competing risks data . . . 75
4.2.2 Clayton's copula model for semi-competing risks data . . . 76
4.2.3 Estimation under Clayton's copula model . . . 77
4.2.3.1 Estimation of the association parameter α . . . 78
4.2.3.2 Inference on the marginal $S_1(\cdot)$ . . . 80
4.2.4 Alternative models . . . 81
4.3 Numerical examples . . . 82
4.3.1 An example based on simulated data . . . 82
4.3.2 Semi-competing risks with interval-censored intermediate event . . . 83

5 Methods for interval-censored data 85

5.1 Univariate interval censoring . . . 86
5.1.1 Noninformativity conditions . . . 86
5.1.2 Nonparametric maximum likelihood estimation . . . 87
5.1.3 Parametric maximum likelihood estimation . . . 88
5.1.4 Competing risks analysis when data is interval-censored . . . 88
5.1.4.1 Illustration: the Spanish Bladder Cancer Study . . . 90
5.2 Bivariate interval-censored data . . . 90
5.2.1 Notation and likelihood function . . . 90
5.2.2 Nonparametric estimation of $F(s,t)$ . . . 92
5.2.3 Estimation of the correlation structure . . . 93
5.2.3.1 Direct estimation of association measures . . . 93
5.2.3.2 The copula approach . . . 94
5.2.4 Final comments . . . 95

6 Interval-Censored Semi-Competing Risks Data **97**

6.1 Notation and model . . . 97

6.1.1 Interval-censored semi-competing risks data . . . 97

6.1.2 Model for $(\boldsymbol{T_1}, \boldsymbol{T_2})$ . . . 99

6.2 Outline of the estimating strategy . . . 99

6.3 The expected concordance and the comparable sample . . . 100

6.3.1 Definition and estimation of the expected concordance . . . 100

6.3.2 The comparable sample . . . 103

6.3.2.1 Conditions of comparability for the interval censoring setting . . . 104

6.3.2.2 The comparable pairs . . . 106

6.4 Estimating equations for α . . . 107

6.4.1 Estimation of α by direct estimation of bias . . . 108

6.4.1.1 Estimation of n_e . . . 110

6.4.2 Estimation of $\boldsymbol{\alpha}$ by inverse probability weighting . . . 111

6.4.2.1 The distribution of $\widetilde{T}_{1ij}|\widetilde{T}_{2ij}$ . . . 112

6.4.2.2 The distribution of $(\widetilde{L}_{ij}, \widetilde{R}_{ij})|\widetilde{T}_{2ij}$ . . . 113

6.5 Estimation of the marginal survivals . . . 114

6.5.1 Estimation of $S_C(c)$, $G_1(l, r|y)$ and $G_2(l, r|y)$ . . . 114

6.5.2 Estimation of $\boldsymbol{S_2(\cdot)}$ and $\boldsymbol{S_T(\cdot)}$ . . . 115

6.5.3 The plug-in estimation of $\boldsymbol{S_1(\cdot)}$ . . . 116

6.6 Estimation algorithm . . . 116

6.6.1 Algorithm . . . 116

7 Asymptotic theory **119**

7.1 Inference on the copula parameter $\boldsymbol{\alpha}$ . . . 119

7.1.1 Estimating equations and U-statistics . . . 119

7.1.2 Consistency of $\widetilde{\alpha}_1$ when $\boldsymbol{S_1(\cdot)}$ and $\boldsymbol{S_2(\cdot)}$ are known . . . 121

7.1.3 Asymptotical distribution of $\widetilde{\alpha}_1$ when $\boldsymbol{S_1(\cdot)}$ and $\boldsymbol{S_2(\cdot)}$ are known . . . 122

7.1.4 Inference on α when $\boldsymbol{S_1(\cdot)}$ and $\boldsymbol{S_2(\cdot)}$ are estimated . . . 124

7.2 Inference on the survival function $\boldsymbol{S_1(\cdot)}$ . . . 124

8 ICSCR analysis of the SBC/EPICURO Study 127

8.1 The recurrence process in bladder cancer . . . 128

8.1.1 Estimation of the association parameter α . . . 130

8.1.2 Estimation of the time to recurrence . . . 130

8.2 The progression process in bladder cancer . . . 133

8.2.1 Estimation of the association parameter α . . . 134

8.2.2 Estimation of the time to progression . . . 134

8.3 Illustration: strongly associated simulated data . . . 135

9 Simulation Study 139

9.1 Simulation scenarios and data generation . . . 140

9.1.1 Parameters defining simulation scenarios . . . 140

9.1.1.1 Determination of the parameters of the marginal distributions . . . 140

9.1.1.2 Percentage of independent censoring . . . 142

9.1.2 Generation of data sets . . . 142

9.2 Evaluation criteria . . . 143

9.2.1 Estimation of the association parameter $\boldsymbol{\alpha}$ . . . 143

9.2.2 Estimation of the marginal survival $\boldsymbol{S_1(\cdot)}$ . . . 145

9.3 Simulation results . . . 145

9.3.1 Results for $\boldsymbol{\alpha}$ . . . 145

9.3.1.1 Exponential marginal distributions . . . 145

9.3.1.2 Weibull Margins . . . 149

9.3.2 Results for $\boldsymbol{S_1(\cdot)}$ . . . 152

9.4 Discussion . . . 154

10 Software contributions 157

10.1 Competing risks . . . 157

10.1.1 A nomogram for competing risks . . . 157

10.1.2 A calibration plot for competing risks . . . 159

10.2 Multi-state models . . . 160

10.2.1 The predictive process . . . 160

10.3 Semi-competing risks for right-censored data . . . 161

10.3.1 Data preparation . . . 162

10.3.2 Estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{S_1(t)}$ . . . 162

10.4 Interval-censored semi-competing risks data . . . 162
10.4.1 Initialize parameters . . . 163
10.4.2 Estimation algorithm . . . 165
10.4.3 Comments on the estimation algorithm . . . 166

11 Discussion and future research **167**
11.1 Modelling the evolution of bladder cancer . . . 167
11.2 Interval-censored semi-competing risks data . . . 169
11.3 Future work . . . 170

Bibliography **173**

Appendix **183**

A The Spanish Bladder Cancer Study **185**
A.1 Cumulative incidence functions for (T_1, C_1) . . . 185
A.2 Cumulative incidence functions for (T_2, C_2) . . . 187
A.3 Cumulative incidence functions for (T_1, C_1^*) . . . 188
A.4 Members of the participating centres . . . 190

B Theoretical Aspects **193**
B.1 Expressions of the cross-sectional ratio $\theta(s,t)$ . . . 193
B.1.1 Equivalence of (4.4) and (4.5) . . . 193
B.1.2 Proof of Proposition 4.1: . . . 194
B.2 Equivalence of Clayton's copula model . . . 195
B.3 The expectation of the concordance indicator . . . 196
B.4 The expected concordance . . . 197
B.5 The comparable sample . . . 200
B.5.1 Case A: when $\boldsymbol{\delta_{1i} = 1}$, $\boldsymbol{\delta_{1j} = 1}$ . . . 200
B.5.2 Case B: $\boldsymbol{\delta_{1i} = 1}$, $\boldsymbol{\delta_{1j} = 0}$, and $y_i < y_j$ . . . 200
B.5.3 Cases C, D and E: When $\boldsymbol{\delta_{1i} = 1}$, $\boldsymbol{\delta_{1j} = 0}$, and $y_i > y_j$ . . . 201
B.5.4 Case F: $\boldsymbol{\delta_{1i} = 0}$, $\boldsymbol{\delta_{1j} = 0}$ . . . 202
B.6 Equivalence on the conditions of comparability . . . 204
B.6.1 Conditions 1 $\Longrightarrow$ Conditions 2: . . . 204
B.6.2 Conditions 2 $\Longrightarrow$ Conditions 1 . . . 206
B.7 U-statistics . . . 208
B.8 Unicity of $U(\alpha) = 0$ . . . 210

C Tables of simulation results **211**

C.1 Estimation of the marginal survival function $S_1(t)$ . . . 211

D R Programmes **223**

D.1 Competing risks analysis with `R` . . . 223

D.1.1 Nonparametric and regression modelling with `R` . . . 223

D.1.1.1 Nonparametric estimation . . . 224

D.1.1.2 Regression modelling . . . 225

D.1.1.3 Prediction . . . 226

D.1.2 Technical details on the construction of a nomogram . . . 227

D.1.3 Function `getCalibrateCIF` . . . 228

D.1.3.1 Function `getEstimates` . . . 229

D.1.3.2 Function `groupCIF` . . . 230

D.2 Multi-state models with `R` . . . 231

D.2.1 Functions `Pi1cr.0` and `Pi1cr.1` . . . 231

D.3 Semi-competing risks analysis with `R` . . . 232

D.3.1 Function `corSCR` . . . 232

D.3.2 Function `margSCR` . . . 233

D.3.3 Internal functions . . . 234

D.4 ICSCR with R . . . 235

D.4.1 Function `algICSCR` . . . 235

D.4.2 Functions `compIC` and `compIC.1` . . . 237

D.4.3 Function `nD` . . . 238

D.4.4 Function `pesC` . . . 239

D.4.5 Function `f.Zij` . . . 239

D.4.6 Function `iS1` . . . 240

D.4.7 Internal functions . . . 240

D.4.8 Functions `rclay.exp` and `fsimulICSCR3` . . . 241

List of Tables

1.1 Patients distribution according to the study variables 13

1.2 Comparison between non-smokers and smokers 14

2.1 Results of Gray's significance test to compare cumulative incidence curves across stratum for Event Free Survival 32

2.2 Cause-specific hazards (CSH) Cox models for (T_1, C_1) 36

2.3 Fine and Gray (FGH) models of the subhazards for (T_1, C_1) 36

2.4 Results of Gray's significance test to compare cumulative incidence curves across stratum for Progression Free Survival 38

2.5 Cause-specific hazards (CSH) Cox models for (T_2, C_2) 41

2.6 Fine and Gray (FGH) models for the subhazards of (T_2, C_2) 41

2.7 Median† time between events 45

2.8 Cause-specific hazards (CSH) Cox models for (T_1, C_1^*) 48

2.9 Fine and Gray (FGH) models for the subhazard for (T_1, C_1^*) 48

3.1 Cox models to fit the multi-state model. 59

3.2 Classification of the risk to progress according to the probability of progression before 60 months (5 years). Baseline prediction vs updated prediction at 12 months. 65

3.3 Patients progressing after 12 months (n=48%). 66

3.4 Patients alive and progression free at 60 months (n=571%). 66

4.1 Estimates of α resulting from semi-competing risks analysis. Imputation methods for the simulated example 84

8.1 Relevant lifetime variables for bladder cancer. . . . 128
8.2 Events of interest (intermediate=recurrence). . . . 129
8.3 Goodness-of-fit tests for the Clayton's model (Fine *et al.*, 2001). . . . 130
8.4 Estimates for α when recurrence is an intermediate event (ICSCR analysis). . . . 131
8.5 Events of interest (intermediate=progression). . . . 134
8.6 Estimates for α when progression is an intermediate event (ICSCR analysis). . . . 134
8.7 Estimates of the probability of progression at 12, 24 and 60 months. . . . 136
8.8 Estimates for $\alpha = 4$ for the simulated set (ICSCR analysis). . . . 136

9.1 Simulation parameters. . . . 141
9.2 Estimation of α: comparison of bias between ICSCR1, ICSCR2 and Midpoint (Exponential margins). . . . 146
9.3 ICSCR estimation of α for a model with Exponential marginals, $T_1 \sim \text{Exp}(\lambda_1)$, $T_2 \sim \text{Exp}(\lambda_2)$. . . . 148
9.4 Estimation of α: comparison of bias between ICSCR1, ICSCR2 and Midpoint (Weibull margins). . . . 150
9.5 ICSCR estimation of α for a model with Weibull marginals, $T_1 \sim \text{Weibull}(\mu_1, \rho_1)$, $T_2 \sim \text{Weibull}(\mu_2, \rho_2)$. . . . 151

C.1a Estimation of $S_1(t)$, for a model with $T_1 \sim \text{Exp}(\lambda_1)$, $T_2 \sim \text{Exp}(\lambda_2)$ and $\alpha = 3$. Narrow intervals. . . . 212
C.1b Estimation of $S_1(t)$, for a model with $T_1 \sim \text{Exp}(\lambda_1)$, $T_2 \sim \text{Exp}(\lambda_2)$ and $\alpha = 3$. Wide intervals. . . . 213
C.1c Estimation of $S_1(t)$, for a model with $T_1 \sim \text{Exp}(\lambda_1)$, $T_2 \sim \text{Exp}(\lambda_2)$ and $\alpha = 5$. Narrow intervals. . . . 214
C.1d Estimation of $S_1(t)$, for a model with $T_1 \sim \text{Exp}(\lambda_1)$, $T_2 \sim \text{Exp}(\lambda_2)$ and $\alpha = 5$. Wide intervals. . . . 215
C.2a Estimation of $S_1(t)$, for a model with $T_1 \sim \text{Weibull}(\mu_1, \rho_1)$, $T_2 \sim \text{Weibull}(\mu_2, \rho_2)$ and $\alpha = 3$. Narrow intervals. . . . 216
C.2b Estimation of $S_1(t)$, for a model with $T_1 \sim \text{Weibull}(\mu_1, \rho_1)$, $T_2 \sim \text{Weibull}(\mu_2, \rho_2)$ and $\alpha = 3$. Wide intervals. . . . 217
C.2c Estimation of $S_1(t)$, for a model with $T_1 \sim \text{Weibull}(\mu_1, \rho_1)$, $T_2 \sim \text{Weibull}(\mu_2, \rho_2)$ and $\alpha = 5$. Narrow intervals. . . . 218
C.2d Estimation of $S_1(t)$, for a model with $T_1 \sim \text{Weibull}(\mu_1, \rho_1)$, $T_2 \sim \text{Weibull}(\mu_2, \rho_2)$ and $\alpha = 5$. Wide intervals. . . . 219

List of Figures

1 Evolution of published papers on competing risks in recent years. 2

2 Illustration of the semi-competing risks situation. 3

1.1 Diagram showing the T-stages of bladder cancer. 10

1.2 Linear model to represent the increasing seriousness of events involved in bladder cancer. 14

1.3 Paths representing the observed disease-related events in the follow-up of bladder cancer. 15

2.1 Individuals at risk for each modelling strategy. 25

2.2 Examples of graphical tools for prediction: nomograms and calibration plots (based on simulated data) 28

2.3 Competing risks structure for disease-related events (EFS) and Death from Other Causes 31

2.4a Cumulative incidence functions for (T_1, C_1) by Gender 32

2.4b Cumulative incidence functions for (T_1, C_1) by Tumour number 33

2.4c Cumulative incidence functions for (T_1, C_1) by Grade 33

2.4d Cumulative incidence functions for (T_1, C_1) by Smoking status 34

2.5 Cumulative incidence functions for (T_1, C_1) across Gender, for smokers and non-smokers. 35

2.6 Competing risks structure for progression of disease (PFS) and Death from Other Causes 37

2.7a Cumulative incidence functions for (T_2, C_2) by Age 38

2.7b Cumulative incidence functions for (T_2, C_2) by Tumour number 39
2.7c Cumulative incidence functions for (T_2, C_2) by Stage 39
2.7d Cumulative incidence functions for (T_2, C_2) by Grade 40
2.8 Nomogram for the predicted probability of prediction at five years accounting for competing risks. 43
2.9 Comparisons between the standard Cox model $\widehat{F}_{Cox}(t)$ and the Fine and Gray model $\widehat{F}_{FGH}(t)$ in terms of prediction of the probability of progression. 44
2.10 Competing risks structure for the time to the first event in the SBC/EPICURO Study 45
2.11 Nonparametric estimates of the the cumulative incidence functions for (T_1, C_1^*) . . . 46

3.1 The illness-death unidirectional multi-state model. 50
3.2 Multi-state model for bladder cancer events. 57
3.3 Multi-state model for bladder cancer events. 57
3.4 Predicted PFS cumulative incidence curves for patients after 24 months after diagnosis. 63
3.5 Risk of progression in the next 3 years given the history at time t after Diagnosis . . 64
3.6 Change in the predicted risk of progression when recurrence occurs. 65

4.1 Semi-competing risks setting for the SBC/EPICURO Study. 70
4.2 Induced protective effect in a competing risks analysis. 70
4.3 Regions of observation in the semi-competing risks framework. 75
4.4 Example of concordant pairs ($\Delta_{ij} = 1$) and discordant pairs ($\Delta_{ij} = 0$). 77
4.5 Examples of comparable and non-comparable pairs: (a) Comparable pair: $(\widetilde{T}_{1ij} = T_{1i}, \widetilde{T}_{2ij} = T_{2i}) \in \mathcal{D}_1$, (b) non-comparable pair: $\widetilde{T}_{1ij}$ is not determined, and $\widetilde{T}_{2ij} = T_{2j} \leq \widetilde{T}_{1ij}$ (c) comparable pair: $(\widetilde{T}_{1ij} = T_{1i}, \widetilde{T}_{2ij} = T_{2i}) \in \mathcal{D}_1$, (d) non-comparable pair: $\widetilde{T}_{1ij} = T_{1i}$ but $\widetilde{T}_{2ij}$ is not observed. 79
4.6 Semi-competing risks analysis for right-censored data: Estimated distribution function vs real distribution function for T_1. 83
4.7 Illustration 1: Impact of imputation methods on the estimation of $1 - S_1(t)$, by treatment arm. 84

5.1 Competing risks analysis with interval-censored data: analysis of the time to the first event occurring, either recurrence (left), or progression/death due to bladder cancer (right). 91
5.2 Examples of bivariate interval-censored data 92

6.1 Different situations of observed interval-censored semi-competing risks data. 99
6.2 Examples of observed pairs of individuals 101
6.3 Determination of Z_{ij} 104

6.4 Cases of comparable pairs. . . . 106
6.5 Excluded pairs $(i, j) \in \mathcal{C}^R \setminus \mathcal{C}^{IC}$, satisfying $O^R_{ij} = 1, O_{ij} = 0$. . . . 109
6.6 Examples of pairs from set D satisfying (i) and (ii) from Proposition 6.5. . . . 111

8.1 Semi-competing risks data for Recurrence and Progression/death. . . . 129
8.2a ICSCR vs CR for Age (intermediate=recurrence) . . . 132
8.2b ICSCR vs CR for Tumour stage (intermediate=recurrence) . . . 132
8.2c ICSCR vs CR for Grade (intermediate=recurrence) . . . 132
8.2d ICSCR vs CR for Smoking status (intermediate=recurrence) . . . 133
8.3 Semi-competing risks data for Progression and Death due to Other Causes. . . . 133
8.4a ICSCR vs CR for Tumour number (intermediate=progression) . . . 135
8.4b ICSCR vs CR for Tumour stage (intermediate=progression) . . . 135
8.5 Interval-censored semi-competing risks analysis: Estimated distribution function vs real distribution function for T_1. . . . 137

9.1 Bivariate data generated following Clayton's copula, with Exponential (top) and Weibull margins (bottom) . . . 144
9.2 Estimation of α: comparison of relative bias between ICSCR1, ICSCR2 and Midpoint (Exponential margins, sample size $n = 500$) . . . 147
9.3 Mean Square Error for the ICSCR estimation of α (sample size $n = 500$, Exponential marginals.) . . . 149
9.4 Estimation of α: comparison of relative bias between ICSCR1, ICSCR2 and Midpoint (Weibull margins, sample size $n = 500$) . . . 150
9.5 Mean Square Error for the ICSCR estimation of α (sample size $n = 500$, Weibull marginals.) . . . 152
9.6 Bias of $S_1(t)$ estimates: comparison of ICSC1, ICSCR2 and Midpoint ($\alpha = 3$, Exponential margins) . . . 153
9.7 Bias of $S_1(t)$ estimates: comparison of ICSC1, ICSCR2 and Midpoint ($\alpha = 3$, Weibull margins) . . . 154

A.1 Cumulative incidence functions for EFS and DOC in the analysis of (T_1, C_1) across age . . . 185
A.2 Cumulative incidence functions for EFS and DOC in the analysis of (T_2, C_2) across the tumour's features: size, stage and grade . . . 186
A.3a Cumulative incidence functions for (T_2, C_2) by Gender . . . 187
A.3b Cumulative incidence functions for (T_2, C_2) by Tumour size . . . 187
A.3c Cumulative incidence functions for (T_2, C_2) by Smoking status . . . 188

A.4 Cumulative incidence functions for (T_1, C_1^*) across the individual's features 189

A.5 Cumulative incidence functions for (T_1, C_1^*) across the tumour's features 190

B.1 Case A: Comparable and not comparable pairs when $\delta_{1i} + \delta_{1j} = 2$. 200

B.2 Case B: Comparable and not comparable pairs when $\delta_{1i} = 1$, $\delta_{1j} = 0$, $y_i < y_j$. 201

B.3 Case C: Comparable and not comparable pairs when $\delta_{1i} = 1$, $\delta_{1j} = 0$, $y_i > y_j$ and $b_i < a_j$. 202

B.4 Case D: Non comparable pairs when $\delta_{1i} = 1$, $\delta_{1j} = 0$, $y_i > y_j$ and $a_i < a_j, b_i > a_j$. . . 203

B.5 Case E: Non comparable pairs when $\delta_{1i} = 1$, $\delta_{1j} = 0$, $y_i > y_j$ and $a_i > a_j$. 203

B.6 Case F: Non comparable pairs when $\delta_{1i} = 0$ and $\delta_{1j} = 0$. 204

B.7 Behavior of function $U\big(\alpha, S_1(\cdot), S_2(\cdot)\big)$ for known $S_1(\cdot)$ and $S_2(\cdot)$. 210

C.1 Bias of $S_1(t)$ estimates: comparison of ICSC1, ICSCR2 and Midpoint ($\alpha = 5$, Exponential margins) 220

C.2 Bias of $S_1(t)$ estimates: comparison of ICSC1, ICSCR2 and Midpoint ($\alpha = 5$, Weibull margins) 221

Introduction

This PhD thesis is concerned with survival analysis in the presence of multiple endpoints and complex censoring patterns. In this context, we propose a new methodology for interval-censored semi-competing risks data.

This work was motivated by the Spanish Bladder Cancer/EPICURO Study, the most important study on bladder cancer ever done in Spain, supervised by Dr. Núria Malats from the *Centro Nacional de Investigaciones Oncológicas* (CNIO). Our contribution in this study was focused on modelling the course of the disease and on identifying prognostic factors of its evolution.

The course of complex diseases such as cancer or HIV-infection is characterized by the occurrence of multiple events on the same patient, for instance, the relapse of the disease, death, or the recovery of the normal levels of a biomarker. These events determine several lifetime or time-to-event endpoints, which can be described by survival analysis techniques. These events can be terminating events, if the follow-up of the individual is stopped by their occurrence, or intermediate events, if the individual continues under observation after their occurrence. The presence of terminating events complicates the analysis since their occurrence prevents from observing the complete course of intermediate events, inducing a possibly dependent censoring.

Appropriate methods are required in this context and, specifically, in this thesis we will focus on competing risks, multi-state models and semi-competing risks. These methodologies will be useful to describe important aspects of bladder cancer course. In particular, two novel contributions to the understanding of bladder cancer result from an appropriate use of competing risks and multi-state models: (1) the characterization of those patients with a high risk of progressing as a first event after diagnosis, and (2) the proposal of a dynamical prognostic model for progression.

In the following section we briefly describe the three methodologies that we consider in this work.

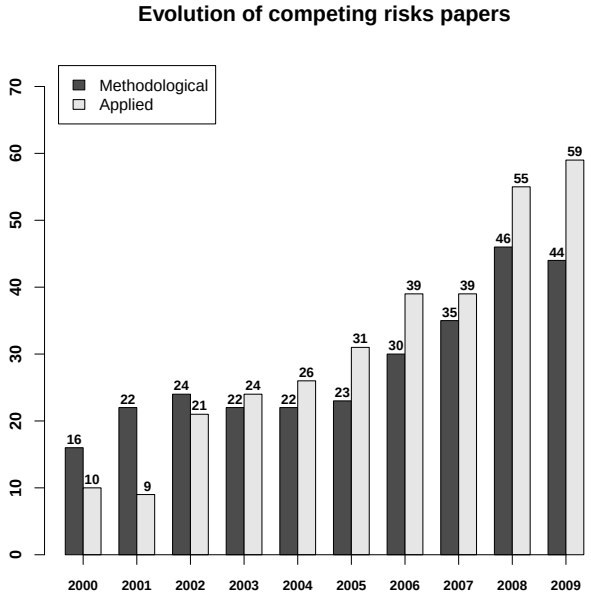


Figure 1: Evolution of published papers on competing risks in recent years.

Competing risks, multi-state models and semi-competing risks

Competing risks arises when we model the time until the first of K possible events, together with the indicator of the type of event observed. Competing risks is a field of survival analysis still in ongoing research. In the last decade (2000-2009), more than 300 methodological papers have appeared in probability and statistical journals.[1] On the other hand, more than 350 papers published in biomedical journals deal with the competing risks problem.[2] Figure 1 shows the increasing trend of published papers on this topic in the last 10 years, both methodological or applied papers, specially the second one. The awareness on the competing risks problem has fortunately gone beyond the statistical community.

In the Spanish Bladder Cancer/EPICURO Study we are interested in the time until the first event is observed, distinguishing between recurrence, progression or death. The characterization of this first event is of paramount clinical importance in order to better target the adequate treatment for each patient. Competing risks could as well help to assess whether the presence of non disease-related deaths have an impact on the observation of the course of bladder cancer.

Multi-state modelling is handled by describing all possible paths that the course of disease could follow, and establishing relationships between the events of interest. In the Spanish Bladder Cancer/EPICURO Study, for instance, a patient after is diagnosed could experience a recurrence, and then die, or he/she could die in remission, that is, before any disease-related event is observed. One

[1] A total of 306 papers identified by a search in Web of Science on May, 18^{th}, with the keyword *competing risks* as a topic, from 2000 to 2009, restricted to subject area *Statistics&Probability*.

[2] A total of 353 papers identified by a search in Pubmed on May, 18^{th}, with the keyword *competing risks* in all fields, from 2000 to 2009, excluding the following statistical journals: Annals of Statistics, Bioinformatics, Biometrical Journal, Biometrics, Biometrika, Biostatistics, Journal of Applied Statistics, Journal of Biopharmaceutical Statistics, Lifetime Data Analysis, Pharmaceutical Statistics, statistics in Medicine and Statistical Methods in Medical Research.

interesting feature of multi-state models is the possibility to make updated predictions according to the occurrence of intermediate events along time. For instance, for the bladder cancer, we will be able to assess the influence that the occurrence of a recurrence has on the posterior risk of p

A , $\mathcal{E}_1$, and a
t

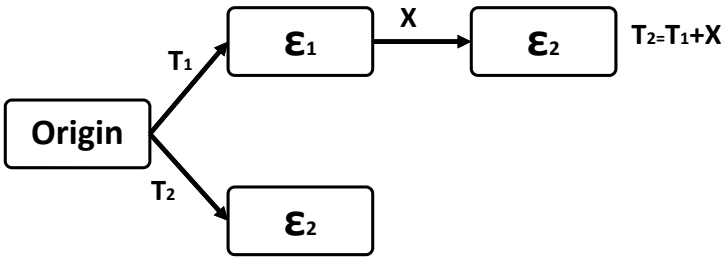


Figure 2: Illustration of the semi-competing risks situation.

This situation is known as semi-competing risk data and differs from the usual multi-state models formulation because the main interest in a semi-competing risk analysis is the marginal distribution of the intermediate event. The study of the marginal law for T_1 is not addressed neither by the competing risks approach nor by the multi-state modelling. While the competing risks approach allows us to analyse the time T to the first between $\mathcal{E}_1$ and $\mathcal{E}_2$, that is, $T = \min(T_1, T_2)$, the multi-state modelling focus on the conditional law of $T_2|T_1$, that is, in how the occurrence of $\mathcal{E}_1$ modifies the risk of $\mathcal{E}_2$.

The difficulty for analyzing the marginal distribution of T_1 comes from the fact that the occurrence of the terminating event prevents the observation of the intermediate event, that is, T_2 dependently censors T_1. Fine *et al.* (2001) first addressed this issue and proposed a strategy for solving this problem by taking advantage of the fact that, for some individuals, T_1 and T_2 are both observed. Thus, if a joint model for (T_1, T_2) is assumed and fit it with data from these individuals, it will be possible to recover the distribution for T_1. Specifically, they proposed a semi-parametric method which assumes a Clayton's copula model to describe the association between the time to the intermediate event and the time to the terminating event.

Fine's method is only valid for right-censored data while, in practice many studies involve interval-censored observations. This is the case, for instance, in most longitudinal studies, where events are evaluated at periodic visits and the time of interest is only known to lie between two consecutive visits. For this reason, the second part of this thesis is devoted to interval-censored semi-competing risk data. Specifically, we propose a new methodology to deal with interval-censored semi-competing risks data, which arises when the time to the intermediate event, T_1, is interval-censored. As an extension of Fine *et al.* (2001) method, we assume a Clayton's copula model to describe the association between T_1 and T_2. Our methodology consists of an iterative estimation algorithm which jointly estimates the association structure of the model and the distribution of the intermediate event.

Structure of the thesis

Part I concerns the modelling of bladder cancer course by means of competing risks and multi-state models techniques.

In Chapter 1 we present with some detail the Spanish Bladder Cancer/EPICURO Study (SBC/EPICURO), together with relevant aspects of the disease necessary to understand the course and the problem being modelled. The events of interest are identified and their corresponding life-time variables defined. We deal with the aspects of the modelling of bladder cancer motivating subsequent chapters, competing risks and multi-state models.

In Chapter 2 we present the state of the art on competing risks methods. These methods are then employed to analyse the data from the SBC/EPICURO Study. We discuss the results, highlighting the differences with standard survival analyses that would ignore the presence of competing risks, specially in two aspects of regression modelling: interpretation of parameters and prediction of future probabilities.

Chapter 3 discusses multi-state models. After a brief summary of the theoretical background for multi-state models, we propose a multi-state approach to model the risk of progression in bladder cancer, which takes into account all the events involved in the evolution of bladder cancer. After modelling each transition, we study the predictive process for progression, which is defined as the probability to progress at a given time u given the history of the individual at the actual time of evaluation, t. The history of the patient is determined, besides baseline characteristics, with the path the patient has followed up to time t. We obtain a dynamic model to make updated predictions on the risk of progression.

Part II is devoted to the problem of interval-censored semi-competing risks data.

In Chapter 4 we present the semi-competing risks problem for right-censored data. First, some issues on bivariate survival data are presented: basic concepts such as the bivariate joint survival functions, as well as measures to assess the dependence between two times of interest. Next, the methodology of Fine *et al.* (2001) is presented in detail: from model specification (Clayton's copula model) to inference on the association structure and the marginal distribution of the intermediate event. Other possible methods are sketched within the chapter.

However, in those situations where the intermediate event is interval-censored, the previous methods are no longer valid, unless some simplification such as midpoint imputation is taken. We need to consider interval censoring into account. So we first provide, in Chapter 5, a state of the art on interval-censored methods, both for univariate and bivariate survival data. Specific methods for interval-censored semi-competing risks data are presented in Chapter 6. Following the ideas of Fine *et al.* (2001), we propose a new estimation algorithm which takes into account the presence of interval-censoring. We deal with the theoretical background of the assumed model and the estimation procedure with some detail. Chapter 7 contains a study on the asymptotical properties of the methods proposed: issues such as consistency and asymptotical behavior are confronted.

The proposed methodology for interval-censored semi-competing risks data is illustrated in Chapter 8 by three examples: one based on simulated data and two taken from the SBC/EPICURO Study. Different approaches to the analysis are considered, some ignoring interval censoring, and some

acknowledging for its presence. We compare the different estimations obtained and interpret the results in the context of bladder cancer.

In Chapter 9, a simulation study is carried out to compare the proposed estimation procedure for interval-censored semi-competing risks data with methods that ignore the presence of interval censoring. In this chapter, we present the design of the simulation study and discuss the most relevant results.

In Chapter 10 we outline the software contributions we have made to implement the methods presented in this work.

This PhD thesis concludes contains the closing Chapter 11, where the main results are summarized and several aspects which remain unsolved or might be approximated differently are addressed.

Part I

Modelling the evolution of bladder cancer

CHAPTER 1

The Spanish Bladder Cancer/EPICURO Study

The study that has motivated the present PhD thesis is the Spanish Bladder Cancer/EPICURO Study, a multicenter study of 1278 newly-diagnosed bladder cancer cases recruited between 1998 and 2001 in 18 Spanish hospitals. The data base includes information on risk factors, genomic DNA, data on the diagnostic and therapeutic processes, and follow-up data, including histological information for recurrences of the tumour. Since recurrences of the tumour remain common among cancer patients, efforts to reduce them are of paramount clinical importance. The availability of prognostic markers that could accurately predict the appearance of new tumoral cells would allow urologists to treat patients more effectively.

The general motivation for this part of the thesis is to model the evolution of disease through the study of the different survival endpoints of interest, and to identify prognostic factors involved in such an evolution. In this chapter, we start presenting the biological background of the bladder cancer disease, and then present the Spanish Bladder Cancer Study in Section 1.1. We describe the cohort under study in Section 1.2. Definition of the events of interests and their corresponding survival endpoints are addressed in Section 1.3. In Section 1.4 we discuss the different censoring patterns arising from the follow-up of the SBC/EPICURO study. We present then our proposed modelling of the course of bladder cancer in Section 1.5.

1.1 Bladder cancer and the SBC/EPICURO Study

1.1.1 Bladder Cancer

Urothelial cell carcinoma of the bladder (UCC) is the fifth most common neoplasm in men in industrialized countries, occurring with a male-to-female ratio of approximately 3:1. Spain is among the countries with highest incidence rate among men (55.0 per 100,000 person-years), but with the

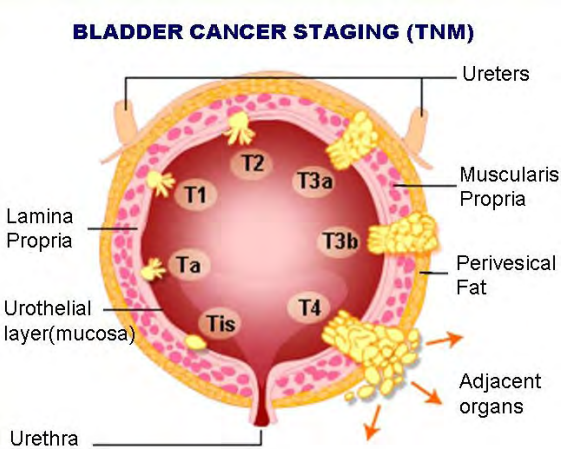


Figure 1.1: Diagram showing the T-stages of bladder cancer.

lowest among women (7.4 per 100,000) (Guey *et al.*, 2009). In addition to male gender, acknowledged risk factors today include high age, tobacco smoking and occupational exposure to carcinogens (Babjuk *et al.*, 2008). The prevalence (persons alive with bladder cancer at any given time) is three to eight times higher than the incidence, making bladder cancer one of the most prevalent neoplasms, and hence, a major burden for all health care systems (Lotan *et al.*, 2008). The overall cause-specific five-year survival rate is about 65%.

Bladder cancer (BC) is a paradigm of a complex disease, both in its etiology and specially in its course. Recurrences of the tumour remain common among cancer patients, with about 40% of the patients experiencing multiple recurrences over many years. The frequency of recurrences forces an strict follow-up of the patients that has a significant impact on the patients' quality of life. Tumours can be classified as either superficial, if the tumour is confined to the lining of the bladder, or invasive, when cancer spread through the lining and invade the muscle wall of the bladder, or spread to nearby organs and lymph nodes. Approximately 75-85% of newly diagnosed cases are classified as superficial or non-muscle invasive (Babjuk *et al.*, 2008).

Superficial tumours can be further classified depending on the depth of invasion or stage into non-invasive papillary carcinoma (Ta), carcinoma in situ or flat tumour (Tis) or tumour invading subepithelial connective tissue (T1). On the other hand, in muscle-invasive tumours, stage can be further classified in muscle-invasive (T2), tumour invading perivesical tissue (T3) or invading nearby organs (T4). Figure 1.1[1] represents graphically the different depth of invasion for each stage in the bladder.

The seriousness of the tumour determine the treatment to give as a first-line therapy. In superficial tumours, the standard treatment is a transurethral resection (TUR) to remove the tumour followed by, in most cases, either immunological therapy, chemotherapy or both. The usual treatment for invasive tumours is cystectomy (removal of the bladder), chemotherapy and radiation therapy. The different course of these two types of cancer suggests the need to perform separate analysis on them. In the present work, we will focus on the course of superficial tumours.

[1]Image from the web site of the College of Medicine, from the University of Oklahoma, `http://www.oumedicine.com/bodycontent.cfm?id$=$2495`.

Because of the existing risk of recurrence for these patients, they need to be routinely monitored by performing a series of cystoscopies to control the appearance of tumoral cells. A cystoscopy is an examination of the bladder tissue in order to determine the presence of malignant agents (Babjuk *et al.*, 2008). The frequency and duration of the tests to be performed has not been established, but some recommendations are available from the European Association of Urology (Babjuk *et al.*, 2008, 2009). For instance, the result of a first cystoscopy at three months after TUR is highly predictive for the relapse of disease, and thus this test is almost mandatory for all types of patients. After this first test, it is advisable to distinguish between low and high risk patients, identified according to the characteristics of the patient and the characteristics of the primary tumour, including number of tumours, size of the largest tumour, stage, grade and concomitant CIS (Babjuk *et al.*, 2008). The recommended follow-up for low-risk patients, if the 3-month results were negative, is to perform a cystoscopy at 9 months and then yearly for 5 years. For high-risk patients, cystoscopies should be repeated every 3 months during two years, every 4 months in the third year, every 6 months until 5 years, and yearly thereafter.

1.1.2 The Spanish Bladder Cancer/EPICURO Study

The Spanish Bladder Cancer/EPICURO Study is the largest bladder cancer case-control study ever done in Spain. Our collaboration in this study was motivated by a joint project held between the *Centro Nacional de Investigaciones Oncológicas* (CNIO), from Madrid, the *Institut Municipal d'Investigació Mèdica* (IMIM), from Barcelona, and the University of Vic (Vic, Barcelona). This project, entitled 'Genetic and Environmental Factors in the Etiology and Prognostic of Bladder Cancer', and funded by the Marató de TV3 Foundation, was aimed at identifying environmental risk factors, genetic susceptibility factors and gene-environment interactions involved in the diagnosis of bladder cancer. In addition, the role of inherited and somatic genetic alterations in the development and progression of bladder cancer, including their prognostic value, was the second main goal of the project. Some results derived from this important study can be found in García-Closas *et al.* (2005), Hernandez *et al.* (2006), Murta-Nascimento *et al.* (2007), Samanic *et al.* (2006) and Guey *et al.* (2009).

Patients were recruited between 1998 and 2001 in 18 Spanish hospitals of five areas in Spain (Asturias, Barcelona, Vallès Occidental/Bages, Alicante and Tenerife). Cases were patients with a newly diagnosed, histologically confirmed, urothelial cell carcinoma of the bladder. Controls were hospital-matched patients according to gender, age within 5-year categories, ethnic origin and region. The end of the follow-up time for the cases is June 30th 2007, so the length of follow-up ranges from 1 month to 117 months (9.8 years).

From the original 1278 cases, 995 were newly-diagnosed with superficial bladder cancer, while the rest 283 patients were diagnosed with invasive bladder cancer. In this thesis, we restrict to the cohort of superficial bladder cancer cases, that is, 995 patients between 22 and 80 years whose tumour is classified in stages Tis, T1 or Ta. Patients with a previous diagnosis of cancer in the urinary system or with bladder tumours that were secondary to other malignancies were excluded.

Clinical and socio-demographic information was obtained from the patients' hospital history. Cases were followed yearly: trained monitors reviewed information on clinical visits and recorded the

events of interest as well as any relevant change in treatment. Telephone interviews were made to expand information on the disease or vital status of the patient.

1.2 Description of the data base

1.2.1 Prognostic factors

The original data set comprised more than 100 variables, including medical and genetic markers. For the present study, we have considered those variables that have been agreed to be risk factors for some bladder cancer survival endpoint as reported in the guidelines for the management of bladder cancer patients (Babjuk *et al.*, 2008):

- **Gender:** Males vs females.
- **Age:** Continuous variable, sometimes categorized for descriptive purposes into younger or equal to 60 years, between 61 and 70 years or older than 70 years.
- **Tumour number or multiplicity:** Categorized between single or multiple tumours.
- **Tumour size:** Size in centimeters of the largest tumour found. Categorized into more or less than 3 centimeters.
- **Stage of the tumour:** Depth of invasion of the tumour. Only superficial tumours are considered in this database, corresponding to stages Ta, T1 or Tis.
- **Tumour grade:** It refers to the grade of differentiation of the cell: well differentiated (Grade 1), moderately differentiated (Grade 2) and poorly differentiated (Grade 3). There can be also neoplasms with low malignant potential (Benign).
- **Smoking status:** Smoking is highly predictive of the development of primary tumours, and can be relevant for the course of the disease. Patients were categorized into smokers, including current and former smokers, and non-smokers, including occasional smokers (defined as patients smoking at least 100 cigarettes in their lifetime but who never smoked regularly).

1.2.2 Descriptive characteristics

Table 1.1 summarizes the distribution of the risk factors across the cohort of superficial cases. The mean age of the 995 patients was 65.7 years (standard deviation 10.0) at the moment of diagnosis of the primary tumour, with women being slightly older (66.9; sd 10.5) than men (65.5; sd 9.9). The cohort was composed mainly by males (87.2%), more than two thirds were older than 60 years, and more than 70% of patients were or had been regular smokers.

As for the characteristics of the tumour, the majority were solitary or single tumours (66.3%), the diameter of the largest being less than 3 cm (57%), more than 80% of them of stage Ta, and benign or grade 1 (42.4%). Notice that the variable size of the tumour was unknown for almost 30% of individuals. An exploratory study comparing the behavior of the three categories (less than 3 cm,

Table 1.1: Patients distribution according to the study variables

Variable	**Categories**	**n(%)**
Gender	Male	868 (87.2%)
	Female	127 (12.8%)
Age (years)	≤60	254 (25.5%)
	61-70	378 (38.0%)
	>70	363 (36.5%)
Tumour number	Single	660 (66.3%)
	Multiple	283 (28.4%)
	Unknown	52 (5.3%)
Tumour size	≤ 3cm	567 (57.0%)
	>3 cm	141 (14.2%)
	Unknown	287 (28.8%)
Stage	Ta	828 (83.2%)
	T1	161 (16.2%)
	Tis	6 (0.6%)
Histological grade	Benign	50 (5.0%)
	GI	374 (37.6%)
	GII	332 (33.4%)
	GIII	239 (24.0%)
Smoking status	Never smoked	117 (11.8%)
	Occasional smokers	38 (3.8%)
	Former smokers	358 (36.0%)
	Current smokers	370 (37.2%)
	Unknown	112 (11.2%)
Total		**995 (100.0%)**

greater or equal than 3 cm and unknown), both with other variables but also with the survival endpoints explained in next section, showed certain evidence that the behavior of the unknown category was similar to the small tumours of less than 3 cm. In order to not losing this large amount of cases, together with feedback from clinicians, we decided to join these two categories in a single one for the analysis. Results did not vary significantly by considering these two categories from ignoring those cases with missing size. Under this conjecture, we are assuming that missing data for this variable was non-ignorable or not at random (Rubin, 1976).

We pay special attention to smoking status. We encounter statistical significant differences (χ^2 test for homogeneity) between smokers and non-smokers both in gender and age. Table 1.2 summarizes these differences. Within smokers, the vast majority are male (96.7%), while in non-smokers, the proportions are quite similar. On the other hand, the mean age in non-smokers is 67.6 years (sd 11.4), while smokers are younger (65.3 years, sd 9.77), showing significant differences (t-Student test, p-value 0.0182). No relationship was found with the rest of clinical covariates (results not shown).

Table 1.2: Comparison between non-smokers and smokers

Variable	Categories	Smoking status		p-Value
		Non-smokers	Smokers	
Total[a]		*155*	*728*	
Gender[b]	Male	68(43.9%)	704 (96.7%)	$< 0.0001^c$
	Female	87(56.1%)	24 (3.3%)	
Age (years)[b]	≤60	28(18.1%)	195 (26.8%)	0.0058^c
	61-70	45 (29.0%)	289 (39.7%)	
	>70	82 (52.9%)	244 (33.5%)	
Age (years) (cont.)	Mean (sd)	67.6 (11.4)	65.3 (9.77)	0.0182^d
	Min-Max	22-80	33-80	

[a] Sample of non missing data.

[b] Percentages computed on the totals of non-missing data.

[c] χ^2 tests for homogeneity.

[d] Student t-test for equal means.

The strong relationship of smoking status with these demographic variables introduces a possible confounding effect that should be appropriately addressed when interpreting the results.

1.3 Lifetime endpoints of interest

1.3.1 Events involved in the evolution of bladder cancer

During his follow-up, a patient with a primary superficial tumour can be diagnosed of a new tumour in the bladder in the form of a recurrence or a progression. Recurrence refers to a new tumour classified as superficial while progression applies when the new tumour is classified as invasive. Furthermore, recurrence and progression may occur more than once in a patient: with any new appearance of the tumour, new treatment is given, and follow-up restarts. A subject may also die due to bladd

A natural se types of failures
is given by tl recurrence of the
tumour that o cancer.

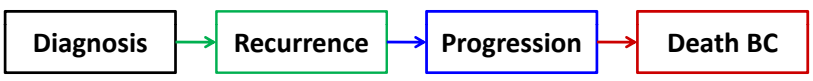


Figure 1.2: Linear model to represent the increasing seriousness of events involved in bladder cancer.

However, this linear order is not observed in practice and a variety of evolutions can be seen in

the study. There are patients who tend to suffer several superficial recurrences of the tumour, but never a progression. Another group of patients develop invasive tumour so rapidly that recurrence is never observed. We adopt the usual terminology that after an invasive tumour is diagnosed, any

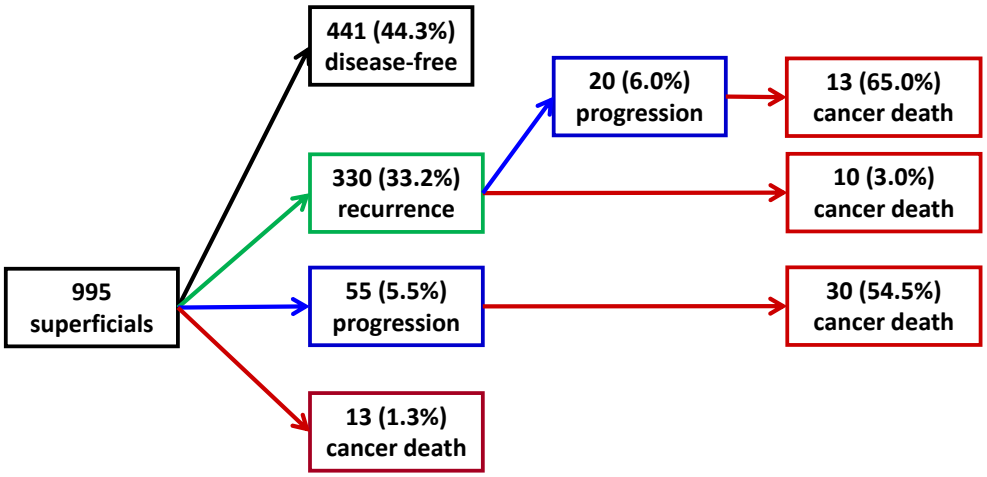


Figure 1.3: Paths representing the observed disease-related events in the follow-up of bladder cancer.

*Percentages at each box are computed over the number of cases of the previous box.
* 212 cases (21.3%) died from other causes.

With respect to the first observed event, the majority are recurrences (33.2%), there are few progressions as a first event (5.5%) and deaths from bladder cancer before any recurrence or progression are scarce (1.3%). Among the patients who suffer at least one recurrence, about 6% progress and 3% die due to bladder cancer. On the other hand, among those patients progressing, a 54.5% dies due to the disease.

1.3.2 Lifetime endpoints considered

Different survival times are usually considered with the goal of characterizing and/or describing different types of events. These survival times are expressed in months:

- **Event Free Survival (EFS) , T_{EFS}**
  Time from diagnosis to relapse (recurrence, progression or death due to bladder cancer).
- **Recurrence Free Survival (RFS), T_{RFS}**
  Time from diagnosis to first recurrence.
- **Progression Free Survival (PFS), T_{PFS}**
  Time from diagnosis to first progression.

- **Disease Specific Survival (DSS), T_{DSS}**
  Time from diagnosis to death due to bladder cancer.

- **Overall survival (OS), T_{OS}**
  Time from diagnosis to death due to bladder cancer or due to any other cause.

Note that, unlike other works, we use the term Event Free Survival for the time from diagnosis to BC-related event, excluding non BC-related deaths. This could also be referred as Relapse Free Survival.

In this work we focus on the two endpoints T_{EFS} and T_{PFS} since they represent the most relevant events of bladder cancer evolution. Event Free Survival is important since it characterizes the most frequent event, and the importance of Progression Free Survival is unquestionable: prediction of the disease progression from superficial to invasive stage would be of great benefit in the management of patients diagnosed with early stage bladder tumours. For those patients dying due to bladder cancer before the diagnosis of a progression, we assign a time to progression equal to the time from diagnosis until death due to bladder cancer.

1.4 Follow-up and censoring patterns

The way individuals are observed in a medical study depends on factors such as the time needed to observe the event of interest, the feasibility of following individuals over time and the mechanism for recording lifetimes and covariate values. It is common in medical studies to follow individuals longitudinally over time, and limitations on the information collected may be imposed by time, cost and other constraints. It is often unfeasible to follow a patient over a long period of time, and many times, the study ends before the event of interest has occurred. Furthermore, a continuous evaluation of the patient is usually unaffordable, and the patient is only observed at scheduled visits. Partial observation of the lifetime variable of interest is known in the survival framework as censoring. This incomplete information of the outcome must be taken into account to derive correct descriptors of the lifetime variable.

In the SBC/EPICURO Study, the observed process, called also inspection process, is as follows. The origin of time is defined as the date that treatment against the primary tumour starts, which we have defined by Diagnose date. Data on the diagnostic process, initial treatment and tumour characteristics were recorded, as well as other characteristics of the patient, such as tobacco and drugs consumption, demographic data, diet and familiar history of disease. When evidence is found of a recurrence or progression, date of new diagnosis is recorded, as well as the date of the previous visit where the subject had still not developed the new tumour. In case of death, the exact date was available. Changes in life habits were also collected annually through a telephonic interview with the patient. The fact that patients are observed intermittently causes some incompleteness on the information available on the exact times of the events of interest producing both right censoring and interval censoring.

Let T be the time of interest, which in our case could be T_{EFS} or T_{PFS}. Right-censoring occurs when only lower bounds on the lifetime are available. The individual is followed up to time C,

when the event of interest has not occurred yet, so T is greater than C. Right-censoring in the SBC/EPICURO data is due to either the end of study, or loss to follow-up. Deaths from other causes will be included as a competing risk event. This type of right-censoring is assumed to be independent.

As we have already mentioned, in the SBC/EPICURO Study, situations of dependent censoring arises. For instance, let T_R be the time to the first recurrence. Given the assumption of the biological model that, after a progression arrives, a recurrence cannot occur, the time to Progression, T_{PFS}, censors the time to recurrence, T_R, $T_{PFS} < T_R$. Given that these two events are closely related, this fact causes dependent censoring.

On the other hand, interval censoring arises in the SBC/EPICURO Study data due to the inspection process, and it could affect both recurrences and progressions. Since the exact time, say T, when a new tumour develops is unknown, the time of interest is known to lie somewhere between them, $L < T \leq R$, where only the previous visit date where the patient was disease-free, L, and the diagnosis date of the new tumour, R, are available. Hence T is said to be censored in the interval $(L, R]$. Therefore, because the occurrence of a recurrence or a progression are interval-censored, T_{EFS} or T_{PFS} are interval-censored when a recurrence or progression occurs (when applicable), exact-censored when death occurs or right-censored, when no event is observed.

The interval censoring problem will be tackled in the second part of the thesis (Chapters 4 to 9). In the first part (Chapters 2 to 3) the interval-censored is reduced by midpoint imputation.

1.5 Modelling the course of bladder cancer

Despite of its complexity and the effort on improving the predictions with the use of biomarkers, bladder cancer is usually modelled with standard survival methods, such as the Cox model or Kaplan-Meier curves. However, more advanced methods such as competing risks and multi-state methods can provide a deeper understanding of the bladder cancer course. Specifically, we will focus on three important aspects: first, the presence of deaths from other causes which prevent the observation of disease-related events. Can they be ignored assuming that they are independent of the disease process? Second, the characterization of the first observed event. Characterizing those patients that progress so rapidly that no recurrence can be observed before their progression is of primary medical interest. And, third, the construction of a dynamic model for the risk of progression that takes into account the history of the patient until a given time. In next Chapters 2 and 3, these questions will be developed and addressed in full detail.

CHAPTER 2

Competing risks analysis of the Spanish Bladder Cancer/EPICURO Study

In this chapter, we analyze Event Free Survival, Progression Free Survival and First Relapse for the SBC/EPICURO study. Most bladder cancer studies treat non bladder cancer deaths as censored observations for the time of interest. This is only appropriate as long as the survival time of interest and the time to non bladder cancer deaths are independent. As we will show next, this is not the case in the Spanish Bladder Cancer Study where non bladder cancer deaths are related to smoking and, in turn, smoking is associated with gender and age, possible risk factors of recurrence and progression. Thus, a competing risk analysis accounting for the type of failure is required.

Though the theoretical basis for competing risks are well developed, its implementation and correct interpretation of results is far from being a straightforward task. For this reason, our emphasis will not only be on the interpretation of the results for a better understanding of the bladder cancer course, but also, on the discussion on how the results of any competing risk analysis should be correctly interpreted.

We start this chapter with a review of competing risks methods in Section 2.1, including model specification, nonparametric methods to estimate the functions of interest and regression modelling. Competing risks models can be used for the identification of risk factors and also as predictive models. However, simple tools for visualizing and validating the predictive probabilities, such as nomograms and calibration plots, are not available for competing risks. We have adapted such tools, available in the statistical software `R` (R Development Core Team, 2009) for standard univariate survival analysis, to competing risks. A brief description of the procedure is given in within Section 2.1.4.

In Sections 2.2 and 2.3 we perform analogous competing risks analysis for Event Free Survival (EFS) and Progression Free Survival (PFS) for the Spanish Bladder Cancer/EPICURO Study accounting

for the competing event Deaths from Other Causes (DOC). In both sections, we first explore nonparametrically the cumulative incidence curves for each of the risk factors considered and then perform a multivariate regression approach in order to explore the joint effect of the variables. We also address the question of whether the presence of the competing event can be ignored by comparing our results with those obtained treating deaths from other causes as independent right-censored observations.

Section 2.4 is devoted to the characterization of the first relapse. We perform a competing risk analysis of the time to the first observed bladder cancer event, distinguishing between recurrence, progression or bladder cancer death and deaths from other causes. This is an unusual approach for modelling bladder cancer events but, as mentioned in the introduction, this modelling is clinically relevant since it will allow identifying those patients with a high risk of fast progressions.

All analyses were done with the statistical software R (R Development Core Team, 2009). In Appendix D.1 we illustrate how to use available functions in this software to perform a competing risks analysis.

2.1 Methods for competing risks

Competing risks data usually arises in studies in which the failure of an individual may be classified into one of k mutually exclusive causes of failure. Examples of competing risks data are found in many fields. In a demographic study we might be interested in analysing mortality distinguishing between leading causes of death: heart disease, cancer, accident... Other instances occur in clinical trials where the endpoint of interest if the first among several events. For instance, in a clinical trial addressed to find the benefits of a new drug to prevent myocardial infarction, patients with coronary heart disease are followed. The failure of interest is myocardial infarction though patients are also at risk of dying for other causes. In reliability analysis, failure may correspond, for example, to breakdown of a mechanical device where there are two causes of failure, vibration or corrosion. Classical survival analysis in this context, which ignores competing risks, may not be appropriate.

A competing risk model is specified through the joint distribution of the time to failure, T, and the cause of failure C. The joint distribution for (T, C) is completely described either by means of the cause-specific hazard functions, representing the rate of occurrence of each cause of failure, or through the cumulative incidence functions, that is, the probability of a subject failing from each cause in the presence of all the competing risks. It has been shown (Pepe and Mori, 1993) that classical survival methods such as the Kaplan-Meier estimate may provide biased results when the competing causes of failure are ignored and treated as right-censored observations. Standard survival analysis techniques rely on the assumption of independence between the failure time and the censoring random variable, which may not be fulfilled by the competing causes.

2.1.1 Model specification

Define, for each individual, the pair (T, C), T being the failure time and C the failure cause. T is assumed to be a continuous and positive random variable, and C takes values in the finite set

$\{1, \ldots, J\}$. It is considered that the individual fails from one and only one cause. For instance, when studying the benefits of a new drug to prevent myocardial infarction, C would take values in $\{1, 2\}$, corresponding to myocardial infarction and death due to other causes, respectively. In the reliability example also two causes of failure are possible, 1 for failures due to vibration and 2 for failures due to corrosion. The joint distribution of (T, C) is completely specified through either the cause-specific hazards, $\lambda_j(t)$, or through the cumulative incidence functions, $F_j(t)$ (Lawless 2003).

While the cause-specific hazard function for the j^{th} cause, $j = 1, \ldots, J$,

$$\lambda_j(t) = \lim_{\Delta t \to 0} \frac{P\big(T < t + \Delta t, C = j | T \geq t\big)}{\Delta t},$$

represents the rate of occurrence of the j^{th} failure, the cumulative incidence function from type j failure, $j = 1, \ldots, J$,

$$F_j(t) = P(T \leq t, C = j), \tag{2.1}$$

corresponds to the probability of a subject failing from cause j in the presence of all the competing risks.

The overall hazard function $\lambda(t)$, defined as the hazard function of T, is obtained summing up all the cause-specific hazards, $\lambda(t) = \sum_{j=1}^{J} \lambda_j(t)$. Denote by $\Lambda_j(t) = \int_0^t \lambda_j(u)du$ and by $\Lambda(t) = \int_0^t \lambda(u)du$ the cumulative cause-specific and overall hazards, respectively. The overall survival function $S(t)$ for T is defined as follows:

$$S(t) = P(T > t) = e^{-\Lambda(t)} = e^{-\sum_{j=1}^{J} \Lambda_j(u)du}.$$

Thus, the survival function can be factorized into the following J functions $S_j(t) = e^{-\Lambda_j(t)}$ as follows

$$S(t) = \prod_{j=1}^{J} e^{-\Lambda_j(t)} = \prod_{j=1}^{J} S_j(t). \tag{2.2}$$

Caution is needed when interpreting functions $S_j(t)$. Despite having the mathematical properties of continuous survivor functions, they are not the survivor functions of any observable random variables. Moreover, $S_j(t) \neq 1 - F_j(t)$.

The distribution function for T is obtained from the cumulative incidence functions through $F(t) = P(T \leq t) = \sum_{j=1}^{J} F_j(t)$, and the marginal distribution of C is:

$$\pi_j = P(C = j) = \lim_{t \to \infty} F_j(t) \qquad j = 1, \ldots, J.$$

The cumulative incidence function for cause j, $F_j(t)$, can be derived from the cause specific hazard $\lambda_j(t)$ and the overall survival function $S(t)$ from the relationship:

$$F_j(t) = \int_0^t \lambda_j(u) S(u) du \qquad j = 1, \ldots, J. \tag{2.3}$$

A different way of describing a competing risks model with J causes of failure is to consider a

failure time T_j for each cause, $j \in \{1, \ldots, J\}$. These times are latent variables corresponding to the hypothetical failure times if the other causes of failure were not present. It has been argued that multivariate models $F(t_1, \ldots, t_J)$ could be specified for the joint distribution of $T_1, \ldots, T_J$ (see Andersen *et al.*, 2002, Kalbfleisch and Prentice, 2002, Lawless, 2003 for further references). However, when all risks are present only $T = \min(T_1, \ldots, T_J)$ can be observed, together with $C = j$, such as $T = T_j$, and an identifiable problem is found (Cox and Oakes, 1984, Tsiatis, 1975). $F(t_1, \ldots, t_J)$ is inestimable solely based on these observations. Two different distributions for $F(t_1, \ldots, t_J)$ may result in the same marginal for (T, C). Only under strong assumptions, such as independence, the multivariate distribution is identifiable. However, this assumption is untestable based solely on observed competing risk data. More details on this discussion are given in Putter *et al.* (2007).

A competing risks model can also be seen as a special case of a multi-state model (Andersen *et al.*, 2002), with one transient state 'Alive' and J absorbing states for each cause of failure. In this framework we can think of as a Markovian process when the goal is to model the transitions between states, through the probabilities of transition, $P_{hj}(s, t)$, that is, the probability of being in state j at t, provided that at time s, the state h was occupied. Note that $P_{0j}(0, t) = P(T \leq t, C = j)$ are the cumulative incidence functions as defined in (2.1), whereas the intensity transition functions are the cause-specific hazards.

2.1.2 Likelihood function

Consider a random sample of n individuals, $(T_1, C_1), \ldots, (T_n, C_n)$, where T_i is the time of failure and C_i is the cause of failure for subject i. For each individual, there exists a non-negative right censoring time V_i. Let $\delta_i = I(T_i \leq V_i)$ be the censoring indicator, and define $\tilde{C}_i = \delta_i C_i$. $\tilde{C}_i$ is the cause of failure for failing individuals or 0 for censored individuals. The observed data for each individual are given by $\{(Y_i = \min(T_i, V_i), \delta_i, \tilde{C}_i),\ i = 1, \ldots, n\}$.

Define by $q(v)$ and $Q(v)$ the density and distribution functions of the censoring variable V. The observed data for an individual i consists of either $(T_i = y_i, C_i = c_i)$ (and thus $V_i \geq y_i$) or $T_i > y_i$ (and thus $V_i = y_i$). The contribution of each individual to the likelihood function is then

$$\left(f_{c_i}(y_i)Q(y_i)\right)^{\delta_i} \left(S(y_i)q(y_i)\right)^{1-\delta_i}.$$

Under the assumption that V is independent of (T, C) and that the supports of T and V are disjoint, the likelihood function for the sample is proportional to

$$\mathcal{L} \propto L = \prod_{i=1}^{n} f_{c_i}(y_i)^{\delta_i} S(y_i)^{1-\delta_i}.$$

Denote by $\delta_{ij} = I(C_i = j)$, where $\delta_i = \sum_{j=1}^{J} \delta_{ij}$. If $\delta_i = 1$, then it exists some j with $\delta_{ij} = 1$. From the factorization of the survival $S(t) = \prod_{j=1}^{J} S_j(t)$ (see (2.2)), and defining $g_j(t) = -S_j'(t) = \lambda_j(t)S_j(t)$, the likelihood function can be rewritten as a product of k components, one for each

failure cause j as follows:

$$L = \prod_{i=1}^{n} \left(\prod_{j=1}^{J} f_j(y_i)^{\delta_{ij}} \right) S(y_i)^{1-\delta_i} = \prod_{i=1}^{n} \left\{ \left(\prod_{j=1}^{J} (\lambda_j(y_i) S(y_i))^{\delta_{ij}} \right) \left(\prod_{j=1}^{J} S_j(y_i)^{1-\delta_i} \right) \right\}$$
$$= \prod_{j=1}^{J} \left(\prod_{i=1}^{n} g_j(y_i)^{\delta_{ij}} S_j(y_i)^{1-\delta_{ij}} \right) \underset{\text{def.}}{=} \prod_{j=1}^{J} L_j. \tag{2.4}$$

This expression provides a factorization of the overall likelihood L in terms of cause-specific likelihoods L_j. This factorization shows how $\lambda_j(t)$ and $\Lambda_j(t)$ are directly estimable from data (Y_i, δ_{ij}), if failure times from other causes are considered as censoring times. In fact, L_j corresponds to the likelihood it would be obtained from this sample, where the corresponding hazard, density and survival functions would be, respectively, $\lambda_j(t)$, $g_j(t)$ and $S_j(t)$. However, it does not exist any observed random variable U_j whose survival function satisfies $S_j(t) = P(U_j > t)$.

2.1.3 Nonparametric estimation

Given the observed data, $\{(y_i, \delta_i, \tilde{c}_i), i = 1, .., n\}$, let $0 < y_1 < \cdots < y_N$ be the ordered distinct observed time points. We denote by d_{ij} the number of subjects failing from cause j at time y_i. The number of subjects failing at time y_i from any cause is obtained by the sum of subjects failing for each cause at y_i, $d_i = \sum_{j=1}^{J} d_{ij}$. We define n_i as the number of individuals at risk at y_i, that is, alive and uncensored just prior to this time.

An estimate of the cause-specific hazard for cause j at time y_i is given by the Nelson-Aalen estimate $\widehat{\lambda}_j(y_i) = \frac{d_{ij}}{n_i}$, and it is 0 at any other time. Hence, the estimator for the cumulative cause-specific hazard function, $\Lambda_j(t) = \int_0^t \lambda_j(u)du$, is given by $\widehat{\Lambda}_j(t) = \sum_{i:y_i \le t} \frac{d_{ij}}{n_i}$, $j = 1, \ldots, J$, and its variance estimated by

$$\widehat{\text{Var}}\left(\widehat{\Lambda}_j(t)\right) = \sum_{i:y_i \le t} \frac{d_{ij}}{n_i^2}.$$

The overall survival function for T can be estimated either by the Kaplan-Meier estimate

$$\widehat{S}_{\text{KM}}(t) = \prod_{i:y_i < t} \left(1 - \frac{d_i}{n_i}\right)^{\delta_i},$$

with variance estimated using Greenwood formula by

$$\widehat{\text{Var}}\left(\widehat{S}_{\text{KM}}(t)\right) = (\widehat{S}_{\text{KM}}(t))^2 \sum_{i:y_i < t} \frac{d_i}{n_i(n_i - d_i)},$$

or as a function of the Nelson-Aalen estimate, that is, $\widehat{S}_{\text{NA}}(t) = \exp\left[-\sum_{j=1}^{J} \widehat{\Lambda}_j(t)\right]$.

A natural non-parametric estimate of the cumulative incidence function $F_j(t)$ is then given by

$$\widehat{F}_j(t) = \int_0^t \widehat{\lambda}_j(u)\widehat{S}(u)du \approx \sum_{i:y_i \le t} \frac{d_{ij}}{n_i} \widehat{S}(y_i^-) \qquad j = 1, \ldots, J, \tag{2.5}$$

where $\widehat{S}(t)$ is indistinctively $\widehat{S}_{\mathrm{KM}}(t)$ or $\widehat{S}_{\mathrm{NA}}(t)$, and, given that it is a step function jumping at y_i, $\widehat{S}(y_i^-)$ is the value of $\widehat{S}$ at the left limit of y_i. The variance of $\widehat{F}_j(t)$ can be approximated by the following expression (Pintilie, 2006):

$$\begin{aligned}\widehat{\mathrm{Var}}\left(\widehat{F}_j(t)\right) = &\sum_{i:y_i\le t}\left\{\left[\widehat{F}_j(t)-\widehat{F}_j(y_i)\right]^2\frac{d_i}{(n_i-1)(n_i-d_i)}\right\}+\sum_{i:y_i\le t}\widehat{S}(t_{i-1})^2\frac{d_{ji}(n_i-d_{ji})}{n_i^2(n_i-1)}\\ &-2\sum_{i:y_i\le t}\left[\widehat{F}_j(t)-\widehat{F}_j(y_i)\right]\widehat{S}(t_{i-1})\frac{d_{ji}(n_i-d_{ji})}{n_i(n_i-d_i)(n_i-1)}.\end{aligned} \tag{2.6}$$

2.1.4 Regression modelling

In competing risks, two different regression modelling strategies are possible: one could model either the cause-specific hazards or the cumulative incidence functions. In the former, each cause-specific hazard is analysed separately by treating individuals failing from other causes as censored observations, as follows from the factorization of the likelihood function (2.4). This approach is appropriate when we are interested in determining factors associated to the risk of a specific cause of failure and it is developed in Section 2.1.4.1. For the latter, the modelling of the cumulative incidence functions is adequate when we want to determine factors associated to the incidence of a given cause, and it is described in Section 2.1.4.2. Finally, in Section 2.1.5 we will discuss how to predict the probability of an event of a specific type to occur in a pre-specified time.

2.1.4.1 Cox proportional hazards model for the cause-specific hazards $\lambda_j(t)$:

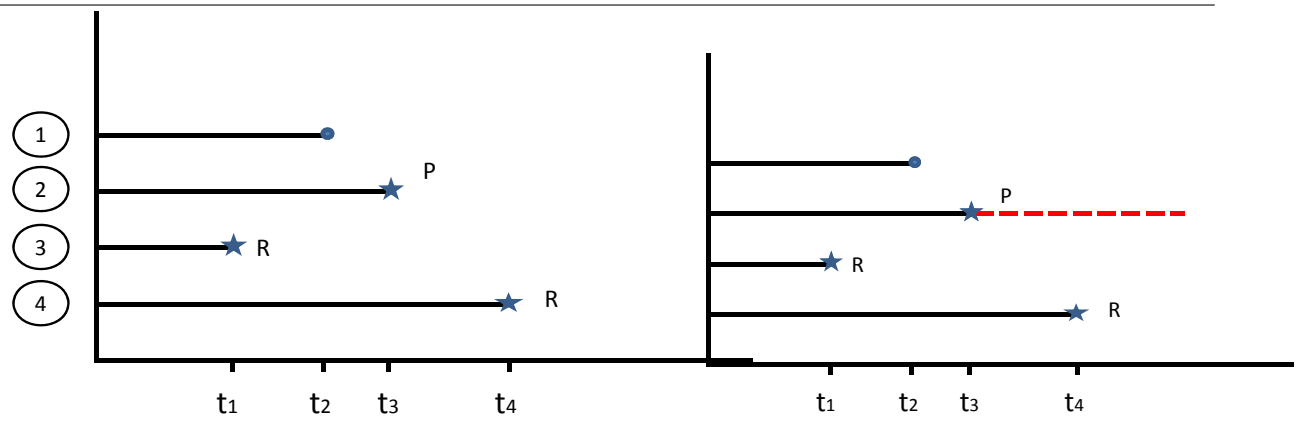
The classical regression analysis of competing risks establishes a Cox proportional hazards (PH) model (Prentice *et al.*, 1978) for each cause-specific hazard:

$$\lambda_j(t|\boldsymbol{Z}) = \lambda_{0j}e^{\boldsymbol{\beta}_j'\boldsymbol{Z}} \qquad j=1,\dots,J,$$

where $\boldsymbol{Z}$ is a $p\times 1$ vector of covariates and $\boldsymbol{\beta_j}$ is a $p\times 1$ vector of regression coefficients for each cause. Each cause of failure is analysed separately, treating individuals failing from other causes as censored observations. The effect of the covariates is assumed to act multiplicatively on an unknown baseline hazard function λ_{0j}. As in classical PH analysis, the validity of the models does not depend on the true form of the baseline hazard, provided the multiplicative form of the model is correct. The PH assumption is a strong one that must be carefully checked for each cause.

Estimation of the regression parameters $\boldsymbol{\beta}_j$ is based on the partial likelihood approach. Given consistent and asymptotically normal estimates $\widehat{\boldsymbol{\beta}}_j$, each cause-specific baseline hazard $\widehat{\Lambda}_{0j}(t)$ can be obtained, for instance, by means of the generalized Nelson-Aalen estimates.

$$\widehat{\Lambda}_{0j}(t) = \sum_{i:t_\ell\le t}\left(\frac{\delta_{ij}}{\sum_{\ell=1}^n Y_\ell(t_i)e^{\widehat{\boldsymbol{\beta}}_j'\boldsymbol{Z}_\ell}}\right) \qquad j=1,\dots,J.$$



(a) Cox model (b) Fine and Gray's model

Figure 2.1: Individuals at risk for each modelling strategy.

2.1.4.2 Fine and Gray's model for the cumulative incidence functions $F_j(t)$:

Consider a new function, the sub-hazard $\gamma_j(t)$ derived from the sub-distribution function:

$$\begin{aligned}\gamma_j(t|\boldsymbol{Z}) &= \lim_{\Delta t \to 0} \frac{Pr\big(T < t + \Delta t, C = j \big|\ \boldsymbol{Z}, \{T \geq t \text{ or } (T < t \text{ and } C \neq j)\}\big)}{\Delta t} \\ &= \frac{f_j(t|\boldsymbol{Z})}{1 - F_j(t|\boldsymbol{Z})} \qquad j = 1, \ldots, J.\end{aligned}$$

This would be the hazard obtained if F_j were a proper distribution. The sub-distribution function is expressed in terms of the sub-hazards as

$$F_j(t|\boldsymbol{Z}) = 1 - \exp\big(-\int_0^t \gamma_j(t|\boldsymbol{Z})\big). \tag{2.7}$$

The conditional expression in the definition of the sub-hazard includes two different scenarios:

i) the event has not occurred at time t,

ii) the event has occurred from a different cause before t.

Thus, the risk set at time t is formed by two types of individuals, corresponding to the two different scenarios. Contrary to the analysis based on the cause-specific hazards, a patient failing from other causes would not be removed from the risk set at his/her time of failure. Figure 2.1 shows a simple example on the differences between these two risk sets. We have four individuals who can experience either an event of type R, or of type P or they can be censored. We have four observation times, from t_1 up to t_4. Inference by means of Cox model for cause R specific hazard proceeds as follows: at t_1 all four individuals are at risk; at t_2 individuals 1,2 and 4 are at risk, because individual 3 has failed before due to cause R; at t_3 only individuals 2 and 4 remain at risk, while at t_4 only individual 4 is at risk. Inference by means of the cumulative incidence for cause R proceeds analogously at times t_1, t_2 and t_3, but differs at time t_4 because individual 2, who experience an event of type P at $t_3 < t_4$ is maintained in the risk set.

Fine and Gray (1999) propose a Cox model for each subhazard, that is

$$\gamma_j(t|\boldsymbol{Z}) = \gamma_{0j}(t)e^{\boldsymbol{\beta}_j'\boldsymbol{Z}}, \qquad j = 1, \ldots, J,$$

where the covariates are linear on a complementary log-log transformed cumulative incidence function. The authors propose a weighted version of the partial likelihood method to estimate the regression coefficients of Cox model. Indeed, if there are N failures at $t_1 < t_2 < \cdots < t_N$, the partial weighted likelihood is defined by

$$\tilde{L}(\boldsymbol{\beta}_j) = \prod_{i=1}^{N} \left(\frac{e^{\boldsymbol{\beta}_j'\boldsymbol{Z}_i}}{\sum_{\ell \in \tilde{R}_i} w_{i\ell} e^{\boldsymbol{\beta}_j'\boldsymbol{Z}_\ell}} \right).$$

Now the risk set for cause j at time t_i is $\tilde{R}_i = \{\ell : t_\ell \geq t_i \text{ or } (t_\ell \leq t_i \text{ and } C \neq j)\}$, where subjects experiencing a competing cause remain in the risk set. The weight $w_{i\ell}$ given to such an individual is $\tilde{G}(t_i)/\tilde{G}(\min(t_\ell, t_i))$, where $\tilde{G}$ is the survivor function for the censoring distribution. An individual satisfying $t_\ell \geq t_i$ is unweighted (i.e. its weight is equal to 1). Maximization of this function provides valid estimates for the coefficients, and inference is derived as for usual partial likelihood, so hypothesis testing on the parameters and selection of the best model can be performed.

Recently, Geskus (2010) has shown that the sandwich type estimator proposed by Fine and Gray to fit their model provide a non-optimal estimate for the standard errors for the coefficients of the model. He proposes two alternative ways to express the estimator of the cumulative incidence functions (2.5): as a weighted cumulative distribution function and as a product limit estimator. These representations permit to make inferences at the cumulative scale by using weighted versions of standard procedures.

2.1.4.3 Other regression models

There are other regression models suitable to fit cause-specific hazards such as parametric alternatives, or other semi-parametric and nonparametric options, such as the additive model (Aalen, 1993, Aalen *et al.*, 2001). Less frequent, there also exist alternatives to Fine and Gray's model when it comes to the modelling of cumulative incidence functions. In the line of their work, more general transformations of the cumulative incidence function are attempted (Fine, 2001). Scheike and Zhang (2008) propose a Cox-Aalen model for the sub-distribution hazards $\gamma_j(t|\boldsymbol{Z})$, and in Scheike *et al.* (2008), they use binomial regression methods to estimate coefficients. Klein and Andersen (2005) and Klein (2006) propose pseudo-values regression models to approximate the sub-distribution functions.

2.1.5 Predictions

The issue of whether to use cause-specific hazards or cumulative incidence functions mainly depends on the scientific question of interest. From a practical point of view, modelling the hazards could be enough if the aim is to identify risk factors. However, the estimation of the cumulative incidence functions is necessary if our goal is focused on prediction. To do so, we could either i) combine the

cause-specific hazards resulting from several Cox models, or ii) fit a Fine and Gray model for the cumulative incidence function.

Assume we have identified the risk factors for the time of interest T by fitting Cox proportional hazards models for each cause-specific hazard, $\widehat{\lambda}_j(t|\boldsymbol{Z})$, $j = 1, \ldots, J$. In order to predict the probability of failing due to cause j before time t_0, that is, the cumulative incidence function for cause j, $F_j(t)$, we plug-in the following estimates for $\Lambda_j(t|\boldsymbol{Z})$ and $S(t|\boldsymbol{Z})$ in equation (2.3):

$$\widehat{\Lambda}_j(t|\boldsymbol{Z}) = \widehat{\Lambda}_{0j}(t)e^{\widehat{\boldsymbol{\beta}}_j'\boldsymbol{Z}} \qquad j = 1, \ldots, k.$$
$$\widehat{S}(t|\boldsymbol{Z}) = \exp\left\{-\sum_{j=1}^{J}\widehat{\Lambda}_{0j}(t)e^{\widehat{\boldsymbol{\beta}}_j'\boldsymbol{Z}}\right\},$$

and the estimate for $F_j(t|\boldsymbol{Z})$ is given by:

$$\widehat{F}_j(t|\boldsymbol{Z}) = \int_0^t \widehat{S}(u|\boldsymbol{Z})d\widehat{\Lambda}_j(u|\boldsymbol{Z}) \approx \sum_{i:y_i\leq t}\delta_{ij}\widehat{S}(y_i|\boldsymbol{Z})d\widehat{\Lambda}_j(y_i|\boldsymbol{Z}). \tag{2.8}$$

The problem with this approach is that no direct estimate for the effect of a covariate in the cumulative incidence function $F_j(t)$ is given. Although the effect of the covariates on the cause-specific hazard $\lambda_j(t|\boldsymbol{Z})$ is directly given by β_j, the effect on the cumulative incidence function $F_j(t)$ combines β_j together with the overall effect on $\widehat{S}(t|\boldsymbol{Z})$. Moreover, it is not possible to test for significant effects on the sub-distribution functions, because some covariates can have a significant effect on the hazard, but not on the cumulative incidence.

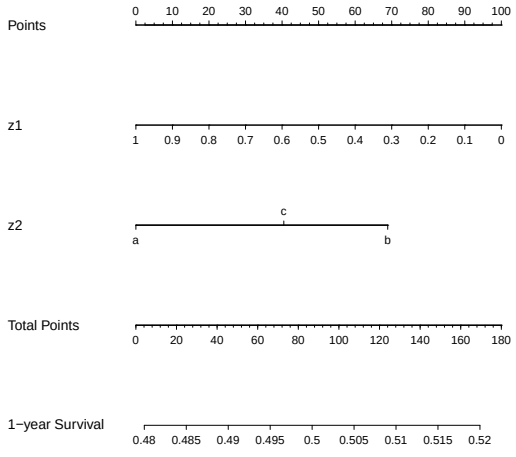
To perform model selection and to obtain estimates for the effects of the covariates on the cumulative incidence functions, as well as to obtain accurate predictions, models based directly on $F_j(t)$, such as Fine and Gray's model, are needed. If the Fine and Gray approach has been used to model the subhazards, $\widehat{\gamma}_j(t|\boldsymbol{Z})$, $j = 1, \ldots, J$, then using expression (2.7) we can obtain estimates for $F_j(t|\boldsymbol{Z})$:

$$\widehat{F}_j(t|\boldsymbol{Z}) = 1 - \exp\big(-\int_0^t \gamma_j(u|\boldsymbol{Z})du\big). \tag{2.9}$$

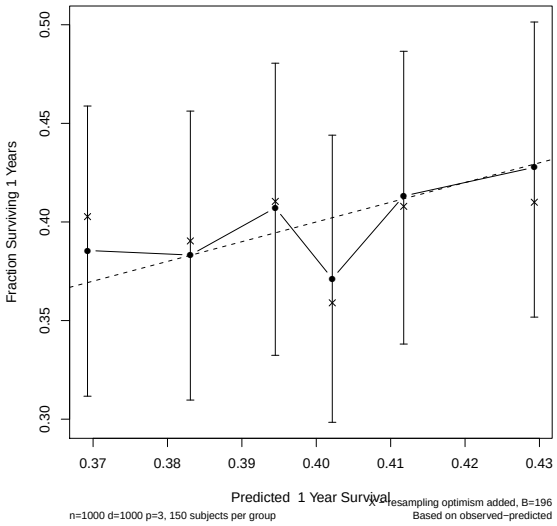
We can use graphical such as the nomogram, to represent this multivariate model, or calibration plots, to assess the predictive validity of the model. In the following, we briefly present these methods in the context of competing risks data.

2.1.5.1 Nomograms for competing risks

Nomograms are graphic representations of a multivariate regression model which provide direct assessment on the predicted probabilities of the event of interest (Harrell *et al.*, 1996). For instance, Figure 2.2(a) represents a multivariate regression model based on simulated data to predict 1-year survival in terms of one continuous covariate, z_1 and one categorical covariate z_2. Each covariate in the model and its values is represented in an horizontal axis, and each of them is assigned a scale of points according to its prognostic significance (in the figure, it is the upper axis labelled Points). The total score for all the variables (Total Points axis) is converted to an estimated probability



(a) Nomogram: predicted 1 year survival.



(b) Calibration plot for 1-year survival

Figure 2.2: Examples of graphical tools for prediction: nomograms and calibration plots (based on simulated data)

or risk (Predicted risk axis). Nomograms outperforms simple classification in groups of risk by providing more tailored predictions for a patient using his/her specific features.

Nomograms have been widely adopted in the context of cancer prognosis (Chun *et al.*, 2006, Iasonos *et al.*, 2008, Karakiewicz *et al.*, 2006, Shabsigh and Bochner, 2006). These graphical tools can be constructed for models involving single responses such as linear models (normally distributed response), Cox proportional hazards models (single time to event endpoint) or logistic regression models (binary response). In all these cases, a direct relationship between the distribution function and the covariates is needed. Function `nomogram` of Frank Harrell's library `Hmisc` construct nomograms for these three examples of models, among others.

In the case of competing risks, the model providing a direct relationship between covariates and probability of an event to occur (cumulative incidence) is Fine and Gray's model. Kattan *et al.* (2003a,b) first proposed a nomogram based on this model for the event of interest, acknowledging for the presence of a competing event. This is achieved by obtaining, for each patient, the predicted probabilities of the event of interest given in expression (2.9), which is derived from Fine and Gray's model, for a fixed time of prediction. Once the predicted probability is correctly computed taking into account the presence of competing risks, the construction of the nomogram is similar to any other regression model, by establishing a linear relationship with a scoring system.

In Chapter 10 we present the technical details and the `R` code based in `Hsmic` routines to construct a competing risks nomogram. We will mimic Kattan's nomogram in the context of the Spanish Bladder Cancer Study later in this Chapter.

2.1.5.2 Calibration curves for competing risks

The goal of an individualized risk prediction model is to predict the outcome as accurately as possible. Calibration is a measure of predictive accuracy regarding how far predictions are from the actual outcomes (Harrell *et al.*, 1996). Calibration is assessed by reviewing the plot of predicted probabilities from the multivariate regression model versus the actual probabilities. Figure 2.2(b) shows a calibrated plot for the simulated model represented in the nomogram of Figure 2.2(a). A perfectly calibrated model would result in a plot where the observed and predicted probabilities fall along the 45-degree line, while the distance between the pairs and the 45-degree line is a measure of the absolute error of the model's prediction.

In the case of competing risks models, we have to compare predicted probabilities for the cause of failure of interest given by Fine and Gray's model and compare them to the nonparametric estimated cumulative incidence functions. The calibration plot can also be constructed based on probabilities resulting from the combination of all Cox cause-specific hazards (recall expression (2.8)).

To obtain this plot, patients are grouped with respect to their predicted probability according to the regression model (for instance, grouped by percentiles), and the mean of the group is compared with the empirical estimation of the cumulative incidence function based in data from that specific group. Bootstrap is then used to obtain confidence intervals for the predicted probability. The implemented procedure to obtain these calibration plots in `R` is detailed in Chapter 10.

2.1.6 Existing software for competing risks

Competing risks analysis can be performed in R by means of the `cmprsk` package (Gray, 2004). This package includes a function for non-parametric estimation of the cumulative incidence functions (function `cuminc`). Function `crr` implements Fine and Gray's model, and predictions under this model can be obtained by means of function `predict.crr`. Recently, It has been shown that standard errors given by the `crr` function are not optimal (Geskus, 2010) and that this model can be implemented by means of a standard Cox proportional hazards model with time-dependent weights. Thus, any software allowing for time-dependent weights could be used to fit this model. Details on his procedure can be found as supplementary material of the paper at `http://www.biometrics.tibs.org/datasets/090931M.zip`.

Cox models for the cause-specific hazards are implemented by the `coxph` function of the `survival` package (Therneau and original R port by Thomas Lumley, 2009). Predictions arising from these Cox models by means of equation 2.8 are not included in the `cmprsk` package, but they are easily implemented in this language. Further packages and functions available in `R` can be found in `http://cran.r-project.org/web/views/Survival.html`.

Regarding to SAS® software, there are not specific procedures designed to perform a competing risks analysis, though we can use existing procedures and web-available macros to implement the methodology. Non-parametric estimates of cumulative incidence functions can be obtained at the web site of the Division of Biomedical Statistics and Bioinformatics of Mayo Clinic (`http://mayoresearch.mayo.edu/mayo/research/biostat/sasmacros.cfm`); and at the web

site of the Division of Biostatistics of the Medical College of Wisconsin (`http://www.biostat.mcw.edu/software/SoftMenu.html`), among others. Cox proportional hazards models for the cause-specific hazards can be fitted by means of the `PHREG` procedure, and estimates for the cumulative incidence functions based on the fitted Cox models can be derived by using the macros written by Rosthøj *et al.* (2004) (available at (`http://staff.pubhealth.ku.dk/~pka/`). No macro nor reference was found to fit Fine and Gray's model using SAS software.

Methods to deal with competing risks analysis are not implemented in the mainstream statistical software SPSS®. However, we can take advantage of the survival analysis facilities of SPSS in order to estimate the cumulative incidence functions non-parametrically or through Cox regression model by combining cause-specific hazard and overall survival estimates by means of expression (2.8). Again, no specific procedure of the SPSS software is available to fit Fine and Gray's model.

Finally, cumulative incidence functions can be estimated using the software Stata® with the module `stcompet.ado` (Coviello and Boggess, 2004). It also implements Fine and Gray's model by its function `stcrreg` (`http://www.stata.com/stata11/stcrreg.html`).

2.1.7 Final comments on competing risks

When competing risks are present, but only one of the causes is of interest, one might be tempted to ignore the presence of competing risks and use standard survival techniques for a single time-to-event endpoint, as if the competing event had no effect on the failure of interest. The strength of the impact of the competing event on the cause of interest will depend on both the proportion of observed competing events and the dependence between the competing causes. Indeed, under different dependent causes of failure, the nonparametric Kaplan-Meier estimator is known to overestimate the real proportion of observed events of a specific type, because it ignores that some of the events of interest will be precluded by the occurrence of the competing event (Pepe and Mori, 1993, Putter *et al.*, 2007). However, the independence assumption cannot be tested based solely on observed data, and therefore, a competing risks analysis is always required.

We can also assess the impact of the competing failures through the joint interpretation of the fitted models, and on the predictions of the probability of the event of interest (which we have already explained in Section 2.1.5). For instance, assume we only have two causes of failure, and hence, that we fit Cox models to their corresponding cause-specific hazards. The model for the cause of interest (say $C = 1$), is identical to the Cox model that we would obtain if we ignored the competing event ($C = 2$). Therefore, when exploring cause-specific hazards, the magnitude of the effects is the same as ignoring competing events, but the important difference is how these estimates (and their corresponding hazard ratios) have to be interpreted. If the competing event is ignored, we will interpret that we obtain the hazard ratios of experiencing the event of interest, while when acknowledging for competing risks, these numbers will be interpreted as the hazard ratios of experiencing the event of interest *before* the competing event. This seemingly unimportant difference is essential for a correct interpretation of the effects and may give rise to strange results, difficult to interpret, such as what has been reported in the literature as 'unexpected protectivity' (Serio, 1997), where the estimated hazard ratios of well-known established risk factors are smaller

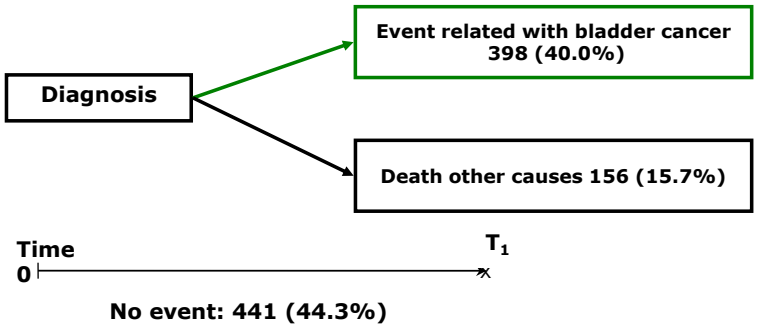


Figure 2.3: Competing risks structure for disease-related events (EFS) and Death from Other Causes

than one, as it is usual for protective factors. In next sections, we will discuss these topics for the competing risks problems found in the Spanish Bladder Cancer Study.

2.2 Analysis of Event Free Survival in the SBC/EPICURO Study

2.2.1 Competing risks for Event Free Survival

In this section, we analyze Event Free Survival taking into account that deaths due to other causes act as a competing risk for disease-related events. Let T_{DOC} be the time from diagnosis to death due to other causes (DOC), non disease-related. The notation for this competing risks problem is the following:

Notation 2.1. *The competing risks situation for the time to the first between EFS or DOC is denoted by* (T_1, C_1) *where* $T_1 = \min(T_{EFS}, T_{DOC})$ *and* C_1 *equals 1 when* $T_1 = T_{EFS}$*, or 2 if* $T_1 = T_{DOC}$*.*

Figure 2.3 represents the competing risks structure when studying disease-related events (EFS). In this situation, T_{DOC} censors T_{EFS}, because the occurrence of DOC prevents the observation of any other event. This censoring can be dependent if independence between the competing events cannot be guaranteed. The percentage of failures due to the competing event is denoted here as the percentage of (possibly) dependent censoring, 15.7% in this case. In many situations the impact of this competing event will be small and a traditional analysis ignoring this fact would provide similar results than the application of more sophisticated approaches for competing risks. Thus, a natural question is whether it is worth the additional work that a competing risk study requires, both in performance and interpretation, or other causes of death can simply be ignored. The degree of impact of the competing event on the analysis will depend on both the proportion of observed competing events and the dependence between the competing causes. We will try to answer this question for Event Free Survival in the Spanish Bladder Cancer/EPICURO Study. To do so, we perform a thoroughly competing risks analysis: we describe the joint distribution (T_1, C_1) by means of a nonparametric description of the cumulative incidence functions in Section 2.2.2, and next by regression modelling in Section 2.2.3. A similar question will be addressed in Section 2.3 for the study of Progression Free Survival.

2.2.2 Prognostic Factors for Event Free Survival. Univariate nonparametric analysis.

We have estimated the cumulative incidence functions for each type of event (EFS and DOC), comparing them across different stratums of the risk factors considered: gender, age, and smoking status at diagnose of the patient, and number, size, grade and stage of the tumour.

We have tested differences between curves by means of a nonparametric test proposed by Gray (1988), which generalizes the log-rank test to compare survival curves for two or more populations for standard right-censored data. The results of these tests are given in Table 2.1. Figures 2.4a to 2.4d contain the plots for those variables that are significant at a 0.05 level for Event Free Survival (gender, tumour number and grade) as well as for smoking status. The cumulative curves for the other covariates can be found in Appendix A.1. We discuss in detail the results for each factor:

Gender: Figure 2.4a shows the cumulative incidence curves of EFS and DOC by gender. The incidence of disease-related events is significatively higher in females than in males (p-value 0.0028).

This finding is very relevant and specific of the Spanish Bladder Cancer Study. In a recent meta-

Table 2.1: Results of Gray's significance test to compare cumulative incidence curves across stratum for Event Free Survival

	Significance test p-values	
Factor	**EFS**	**DOC**
Gender (Female/Male)	0.0028	0.0002
Age (<=60/61-70/>70)	0.6851	<0.0001
Tumour number (Multiple/Single)	<0.0001	0.0302
Tumour size (>= 3cm/<3cm)	0.0560	0.7431
Stage (T1+Tis/Ta)	0.1482	0.0083
Grade (G1+Benign/G2/G3)	0.0103	0.0616
Smoking status (smoker/non-smoker)	0.1160	0.0003

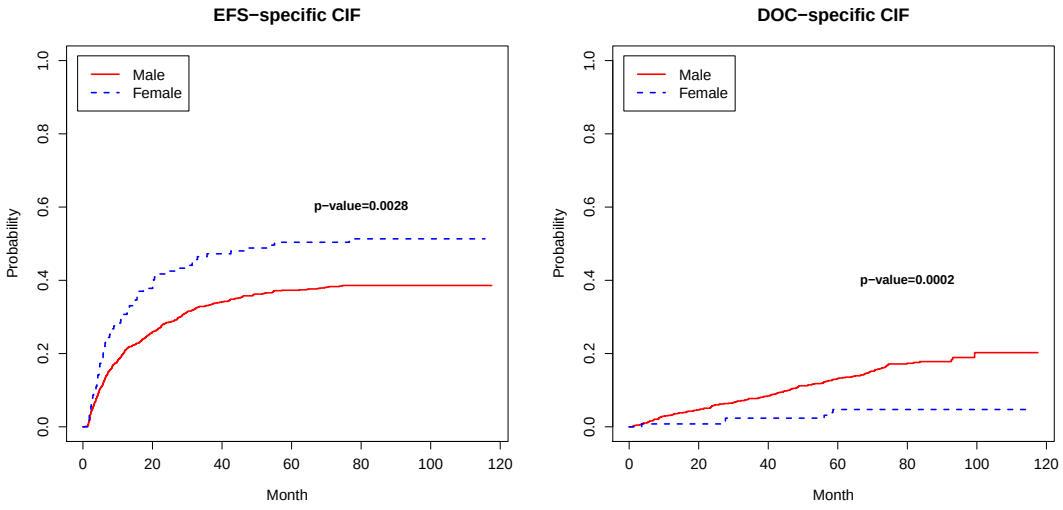


Figure 2.4a: Cumulative incidence functions for (T_1, C_1) by Gender

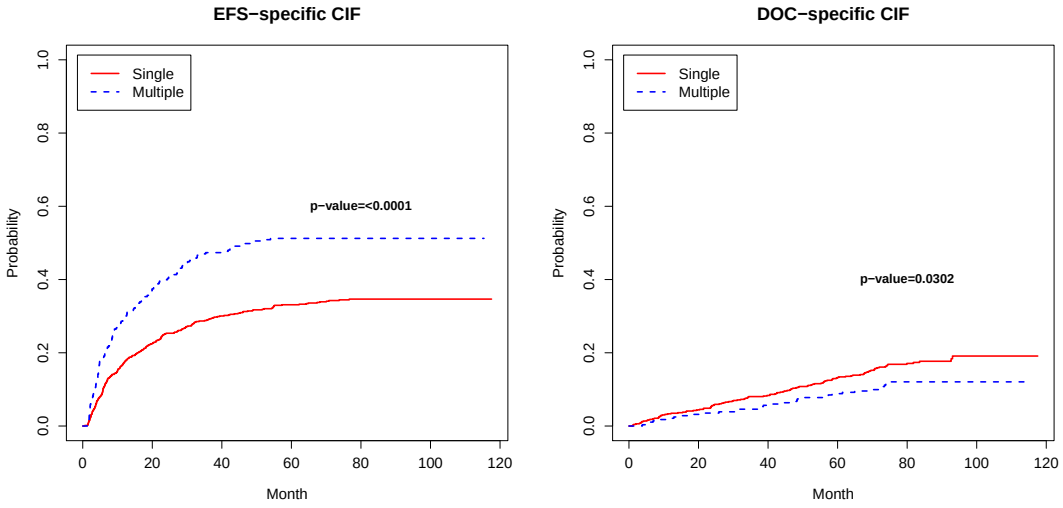


Figure 2.4b: Cumulative incidence functions for (T_1, C_1) by Tumour number

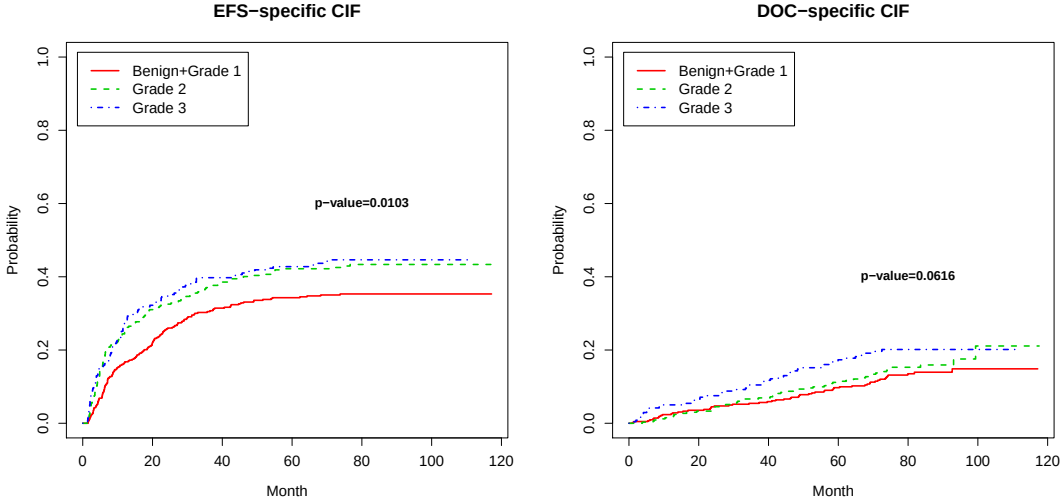


Figure 2.4c: Cumulative incidence functions for (T_1, C_1) by Grade

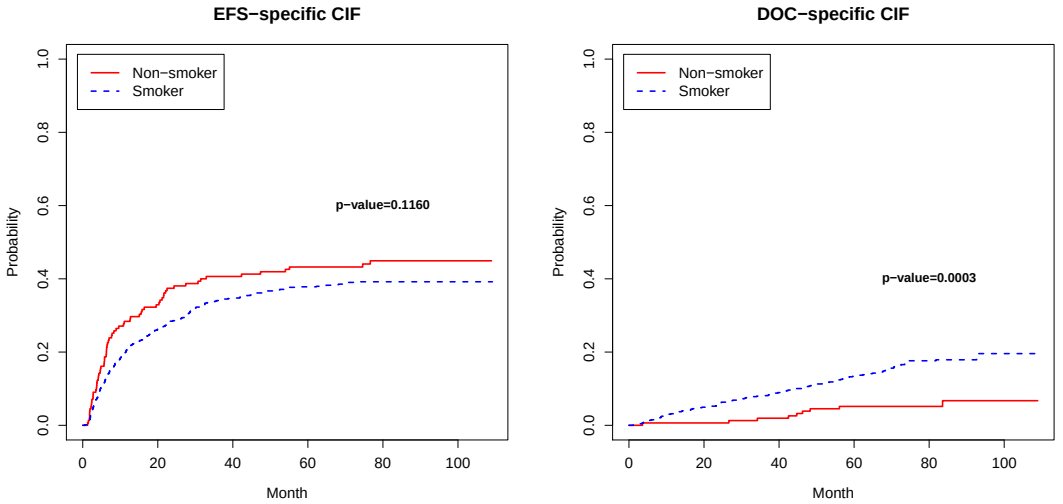


Figure 2.4d: Cumulative incidence functions for (T_1, C_1) by Smoking status

analysis to identify prognostic factors for recurrence and progression in non-muscle invasive bladder cancer, none of the 5 referred studies that included gender in the analysis found this factor as a significant prognostic factor for recurrence (van der Aa *et al.*, 2009). Given that the incidence of deaths due to other causes (DOC) is higher in males than in females (p-value 0.0002), one could speculate on what would happen if deaths from other causes would be prevented, and whether the apparent effect of gender on EFS would be manifest then. In general, this question could not be answered in this framework, because the latent marginal distributions are unidentifiable from competing risks data.

Nonetheless, as described in Section 1.2.2, the majority of patients in our sample are males (87.2%), and among them, 91% are smokers. The strong relationship of smoking status and gender has to be addressed appropriately since it can induce a confounding effect. In particular, we observe in Figure 2.4d that the cumulative incidence of EFS is higher in non-smokers (though non-significantly). We wonder if this is a genuine behaviour of smoking or, since most smokers are men and non-smokers are women, this is just reflecting the differences in EFS by gender. Moreover, similar curves for DOC are found for males (Figure 2.4a) and smokers (Figure 2.4d).

In order to clarify this possible confusion we reanalyze EFS and DOC by gender stratifying by smoking status (Figure 2.5). Within non-smokers, few competing events are present: this fact suggests that the observed differences in EFS between males and females are genuine and not caused by the effect of the competing event. Within smokers, more competing deaths are observed but the cumulative incidence curves of EFS for males and females are similar to those of non-smokers. These explorations suggest that differences in EFS are explained by gender while no differences are manifest by smoking status. They also suggest that the observed differences in mortality before relapse by gender are actually due to smoking. The multivariate regression analysis of next section also confirms these conclusions.

Age: There are no significant differences in EFS across stratum, but there are differences in DOC, where the older the patient, the higher the probability of dying for other causes than to develop disease (p-value$<$0.0001).

Tumour number: Multiple tumours (p-value$<$0.0001) exhibit a higher incidence of disease-related events than single tumours (Figure 2.4b). Consequently, patients with single tumours have more opportunities to be observed dying from other causes before a relapse (p-value 0.0302).

Grade: Grade 2 and 3 tumours (0.0103) have higher rates of relapse than Benign/Grade 1 tumours (Figure 2.4c).

Size and Stage: There are no significant differences in size or in stage.

2.2.3 Multivariate regression model for Event Free Survival.

After the univariate analysis of each covariate we proceed with the analysis of their joint effect on Event Free Survival. We present in Tables 2.2 and 2.3 the hazard ratios (HR), the 95% confidence

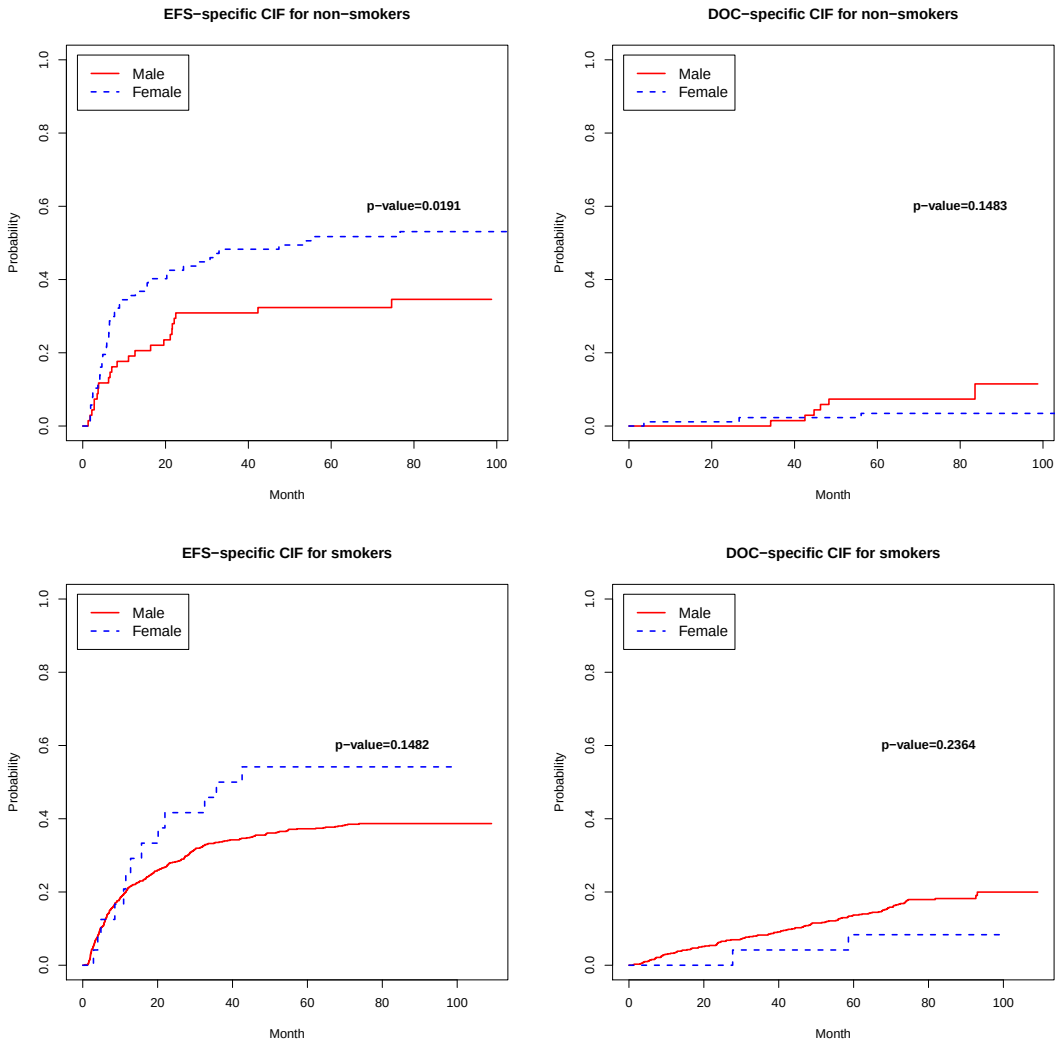


Figure 2.5: Cumulative incidence functions for (T_1, C_1) across Gender, for smokers and non-smokers.

intervals of the HR (CI95%) and the p-values of the fitted regression models for (T_1, C_1): proportional hazards Cox models for the cause-specific hazards of EFS and DOC, referred in the following as CSH models, and Fine and Gray's model for the subhazard of the cumulative incidence function of EFS, to which we will refer as FGH model.

Both approaches confirm gender, multiplicity and grade (Grade 2 vs Grade 1 plus Benign tumours) as the most important prognostic factors for Event Free Survival. Multiplicity and grade are common prognostic factors for recurrence in the literature (Babjuk *et al.*, 2008, van der Aa *et al.*, 2009) while, as mentioned before, the effect of gender is specific of the SBC/EPICURO study and raises the question of why this higher risk of recurrence among Spanish women. Tumour size and Grade 3 vs G1+Benign are also well established prognostic factors for recurrence that, in our study, do not yield strictly significance. However, with a p-value of 0.07, our results are in the line of most studies, suggesting the same tendency: larger and G3 primary tumours are more likely to relapse during the patient's follow-up.

When it comes to smoking status, it is interesting to recover the discussion of the previous section. Indeed, note how the obtained hazard ratios, though not significant for both models for EFS, are

Table 2.2: Cause-specific hazards (CSH) Cox models for (T_1, C_1)

	EFS			DOC		
Factor	**HR**	**CI95%**	**p-value**	**HR**	**CI95%**	**p-value**
Gender (Female vs Male)	1.701	(1.172, 2.469)	0.005	0.579	(0.195, 1.716)	0.324
Age (by year)	0.999	(0.988, 1.010)	0.837	1.087	(1.059, 1.116)	0.000
Tumour number (Multiple vs single)	1.566	(1.248, 1.965)	0.000	0.632	(0.408, 0.978)	0.040
Tumour size ($\geq$ 3cm vs $\leq$ 3 cm)	1.298	(0.970, 1.737)	0.079	1.275	(0.786, 2.069)	0.326
Stage (T1Tis vs Ta)	0.971	(0.677, 1.392)	0.873	2.003	(1.121, 3.578)	0.019
Grade (G2 vs G1+Benign)	1.383	(1.069, 1.790)	0.013	1.180	(0.776, 1.795)	0.439
Grade (G3 vs G1+Benign)	1.374	(0.969, 1.947)	0.075	1.104	(0.614, 1.987)	0.741
Smoker vs Non-smoker†	1.149	(0.808, 1.634)	0.439	2.654	(1.164, 6.056)	0.020

†Smoker includes current and former smokers. Non-smoker includes occasional smokers.

Table 2.3: Fine and Gray (FGH) models of the subhazards for (T_1, C_1)

	EFS			DOC		
Factor	**HR**	**CI95%‡**	**p-value**	**HR**	**CI95%‡**	**p-value**
Gender (Female vs Male)	1.718	(1.182, 2.497)	0.005	0.498	(0.168, 1.478)	0.209
Age (by year)	0.996	(0.985, 1.007)	0.497	1.080	(1.054, 1.107)	< 0.001
Tumour number (Multiple vs single)	1.613	(1.286, 2.022)	< 0.001	0.519	(0.335, 0.803)	0.003
Tumour size ($\geq$ 3cm vs $\leq$ 3 cm)	1.280	(0.956, 1.713)	0.098	1.111	(0.685, 1.801)	0.670
Stage (T1Tis vs Ta)	0.934	(0.651, 1.341)	0.711	1.655	(0.945, 2.898)	0.078
Grade (G2 vs G1+Benign)	1.373	(1.061, 1.777)	0.016	1.054	(0.692, 1.605)	0.806
Grade (G3 vs G1+Benign)	1.346	(0.949, 1.911)	0.096	1.106	(0.626, 1.955)	0.728
Smoker vs Non-smoker†	1.108	(0.779, 1.577)	0.568	2.546	(1.113, 5.821)	0.027

†Smoker includes current and former smokers. Non-smoker includes occasional smokers.

‡FGH model fitted according to Geskus (2010).

greater than one in both cases, showing that smokers are at higher risk of a BC event than non-smokers. Recall from Figure 2.4d how non-smokers exhibit a higher probability of EFS-specific failures. When we adjust for other covariates, the more natural direction of effect is recovered, probably corrected by the inclusion of the confounding effect of gender.

Apart from the identification of the prognostic factors, it is interesting to look at how the presence of a competing event affects the coefficients of both the CSH model and the FGH model. Note that the coefficients and p-values in the first two columns of Table 2.2, corresponding to a Cox model for the cause-specific hazard (CSH model) for EFS, are identical to the coefficients of the standard Cox model that we would obtain if we ignore the competing event DOC by treating these deaths as censored observations. Instead, the coefficients of the Fine and Gray model (FGH model) include the effect of the competing event. In this study, the coefficients and significance from both models are fairly identical and the same conclusions for Event Free Survival would be driven if competing risks would have been ignored. As we will see in next section when analyzing Progression Free Survival, both approaches not always agree, and the competing event can affect the significance of some prognostic factors.

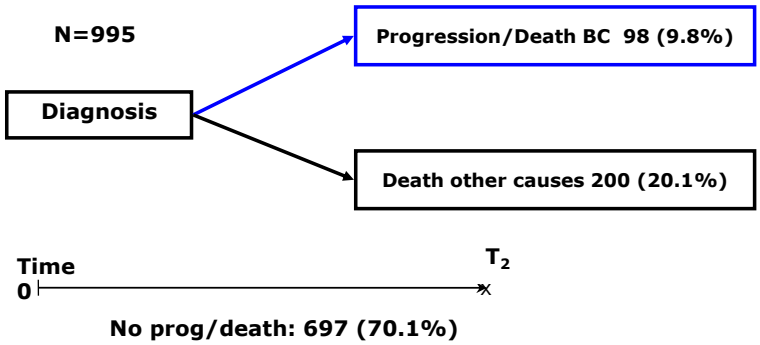


Figure 2.6: Competing risks structure for progression of disease (PFS) and Death from Other Causes

The validity of the fitted Cox models was checked by assessing their residuals: procedures based on score residuals were used to assess the proportional hazards assumption, and martingale residuals were used to test whether a transformation was needed to deal with the continuous effect of age. The same checking was performed when analysing Progression Free Survival (Section 2.3) and time to first event (Section 2.4).

2.3 Analysis for Progression Free Survival in the SBC/EPICURO Study

2.3.1 Competing risks for Progression Free Survival

In this section, we focus on Progression Free Survival taking into account that non disease-related deaths act as a competing event for progression of disease. This competing risks problem is formally defined as:

Notation 2.2. *The competing risks situation for the time to the first between PFS or DOC is denoted by* (T_2, C_2) *with* $T_2 = \min(T_{PFS}, T_{DOC})$ *and* C_2 *equals 1 if* $T_2 = T_{PFS}$*, 2 otherwise.*

Figure 2.6 represents the competing risks structure for progression of disease (PFS) and DOC. Again, T_{DOC} censors T_{PFS}, because the occurrence of DOC prevents the observation of progression. The percentage of (possibly) dependent censoring in this case is 20.1%.

In the following, we address the question of whether the competing risk has an impact on the estimation of Progression Free Survival, or, on the contrary, it could have been ignored. Again, we first explore nonparametrically the covariates with their cumulative incidence functions (Section 2.3.2), and then perform a regression modelling to identify the risk profile for PFS (Section 2.3.3). We present, in addition, an illustration of the computation of predictive probabilities of progression in Section 2.3.4.

2.3.2 Prognostic factors for Progression Free Survival. Nonparametric analysis.

We have estimated the cumulative incidence functions for each type of event (PFS or DOC), comparing them across different stratums of the risk factors considered: gender, age, and smoking

Table 2.4: Results of Gray's significance test to compare cumulative incidence curves across stratum for Progression Free Survival

	Significance test p-values	
Factor	**PFS**	**DOC**
Gender (Female/Male)	0.4101	0.0011
Age (<=60/61-70/>70)	0.0030	<0.0001
Tumour number (Multiple/Single)	0.0014	0.6375
Tumour size (>= 3cm/<3cm)	0.3350	0.8394
Stage (T1+Tis/Ta)	<0.0001	0.4814
Grade (G1+Benign/G2/G3)	<0.0001	0.1130
Smoking status (smoker/non-smoker)	0.3478	0.0020

status at diagnose of the patient, and tumour number, size, grade and stage of the tumour. The results of the nonparametric test proposed by Gray (1988) are given in Table 2.4. Figures 2.7a to 2.7d contain the plots for those variables significant at a 0.05 level for Progression Free Survival (age, tumour number, stage and grade). The cumulative curves for the other covariates can be found in Appendix A.2. In the following we discuss the results for each factor.

Gender: There are no significant differences between males and females in the incidence of progression of disease (p-value 0.4101). Though still the curve for females is higher than the curve for males, if we take into account the results for EFS, where differences existed, it seems that females have a higher incidence of recurrence than males. This fact will be checked when characterizing the time to the first event in Section 2.4.

Age: Age results in significant differences (Figure 2.7a): older patients are more incident not only in deaths from other causes (p-value< 0.0001) but also on progression of disease (p-value 0.0030).

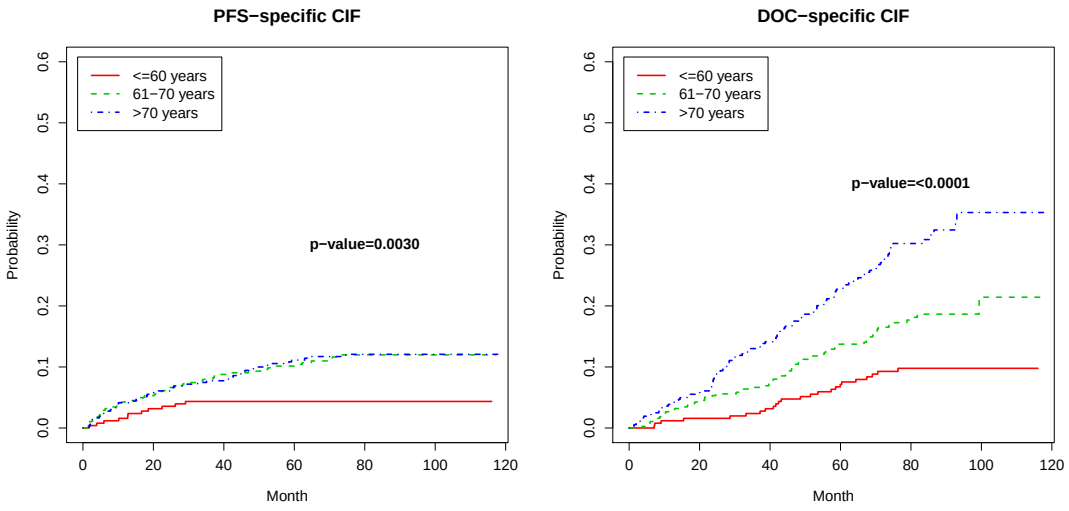


Figure 2.7a: Cumulative incidence functions for (T_2, C_2) by Age

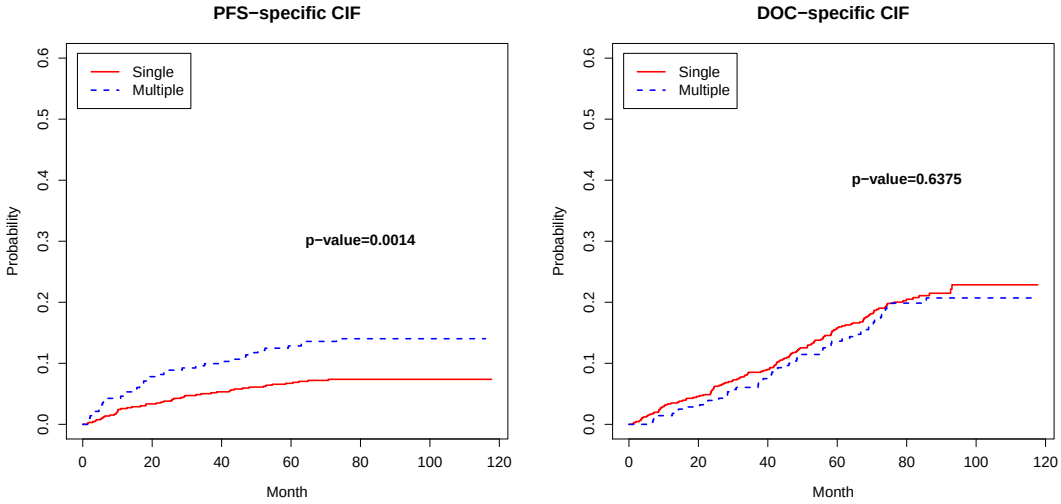


Figure 2.7b: Cumulative incidence functions for (T_2, C_2) by Tumour number

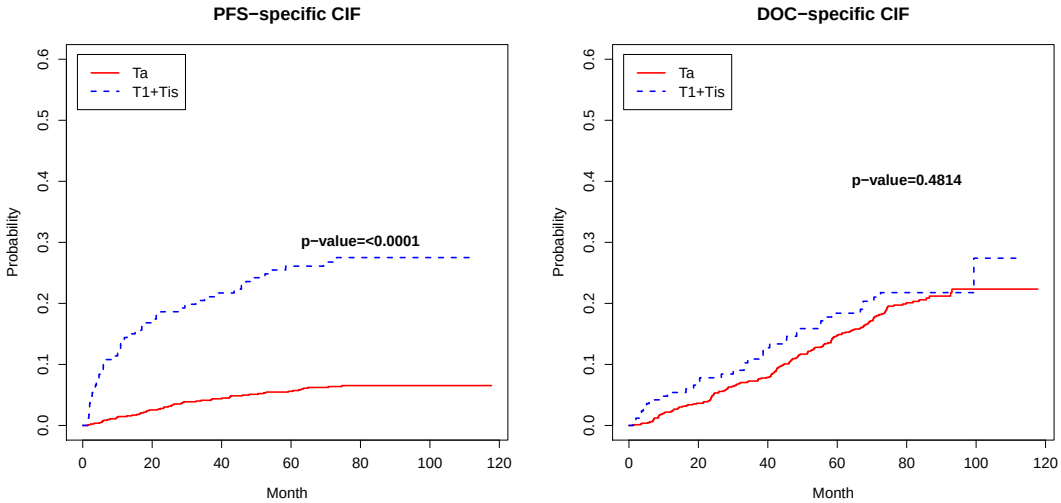


Figure 2.7c: Cumulative incidence functions for (T_2, C_2) by Stage

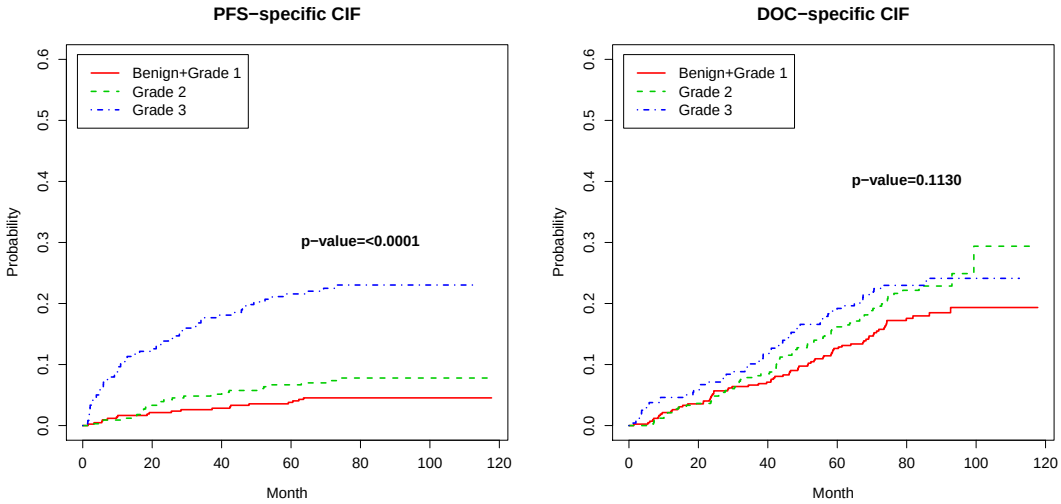


Figure 2.7d: Cumulative incidence functions for (T_2, C_2) by Grade

Tumour number: Multiple tumours (p-value 0.0014) have a higher incidence of progression of disease (Figure 2.7b).

Size: There are no significant differences in size of the tumour.

Stage: Figure 2.7c reveals that patients with stage T1 or Tis tumours have a significantly higher incidence of progression of disease (p-value < 0.0001) before dying from other causes. No differences in the incidence of DOC where found.

Grade: Grade 3 tumours have higher rates of progression of disease than Benign/Grade 1 and Grade 2 tumours (Figure 2.7d, p-value < 0.0001).

Smoking status: There are no significant differences between smokers and non-smokers. In numerical terms, though, non-smokers present a higher incidence of progression than smokers (see Appendix A.2), probably justified by the presence of more non BC deaths in smokers (p-value 0.0020).

2.3.3 Multivariate regression model for Progression Free Survival

Now we explore the joint contribution of the covariates on Progression Free Survival. That is, we fit regression models for (T_2, C_2), by applying the methods presented in Section 2.1.4. In Tables 2.5 and 2.6 we show, respectively, the estimated hazard ratios for the Cox Cause-Specific Hazard (CSH) and the Fine and Gray Hazard (FGH) models for PFS and DOC failures.

Both modelling approaches provide similar results for the considered covariates and confirm that the main prognostic factors for progression in the SBC/EPICURO Study are age, tumour number, stage and grade 3. These are well established prognostic factors in most bladder cancer studies (van der Aa *et al.*, 2009). This agreement also applies in the magnitude of the effect of the prognostic factors. The most important effect is provided by grade 3 (HR=3.25, p-value=0.0018), followed by stage (HR=2.33, p-value=0.0049). Unlike other studies, we do not find significance for the effect of tumour size, though the magnitude of the effect is on the expected direction (HR=1.36, p-value=0.28).

Though smoking status resulted not significant, the direction of the effect goes in the 'natural' sense: the hazard ratio of 1.18 suggests that smokers have a higher incidence of progression than non-smokers. Again, the adjusted analysis has corrected the unexpected direction of the smoking effect in the univariate analysis, where non-smokers exhibited a higher progression rate.

The significance and the magnitude of the effects of interest can be affected by the competing cause of failure. To merely illustrate this fact, because the coefficients for PFS in both CSH and FGH models are very similar, observe the results for age in this study. Age is a strong risk factor for death from other causes (HR= 1.0868, p-value< 0.0001). According to the CSH Cox model, age is also a significant risk factor for progression (HR= 1.0364, p-value= 0.0163), however, according to the FGH model, age is at the limit of significance (HR= 1.029; p = 0.048) at the usual 0.05 level.

Table 2.5: Cause-specific hazards (CSH) Cox models for (T_2, C_2)

	PFS			DOC		
Factor	**HR**	**CI95%**	**p-value**	**HR**	**CI95%**	**p-value**
Gender (Female vs Male)	1.344	(0.601, 3.004)	0.471	0.764	(0.343, 1.702)	0.510
Age (by year)	1.036	(1.007, 1.067)	0.016	1.087	(1.063, 1.111)	< 0.001
Tumour number (Multiple vs single)	1.548	(0.970, 2.470)	0.067	0.813	(0.574, 1.152)	0.244
Tumour size ($\geq$ 3cm vs $\leq$ 3 cm)	1.367	(0.771, 2.423)	0.284	1.117	(0.732, 1.705)	0.608
Stage (T1Tis vs Ta)	2.334	(1.294, 4.212)	0.005	1.219	(0.736, 2.017)	0.442
Grade (G2 vs G1+Benign)	1.204	(0.587, 2.468)	0.613	1.158	(0.807, 1.662)	0.427
Grade (G3 vs G1+Benign)	3.258	(1.555, 6.830)	0.002	1.324	(0.824, 2.128)	0.247
Smoker vs Non-smoker†	1.182	(0.553, 2.526)	0.666	2.157	(1.116, 4.169)	0.022

†Smoker includes current and former smokers. Non-smoker includes occasional smokers.

Table 2.6: Fine and Gray (FGH) models for the subhazards of (T_2, C_2)

	PFS			DOC		
Factor	**HR**	**CI95%‡**	**p-value**	**HR**	**CI95%‡**	**p-value**
Gender (Female vs Male)	1.344	(0.598, 3.022)	0.475	0.744	(0.337, 1.645)	0.465
Age (by year)	1.029	(1.000, 1.059)	0.048	1.083	(1.059, 1.107)	< 0.001
Tumour number (Multiple vs single)	1.685	(1.058, 2.683)	0.028	0.766	(0.542, 1.084)	0.133
Tumour size ($\geq$ 3cm vs $\leq$ 3 cm)	1.362	(0.768, 2.415)	0.291	1.095	(0.717, 1.673)	0.675
Stage (T1Tis vs Ta)	2.240	(1.238, 4.053)	0.008	0.984	(0.590, 1.640)	0.950
Grade (G2 vs G1+Benign)	1.198	(0.584, 2.456)	0.622	1.183	(0.825, 1.697)	0.361
Grade (G3 vs G1+Benign)	3.147	(1.498, 6.613)	0.002	1.271	(0.787, 2.053)	0.326
Smoker vs Non-smoker†	1.153	(0.536, 2.478)	0.716	2.117	(1.102, 4.066)	0.024

†Smoker includes current and former smokers. Non-smoker includes occasional smokers.

‡FGH model fitted according to Geskus (2010).

This result is indicating that though age is a risk factor for progression, we will actually observe similar rates of progressions in older people than in younger, because older people tend to die for other causes earlier, before being able to experience the bladder cancer progression.

Thus, the presence of a competing event may modify the effects on the event of interest and their significance and should be appropriately addressed for a correct interpretation of results. Furthermore, the existence of a competing cause of failure may have an important impact on the estimated probability of the effect of interest. In next section, we show how to obtain accurate predictions of the probability of progression while taking into account the competing event (DOC) and compare the predictions with the ones obtained if ignoring competing risks.

2.3.4 Prediction of the probability of progression

Given the results of the multivariate regression analysis for PFS, we will describe three possible strategies for estimating the probability of progression, one that ignores the existence of a competing event and two alternatives that take competing risks into account.

Denote by $\widehat{\boldsymbol{\beta}}$ the coefficients for PFS of the CSH model, by $\widehat{\boldsymbol{\gamma}}$ the coefficients for DOC of the CSH model and by $\widehat{\boldsymbol{\beta}}^*$ the coefficients for PFS of the FGH model. Let $\boldsymbol{Z}$ be the vector of prognostic factors.

(a) The first method, denoted by Cox, is the standard Cox model ignoring competing risks that estimates the probability of the events using exclusively the coefficients $\widehat{\boldsymbol{\beta}}$ for the cause-specific hazard of PFS:

$$\widehat{F}_{Cox}(t|\widehat{\boldsymbol{\beta}}, \boldsymbol{Z}) = \int_0^t \widehat{\lambda}_{PFS}(u|\widehat{\boldsymbol{\beta}}, \boldsymbol{Z}) \exp\{-\widehat{\Lambda}_{PFS}(u|\widehat{\boldsymbol{\beta}}, \boldsymbol{Z})\} du,$$

where $\widehat{\lambda}_{PFS}(u|\widehat{\boldsymbol{\beta}}, \boldsymbol{Z}) = \widehat{\lambda}_{PFS,0}(u) \exp\{\widehat{\boldsymbol{\beta}}' \boldsymbol{Z}\}$ is the CSH model for PFS, and $\widehat{\Lambda}_{PFS}(t|\widehat{\boldsymbol{\beta}}, \boldsymbol{Z})$ is its corresponding cause-specific cumulative hazards.

(b) The second approach, denoted by CSH, explicitly combines the results of the estimated cause-specific hazards of all competing events, in our case, β for PFS and γ for DOC:

$$\widehat{F}_{CSH}(t|\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \boldsymbol{Z}) = \int_0^t \widehat{\lambda}_{PFS}(u|\widehat{\boldsymbol{\beta}}, \boldsymbol{Z}) \exp\{-(\widehat{\Lambda}_{PFS}(u|\widehat{\boldsymbol{\beta}}, \boldsymbol{Z}) + \widehat{\Lambda}_{DOC}(u|\widehat{\boldsymbol{\gamma}}, \boldsymbol{Z}))\} du,$$

where $\widehat{\Lambda}_{DOC}(t|\widehat{\boldsymbol{\beta}}, \boldsymbol{Z})$ is the cause-specific cumulative hazard corresponding to the model fitted for DOC.

(c) The third approach, denoted by FGH, uses the coefficients $\widehat{\boldsymbol{\beta}}^*$ from Fine-Gray model, that already were obtained taking into account the existence of competing events:

$$\widehat{F}_{FGH}(t|\widehat{\boldsymbol{\beta}}^*, \boldsymbol{Z}) = 1 - \exp\big(-\int_0^t \widehat{\gamma}_{PFS}(u|\widehat{\boldsymbol{\beta}}^{*'} \boldsymbol{Z}) du\big),$$

where $\widehat{\gamma}_{PFS}(u|\widehat{\boldsymbol{\beta}}^{*'} \boldsymbol{Z})$ is the subhazard derived from the cumulative incidence function for PFS.

Though the second and third approach give similar results, the FGH approach is more handy, because it involves only one coefficient per variable. Moreover, it establishes a direct relationship between covariates and probability of progression, and thus it allows for graphical representations, such as the nomogram in Figure 2.8. With this nomogram, and following a scoring system, one can easily obtain the predicted probability of progression before 5 years taking into account the presence of a competing event.

We remark here that, though in the bladder cancer data approaches (b) and (c) lead to similar results, this is not true in general. Both approaches rely on different proportional hazards model which may not hold at the same time or may provide different results in practice. Indeed, suppose we are interested in obtaining the predicted probability of progressing before 5 years for a women, aged 72 years-old, with a solitary tumour of more than 3 cm of diameter, classified as T1 Grade II, and non-smoker. First, we locate the patient's gender on the Gender axis, and draw a straight line up to the points to determine how many points towards progression a female should receive: it results in 27 points (approximately). Then locate the patient's age on the age axis, and, again, draw

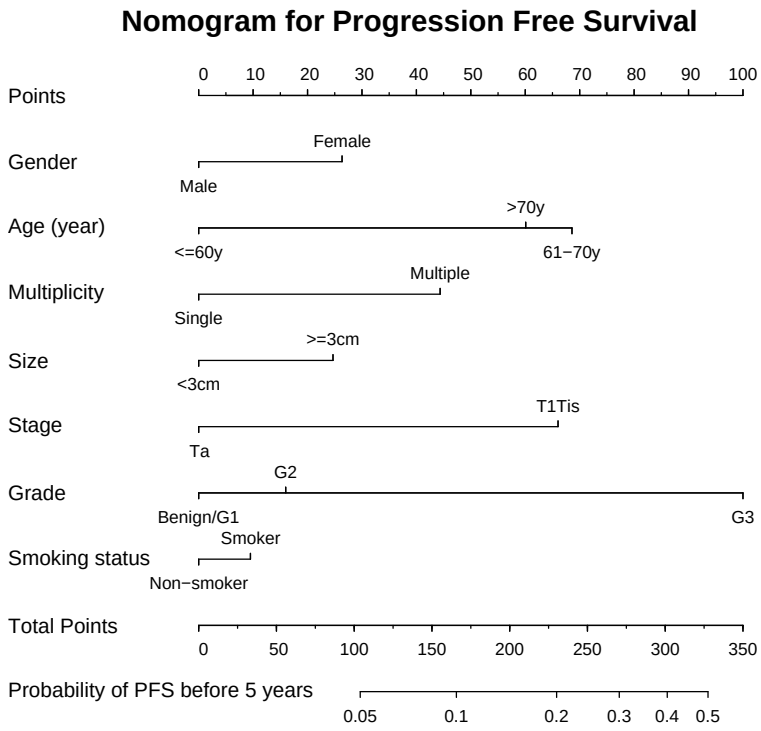
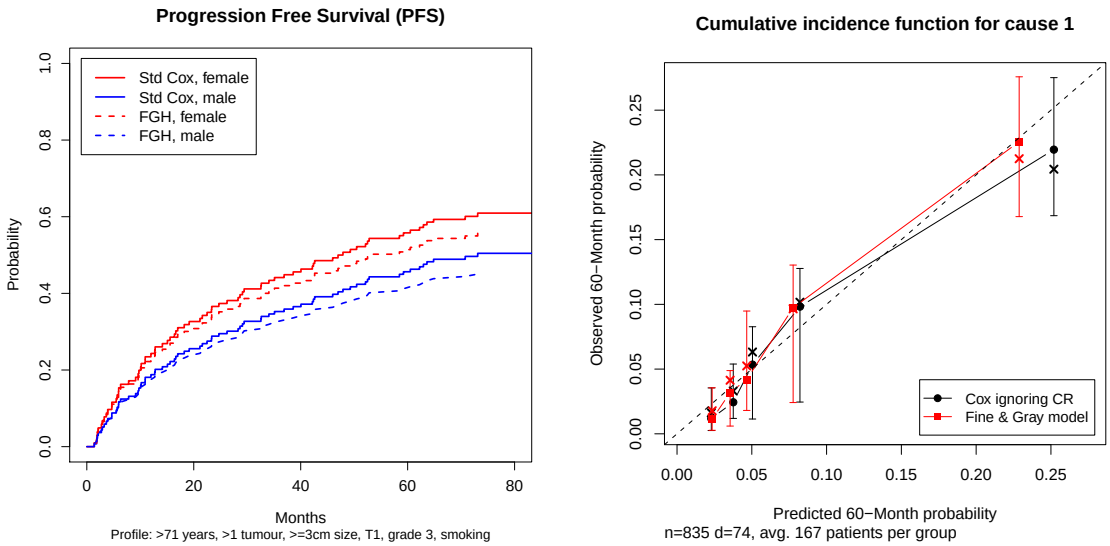


Figure 2.8: Nomogram for the predicted probability of prediction at five years accounting for competing risks.

a straight line up to the points' axis to obtain the corresponding punctuation: 60 points. Repeat this process for each of the remaining axes, drawing a straight line each time to the points axis: for a solitary tumour, 0 points; for size greater than 3 cm, 25 points; for stage T1, 67 points; for grade 2, 15 points; for non-smoker, 0 points. Now we sum the points received for each prognostic factor (194 points), and locate this score on the total points axis. Then, we draw a straight line down from total points to the 5-years probability of PFS axis to obtain the patients' predicted probability to progress within five years, which, for 194 points, it corresponds to a probability of 0.14 approximately. Details on system to assign the punctuation in a nomogram is given in Appendix D.1.2.

We center now our attention on exploring the differences between predictions from the standard Cox approach ignoring competing risks ($\widehat{F}_{Cox}(t)$) and predictions from the Fine and Gray approach ($\widehat{F}_{FGH}(t)$).

In Figure 2.9(a) we compare $\widehat{F}_{FGH}(t)$ (dashed line) with $1 - \widehat{S}_{Cox}(t)$ (solid line) for progression for a smoking person older than 71 years, with multiple tumours, size of the main tumour being larger than 3 cm, stage T1 and Grade 3, by gender (red for female and blue for male). It can be observed that predictions based on the Cox model, ignoring competing risks, slightly overestimate the probability of observing the event which is correctly estimated with the Fine-Gray approach. In particular, if we focus on predictions of the probability of progression during the first 5 years from diagnosis, the predicted probability of progression for the considered individual profile is larger when the competing risk from other causes of death is ignored (for males, 0.442 vs 0.412, for females, 0.544 vs 0.514). However, the magnitude of the overestimation is small and probably



(a) Predicted probability of progression ignoring CR (b) Calibration curves for $\widehat{F}_{Cox}(t)$ and $\widehat{F}_{FGH}(t)$.

Figure 2.9: Comparisons between the standard Cox model $\widehat{F}_{Cox}(t)$ and the Fine and Gray model $\widehat{F}_{FGH}(t)$ in terms of prediction of the probability of progression.

both models would yield to similar conclusions.

In order to further evaluate the impact of ignoring competing risks we can compare the performance of both predictive models, as proposed in Wolbers *et al.* (2009), with the usual measures for predictive ability, appropriately adapted to the competing risks framework. In particular, the calibration plot provides a tool for visualizing the agreement between predicted and observed events. Figure 2.9(b) is the calibration plot for predictions of the probability of progression during the first 5 years from Fine and Gray model (red) and Cox model ignoring competing risks (black). Departures from the diagonal indicate worse predictions, and this case, the standard Cox model performs slightly worse than the Fine and Gray's approach, specially for the higher predictions.

2.4 Characterization of the first relapse

2.4.1 Motivation

The characterization of the first relapse after diagnosis is also a relevant issue. Though all bladder cancer patients follow similar follow-up protocols, the aggressiveness of the disease is very different and some patients with a superficial primary tumour have a first relapse in the form of an invasive tumour. Characterizing the risk factors for these first-event progressions is very important in order to define a more strict follow-up for these patients. Moreover, if identified, these patients could be treated more aggressively.

Table 2.7 displays the median transition times from diagnosis to each event of interest. The median time to develop a recurrence or a progression as first event is similar, equal to 10.33 and 10.04 months, respectively. This fact discards the possibility that these first event progressions are due to an inappropriate follow-up, and suggests that there exist distinct courses and/or aggressiveness

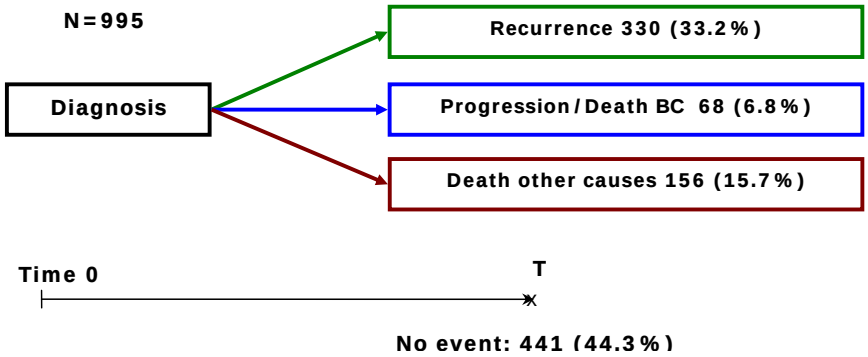


Figure 2.10: Competing risks structure for the time to the first event in the SBC/EPICURO Study

of the tumour development.

Table 2.7: Median† time between events

Transition	Median	Q25%	Q75%
Diagnosis $\longrightarrow$ Recurrence	10.33	4.500	22.99
Diagnosis $\longrightarrow$ Progression	10.04	3.780	25.46
Diagnosis $\longrightarrow$ Death BC	19.15	15.080	32.56
Diagnosis $\longrightarrow$ No event	82.73	75.530	91.07

†Median times computed among those patients with observed events.

Again, a competing risk analysis is the appropriate tool for this problem:

Notation 2.3. *The competing risks situation for the time to the first between recurrence, progression/death from BC or DOC is defined by* (T_1, C_1^*)*, where* $T_1 = \min(T_R, T_{PFS}, T_{DOC})$*, is the time to the occurrence of the first event observed,* T_R *is the time to the first recurrence, and* C_1^* *equals 1 if* $T_1 = T_R$*, 2 if* $T_1 = T_{PFS}$ *and 3 is* $T_1 = T_{DOC}$*.*

In the SBC/EPICURO Study about 33% of the patients experienced a recurrence as a first relapse, 7% were progressions or BC-deaths and 16% died from other causes (see Figure (2.10)). In the following section, we provide the risk profile of patients that will progress as a first event.

2.4.2 Prognostic factors for the first event

We first obtain the nonparametric estimates of the cumulative incidence functions specific of each cause of failure. Figure 2.11 shows the estimated cumulative incidence functions for each kind of failure. We observe that the cumulative incidence of experiencing a recurrence is higher than the incidence of experiencing a progression or death. Notice that the risk of dying from other causes increases more rapidly than the risk of progressing along with time; this is due to the risk of dying of other causes increasing with age in our cohort. Progression or death due to bladder cancer as

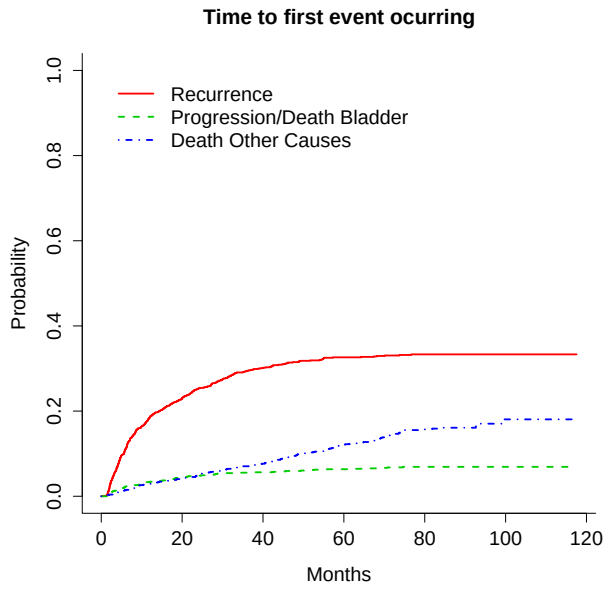


Figure 2.11: Nonparametric estimates of the the cumulative incidence functions for (T_1, C_1^*)

a first event has a low cumulative incidence of occurring. The curves across stratums of the risk factors considered can be found in Appendix A.3.

Regression models are used in order to identify prognostic factors which characterize and differentiate patients who progress from those experiencing a recurrence as a firsts event. In Table 2.8 we present the Cox proportional hazards model fitted to our data to describe the cause-specific hazards for each cause. In Table 2.9 we present the results for the Fine and Gray models for the subhazard of the cumulative incidence functions. Both approaches give similar results, and we will use mainly FGH since it implicitly contains the effects of the other competing events and it informs on cumulative incidence functions, that is, on the incidences that we will actually observe.

Apart from acknowledging the usual prognostic factors for progression (age, multiplicity, stage and grade 3), though only stage is significant, the main finding of this analysis is gender. Gender is not a prognostic factor for progression but it turns up for progression as a first observed event with women having more than twice the risk of men of having a progression before any other event (FGH HR=2.269, p-value=0.075). This cannot be explained by the fact that males could die more frequently due to other causes as a first event than women since the effect of gender in DOC is not significant (FGH HR: 0.498, p-value=0.209).

Notice that the issue of correct interpretation of the coefficients is also present here. Regarding recurrence as a first event, both CSH and FGH models give a hazard ratio for stage lower than one (0.639 and 0.554, respectively). An incorrect interpretation of these results will conclude that tumours with deeper invasion of the bladder protect against recurrences, while the correct explanation for the obtained hazard ratios is that less T1/Tis tumours experience recurrence because progression or death due to bladder cancer occur first (CSH HR: 3.025, FGH HR: 3.123).

To conclude, individuals at higher risk of observing a progression *as a first event* are females, with stage T1/Tis and Grade 3 tumours. A more accurate follow-up should be planned for these patients in order to prevent the aggressive course of the tumour.

Table 2.8: Cause-specific hazards (CSH) Cox models for (T_1, C_1^*)

	Recurrence			Pr+DBC			DOC		
Factor	**HR**	**CI95%**	**p-value**	**HR**	**CI95%**	**p-value**	**HR**	**CI95%**	**p-value**
Gender (Female vs Male)	1.580	(1.050, 2.379)	0.028	2.240	(0.922, 5.443)	0.075	0.579	(0.195, 1.716)	0.324
Age (by year)	0.995	(0.983, 1.006)	0.359	1.035	(0.997, 1.074)	0.069	1.087	(1.059, 1.116)	< 0.001
Tumour number (Multiple vs Single)	1.554	(1.213, 1.990)	< 0.001	1.678	(0.930, 3.027)	0.086	0.632	(0.408, 0.978)	0.040
Tumour size ($\geq$3cm vs <3cm)	1.305	(0.950, 1.794)	0.100	1.342	(0.642, 2.806)	0.435	1.275	(0.786, 2.069)	0.326
Stage (T1/Tis vs Ta)	0.639	(0.406, 1.005)	0.053	3.025	(1.396, 6.553)	0.005	2.003	(1.121, 3.578)	0.019
Grade (G2 vs G1+Benign)	1.462	(1.120, 1.908)	0.005	0.825	(0.302, 2.256)	0.708	1.180	(0.776, 1.795)	0.439
Grade (G3 vs G1+Benign)	1.150	(0.776, 1.703)	0.487	2.870	(1.089, 7.568)	0.033	1.104	(0.614, 1.987)	0.741
Smoker vs Non-smoker†	1.060	(0.727, 1.544)	0.763	1.751	(0.674, 4.550)	0.251	2.654	(1.164, 6.056)	0.020

†Smoker includes current and former smokers. Non-smoker includes occasional smokers.

Table 2.9: Fine and Gray (FGH) models for the subhazard for (T_1, C_1^*)

	Recurrence			Pr+DBC			DOC		
Factor	**HR**	**CI95%‡**	**p-value**	**HR**	**CI95%‡**	**p-value**	**HR**	**CI95%‡**	**p-value**
Gender (Female vs Male)	1.541	(1.021, 2.327)	0.040	2.269	(0.912, 5.643)	0.078	0.498	(0.168, 1.478)	0.209
Age (by year)	0.992	(0.981, 1.004)	0.179	1.033	(0.997, 1.072)	0.076	1.080	(1.054, 1.107)	< 0.001
Tumour number (Multiple vs Single)	1.554	(1.214, 1.990)	< 0.001	1.696	(0.946, 3.040)	0.076	0.519	(0.335, 0.803)	0.003
Tumour size ($\geq$ 3cm vs <3cm)	1.269	(0.923, 1.745)	0.143	1.270	(0.608, 2.654)	0.524	1.111	(0.685, 1.801)	0.670
Stage (T1Tis vs Ta)	0.554	(0.350, 0.875)	0.011	3.123	(1.447, 6.738)	0.004	1.655	(0.945, 2.898)	0.078
Grade (G2 vs G1+Benign)	1.467	(1.124, 1.914)	0.005	0.724	(0.265, 1.977)	0.529	1.054	(0.692, 1.605)	0.806
Grade (G3 vs G1+Benign)	1.073	(0.722, 1.594)	0.726	2.667	(1.019, 6.979)	0.046	1.106	(0.626, 1.955)	0.728
Smoker vs Non-smoker†	1.014	(0.694, 1.481)	0.944	1.802	(0.667, 4.871)	0.245	2.546	(1.113, 5.821)	0.027

†Smoker includes current and former smokers. Non-smoker includes occasional smokers.

‡FGH model fitted according to Geskus (2010).

CHAPTER 3

Multi-state models: a dynamical model for the risk of progression

In this chapter we use the multi-state modelling approach to derive a dynamical model for the risk of progression. This approach will allow updating the prognostic of an individual according to his evolution during his follow-up.

Most prognostic models for bladder cancer progression are static in the sense that predictions are based on the baseline characteristics of the patient and the tumour. These approaches are incomplete since they do not allow for including potential informative events, such as recurrence, which happens during patient's follow-up. Some models include this important information by including the rate of previous recurrences as an additional prognostic factor. This is only valid for modelling secondary tumours that enter the study with a history of previous recurrences. Once the patient is included in the new study, and the follow-up is started, the new observed recurrences cannot be added to the rate of recurrences or treated as a baseline covariate; they have to be treated as a time varying covariate that only affects the risk of progression of a patient after the recurrence has occurred.

Besides providing a clear graphical representation of the process under study, a multi-state model permits to obtain a complete picture of it by modelling the different paths a patient can follow along the multi-state. It permits to establish relationships between events of interest and, in addition, it is useful to obtain updated predictions of a final outcome conditioning on which intermediate states a patient has visited. The literature in multi-state models is extensive and in Section 3.1 we will review the most important elements for building and using multi-state models: how to specify such models (Section 3.1.1), how to perform regression modelling and identify regression factors of interest (Section 3.1.2) and, finally, how to summarize the information of each path to build a single prediction of some future outcome (Section 3.1.3).

Next, a multi-state model for the course of bladder cancer is built. In Section 3.2 we construct the multi-state model and adjust adequate models for each transition. In Section 3.3, a predictive process is obtained with which we will obtain dynamical predictions of the risk of progression.

3.1 Review of multi-state models

Multi-state models generalize competing risks models to acknowledge the presence of intermediate events. Indeed, competing risks could be represented as a multi-state model with one initial state and J mutually exclusive absorbing states. Often, the course of complex time-varying processes involve several events of interest that can be intermediate, in the sense that they occur during follow-up and they do not prevent the observation of some final endpoint of interest.

The most simple of this multi-state models is the unidirectional illness-death model represented in Figure 3.1: individuals start at a healthy state (State 0) and they may become ill (State 1) and afterwards die (State 2), or they may die from the healthy state (direct arrow from State 0 to State 2). In this unidirectional model, individuals cannot recover from illness: this model represents for instance chronic diseases such as HIV.

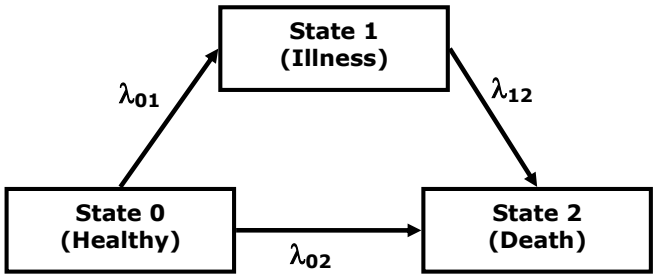


Figure 3.1: The illness-death unidirectional multi-state model.

For the sake of simplicity we will use the illness-death model for introducing the main theoretical aspects of multi-state models. Moreover, we will focus on regression modelling and we will restrict to the use of Cox proportional hazards and extensions of it. This review does not pretend to be exhaustive: we focus on the relevant and practical issues of this methodology that will help us understand the course of a complex disease such as bladder cancer. We mainly follow the references from Klein and Moeschberger (1997, ch.5) and Putter *et al.* (2007), which summarize and complete the readings from Andersen and Keiding (2002), Hougaard (1999) or the different papers about the modelling of bone marrow transplantation published by John Klein and others (Keiding *et al.*, 2001, Klein and Shu, 2002, Klein *et al.*, 1994, 2001a,b).

A Bayesian approach, out of the scope of this work, can be found in Kneib and Hennerfeind (2008). An updated review on the literature can be found in Meira-Machado *et al.* (2009), which includes a useful summary on existing software to deal with multi-state models (see Section 3.1.4 for details). Advanced methods in multi-state models are currently appearing in statistical journals due to the concern for the modelling of complex diseases (Aalen, 2010). These papers include new

tools to deal with distinct elements arising from the observation of the dynamics o the disease: for instance, interval-censored data (Foucher *et al.*, 2010), current-status data (Lan and Datta, 2010), or nonignorable inspection processes (Chen *et al.*, 2010, Sweeting *et al.*, 2010).

3.1.1 Model specification

Multi-state models may contain initial states, final or absorbing states and intermediate or transient states. Initial states represent the state in which the individual remains from the origin of time until an event occurs. For instance, the healthy state in the illness-death model. Final states typically represent an endpoint: the individual remains in this state after entering. A typical example is death. Transient states represent events that occur during the course of disease and might modify the risk of some final endpoint. In the illness-death model, illness is the transient state.

A multi-state model is characterized by describing all of its transitions. Transitions are depicted by arrows between states, and represent the occurrence of an event which determines the passing from one state to another. The hazard rate (or intensity) of the transition rs is

$$\lambda_{rs}(t) = \lim_{\Delta t \to 0} \frac{P\big(t \leq T_{rs} < t + \Delta t | T_{rs} \geq t\big)}{\Delta t},$$

where T_{rs} represents the time of entering state r coming from state s. The cumulative hazard for transition rs is given by

$$\Lambda_{rs}(t) = \int_0^t \lambda_{rs}(u) du.$$

In the illness-death model, with three states, one has to characterize the following three transitions: transition 01 representing from the healthy state to illness, transition 02 from the healthy state to death, and transition 12 from the illness state to death. We will use the 'clock forward' time scale (Putter *et al.*, 2007), in which the time points always refer to the time since the individual entered the initial state.

A common assumption to simplify the model is the Markov assumption, which states that the future evolution of the process studied only depends on the state at time t, that is, the history of the individual is summarized by the state at time t. The model could also be assumed to be semi-Markovian if the future depends not only in the present state but also on the time since entry on this state.

3.1.2 Regression modelling

In this section, we consider distinct modellings of the transition intensities for a multi-state model. We summarize three approaches, two assuming the Markov property and a third relaxing this hypothesis into a more flexible semi-Markov assumption. Let $\boldsymbol{Z}$ be a vector of fixed covariates, that is, covariates defined at the origin of time, $t = 0$, which are not time-dependent.

The first approach fits a Cox proportional hazards model for all the transitions in the model. For

a model with p states and k transitions, we would fit k models to the hazards of each transition:

$$\lambda_{rs}(t|\boldsymbol{Z}) = \lambda_{rs,0}(t)\exp\{\boldsymbol{\beta}'_{rs}\boldsymbol{Z}\}$$

where $\lambda_{rs,0}(t)$ is the baseline hazard for transition rs and $\boldsymbol{\beta}_{rs}$ are the vector of regression coefficients measuring the effect of the covariates $\boldsymbol{Z}$ on the transition intensity. In the illness death model, we would fit the following three models

$$\begin{aligned}\lambda_{01}(t|\boldsymbol{Z}) &= \lambda_{01,0}(t)\exp\{\boldsymbol{\beta}'_{01}\boldsymbol{Z}\}\\ \lambda_{02}(t|\boldsymbol{Z}) &= \lambda_{02,0}(t)\exp\{\boldsymbol{\beta}'_{02}\boldsymbol{Z}\}\\ \lambda_{12}(t|\boldsymbol{Z}) &= \lambda_{12,0}(t)\exp\{\boldsymbol{\beta}'_{12}\boldsymbol{Z}\}.\end{aligned}$$

The second approach is based on the idea of modelling endpoints of interest instead of all possible transitions. It consists in assuming that some of the transitions have proportional baseline hazards. For instance, in the illness-death model, we may assume that transitions 02 and 12 are proportional. Let T_I be the time from origin until the individual reaches the illness state, and T_D the time it takes until death. Modelling transition 01 is equivalent to model T_I, the time until the intermediate event (illness) occurs, that is,

$$\lambda_{01}(t|\boldsymbol{Z}) = \lambda_I(t|\boldsymbol{Z}) = \lambda_{I,0}(t)\exp\{\boldsymbol{\beta}'_I\boldsymbol{Z}\}, \tag{3.1}$$

where $\lambda_I(t)$ is the hazard function of T_I. This model, in fact, is the same as we would obtain from the first approach. Transitions 02 and 12 result from modelling the hazard of the time to the final event, T_D, including time to the intermediate event as a time-dependent covariate. This is a way to ensure the proportionality of the transitions. Indeed, let

$$I(t) = \begin{cases} 0 & \text{if } t < T_I \\ 1 & \text{if } t \geq T_I \end{cases},$$

be the time-dependent covariate, zero if at the time point t the individual has not left the healthy state, one if by time t the individual is already at the illness state. Both transitions 02 and 12 can be derived from the model:

$$\lambda_D(t|\boldsymbol{Z}, I(t)) = \lambda_{D,0}(t)\exp\{\boldsymbol{\beta}'_D\boldsymbol{Z} + \delta I(t)\}, \tag{3.2}$$

so to recover the transition-specific hazards,

$$\begin{aligned}\lambda_{02}(t|\boldsymbol{Z}) &= \lambda_D(t|\boldsymbol{Z}, I(t)=0) = \lambda_{D,0}(t)\exp\{\boldsymbol{\beta}'_D\boldsymbol{Z}\},\\ \lambda_{12}(t|\boldsymbol{Z}) &= \lambda_D(t|\boldsymbol{Z}, I(t)=1) = \lambda_{D,0}(t)\exp\{\boldsymbol{\beta}'_D\boldsymbol{Z} + \delta\}.\end{aligned}$$

The second approach provides more parsimonious models than the first one: in the illness-death model, for three states with three transitions, only two Cox models are needed. More generally, for a multi-state model with p states and $k > p$ transitions, we could analyse the multi-state model with $p-1$ Cox proportional hazards models, one for each endpoint representing the arrival to a

non-initial state.

Another advantage of this approach is that the coefficient δ quantify the effect of the intermediate event on the risk of the final endpoint death. Naturally, some evidence is needed to assume the proportionality between transitions. An informal check of this condition might be obtained by fitting models for all transitions according to the first approach and then plot and compare the estimated baseline hazards for each transition.

The third approach assumes proportionality of hazards but also that the multi-state model is semi-Markov, by explicitly including in the model the time until the intermediate event, $T_I = t_I$, as a time-dependent covariate:

$$I_1(t) = \begin{cases} 0 & \text{if } t < t_I \\ t_I & \text{if } t \geq t_I \end{cases} .$$

Therefore, the model for transition 01 remains the same as in (3.1), but the model for T_D is now:

$$\lambda_D(t|\boldsymbol{Z}, I(t), I_1(t)) = \widetilde{\lambda}_{D,0}(t) \exp\{\widetilde{\boldsymbol{\beta}}_D' \boldsymbol{Z} + \lambda I(t) + \gamma I_1(t)\}, \tag{3.3}$$

and the transition-specific hazards are given by

$$\begin{aligned} \lambda_{02}(t|\boldsymbol{Z}) &= \lambda_D(t|\boldsymbol{Z}, I(t) = 0, I_1(t) = 0) = \widetilde{\lambda}_{D,0}(t) \exp\{\widetilde{\boldsymbol{\beta}}_D' \boldsymbol{Z}\}, \\ \lambda_{12}(t|\boldsymbol{Z}) &= \lambda_D(t|\boldsymbol{Z}, I(t) = 1, I_1(t) = t_I) = \widetilde{\lambda}_{D,0}(t) \exp\{\widetilde{\boldsymbol{\beta}}_D' \boldsymbol{Z} + \lambda + \gamma t_I\}. \end{aligned}$$

3.1.3 Predictive Process

The modelling of all transitions permits us to identify the risk factors associated to each hazard, but also to go one step further, and obtain updated predictions based on the history of disease described for each patient (Andersen and Keiding, 2002, Klein and Moeschberger, 1997, Klein *et al.*, 1994). To do so, we need to summarize the information included in all the transitions involved in the specific history of the patient.

In the present section we deal with the problem of making predictions at time t of future events that may occur between t and u, $t < u$, given the history of the patient at the instant t. The history of the patient, besides baseline covariates, is given by the path of intermediate events followed until instant t. In other words, it is determined by the observed course of disease. Let's denote by $H(t)$ the history of the patient at instant t.

In the illness-death model, for instance, we are particularly interested in two histories. Firstly, we denote by $H_0(t)$ the history of a patient who, at time t remains at state 0 (healthy state), so the endpoints of interest, illness and death have not occurred by that time. More formally:

$$H_0(t) = \{T_I > t, T_D > t\}.$$

Secondly, we consider $H_1(t)$, the history of a patient who, by time t, is alive and has already moved

to the illness state at time r

$$H_1(t,r) = \{T_I = r, T_D > t, r \leq t\}.$$

In a Markov model, the history of the patient is summarized by the present state, that is, the history of the patient reduces to

$$H_1(t) = \{T_I \leq t, T_D > t\},$$

and it does not depend on the exact time at which the illness state was entered.

We define the predictive process $\pi(u,t)$ as the probability of an event to occur by time u given the baseline covariates and the history of the patient at time t. In the case of the illness-death model, this predictive process is defined by

$$\pi(u,t) = P(t < T_D \leq u \mid \boldsymbol{Z}, H(t)).$$

In a multi-state model with no recurrent events, we can obtain explicit expressions for this predictive model depending only on the hazards of the transitions λ_{rs}. We illustrate this fact with the illness-death model.

We first compute the predictive process for the history of a patient who, by time t, has already experienced the intermediate event. We consider the case of a Markov model, that is

$$\pi_1(u,t) = P(t < T_D \leq u \mid \boldsymbol{Z}, H_1(t))$$

where $H_1(t) = \{T_I \leq t, T_D > t\}$. This probability can be further developed by

$$\begin{aligned} \pi_1(u,t) &= P(t < T_D \leq u \mid \boldsymbol{Z}, T_I \leq t, T_D > t) \\ &= \int_t^u \lambda_{12}(s|\boldsymbol{Z}) \exp\Big\{ - \big[\Lambda_{12}(s|\boldsymbol{Z}) - \Lambda_{12}(t|\boldsymbol{Z})\big)\Big\} ds. \end{aligned} \tag{3.4}$$

This expression is obtained integrating out, between all possible values of $T_D = s$ between t and u, the risk of entering the death state at t ($\lambda_{12}(s|\boldsymbol{Z})$) multiplied by the probability of remaining in the illness state and thus not dying between t and s ($\exp\{-[\Lambda_{12}(s|\boldsymbol{Z}) - \Lambda_{12}(t|\boldsymbol{Z})]\}$).

Now we compute the predictive process for the history of a patient who by time t is alive and at the healthy state, $\pi_0(u,t) = P(t < T_D \leq u \mid \boldsymbol{Z}, H_0(t))$, which can be expressed by

$$\begin{aligned} \pi_0(u,t) &= P(t < T_D \leq u \mid \boldsymbol{Z}, T_I > t, T_D > t) \\ &= \int_t^u \exp\{-(\Lambda(s|\boldsymbol{Z}) - \Lambda(t|\boldsymbol{Z}))\} \left[\lambda_{02}(s|\boldsymbol{Z}) + \lambda_{01}(s|\boldsymbol{Z})\pi_1(s,t)\right] ds. \end{aligned} \tag{3.5}$$

We explain with some detail this expression. Transitions 01 and 02 are competing risks, because are mutually exclusive. This competing risks problem involves the time to the first observed event, $T = \min(T_I, T_D)$, together with the type of failure, D or I. The cause-specific hazards for each failure are the hazard transitions λ_{01} and λ_{02}. Therefore, the hazard function for variable T is

obtained by

$$\lambda(t) = \lambda_{01}(t) + \lambda_{02}(t),$$

the corresponding cumulative hazard

$$\Lambda(t) = \Lambda_{01}(t) + \Lambda_{02}(t),$$

and the survival function for T is equal to

$$S(t) = P(T > t) = \exp\{-(\Lambda_{01}(t) + \Lambda_{02}(t))\}.$$

The predictive process (3.5) can be hence rewritten in terms of T:

$$\pi_0(u, t \mid \boldsymbol{Z}) = P(t < T_D \leq u \mid \boldsymbol{Z}, T > t) = \int_t^u \frac{S(s|\boldsymbol{Z})}{S(t|\boldsymbol{Z})} \left[\lambda_{02}(s|\boldsymbol{Z}) + \lambda_{01}(s|\boldsymbol{Z})\pi_1(s, t)\right] ds. \quad (3.6)$$

The first summand in the previous expression represents the instant probability of entering state 2 directly from state 0: it is the product between the risk of dying at s ($\lambda_{02}(s|\boldsymbol{Z})$) and the probability that a patient is alive and healthy at s, given he was alive and healthy at $t < s$ ($S(s)/S(t)$). The second summand represents the instant probability of entering state 2 after having entered into state 1 at s, given that at t the patient was in state 0: it is the product between the risk of entering the illness state at s ($\lambda_{01}(t|\boldsymbol{Z})$), the probability of dying between s and u ($\pi_1(s, t)$) and the probability that a patient is alive and healthy at s, given he was alive and healthy at $t < s$ ($S(s)/S(t)$).

The predictive processes $\pi_0(u, t)$ and $\pi_1(u, t)$ can be estimated with the estimated risk factors obtained from the modelling approach chosen. For instance, for the second approach, the Markov proportional hazards approach, we fit models (3.1) and (3.2) to obtain:

$$\begin{aligned}
\widehat{\lambda}_{01}(t|\boldsymbol{Z}) &= \widehat{\lambda}_{I,0}(t) \exp\{\widehat{\boldsymbol{\beta}}_I' \boldsymbol{Z}\} \\
\widehat{\lambda}_{02}(t|\boldsymbol{Z}) &= \widehat{\lambda}_{D,0}(t) \exp\{\widehat{\boldsymbol{\beta}}_D' \boldsymbol{Z}\} \\
\widehat{\lambda}_{12}(t|\boldsymbol{Z}) &= \widehat{\lambda}_{D,0}(t) \exp\{\widehat{\boldsymbol{\beta}}_D' \boldsymbol{Z} + \widehat{\delta}\}.
\end{aligned}$$

Then, the predictive processes (3.4) and (3.5) are estimated by:

$$\begin{aligned}
\widehat{\pi}_1(u, t|\boldsymbol{Z}) &= \sum_{t<t_i<u} \widehat{\lambda}_{12}(t_i|\boldsymbol{Z}) \exp\Big\{ - \big[\widehat{\Lambda}_{12}(t_i|\boldsymbol{Z}) - \widehat{\Lambda}_{12}(t|\boldsymbol{Z})\big)\Big\} \\
\widehat{\pi}_0(u, t|\boldsymbol{Z}) &= \sum_{t<t_i<u} \exp\{-(\widehat{\Lambda}(t_i|\boldsymbol{Z}) - \widehat{\Lambda}(t|\boldsymbol{Z}))\} \left[\widehat{\lambda}_{02}(t_i|\boldsymbol{Z}) + \widehat{\lambda}_{01}(t_i|\boldsymbol{Z})\widehat{\pi}_1(t_i, t|\boldsymbol{Z})\right].
\end{aligned}$$

with $\widehat{\Lambda}(s|\boldsymbol{Z}) = \widehat{\Lambda}_{01}(s|\boldsymbol{Z}) + \widehat{\Lambda}_{02}(s|\boldsymbol{Z})$, the sum of Breslow estimates of the baseline cumulative hazards for models (3.1) and (3.2).

3.1.4 Existing software for for multi-state models

In this Section we present a brief review on existing software, summarizing the completer reviews found in Putter *et al.* (2007) and Meira-Machado *et al.* (2009): we refer to them for further details.

Estimation of the models for each transition presented in this work can be done in R, SAS and Stata using the counting process notation defined by Therneau and Grambsch (2000), by essentially employing Cox proportional hazards models fitted to data sets where all the transitions made by a single individual are specified. SPSS does not allow for left truncation, thus only Cox models with time-dependent covariates assuming proportionality between transitions can be fitted (Putter *et al.*, 2007).

Several `R` packages exists which permit to fit multi-state models under several assumptions and to obtain summary information from the modelling of all transition hazards: `msm` (Jackson, 2009), `mstate` (de Wreede *et al.*, 2010), `changeLOS` (Wangler and Beyersmann, 2009) or `tdc.msm` (Meira-Machado *et al.*, 2007). These packages permit dynamically predict the evolution of disease according to the past history of disease.

Specific procedures to implement predictive processes or summary information on multi-state models are not readily available in SAS, Stata nor SPSS. SPSS would be the most limited software for the simplicity of the models it can manage to fit. Hui-Min *et al.* (2004) published a sAS macro program for estimating the transition parameters in multi-state homogeneous or non-homogeneous Markov. This macro permits to incorporate covariates, or derive transition probabilities, among other features.

3.2 A dynamical model for the SBC/EPICURO Study

3.2.1 The complete picture of the disease

Classical univariate methods, focused on a particular event of interest, and the competing risks approach presented before, dealing with different causes of failure, provides valuable information on different aspects of the bladder cancer course. However, a global description of the whole complexity of the process is still lacking and can be approached by a multi-state modelling.

A multi-state model starts with a graphical representation of the different possible events (states) linked with arrows representing the different paths (transitions) between events. The bladder cancer course can be described with a multi-state model with five states: Diagnosis, Recurrence, Progression, Death due to Bladder Cancer (Death BC in the figure) and Death due to Other Causes (Death OC in the figure) (Figure 3.2):

A patient diagnosed with a primary tumour remains in the Diagnosis state as long as no other event is observed or until the end of follow-up. The state Recurrence is reached after the first recurrence occurs, and the end of the stay is determined by the occurrence of progression or death, or by the end of follow up. When progression of the tumour occurs, we enter the third state Progression. From Progression, a patient can die due to disease or recover and die due to other causes. Both

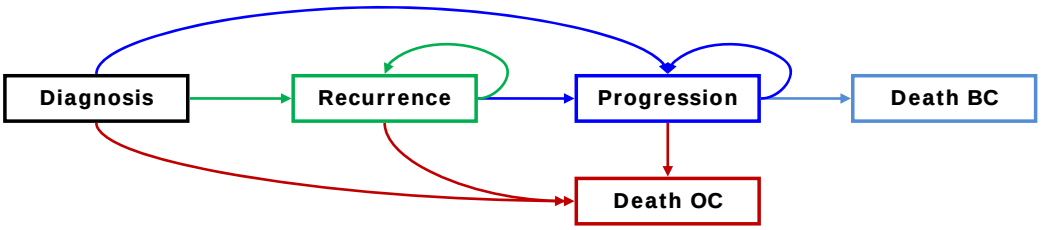


Figure 3.2: Multi-state model for bladder cancer events.

death states are absorbing states. A patient can experience several recurrences and progressions and this is considered by a recursive transition within these states.

To simplify the problem, we have considered only the first recurrence and the first progression as intermediate events, ignoring second, third,...recurrences or progressions. Not only the problem is simplified, but also different modellings including distinct aspects of the recurrence process (number of recurrences, for instance) indicated us that the first recurrence has the greater impact on the risk of progression, and adding posterior recurrences add no valuable information on the progression process. Therefore, the final multi-state model we are going to analyse is the following (Figure 3.3):

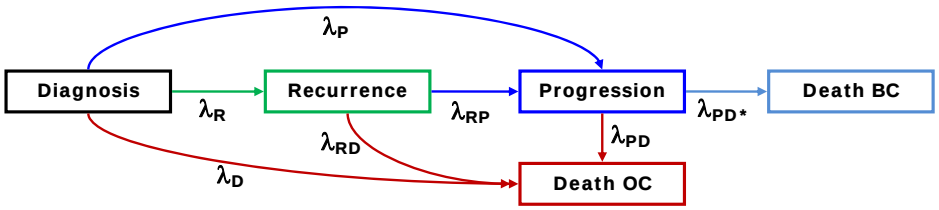


Figure 3.3: Multi-state model for bladder cancer events.

In each arrow we have written the hazard function λ specific for the transition between two consecutive states. In the following Section, the adjustment of this model is undertaken, following the strategies presented in the methods section.

3.2.2 Model fitting

To adjust the multi-state model given in Figure 3.3, we need to specify a model for each transition. For the SBC/EPICURO data we chose the second modelling approach presented in Section 3.1.2: we assume a Markov multi-state model with proportional hazards of several transitions. We take this approach instead of the first for several reasons. First, this approach employs all available individuals for fitting the models involved, while if we modelled one by one the transitions, only individuals at risk for each transitions will be used. Second, we will be able to explicitly quantify the effect of recurrence on progression. Third, more parsimonious models are obtained. The markovian assumption has been checked by fitting model (3.3). No significant evidence for a semi-markovian

process was found: the coefficient γ in this model, which quantified the effect of the time being recurrence-free, was statistically non-significant.

The competing risks endpoints involved in this multi-state model are three: the time to the first event observed distinguishing between different types, (T_1, C_1^*), given in Notation 2.3; the time until the first between PFS or death, (T_2, C_2), given in Notation 2.2; the time until death distinguishing between deaths due to bladder cancer or due to other causes, (T_D, C_D), where $C_D = BC$ (if due to bladder cancer) or $C_D = DOC$ (if due to other causes). Consider the following time-dependent covariates corresponding to each intermediate event:

$$R(t) = \begin{cases} 1 & \text{if } T_1 \leq t, C_1^* = R \\ 0 & \text{if } T_1 > t \text{ or } T_1 \leq t, C_1^* \neq R \end{cases},$$

for recurrences, and

$$P(t) = \begin{cases} 1 & \text{if } T_2 \leq t, C_2 = PFS \\ 0 & \text{if } T_2 > t \text{ or } T_2 \leq t, C_2 = DOC \end{cases}$$

for progressions. The multi-state model can be analysed with the following four Cox proportional hazards (PH) models:

- **Model 1:** Cox PH model for the risk to recurrence, $(T_1, C_1^* = \text{Recurrence})$:

$$\lambda^1(t|\boldsymbol{Z}) = \lambda_0^1(t) \exp\{\theta_0 + \theta_1 Z_1 + \ldots \theta_k Z_k\}. \tag{3.7}$$

- **Model 2:** Cox PH model for the risk to progression, $(T_2, C_2 = PFS)$:

$$\lambda^2(t|\boldsymbol{Z}, R(t)) = \lambda_0^2(t) \exp\{\beta_0 + \beta_1 Z_1 + \ldots \beta_k Z_k + \delta R(t)\}. \tag{3.8}$$

- **Model 3:** Cox PH model for the risk to death due to bladder cancer, $(T_D, C_D = BC)$:

$$\lambda^3(t|\boldsymbol{Z}) = \lambda_0^3(t) \exp\{\gamma_0^b + \gamma_1^b Z_1 + \ldots \gamma_k^b Z_k + \lambda^b R(t) + \nu^b P(t)\}, \tag{3.9}$$

- **Model 4:** Cox PH model for the risk to death due to other causes, $(T_D, C_D = DOC)$:

$$\lambda^4(t|\boldsymbol{Z}) = \lambda_0^4(t) \exp\{\gamma_0^o + \gamma_1^o Z_1 + \ldots \gamma_k^o Z_k + \lambda^o R(t) + \nu^o P(t)\}. \tag{3.10}$$

Table 3.1 summarizes the fitting of the 4 above models for the SBC/EPICURO data. We have used all baseline covariates for the 4 models to provide better distinction and interpretation between different individual profiles. In other instances, only a subset of the baseline covariates could be considered.

When we look at the estimated hazard ratios for Model 2, we observe that the effects for the baseline covariates $\boldsymbol{Z}$ are similar to the obtained in the model without time-dependent covariates (in Table 2.5). However, the time-dependent effect of recurrence resulted significant for PFS (p-value 0.0049). We obtain a HR of $e^{\delta} = 2.0746$, indicating the increase of risk of progression for

Table 3.1: Cox models to fit the multi-state model.

	Mod.1 Rec			Mod.2 PFS		
Factor	**HR**	**CI95%**	**p-value**	**HR**	**CI95%**	**p-value**
Gender (Female vs Male)	1.580	(1.050, 2.379)	0.028	1.301	(0.574, 2.948)	0.528
Age (by year)	0.995	(0.983, 1.006)	0.359	1.038	(1.008, 1.069)	0.012
Tumour number (Multiple vs Single)	1.554	(1.213, 1.990)	< 0.001	1.472	(0.922, 2.350)	0.106
Tumour size($\geq$3cm vs <3cm)	1.305	(0.950, 1.794)	0.100	1.339	(0.755, 2.374)	0.318
Stage (T1+Tis vsTa)	0.639	(0.406, 1.005)	0.053	2.435	(1.354, 4.379)	0.003
Grade (G2 vs G1+Benign)	1.462	(1.120, 1.908)	0.005	1.111	(0.540, 2.285)	0.775
Grade (G3 vs G1+Benign)	1.150	(0.776, 1.703)	0.487	3.213	(1.541, 6.699)	0.002
Smoker vs non-smoker	1.060	(0.727, 1.544)	0.763	1.182	(0.545, 2.567)	0.672
Recurrence (Yes vs No)†	–	–		2.075	(1.247, 3.451)	0.005
	Mod.3 DBC			**Mod.4 DOC**		
Factor	**HR**	**CI95%**	**p-value**	**HR**	**CI95%**	**p-value**
Gender (Female vs Male)	1.479	(0.546, 4.002)	0.441	0.848	(0.403, 1.784)	0.663
Age (by year)	1.055	(1.014, 1.099)	0.009	1.084	(1.061, 1.108)	< 0.001
Tumour number (Multiple vs Single)	0.987	(0.544, 1.792)	0.966	0.825	(0.590, 1.154)	0.262
Tumour size ($\geq$3cm vs <3cm)	0.912	(0.438, 1.899)	0.807	1.046	(0.690, 1.585)	0.833
Stage (T1+Tis vs Ta)	1.137	(0.524, 2.465)	0.745	1.197	(0.737, 1.943)	0.468
Grade (G2 vs G1+Benign)	0.755	(0.300, 1.900)	0.551	1.210	(0.848, 1.726)	0.294
Grade (G3 vs G1+Benign)	1.972	(0.794, 4.897)	0.143	1.308	(0.821, 2.085)	0.258
Smoker vs non-smoker	1.321	(0.505, 3.455)	0.570	2.077	(1.108, 3.895)	0.023
Recurrence (Yes vs No)†	2.776	(1.453, 5.306)	0.002	0.980	(0.696, 1.380)	0.908
Progression (Yes vs No)‡	35.121	(18.341,67.252)	< 0.001	1.372	(0.686, 2.747)	0.371

†Time-dependent covariate $R(t)$.
‡Time-dependent covariate $P(t)$.

a patient if suffering a recurrence. Model 1 coincides with the cause-specific model for recurrence we performed for the time to the first observed event (Table 2.8): it specifies the transition from Diagnosis to Recurrence. On the other hand, the analysis of the different causes of death shows that, as expected, recurrence (p-value 0.0020) and above all, progressions (p-value < 0.0001), are highly predictive of death due to bladder cancer (Model 3). As well as expected, recurrence and progression do not predict death due to other causes.

Now, the hazards between transitions defined in the multi-state model in Figure 3.3 are obtained from models (3.8), (3.7), (3.9) and (3.10) as follows:

Transitions from Diagnosis:

$$\lambda_R(t|\boldsymbol{Z}) = \lambda^1(t|\boldsymbol{Z}) = \lambda_0^1(t)\exp\{\boldsymbol{\theta}'\boldsymbol{Z}\}$$
$$\lambda_P(t|\boldsymbol{Z}) = \lambda^2(t|\boldsymbol{Z}, R(t)=0) = \lambda_0^2(t)\exp\{\boldsymbol{\beta}'\boldsymbol{Z}\}$$
$$\lambda_D(t|\boldsymbol{Z}) = \lambda^4(t|\boldsymbol{Z}, R(t)=0, P(t)=0) = \lambda_0^4(t)\exp\{\boldsymbol{\gamma}^{o\prime}\boldsymbol{Z}\}$$

Imagine a patient who, by time t is alive and satisfying $R(t) = 0$ and $P(t) = 0$: the patient is hence at risk of experiencing any of the events, and thus we can compute his risk of recurrence from Model 1 ($\lambda_R(t|\boldsymbol{Z})$), his risk of progression from Model 2 with $R(t) = 0$ ($\lambda_P(t|\boldsymbol{Z})$) and his risk of death due to other causes from model 4 with $R(t) = 0$ ($\lambda_D(t|\boldsymbol{Z})$).

Transitions from Recurrence:

$$\lambda_{RP}(t|\boldsymbol{Z}) = \lambda^2(t|\boldsymbol{Z}, R(t) = 1) = \lambda_0^2(t)\exp\{\boldsymbol{\beta}'\boldsymbol{Z} + \delta\}$$
$$\lambda_{RD}(t|\boldsymbol{Z}) = \lambda^4(t|\boldsymbol{Z}, R(t) = 1, P(t) = 0) = \lambda_0^4(t)\exp\{\boldsymbol{\gamma}^{o'}\boldsymbol{Z} + \lambda^o\}$$

If by time t, recurrence has already occurred ($R(t) = 1$), but progression not ($P(t) = 0$), then the patient has moved to state Recurrence, and therefore, he's no longer at risk of recurrence. We can compute, though, his risk of progression from Model 2 with $R(t) = 1$ ($\lambda_{RP}(t|\boldsymbol{Z})$), or his risk of death from Model 4 with $R(t) = 1$ ($\lambda_{RD}(t|\boldsymbol{Z})$).

Transitions from Progression:

$$\lambda_{PD}(t|\boldsymbol{Z}, R(t) = 0) = \lambda^4(t|\boldsymbol{Z}, R(t) = 0, P(t) = 1) = \lambda_0^4(t)\exp\{\boldsymbol{\gamma}^{o'}\boldsymbol{Z} + \nu^o\}$$
$$\lambda_{PD}(t|\boldsymbol{Z}, R(t) = 1) = \lambda^4(t|\boldsymbol{Z}, R(t) = 1, P(t) = 1) = \lambda_0^4(t)\exp\{\boldsymbol{\gamma}^{o'}\boldsymbol{Z} + \lambda^o + \nu^o\}$$
$$\lambda_{PD^*}(t|\boldsymbol{Z}, R(t) = 0) = \lambda^3(t|\boldsymbol{Z}, R(t) = 0, P(t) = 1) = \lambda_0^3(t)\exp\{\boldsymbol{\gamma}^{b'}\boldsymbol{Z} + \nu^b\}$$
$$\lambda_{PD^*}(t|\boldsymbol{Z}, R(t) = 1) = \lambda^3(t|\boldsymbol{Z}, R(t) = 1, P(t) = 1) = \lambda_0^3(t)\exp\{\boldsymbol{\gamma}^{b'}\boldsymbol{Z} + \lambda^b + \nu^b\}$$

If by time t, progression has already occurred ($P(t) = 1$) but recurrence not ($R(t) = 0$), the patient has moved from the Diagnosis state to the Progression state directly, and he is at risk of death, due to bladder cancer or due to any cause. Therefore, we can compute both risks from Models 3 and 4 with $R(t) = 0$ and $P(t) = 1$ ($\lambda_{PD}(t|\boldsymbol{Z}, R(t) = 0)$ and $\lambda_{PD^*}(t|\boldsymbol{Z}, R(t) = 0)$). On the contrary, if by time t both recurrence and progression have occurred, then the risk of patient of dying due to disease or due to other causes is obtained from Models 3 and 4 by considering $R(t) = 1$ and $P(t) = 1$ ($\lambda_{PD}(t|\boldsymbol{Z}, R(t) = 1)$ and $\lambda_{PD^*}(t|\boldsymbol{Z}, R(t) = 1)$).

For simplicity, in the following section we omit, from the intensity transitions described above, the dependency from $\boldsymbol{Z}$, though it is implicitly assumed.

3.3 Predictive process of the risk of progression

Now we turn to the construction of the predictive process for the risk of progression. This process will permit us to dynamically predict the occurrence of a future progression: we will be able to update our predictions based on the history of the patient up to the time from where we want to predict. For instance, we will be able to compare the predicted probability of progression at time u of two patients, both alive by time $t < u$, one having experienced a recurrence before t, and the other not. This feature of multi-state models allow us to enrich our analysis by obtaining more

accurate predictions based on up-to-date information, and not only on baseline information, whose impact at long-term is often diminished with time.

We define the predictive process for progression by

$$\pi(u,t) = P(t < T_2 \leq u, C_2 = PFS \mid H(t)),$$

which is the probability that an event of type PFS occurs between t and u given the history $H(t)$ of the patient at the instant t. We consider two type of histories:

- Patients who are alive and without events at time t,

  $$H_0(t) = \{T_1 > t\},$$

  so the patient is at risk of all events.

- Patients alive and with at least one recurrence at time t,

  $$H_1(t) = \{T_2 > t, R(t) = 1\} = \{T_2 > t, T_1 \leq t, C_1^* = R\},$$

  so the patient is at risk of progression or death.

Let's denote by $\pi_1(u,t)$ the predictive process corresponding to the history $H_1(t)$. This process has the following expression:

$$\begin{aligned}
\pi_1(u,t) &= P(t < T_2 \leq u, C_2 = PFS \mid H_1(t)) \\
&= P(t < T_2 \leq u, C_2 = PFS \mid T_2 > t, R(t) = 1) \\
&= \int_t^u \frac{f_{RP}(s)}{S_2(t)} ds = \int_t^u \frac{\lambda_{RP}(s) S_2(s)}{S_2(t)} ds \\
&= \int_t^u \exp\{-(\Lambda_2(s) - \Lambda_2(t))\} \lambda_{RP}(s) ds,
\end{aligned} \tag{3.11}$$

where $S_2(t) = P(T_2 > t)$ is the overall survival function for time T_2 and $\Lambda_2(t)$ is the overall cumulative hazard function. As explained in Section 2.1.1 when introducing competing risks, the overall cumulative hazard can be expressed as the sum of the cause-specific cumulative hazards of the competing causes for T_2, PFS and DOC:

$$\Lambda_2(t) = \Lambda_{RP}(t) + \Lambda_{RD}(t).$$

The expression of this predictive process (3.11) is obtained integrating out s between t and u (see the first integral) the probability of moving to the Progression state at time s from the state Recurrence (f_{RP}) given that at time t the patient was alive and without recurrence ($1/S_2(t)$)

We consider now the predictive process $\pi_0(u,t)$ for the history $H_0(t)$:

$$\begin{aligned}
\pi_0(u,t) &= P(t < T_2 \leq u, C_2 = PFS \mid H_0(t)) \\
&= P(t < T_2 \leq u, C_2 = PFS \mid T_1 > t)
\end{aligned}$$

$$
\begin{aligned}
&= \int_t^u \left[\frac{f_P(s)}{S_1(t)} + \frac{f_R(s)}{S_1(t)} \pi_1(u,s) \right] ds \\
&= \int_t^u \left[\frac{\lambda_P(s) S_1(s)}{S_1(t)} + \frac{\lambda_R(s) S_1(s)}{S_1(t)} \pi_1(u,s) \right] ds \\
&= \int_t^u \frac{S_1(s)}{S_1(t)} \left[\lambda_P(s) + \lambda_R(s) \pi_1(u,s) \right] ds \\
&= \int_t^u \exp\{-(\Lambda_1(s) - \Lambda_1(t))\} \left[\lambda_P(s) + \lambda_R(s) \pi_1(u,s) \right] ds, \qquad (3.12)
\end{aligned}
$$

where $S_1(t) = P(T_1 > t)$ is the overall survival function for time T_1 and $\Lambda_1(t)$ is the overall cumulative hazard function. This function can be expressed in terms of the cause-specific cumulative hazards of the competing causes for T_1, recurrence, progression and death due to other causes:

$$\Lambda_1(t) = \Lambda_R(t) + \Lambda_P(t) + \Lambda_D(t).$$

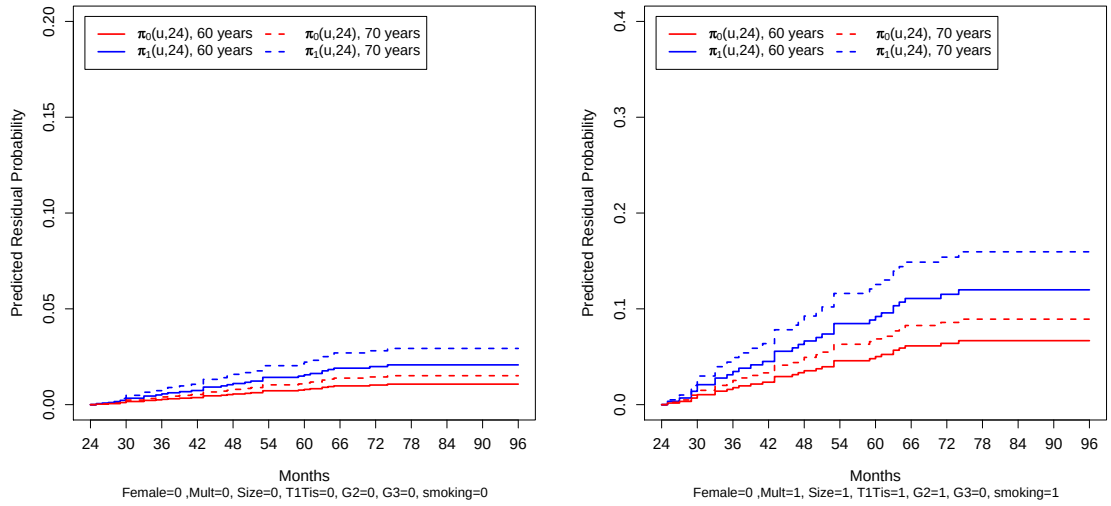
The expression of this predictive process (3.11) is obtained integrating out s between t and u (see the first integral) the probability for a patient of moving to the Progression state at time s given that at instant t the patient was alive and without events $(1/S_1(t))$. For this, he has two options: (i) he can go directly from Diagnosis, and thus the probability of progressing at instant s from the state Diagnosis is $f_P(s)$, or (ii) he can have first a recurrence, and thus the probability of progression at s is the probability of experiencing a recurrence at time s $(f_R(s))$ multiplied by the probability of progressing at u given that at s a recurrence has occurred $(\pi_1(u,s))$.

The predictive process depends on the time t at which the history is known, and the point u at which we wish to make a prediction. By fixing or varying t and u we get different insights into the problem. For instance, if we fix the point at which the history of the patient is known, t, and vary the time at which the predictions are made, u, we obtain the residual cumulative incidence function, that is, the residual probability of progression once the information of the patient up to time t is known. On the other hand, another perspective is obtained if we vary the time t at which we assess the history of the patient, and compute the predicted probability of progression at time $u = t + \Delta$, with Δ fixed, which would give us, for each time t, and given the history of the patient at this time t, the probability of progression in the next Δ months. In the following, we present two numerical examples of these quantities.

3.3.1 Residual cumulative incidence of progression

We first consider the predictive process when the history of the patient is known at $t = 24$ months (2 years) and we vary u, the time at which we want to make the predictions, from 24 to 96 months (8 years). This corresponds to the predicted residual probability of progression for patients 2 years after the diagnosis of the primary tumour. The predictive process for an individual alive at 24 months, who has already experienced a recurrence is

$$\pi_1(u, 24) = P(24 < T_2 \leq u \mid T_2 > 24, R(24) = 1),$$



(a) Low risk profile.

(b) High-risk profile.

Figure 3.4: Predicted PFS cumulative incidence curves for patients after 24 months after diagnosis.

and, on the other hand, the predictive process for an individual for whom no event has been observed by 24 months is

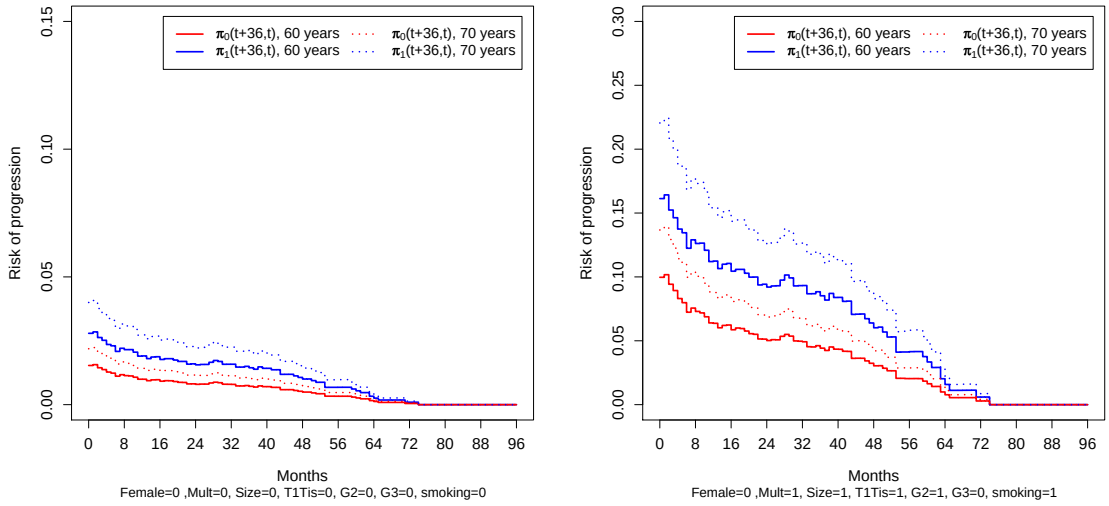
$$\pi_0(u, 24) = P(24 < T_2 \leq u \mid T_1 > 24, R(24) = 0).$$

In Figure 3.4 we plot the predicted residual cumulative incidence curves at 24 months after diagnosis, for males of 60 years or 70 years, with a low-risk profile (Fig. 3.4(a): single tumours, size less than 3cm, Ta, Grade 1 or Benign, non-smoker) or a high-risk profile (Fig. 3.4(b): multiple tumours, size of the largest greater than 3 cm, T1 or Tis, Grade 2, smoker). Notice the different scales chosen for each plot in order to highlight differences between curves. In both plots, it can be observed the effect of age as a fix risk factor, because older patients (70 years, dashed lines) have higher predicted incidence than younger (60 years, solid lines).

If we fix age, there are differences between the curves for those patients who have already suffered at least one recurrence (blue lines) with respect to the ones who have not experienced any (red lines). For instance, for a 60 years old low-risk profile having at baseline a predicted probability of progressing in the first 5 years equal to 0.010, his updated probability after 2 years of follow-up is $\pi_1(60, 24) = 0.016$ if recurrence has occurred ($R(24) = 1$) and $\pi_0(60, 24) = 0.008$ if no events have occurred ($R(24) = 0$). These differences are even more remarkable for high-risk patients: the predicted probability of progression at baseline within the next 5 years was 0.076 and, if $R(24) = 1$, the residual probability 60 months is 0.086, and if $R(24) = 0$, it is 0.048.

3.3.2 Conditional risk of progression

Another analysis of the predictive process can be done by varying the time t at which we assess the history of the patient. For instance, we can compute the predicted risk of progression at time $u = t + 36$ that, given the history at time t, provides the probability of progressing in the following



(a) Low risk profile.

(b) High-risk profile.

Figure 3.5: Risk of progression in the next 3 years given the history at time t after Diagnosis

3 years (36 months). We consider the predictive process for a patient with $R(t) = 1$: alive at t months, who has already experienced a recurrence by then,

$$\pi_1(t+36, t) = P(t < T_2 \leq t+36 \mid T_2 > t, R(t) = 1),$$

and on the other hand, the predictive process for an individual for whom no event has been observed by t months,

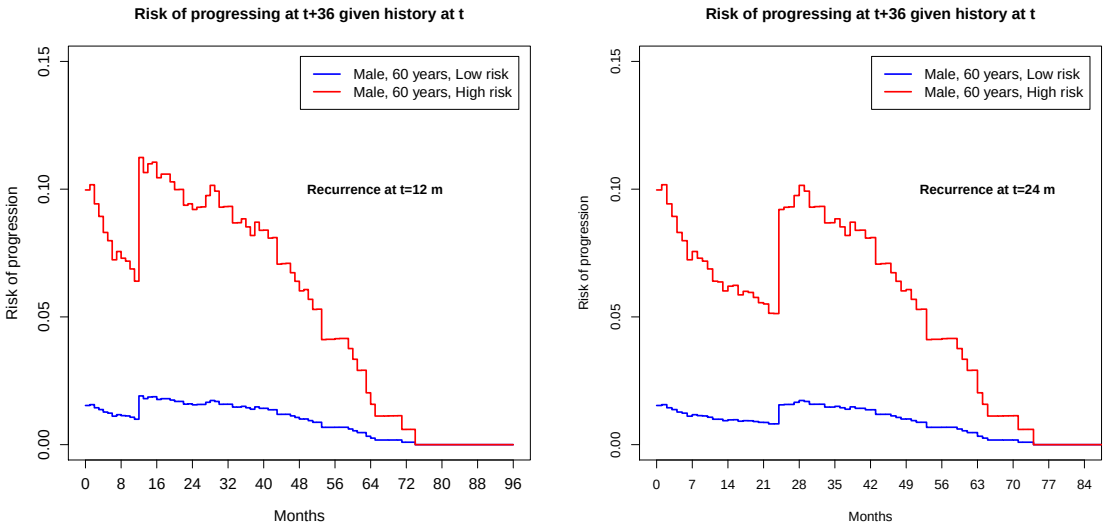
$$\pi_0(t+36, t) = P(t < T_2 \leq t+36 \mid T_1 > 36, R(36) = 0).$$

Figure 3.5 depicts such probabilities. As it was expected, the risk of progression diminishes as time goes by and patient remains in the same state. In Figure 3.6 we observe how, after a patient experiences a recurrence, his individual risk curve π_0 jumps to the risk curve π_1 indicating an update in his predicted risk of progression.

3.3.3 Updated classification in risk groups

The construction of a predictive process based on a multi-state model allows henceforth to make dynamical predictions on different courses of disease. It is clear that the occurrence of an intermediate event (recurrence) at a certain time during follow up changes the prognostic of the patient. And the non-occurrence of the intermediate event during a reasonable period of time also changes the prognostic of the patient: we have seen in the last figure how the risk of progression diminishes. Therefore, by using the last updated information, we can reclassify patients into new risk categories, and thus obtain a more accurate classification of patients.

For instance, Table 3.2 contains the classification of the patients according to their risk of progressing before 60 months (5 years). This classification is based on the predicted probability of



(a) Recurrence at t=12 months.

(b) Recurrence at t=24 months.

Figure 3.6: Change in the predicted risk of progression when recurrence occurs.

progressing before 60 months made at baseline (rows of the table), and it is compared with the classification made according to the predicted probability of progressing before 60 months given the history of the patient at 12 months. The cuts for the classification for both predictions are obtained from quartiles Q2 and Q3, resulting in:

- Low risk: probability of progressing before 60 months ≥ 0.025.
- Medium risk: probability of progressing before 60 months between 0.025 and 0.050.
- High risk: probability of progressing before 60 months above 0.050.

Table 3.2: Classification of the risk to progress according to the probability of progression before 60 months (5 years). Baseline prediction vs updated prediction at 12 months.

		Updated prediction given H(12)			
		Low Risk	**Medium Risk**	**High Risk**	**Total**
Baseline prediction	**Low Risk**	321	71	11	403 (51.4%)
	Medium Risk	74	95	39	208 (26.5%)
	High Risk	0	11	162	173 (22.1%)
	Total	395 (50.4%)	177 (22.6%)	212 (27.0%)	784†(100.0%)

†Patients alive and at risk of progression at 12 months, with non-missing covariates.

In the table we observe the changes in the predicted risk categories when the predictions are updated after 12 months of follow-up. Of the 408 patients at low risk at baseline, 321 (80%) remains at low risk after 12 months of follow-up, and the rest are reclassified as medium or high risk of progression. In the medium risk group at baseline, 74 (35%) are still as low risk and 39 (19%) as high risk. Of the 173 patients in the high risk group at baseline, 162 (94%) are still at high risk after 12 months, while 11 are reclassified as medium risk.

To evaluate if this new reclassifications are beneficial for the patients, we consider first those patients that suffered a progression after 12 months (a total of 48 individuals). Table 3.3 reflects the updated predictions for this subgroup of patients. For them, any change to a higher risk category is beneficial. Six over 14 patients initially classified as low risk are reclassified in a higher risk category; 6 over 12 in the medium risk category are reclassified as high risk and one patient initially classified as high risk is reclassified as medium risk. Consequently, a total of 12 patients (25%) would benefit from the update while only 1 (2%) would be erroneously reclassified in a lower risk category.

Table 3.3: Patients progressing after 12 months (n=48%).

		Updated prediction given H(12)			
		Low Risk	**Medium Risk**	**High Risk**	**Total**
Baseline prediction	**Low Risk**	8	3	3	14 (29.2%)
	Medium Risk	0	6	6	12 (25.0%)
	High Risk	0	1	21	22 (45.8%)
	Total	8 (16.7%)	10 (20.8%)	30 (62.5%)	48†(100.0%)

†Patients alive and at risk of progression at 12 months, with non-missing covariates.

Evaluating the benefits of the reclassifications in the rest of individuals is complex because this group contains both individuals alive and progression free at the end of follow-up and patients that have died because other causes during follow-up. For this reason we restrict now the analysis to those individuals alive and progression-free at 60 months, a total of 571 patients (Table 3.4). For these patients, any change to a higher risk category is incorrect or unnecessary. This happened to a total of 74 patients (13%). There are 62 patients (10.9 %) that benefit from the acquired information by being reclassified in a lower risk category.

Table 3.4: Patients alive and progression free at 60 months (n=571%).

		Updated prediction given H(12)			
		Low Risk	**Medium Risk**	**High Risk**	**Total**
Baseline prediction	**Low Risk**	269	49	4	322 (56.4%)
	Medium Risk	52	63	21	136 (23.8%)
	High Risk	0	10	103	113 (19.8%)
	Total	321 (56.2%)	122 (21.4%)	128 (22.4%)	571†(100.0%)

†Patients alive and at risk of progression at 12 months, with non-missing covariates.

In addition, if we consider the high risk category versus the low and medium categories together, we can obtain from Tables 3.3 and 3.4 an approximate measure of the sensitivity and specificity of both approaches. By updating the predictions after 12 months of follow-up, the sensitivity is increased from 45.8% to 62.5% while the specificity is hardly reduced, 80.2% versus 77.6%.

All in all, and taking into account the bad prognosis after progression, we can assert a clear benefit of the new reclassification that uses the information on recurrence during the first year after diagnosis.

Part II

Interval-censored semi-competing risks data

CHAPTER 4

Methods for semi-competing risks data

In this Chapter, we present the problem of semi-competing risks for right-censored data (Fine *et al.*, 2001). A semi-competing risks situation can be described as a bivariate survival situation with dependent censoring. In this setting, individuals are at risk of experiencing two events, $\mathcal{E}_1$ and $\mathcal{E}_2$, one of them intermediate, and the other terminating, in such a way that the occurrence of the terminating event $\mathcal{E}_2$ precludes the observation of the intermediate one, $\mathcal{E}_1$, and thus the time until the terminating event, T_2, dependently censors the time until the intermediate event, T_1. The goal of the semi-competing risks methodology is to recover and characterize the marginal distribution of T_1 through the characterization of the association structure between T_1 and T_2.

Semi-competing risks data is also encountered in the Spanish Bladder Cancer Study. Recurrence is an intermediate event for both progression and death, and we may be interested in the marginal recurrence process or in evaluating the influence of factors in the marginal recurrence process. This situation is depicted in Figure 4.1 where T_1 is the marginal lifetime to recurrence and T_2 is the time to the first occurring event, progression or death.

By a competing risks analysis, we were able to characterize T, the minimum between T_1 and T_2 but not the marginal law of T_1. Indeed, the dependent censoring of T_2 on T_1 generates an identifiability problem (Tsiatis, 1975) on the distribution of T_1. However, under semi-competing risks data we can take advantage of that for some patients, since both events can be observed on the same individual. Hence, using these patients, we can explore the dependency between T_1 and T_2, fit a model for their joint distribution and derive the marginal distribution of T_1.

The discussion about whether or not the distribution of the intermediate event is clinically meaningful has been in the past controversial (Fine *et al.*, 2001, Jiang *et al.*, 2003, Wang, 2003). There are situations where it is difficult to interpret the meaning of the assumed marginal or latent distribution, as the process that would be observed if terminating events, such as death, would be avoided. In other clinical studies, its interpretation is more clear: Jiang *et al.* (2003) present an

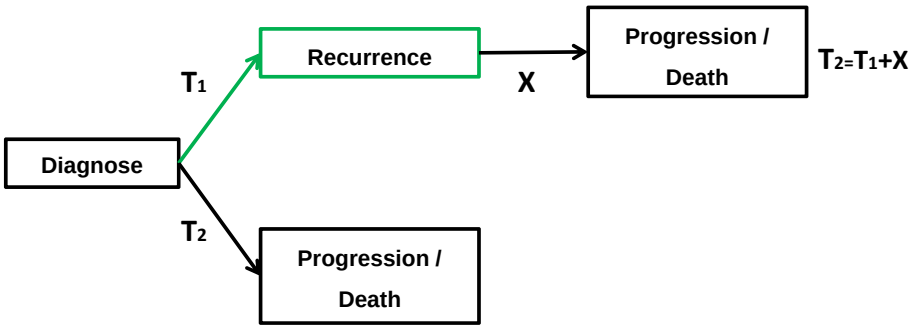


Figure 4.1: Semi-competing risks setting for the SBC/EPICURO Study.

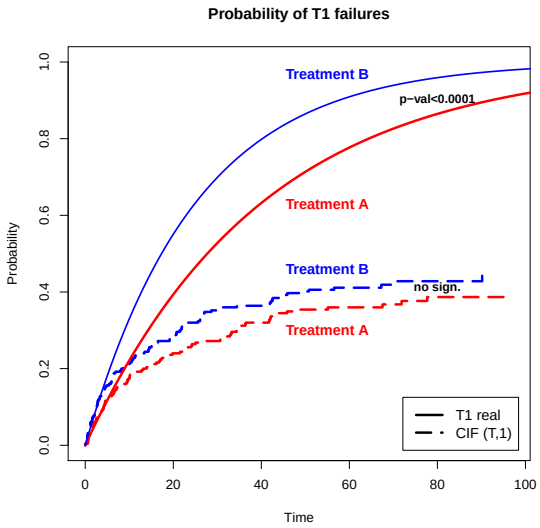


Figure 4.2: Induced protective effect in a competing risks analysis.

HIV study where the intermediate event is virological failure and the terminating event is dropout due to adverse effects of treatment and thus, informative of the process being described. In this case, the clinical interest relies on the virological failure and on the effect the given treatment has on it.

Semi-competing risks can also be useful in clinical trials where the goal is to evaluate the effect of a treatment. In this kind of studies, the presence of a competing risk event complicates the analysis since the effect of the treatment can not be evaluated on the marginal distribution of interest but only on the cause-specific hazard or the cumulative incidence hazard. This competing risks analysis may provide regression coefficients that do not reflect the real effect of the treatment on the marginal distribution.

In Figure 4.2 we represent this phenomenon through a simulated example that mimics a clinical trial for proving the effectiveness of a treatment A over a treatment B. We have simulated bivariate correlated data of (T_1, T_2) in the two categories of treatment, A and B, such that treatment A reduces significantly the proportion of events in both T_1 and T_2 (HR 1.6 for both times, p-value<0.0001). In the figure, we represent the marginal distribution functions for T_1, $P(T_1 \leq t)$,

in both treatment arms (solid lines). The competing risks analysis of this data analyzes $T = \min(T_1, T_2)$ and do not provide significant differences between treatment arms at the cumulative incidence levels (Fine and Gray model, dashed lines in Figure 4.2, HR 1.1, p-value 0.5). The presence of a competing risk makes the cumulative incidence functions to be far from the true marginal distributions and, in addition, the effect of treatment in the marginal scale cannot be recovered. As we will show later in this Chapter, when we will go back to this example, if the situation is not truly competing risks but instead, one of the events is not terminating, we will be able to recover the marginal distribution and the real effect of the treatment using the semi-competing risk methodology.

The present chapter is organized as follows. Before directly addressing semi-competing risks methods, in Section 4.1 we provide basic concepts of bivariate survival data. Section 4.2 is devoted to right-censored semi-competing risks data. After introducing the problem in Section 4.2.1, we present one specific model based on a copula approach and we explore in detail the estimation process (Sections 4.2.2 and 4.2.3). In Section 4.3, we conclude with some numerical examples where we apply the reviewed semi-competing risks methods.

4.1 Concepts of bivariate survival data

4.1.1 Notation

Let T_1 and T_2 be two survival times corresponding to events $\mathcal{E}_1$ and $\mathcal{E}_2$, respectively, with marginal survival functions $S_1(s) = P(T_1 > s)$, $S_2(t) = P(T_2 > t)$ and bivariate survival function $S(s,t) = P(T_1 > s, T_2 > t)$.

The marginal laws can be recovered from the bivariate as follows: $S_1(s) = S(s,0)$ and $S_2(t) = S(0,t)$. The relationship between the joint survival function and the joint distribution function, $F(s,t) = P(T_1 \leq s, T_2 \leq t)$ is given by

$$F(s,t) = S(s,t) - S_1(s) - S_2(t) + 1. \tag{4.1}$$

Similarly, the marginal distribution functions for T_1 and T_2 are given by $F_1(s) = 1 - S_1(s)$ and $F_2(t) = 1 - S_2(t)$, and we can rewrite equation 4.1 in terms of the marginal distributions:

$$F(s,t) = S(s,t) - 1 + F_1(s) + F_2(t).$$

The joint density, hazard and cumulative hazard functions for (T_1, T_2) as well as the relationship between S and Λ are given by

$$\begin{aligned}
f(s,t) &= \frac{\partial^2 S(s,t)}{\partial s \partial t}, \\
\lambda(s,t) &= \lim_{\max(\delta_1,\delta_2)\to 0+} \frac{P\left(s \leq T_1 \leq s+\delta_1, t \leq T_2 \leq t+\delta_2 | T_1 > s, T_2 > t\right)}{\delta_1 \delta_2} = \frac{f(s,t)}{S(s,t)}, \\
\Lambda(s,t) &= \int_0^s \int_0^t \lambda(u,v) du dv, \text{ and} \\
S(s,t) &= S_1(s) S_2(t) e^{\Lambda(s,t)}.
\end{aligned}$$

4.1.2 Measures of dependence

The dependence between T_1 and T_2 can be described through measures of concordance such as Spearman's rank correlation ρ_S, or Kendall's coefficient of concordance, $\tau_{\mathcal{K}}$.

Spearman's ρ_S for T_1 and T_2 is defined by

$$\rho_S(T_1, T_2) = \rho(F_1(T_1), F_2(T_2)) = \frac{\text{cov}(F_1(T_1), F_2(T_2))}{\sqrt{\text{var}(F_1(T_1))\text{var}(F_2(T_2))}} \tag{4.2}$$

where ρ is the linear correlation coefficient (Pearson correlation) between two random variables. That is, ρ_S is the linear correlation coefficient between $F_1(T_1)$ and $F_2(T_2)$, and thus a measure of rank correlation. The use of the linear correlation coefficient ρ directly on (T_1, T_2) is not appropriate since this measure is restricted to random variables with bivariate normal distribution, which is unusual in the context of lifetime random variables.

On the other hand, Kendall's coefficient of concordance for two pairs (T_{1i}, T_{2i}) and (T_{1j}, T_{2j}) is defined by

$$\begin{aligned} \tau_{\mathcal{K}} &= \mathrm{E}\big[\text{sgn}((T_{1i} - T_{1j})(T_{2i} - T_{2j}))\big] \\ &= P\big((T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0\big) - P\big((T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0\big) \\ &= 4 \int\int S(u, v) f(u, v) du dv - 1, \end{aligned} \tag{4.3}$$

where (T_{1i}, T_{2i}) and (T_{1j}, T_{2j}) are two independent and identically distributed random vectors and sgn is the sign function, that is, $\text{sgn}(x) = 1$ when $x > 0$, -1 when $x < 0$ and 0 for $x = 0$. This coefficient measures the difference between the probability for a pair being concordant and a pair being discordant. In the following, we will mainly use this measure of global dependence instead of Spearman's ρ_S for its intuitive interpretation in terms of concordance.

A measure of local dependence could be given the ratio

$$\theta(s, t) = \frac{S(s, t) \frac{\partial^2 S(s,t)}{\partial s \partial t}}{\frac{\partial S(s,t)}{\partial s} \frac{\partial S(s,t)}{\partial t}} \qquad \forall (s, t), \tag{4.4}$$

referred to as *cross-ratio function* $\theta(s, t)$ by Oakes (1989). This measure gives information on the strength of the local dependence at the point (s, t). A natural interpretation is given in terms of conditional hazards by

$$\theta(s, t) = \frac{\lambda_2(t|T_1 = s)}{\lambda_2(t|T_1 > s)} \qquad \forall (s, t), \tag{4.5}$$

where $\lambda_2(t|A)$ is the hazard function of T_2 given that event A occurs. We can also express the cross-ratio function in terms of $\lambda_1(s|A)$, the hazard function of T_1 given event A:

$$\theta(s, t) = \frac{\lambda_1(s|T_2 = t)}{\lambda_1(s|T_2 > t)},$$

because of the symmetric role of T_1 and T_2. The equivalence between (4.4) and (4.5) is shown in

Appendix B.1.

A conditional version of $\tau_{\mathcal{K}}$, related to the cross-ratio, is given by $\tau^*(s,t)$:

$$\tau^*(s,t) = E[\text{sgn}(T_{1i} - T_{1j})(T_{2i} - T_{2j})|\widetilde{T}_{1ij} = s, \widetilde{T}_{2ij} = t] = \frac{\theta(s,t) - 1}{\theta(s,t) + 1} \quad \forall (s,t).$$

The relationship is justified by the following proposition.

Proposition 4.1. *The cross-sectional ratio can be expressed in terms of conditional probabilities of concordance and discordance (Oakes, 1989):*

$$\theta(s,t) = \frac{P((T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0|\widetilde{T}_{1ij} = s, \widetilde{T}_{2ij} = t)}{P((T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0|\widetilde{T}_{1ij} = s, \widetilde{T}_{2ij} = t)}, \tag{4.6}$$

where $\widetilde{T}_{1ij} = \min(T_{1i}, T_{1j})$ *and* $\widetilde{T}_{2ij} = \min(T_{2i}, T_{2j})$.

The proof of this proposition can be found in B.1.

4.1.3 Copula models for bivariate survival data

Copulas have become a popular tool when bivariate dependence is of interest, because they allow to model separately the marginal distributions and the association structure. Copulas methods assume that the marginal distributions do not depend on the dependence structure. Exhaustive revisions of the copula approach can be found in Nelsen (2006), Georges *et al.* (2001), Trivedi and Zimmer (2007), Joe (1997) or Hougaard (2000).

Definition 4.1. *A copula* $C_{\boldsymbol{\alpha}}(u,v)$ *is a continuous bivariate function defined as*

$$C_{\boldsymbol{\alpha}} : [0,1] \times [0,1] \longrightarrow [0,1],$$

such that is non-decreasing on each component of (u,v) *and satisfies* $C_{\boldsymbol{\alpha}}(u,0) = C_{\boldsymbol{\alpha}}(0,v) = 0$*, and* $C_{\boldsymbol{\alpha}}(u,1) = u$*,* $C_{\boldsymbol{\alpha}}(1,v) = v$*. The functional form of* $C_{\boldsymbol{\alpha}}(u,v)$ *depends on a vector of parameters* $\boldsymbol{\alpha}' = (\alpha_1, \ldots, \alpha_p)$.

Let T_1 and T_2 be two non-negative random variables with marginal survival functions $S_1(s)$ and $S_2(t)$. Let $C_\alpha(u,v)$ be a copula function for $0 \leq u \leq 1$ and $0 \leq v \leq 1$ and where α measures the association between T_1 and T_2. We assume that the joint survival function of (T_1, T_2), $S(s,t)$, can be written as a function of the marginals S_1, S_2 and the parameter α through the following expression

$$S(s,t) = C_\alpha\big(S_1(s), S_2(t)\big). \tag{4.7}$$

It is satisfied that $S(s,\infty) = 0$, $S(\infty,t) = 0$ and

$$S(s,0) = C_\alpha(S_1(s), 1) = S_1(s)$$
$$S(0,t) = C_\alpha(1, S_2(t)) = S_2(t).$$

Sklar's canonical representation (see Nelsen (2006), for instance) guarantees that $S(s,t)$ defined as in (4.7) is indeed a joint survival function. We adopt the notation of survival copulas, where the copula function relates the joint survival with its marginal survivals, instead of the more common use that relates the joint distribution with its margins.

Following (4.2) and (4.3), Spearman's ρ_S and Kendall's $\tau_{\mathcal{K}}$ can both be expressed in terms of the copula function:

$$\begin{aligned}\rho_S &= 12\int_0^1\int_0^1 \{C_\alpha(u,v)-uv\}dudv = 12\int_0^1\int_0^1 C_\alpha(u,v)dudv - 3\\ \tau_{\mathcal{K}} &= 4\int_0^1\int_0^1 C_\alpha(u,v)dC_\alpha(u,v) - 1.\end{aligned} \tag{4.8}$$

Definition 4.2. *Let ϕ_α be a decreasing convex function defined in $(0,1]$ such that $\phi_\alpha(1)=0$. An Archimedean copula function of $(u,v)\in[0,1]^2$ is given by*

$$C_\alpha(u,v) = \phi_\alpha^{-1}\{\phi_\alpha(u)+\phi_\alpha(v)\}, \tag{4.9}$$

The cross-ratio (4.5) of an Archimedean copula model only depends on (s,t) through $S(s,t)$, that is $\theta(s,t)=\theta_\alpha\{S(s,t)\}$, where

$$\theta_\alpha(v) = -v\frac{\phi''_\alpha(v)}{\phi'_\alpha(v)}.$$

From Proposition 4.1, a new expression for $\tau_{\mathcal{K}}$ is obtained:

$$\tau_{\mathcal{K}} = \mathrm{E}\left[\frac{\theta(\widetilde{T}_{1ij},\widetilde{T}_{2ij})-1}{\theta(\widetilde{T}_{1ij},\widetilde{T}_{2ij})+1}\right] = \mathrm{E}\left[\frac{\theta_\alpha\big[S(\widetilde{T}_{1ij},\widetilde{T}_{2ij})\big]-1}{\theta_\alpha\big[S(\widetilde{T}_{1ij},\widetilde{T}_{2ij})\big]+1}\right].$$

Definition 4.3. *Clayton's copula (1978) is a special case of Archimedean copula with $\phi_\alpha(x) = (x^{1-\alpha}-1)/(\alpha-1)$. The copula functions is explicitly given by*

$$C_\alpha(u,v) = \{u^{1-\alpha}+v^{1-\alpha}-1\}^{1/(1-\alpha)}, \tag{4.10}$$

with $\alpha>1$, $(u,v)\in[0,1]^2$. Another parametrization with $\theta=\alpha-1$ is often given instead.

Model (4.10) is valid for positive associations between times. As $\alpha\to 1^+$, $S(s,t)\to S_1(s)S_2(t)$, corresponding to independence between T_1 and T_2, and if $\alpha\to\infty$, $S(s,t)\to\min\{S_1(s),S_2(t)\}$, the bivariate distribution exhibiting maximal association between T_1 and T_2 (Oakes, 1982). Kendall's tau under this model is equal to $\tau_{\mathcal{K}}=\frac{\alpha-1}{\alpha+1}$.

Among the many possible choices for a copula we are choosing Clayton's archimedean copula because it has many interesting features. First, it is equivalent to a model with constant cross-ratio (Clayton, 1978). Second, it is equivalent to the gamma frailty model (Oakes, 1989), and third, if T_1 and T_2 are the lifetimes corresponding to events $\mathcal{E}_1$ and $\mathcal{E}_2$, then $\lambda_2(t|T_1=s)=\alpha\lambda_2(t|T_1>s)$. That is, the conditional hazard for $\mathcal{E}_2$ given the occurrence of T_1 is α times the conditional hazard for $\mathcal{E}_2$ given that $\mathcal{E}_1$ will occur after s. This relationship gives an intuitive interpretation of parameter α.

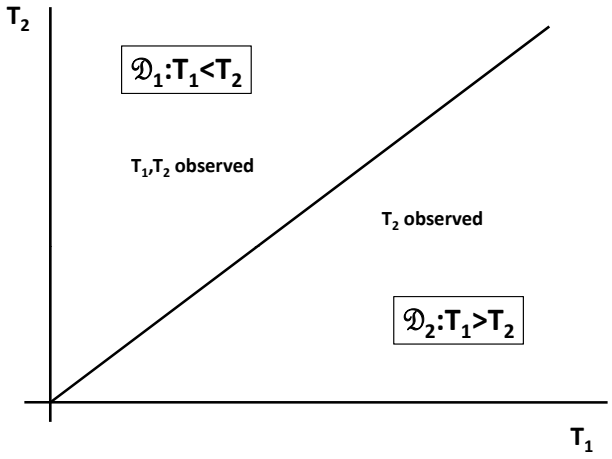


Figure 4.3: Regions of observation in the semi-competing risks framework.

4.2 Semi-competing risks

4.2.1 Semi-competing risks data

As we have briefly discussed in the introduction of this chapter, semi-competing risks data arises as a consequence of having two events of interest in such a way that one of them, $\mathcal{E}_2$, prevents the observation of the other, $\mathcal{E}_1$, and thus, T_2 might censor T_1, but not viceversa.

Let C be a censoring time independent of (T_1, T_2), and define $X = \min(T_1, T_2, C)$, $\delta_1 = I(T_1 < \min(T_2, C))$, $Y = \min(T_2, C)$ and $\delta_2 = I(T_2 < C)$, where $I(.)$ is the indicator function. Now, a sample of semi-competing risks data is denoted by $\mathfrak{D}^{\mathfrak{r}}$, the sample of 4-dimensional vectors on n independent individuals

$$\mathfrak{D}^{\mathfrak{r}} = \{(X_i, \delta_{1i}, Y_i, \delta_{2i}), i = 1, \cdots, n\}.$$

In this setup, while the distribution of T_2 is only subject to independent right-censoring, and hence it can be observed in all the plane defined by (T_1, T_2) and can be consistently estimated only from observed data, the distribution of (T_1, T_2) is only nonparametrically identifiable in the upper wedge of the plane, that is, in the region $\mathcal{D}_1 = \{(s, t)|s \leq t\}$, where both events are observed (Figure 4.3).

An appropriate nonparametric estimator for the joint survival function $S(s, t)$ is given by

$$\widehat{S}(s,t) = \frac{1}{\widehat{G}(t)} \left\{ \frac{1}{n} \sum_{i=1}^{n} I(X_i > s, Y_i > t) \right\},$$

where $G(t) = P(C > t)$ is the survival function of the censoring time C, and $\widehat{G}(t)$ is its Kaplan-Meier estimator evaluated from observed data $\{(Y_i, 1 - \delta_{2i}), i = 1, \ldots, n\}$. It can be proved that $\widehat{S}(s, t)$ is uniformly consistent for $S(x, y)$ for $0 \leq x \leq y \leq \tau$, where $P(\min(T, C) > \tau) > 0$ (that is, $S_T(\tau)G(\tau) > 0$) (Lin and Ying, 1993). This estimator is derived from the following proposition.

Proposition 4.2. *In $\mathcal{D}_1 = \{(s, t)|s \leq t\}$, we can rewrite the joint survival function as*

$$S(s,t) = \frac{P(X > s, Y > t)}{G(\max(s, t))}.$$

Proof. Indeed, for $s \leq t$

$$\begin{aligned} P(X > s, Y > t) &= P(\min(T_1, \min(T_2, C)) > s, \min(T_2, C) > t) \\ &= P(T_1 > s, T_2 > s, C > s, T_2 > t, C > t) \\ &\underset{C \perp (T_1, T_2)}{=} P(T_1 > s, T_2 > s, T_2 > t) P(C > s, C > t) \\ &\underset{\max(s,t)=t}{=} P(T_1 > s, T_2 > t) P(C > t) = S(s,t) G(t). \end{aligned}$$

The symbol $\perp$ indicates independence. □

Since X and Y are observation times, it is natural to estimate $P(X > s, Y > t)$ by the empirical survival function $n^{-1} \sum_{i=1}^{n} I(X_i > s, Y_i > t)$.

Corollary 4.1. *$S(s, 0)$ cannot be recovered from observed data, and thus $S_1(s) = S(s, 0)$, is not empirically identifiable.*

Proof. Indeed, for $s > 0$, from the previous expression,

$$\begin{aligned} S(x, 0) = P(X > s, Y > 0) &= P(\min(T_1, \min(T_2, C)) > s, \min(T_2, C) > 0) \\ &= P(T_1 > s, T_2 > s, C > s, T_2 > 0, C > 0) \\ &\underset{C \perp (T_1, T_2)}{=} P(T_1 > s, T_2 > s) P(C > s) = S(s, s) G(s). \end{aligned}$$

□

To recover the marginal distribution $S_1(s)$, a model for the joint survival function is needed.

4.2.2 Clayton's copula model for semi-competing risks data

The proposal of Fine *et al.* (2001) is to posit a Clayton's copula model (4.10) to describe the association between T_1 and T_2 in the upper wedge $\mathcal{D}_1$. This copula model expresses the joint survival function $S(s, t)$ in terms of the marginal survival functions of T_1 and T_2, $S_1(s) = P(T_1 > s)$ and $S_2(t) = P(T_2 > t)$, respectively, by

$$S(s, t) = P(T_1 > s, T_2 > t) = \left(S_1(s)^{1-\alpha} + S_2(t)^{1-\alpha} - 1 \right)^{1/(1-\alpha)}. \tag{4.11}$$

The specification of the copula model only in the upper wedge $\mathcal{D}_1$ implies that not all the properties of this model are valid. In particular, the relationship of the association parameter with Kendall's τ, $\tau_{\mathcal{K}} = (\alpha - 1)/(\alpha + 1)$, does not hold. Moreover, though the general Clayton's copula model is equivalent to the gamma frailty model (Clayton, 1978), when restricted to $\mathcal{D}_1$, the association parameter α is no longer interpretable as the variance of the gamma random variable (Jiang *et al.*, 2005a). On the contrary, Day *et al.* (1997) showed that the property of the cross-ratio being constant and equal to α is valid even restricted to the upper wedge,

$$\theta(s, t) = \alpha \quad \forall (s, t) \in \mathcal{D}_1.$$

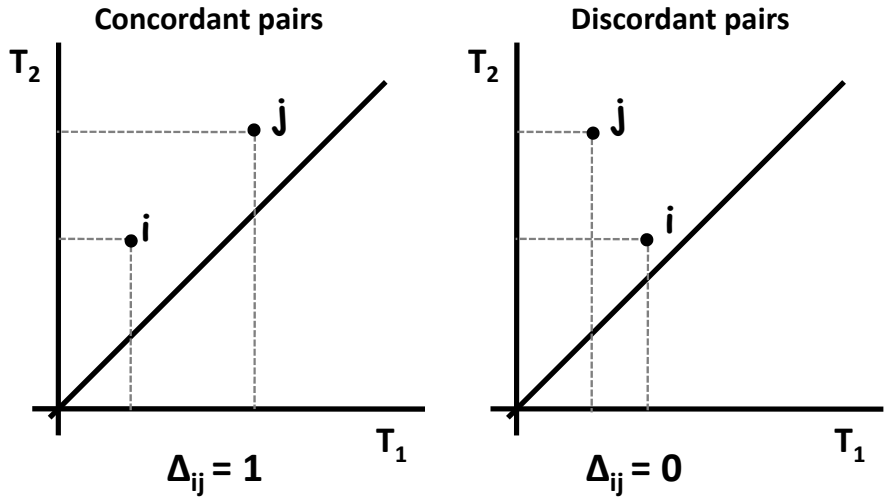


Figure 4.4: Example of concordant pairs ($\Delta_{ij} = 1$) and discordant pairs ($\Delta_{ij} = 0$).

Some hints on the proof are found in Appendix B.2.

Let $f(s,t) = \partial^2 S(s,t)/\partial s \partial t$ be the density function of model (4.11). The joint distribution on the lower wedge $\mathcal{D}_2 = \{(s,t)|s > t\}$ is unspecified, but its joint density function $f_{\mathcal{D}_2}$ must satisfy

$$S(s,t) = \int_t^\infty \int_s^v f(u,v)dudv + \int_t^\infty \int_v^\infty f_{\mathcal{D}_2}(u,v)dudv.$$

There are infinite solutions to the previous integral equation, but it is difficult to find one which is valid. Fine *et al.* (2001) propose a class of joint distributions satisfying that $S_1(s) = S(s,0)$ in the lower wedge. We assume, in the rest of this work, that this condition holds, and thus $S_1(s)$ and $S_2(t)$ are interpreted as the marginal survival functions of T_1 and T_2, respectively.

4.2.3 Estimation under Clayton's copula model

In the usual bivariate survival framework, estimation of the association parameter α can be obtained by maximization of a pseudolikelihood derived from inserting consistent estimates for $S_1(\cdot)$ and $S_2(\cdot)$ into the likelihood function (Shih, 1998). One can use their corresponding Kaplan-Meier estimates, and a consistent estimate for α is obtained. In the case of semi-competing risks data, this strategy is not valid: since the marginal distribution of T_1 is empirically non-identifiable, without an estimator of α, a consistent estimate of $S_1(\cdot)$ may not exist. Therefore, the proposal of Fine and colleagues is to obtain a closed-form estimate of α based on a concordance measure between pairs of individuals, which does not depend on $S_1(\cdot)$ nor on $S_2(\cdot)$. Then, an estimate for $S_1(\cdot)$ is proposed, based on consistent estimates of α, $S_2(\cdot)$ and $S_T(\cdot)$, the survival function of $T = \min(T_1, T_2)$, the time to the first event occurring.

4.2.3.1 Estimation of the association parameter α

A consistent estimator for the association parameter α is obtained based on the concordance indicator between two pairs (T_{1i}, T_{2i}) and (T_{1j}, T_{2j}) of individuals i and j:

$$\Delta_{ij} = I((T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0).$$

Figure 4.4 shows an example of concordant pairs ($\Delta_{ij} = 1$) and discordant pairs ($\Delta_{ij} = 0$).

In the presence of right-censoring, the concordance indicator between two individuals cannot always be determined. It is necessary that the minimum T_1 and the minimum T_2 of the two individuals are exactly observed in the observable region $\mathcal{D}_1$. This condition is formally expressed in the following definition of comparable pairs:

Definition 4.4. *A pair (i, j) is said to be comparable if*

$$\widetilde{T}_{1ij} < \widetilde{T}_{2ij} < \widetilde{C}_{ij}$$

where where $\widetilde{T}_{1ij} = \min(T_{1i}, T_{1j})$, $\widetilde{T}_{2ij} = \min(T_{2i}, T_{2j})$ and $\widetilde{C}_{ij} = \min(C_i, C_j)$.

Let $O_{ij}^R = I\left(\widetilde{T}_{1ij} < \widetilde{T}_{2ij} < \widetilde{C}_{ij}\right)$ be the indicator to determine the comparable sample. Then the set of comparable pairs for right-censored semi-competing risks data is denoted by

$$\mathcal{C}^R = \{(i, j) \in C_{n,2} | O_{ij}^R = 1\}, \tag{4.12}$$

where $C_{n,2}$ is the set of all $\binom{n}{2}$ combinations of 2 integers (i, j), $i < j$ chosen from $(1, 2, \ldots, n)$.

Figure 4.5 shows examples of comparable and non-comparable pairs.

Proposition 4.3. *Under Clayton's copula model, it is satisfied that $E[\Delta_{ij}] = \alpha/(1 + \alpha)$.*

The proof of this proposition can be found in Appendix B.3. This property is true even when the model is only assumed in the upper wedge, thanks to the fact that the cross-ratio function or predictive hazard ratio $\theta(s, t)$ is constant. Given a sample of semi-competing risks data $\{(X_i, \delta_{1i}, Y_i, \delta_{2i}), i = 1, \ldots, n\}$, an estimate $\widehat{\alpha}$ is obtained as the root of the estimating equation

$$U^R(\alpha) = \binom{n}{2} \sum_{i<j} W(\widetilde{X}_{ij}, \widetilde{Y}_{ij}) O_{ij}^R \left\{ \Delta_{ij} - \frac{\alpha}{\alpha + 1} \right\} = 0, \tag{4.13}$$

where $W(u, v)$ is a weight random function satisfying $\sup_{u,v} |W(u, v) - \widetilde{W}(u, v)| \to 0$ in probability, where $\widetilde{W}(u, v)$ is a deterministic and bounded function for (u, v) in the support of $(\widetilde{X}_{ij} = \min(X_i, X_j), \widetilde{Y}_{ij} = \min(Y_i, Y_j))$. Fine *et al.* (2001) propose the use of function

$$W_{a,b}^{-1}(x, y) = n^{-1} \sum_{i=1}^{n} I(X_i \geq \min(a, x), Y_i \geq \min(b, y)), \tag{4.14}$$

where a and b are constants. When $a = b = 0$, it corresponds to $W = 1$. When $a = b = \infty$, the contribution of each pair is weighted by the size of the risk set at their observation times, and it results the same estimator as in Oakes (1986) for the case of bivariate survival data.

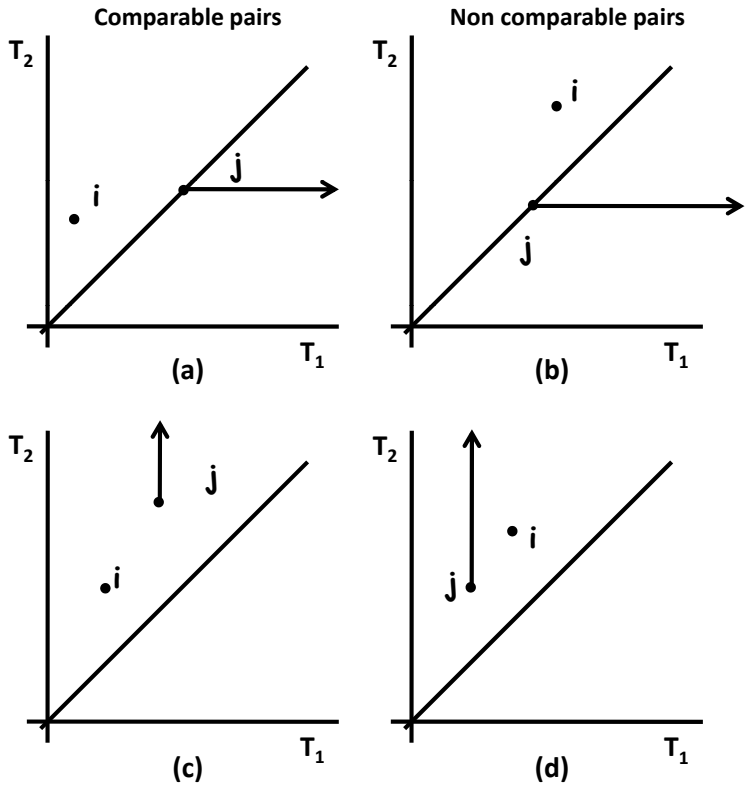


Figure 4.5: Examples of comparable and non-comparable pairs: (a) Comparable pair: $(\widetilde{T}_{1ij} = T_{1i}, \widetilde{T}_{2ij} = T_{2i}) \in \mathcal{D}_1$, (b) non-comparable pair: $\widetilde{T}_{1ij}$ is not determined, and $\widetilde{T}_{2ij} = T_{2j} \leq \widetilde{T}_{1ij}$ (c) comparable pair: $(\widetilde{T}_{1ij} = T_{1i}, \widetilde{T}_{2ij} = T_{2i}) \in \mathcal{D}_1$, (d) non-comparable pair: $\widetilde{T}_{1ij} = T_{1i}$ but $\widetilde{T}_{2ij}$ is not observed.

Proposition 4.4. *$U^R(\alpha)$ is a unbiased zero-mean random variable, $E[U^R(\alpha)] = 0$.*

Proof. We compute the expectation of $U^R(\alpha)$:

$$\begin{aligned}
\mathrm{E}\left[U^R(\alpha)\right] &= \binom{n}{2} \sum_{i<j} \mathrm{E}\left[W(\widetilde{X}_{ij}, \widetilde{Y}_{ij}) O_{ij}^R \left\{\Delta_{ij} - \frac{\alpha}{\alpha+1}\right\}\right] \\
&= \binom{n}{2} \sum_{i<j} \mathrm{E}\left[\mathrm{E}\left[W(\widetilde{X}_{ij}, \widetilde{Y}_{ij}) O_{ij}^R \left\{\Delta_{ij} - \frac{\alpha}{\alpha+1}\right\} \mid \widetilde{X}_{ij}, \widetilde{Y}_{ij}\right]\right] \\
&\underset{(i)}{=} \binom{n}{2} \sum_{i<j} \mathrm{E}\left[W(\widetilde{X}_{ij}, \widetilde{Y}_{ij}) \mathrm{E}\left[O_{ij}^R \mid \widetilde{X}_{ij}, \widetilde{Y}_{ij}\right] \mathrm{E}\left[\Delta_{ij} - \frac{\alpha}{\alpha+1} \mid \widetilde{X}_{ij}, \widetilde{Y}_{ij}\right]\right] \\
&\underset{Prop.4.3}{=} \binom{n}{2} \sum_{i<j} \mathrm{E}\left[W(\widetilde{X}_{ij}, \widetilde{Y}_{ij}) \mathrm{E}\left[O_{ij}^R \mid \widetilde{X}_{ij}, \widetilde{Y}_{ij}\right] \cdot 0\right] = 0.
\end{aligned}$$

(i) O_{ij} and Δ_{ij} are independent conditionally on $\widetilde{X}_{ij}$ and $\widetilde{Y}_{ij}$ (Oakes, 1986). □

The estimate $\widehat{\alpha}$, the unique solution of equation (4.13), is explicitly given by

$$\widehat{\alpha} = \frac{\sum_{i<j} W(\widetilde{X}_{ij}, \widetilde{Y}_{ij}) O_{ij}^R \Delta_{ij}}{\sum_{i<j} W(\widetilde{X}_{ij}, \widetilde{Y}_{ij}) O_{ij}^R (1 - \Delta_{ij})}, \tag{4.15}$$

and is shown to be strongly consistent for α, and asymptotically normal (Fine *et al.*, 2001). An intuitive interpretation of parameter α is derived from expression (4.15): if $W(u,v) = 1$ for any (u,v), then α is the ratio between concordant and discordant pairs.

4.2.3.2 Inference on the marginal $S_1(\cdot)$

The marginal survival function of T_1 evaluated at time s, $S_1(s)$ can be isolated from Clayton's copula equation (4.11) with $s = t$ and be expressed as a function of $S(s,s)$, $S_2(s)$ and α. In addition, the bivariate survival S when $s = t$ can be expressed in terms of $T = \min(T_1, T_2)$:

$$S(s,s) = P(T_1 > s, T_2 > s) = P(\min\{T_1, T_2\} > s) = P(T > s) = S_T(s). \tag{4.16}$$

Thus, $S_1(s)$ can be expressed as a function of $S_T(s)$, $S_2(s)$ and α. Specifically:

$$S_1(s; S_T(s), S_2(s), \alpha) = g(S_T(s), S_2(s), \alpha), \tag{4.17}$$

where $g(a,b,c) = (a^{1-c} - b^{1-c} + 1)^{1/(1-c)}$. Function g is continuous and has bounded derivatives.

Consider the following estimates for α, $S_2(s)$ and $S_T(s)$: $\widehat{\alpha}$ given in (4.15), and the Kaplan-Meier estimates $\widehat{S}_T(s)$ and $\widehat{S}_2(s)$ based, respectively, on $\{(X_i, \delta_{Ti} = \delta_{1i} + (1-\delta_{1i})\delta_{2i}), i = 1, \ldots, n\}$, and $\{(Y_i, \delta_{2i}), i = 1, \ldots, n\}$. Then we can plug-in the estimates instead of the theoretical values in expression (4.17) to obtain an estimator of $S_1(\cdot)$:

$$\widehat{S}_1(s) = g(\widehat{S}_T(s), \widehat{S}_2(s), \widehat{\alpha}). \tag{4.18}$$

$\widehat{S}_T(s)$ and $\widehat{S}_2(s)$ are known to be strongly consistent in $s \in [0, \tau]$, with τ satisfying $\widehat{S}_T(\tau)\widehat{G}(\tau) > 0$ with $\widehat{S}_T$ as defined above, and $\widehat{G}$ the Kaplan-Meier estimate of $G(t) = P(C > t)$ as defined in page 75. Since $\widehat{\alpha}$ is consistent, the following theorem ensures the strong consistency of $\widehat{S}_1(s)$ in $s \in [0, \tau]$:

Theorem 4.1. *Continuous mapping theorem*

(a) Let $\{X_n\}$ be a sequence of random variables, converging in probability to a, $X_n \xrightarrow{P} a$, and g a continuous function. Then $g(X_n) \xrightarrow{P} g(a)$.

(b) Let $\{X_n\}$ be a sequence of random variables, converging in distribution to the random variable X, $X_n \xrightarrow{D} X$, and g a continuous function. Then $g(X_n) \xrightarrow{P} g(X)$.

(c) Let $\{W_n\}_n$ be a consistent sequence of estimators of θ. Let g be a continuous function defined in Θ. Then, $\{g(W_n)\}_n$ is a consistent sequence for $g(\theta)$.

Moreover, it can be be shown that $n^{1/2}(\widehat{S}_1(t) - S_1(t))$ converges weakly to a Gaussian process for $t \in [0, \tau]$. The proposed estimator $\widehat{S}_1(s)$ is a step function which jumps at the observed values of T_1 and T_2 such that $\widehat{S}_T(s)^{1-\widehat{\alpha}} - \widehat{S}_2(t)^{1-\widehat{\alpha}}$ jumps. It may be non-monotone or not well defined: in finite samples $\widehat{S}_T(s)$ may be greater than $\widehat{S}_2(t)$, although $S_T(s) \leq S_2(s)$ for all s. In addition, $\widehat{\alpha}$ might be less than one.

To address this issue, the authors propose the isotonic estimate $\widehat{S}_1^*(s) = \min_{s \leq t}\{\widehat{S}_1(s)\}$, where $t \in [0, t^*]$, with t^* satisfying

$$t^* \leq \max\{s : \widehat{S}_T(s)^{1-\widehat{\alpha}} - \widehat{S}_2(t)^{1-\widehat{\alpha}} > -1, 0 \leq \widehat{S}_1(u) \leq 1, u \leq s\}.$$

4.2.4 Alternative models

The concept of semi-competing risk is introduced by Fine *et al.* (2001), however, as a special bivariate distribution with dependent censoring has been studied in a large number of papers. Day *et al.* (1997) use Clayton's copula to model the joint survival function of (T_1, T_2) in the observable region $\mathcal{D}_1$, though the estimation of the copula parameter is achieved modifying the likelihood proposed originally by Clayton (1978), whilst the proposal of Fine extends the methods in Oakes (1982, 1986). We refer to the original paper by Fine *et al.* (2001) and to a nice summary by Jiang *et al.* (2003) to deepen on this particular work.

Wang (2003) propose a more general model allowing the dependence structure to vary with time, by the use of the Archimedean copula family (4.9), stating the relationship

$$S(s,t) = \phi_\alpha^{-1}\left[\phi_\alpha\{S_1(s)\} + \phi_\alpha\{S_2(t)\}\right].$$

Complex estimating equations are proposed to obtain the dependence parameter α. Lakhal *et al.* (2008) chooses as well the Archimedean family to model the association between times, but the estimation procedures are based in the concordance indicator. In fact, they propose a common framework in which the estimates given by Day *et al.* (1997), Fine *et al.* (2001) and Wang (2003) can be included.

When it comes to the estimation of the marginal $S_1(s)$, some criticism has been done to the simple plug-in proposal of Fine's. This estimator cannot be considered as a generalization of the product-limit estimator, since it jumps outside the observed times X, and monotonicity is not granted, which is the reason why $\widehat{S}_1^*$ must be defined. Alternatives have been proposed: Jiang *et al.* (2005a) obtain an estimate satisfying pseudo-self consistent equations, while Lakhal *et al.* (2008) use the so-called copula-graphic estimator proposed by Zheng and Klein (1995) in the context of competing-risks. Both estimators are shown to perform better than the original plug-in estimate, but similar between them. In the following chapters when accounting for interval censoring, we adopt the plug-in estimator, and we explore how other estimates can be extended.

The presence of covariates is indeed a relevant issue and Ghosh (2006), under Clayton's copula model, proposes a class of rank statistics to test whether the association between T_1 and T_2 remains constant across strata for a discrete covariate Z. Hsieh *et al.* (2008) also considers a single discrete covariate, but proposes flexible copula models for each stratum of the covariate. Peng and Fine (2007) propose a very general framework of regression modelling in semi-competing risks, where the effect of the covariates on the marginal $S_1(\cdot)$ are included via a functional regression model, and a general time-dependent copula model is adopted. Non-linear estimating equations are derived for both the dependence model and the marginal of T_1.

The extension of semi-competing risks models to left-truncation data is proposed in Jiang *et al.*

(2005b) and in Peng and Fine (2006). To our knowledge, no work has been done to extend semi-competing risks data in the presence of interval censoring.

4.3 Numerical examples

4.3.1 An example based on simulated data

We present an example of simulated data to illustrate a right-censored semi-competing risks analysis. This example completes the one presented in the motivation of this Chapter, thought as to mimic a clinical trial for proving the effectiveness of a treatment A over a treatment B.

Bivariate correlated data of (T_1, T_2), corresponding to the times until the intermediate $\mathcal{E}_1$ and terminating $\mathcal{E}_2$ events, was simulated following Clayton's copula 4.11 model to describe the association structure. For both treatment arms, a copula parameter of $\alpha_A = \alpha_B = 4$ was chosen, but marginal distributions were taken in such a way that treatment A reduced significantly the proportion of events in both T_1 and T_2 (HR 1.6 for both times, p-value<0.0001). This was achieved by picking up Exponential distributions for both T_1 and T_2 with rates $\lambda_1^A = \lambda_2^A = 0.025$ ($E[T_1] = E[T_2] = 40$), for treatment A, and rates $\lambda_1^B = \lambda_2^B = 1.6\lambda_2^A = 0.4$ for treatment B. Data for 500 individuals was generated, half at each treatment arm. In Figure 4.2 the distribution function of the marginals at each treatment arm is depicted. An independent censoring variable C was generated following a Uniform distribution, $C \sim U[0, 200]$, providing 18.4% of non-informative censoring in our data.

This way, we obtained a sample of right-censored semi-competing risks data of 500 individuals,

$$\{(X_i, Y_i, \delta_{1i}, \delta_{2i}, Z_i), i = 1, .., 500\},$$

where $X_i = \min(T_{1i}, T_{2i}, C_i)$, $Y_i = \min(T_{2i}, C_i)$, $\delta_{1i} = I(T_{1i} < \min(T_{2i}, C_i))$, $\delta_{2i} = I(T_{2i} \leq C_i)$ and Z_i is the treatment indicator.

A competing risks approach would result from analysing the sample $X = \min(T, C)$, where $T = \min(T_1, T_2)$, together with an indicator of the type of the first event observed. Within our data, 37.6% of patients experience $\mathcal{E}_1$ as a first event, while 46.1% experience $\mathcal{E}_2$ as a first event. The rest are censored observations. We already showed in the opening of this Chapter that the cumulative incidence functions for $(T, 1)$ were far from the true marginal distribution T_1 at each treatment arm (Figure 4.2). Moreover, this analysis did not provide significant differences between treatment arms at the cumulative incidence levels.

In the semi-competing risks framework, however, we analyse the bivariate distribution to estimate the association between the times, correct the existing dependent censoring and recover the marginal scale. The dependence censoring is quantified by individuals for whom the terminating event is observed to occur before the intermediate event ($\delta_1 = 0$, $\delta_2 = 1$). This corresponds to 234 individuals (46.8%): this simulated data set exhibits a large percentage of dependent censoring.

The estimates for α obtained by Fine's method for right-censored data are $\widehat{\alpha}_A = 4.12(0.42)$ for treatment A and $\widehat{\alpha}_B = 3.87(0.29)$ for treatment B (standard deviations of the estimators in parentheses). Both estimates are reasonably close to the true value $\alpha = 4$. Figure 4.6 shows the estimates

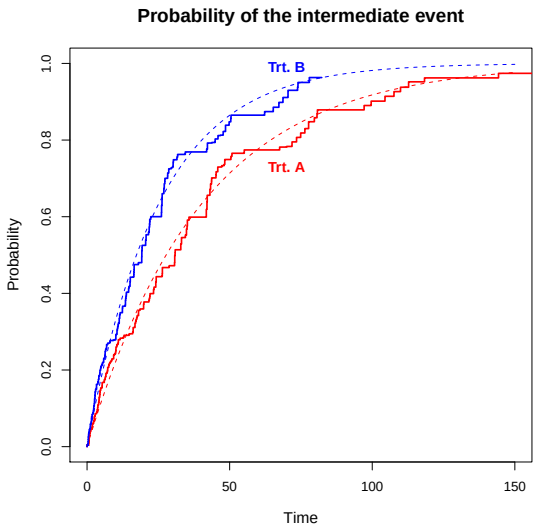


Figure 4.6: Semi-competing risks analysis for right-censored data: Estimated distribution function vs real distribution function for T_1.

for $1 - S_1(t)$ at each treatment arm. We see that the estimators appropriately approximate the marginal distributions. Most importantly, the distance between estimators reflect the marginal effect of treatments. Therefore, a semi-competing risks analysis permits to recover the significant marginal effect.

4.3.2 Semi-competing risks with interval-censored intermediate event

An additional complication in the analysis of semi-competing risks data is the possibility that the intermediate event is interval-censored. This situation is frequent in medical studies, where the status of the disease or the event of interest are assessed only at scheduled visits. The simplest approach to deal with this kind of censoring is to convert the data into right-censored semi-competing risks by imputation of a single point. A common choice is to assign the midpoint of the censoring interval to T_1, $T_1 = L + \frac{R-L}{2}$. However, other choices are possible, for instance, whenever interval censoring is present, to impute the left extreme of the interval, $T_1 = L$, or the right extreme of the interval $T_1 = R$.

We use the above simulated example to explore the impact of the imputation choice (left, midpoint or right imputation) in the estimators obtained. With this aim, we have generated intervals of observation for T_1: whenever T_1 belonged to the upper wedge $\mathcal{D}_1$ of the plane $T_1 \times T_2$, we generated L and R such that $T_1 \in (L, R]$, $R < \infty$ (the method to generate such intervals is explained in Chapter 9).

The results for the estimation of α are presented in Table 4.1, and the estimators for $1 - S_1(t)$ are presented in Figure 4.7.

We note in this example that the choice of the imputed value has a great impact on the results, giving very different values for the estimate of α. Left imputation underestimates the true value of α, while right imputation greatly overestimates the true value of α. This pattern is meaningful

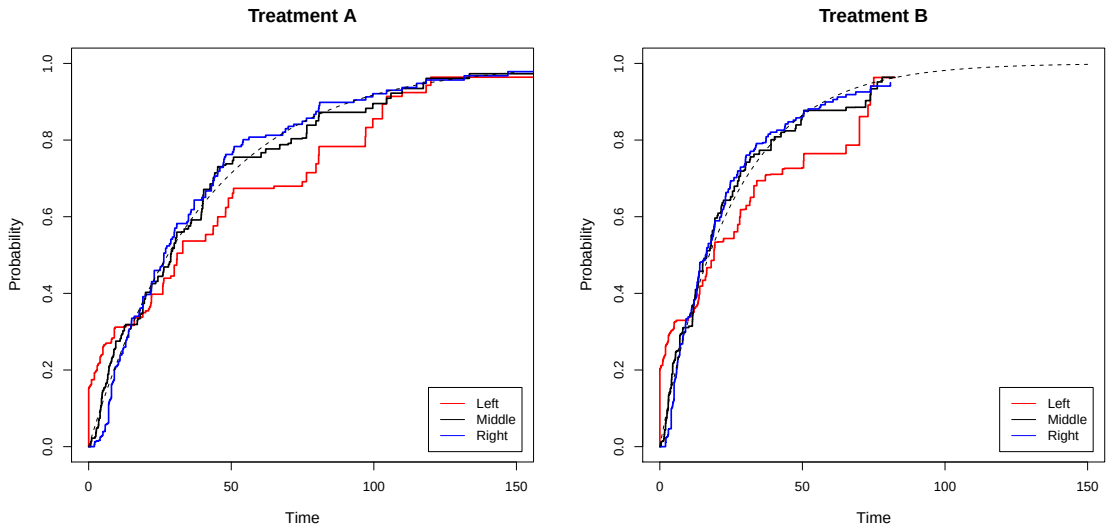


Figure 4.7: Illustration 1: Impact of imputation methods on the estimation of $1 - S_1(t)$, by treatment arm.

Table 4.1: Estimates of α resulting from semi-competing risks analysis. Imputation methods for the simulated example

	$\widehat{\alpha}(SE)$	
Method	**Trt. A**	**Trt. B**
Left imputation	2.55 (0.27)	2.34 (0.19)
Midpoint imputation	3.93 (0.41)	4.07 (0.34)
Right imputation	7.61 (1.17)	12.30 (2.67)

[a] SE: standard error.

[b] Fine's method with $W_{\infty,\infty}$ (see Section 4.2.3.1).

since the larger the imputed value is, the more we favor positive association between T_1 and T_2. In this example, the estimates obtained by midpoint imputation are the best for both, α and $S_1(t)$. The different results obtained of the different imputation schemes and the fact that none of this imputations takes into account the uncertainty contained in the censoring interval suggest the need for a specific method, as the one we propose in Chapter 6, that appropriately deals with interval-censored semi-competing risks data. Before, next Chapter is a review of some methods available in the literature of interval censoring

CHAPTER 5

Methods for interval-censored data

Interval censoring appears when the time of interest is not exactly observed, but is known to occur within an observed interval of time. This censoring pattern is frequent in longitudinal studies in many areas of medical research, where the occurrence of the event can often be recorded only at periodic follow-ups. Interval censoring may also arise when an individual misses one or more scheduled visits, and when returns, its disease status has changed. Examples of interval-censored data are found, in particular, in dentistry studies when analysing time to caries development or to emergence of a tooth, where the event occurs between consecutive visits of a patient. Another example is found in the evolution of an individual after diagnosis of a complex disease such as cancer or HIV, where biological markers are used to asses the relapse of disease. Relapse is defined as the time until exceeding a determined threshold of a biological marker. If such a marker is not obtained continuously on time, interval censoring between consecutive measurements might arise.

Despite the fact that there exists specific methodology to tackle with interval-censored data, it is still common practice to simplify the problem and use imputation approaches to reduce to the right-censoring framework. Though they might obtain biased estimates, or underestimate the variability of point estimates (Sun, 2006), their use is widespread, one of the reasons for this being the lack of software and/or implemented routines to perform specific analysis accounting for interval-censored data. In order to invert this tendency and make them available to a more general audience, the recent tutorial by Gomez *et al.* (2009) is focused on the implementation of interval censoring methods in `R`.

Section 5.1 provides the most relevant procedures of univariate interval-censored data, without intending to be exhaustive. Other reviews on this topic are: Gómez *et al.* (2004), Lesaffre *et al.* (2005), Sun (2006), and recently, Zhang and Sun (2010). We focus on maximum-likelihood estimation methods to estimate nonparametrically the survival function (Section 5.1.2), and to perform regression modelling via parametric models (Section 5.1.3). We review the literature available ad-

dressing the problem of competing risks when interval-censored data is present in Section 5.1.4, illustrating the methodologies with the data from the Spanish Bladder Cancer/EPICURO Study.

Section 5.2 is devoted to the existing references dealing with bivariate interval-censored data. The general framework is defined in Section 5.2.1. In the analysis of bivariate interval-censored data, two difficulties arise: (i) the presence of interval censoring, and (ii) the treatment of the correlation structure between the times of interest. In Sections 5.2.2 and 5.2.3 we focus, respectively, on existing methods to deal with nonparametric estimation of the bivariate distribution function, and on estimation of the correlation structure. Issues on multi-state models or regression modelling are not covered here (see, for instance, Sun (2006) or Cook *et al.* (2008)). In section 5.2.4, we expose the difficulties to apply the presented existing methods when dependent censoring (and thus, a structure of semi-competing risks data) is present.

5.1 Univariate interval censoring

The notation used throughout this work is the following. Let T be a nonnegative random variable representing the time of interest, and let L and R be the times such that $T \in (L, R]$ with probability one. In addition, $L \leq R$ almost surely. We use the convention that $L = R$ means an exact observation, and $R = \infty$ represents a right-censored observation (Sun, 2006).

5.1.1 Noninformativity conditions

The majority of methods for interval-censored data are based on the assumption that the censoring mechanism is noninformative. Noninformativeness means that the mechanism that generates the censoring is noninformative for the underlying variable of interest, T. In other words, the observed interval $(L, R]$ carries no further information on the survival time T other than the fact that T lies in the interval $(L, R]$. This assumption permits evaluating censored data without modelling the censoring process.

We adopt the noninformativity conditions in Oller *et al.* (2004) where three equivalent characterizations of noninformativeness are given, conditions ensuring that the censoring mechanism cannot affect the distribution of T; moreover, they provide a constant-sum property which ensures that the inference process can omit the randomness of the intervals. In Oller *et al.* (2007), the authors study the relevance of this constant-sum property in the identifiability of the lifetime distribution. For technical details, we refer the reader to the original papers.

These properties describing non-informativeness guarantee that the contribution to the likelihood function of an individual with observed interval $(\ell, r]$,

$$\int_L^R f_{T,L,R}(t, \ell, r)dt = P(T \in (L, R], L \in d\ell, R \in dr), \tag{5.1}$$

is proportional to $P(T \in (\ell, r])$, that is, the probability that T belongs to $(\ell, r]$ ignoring the censoring mechanism. This probability is denoted as simplified likelihood.

5.1.2 Nonparametric maximum likelihood estimation

Consider first the problem of estimating nonparametrically the survival function of the lifetime of interest, $S(t) = P(T > t)$. Let $\{(\ell_i, r_i]; i = 1, \ldots, n\}$ be a sample of interval-censored data from n independent individuals, where the time of interest for the i^{th} individual, T_i, belongs to the observed interval $(\ell_i, r_i]$. Under non-informative censoring, the likelihood function corresponding to this sample is

$$\mathfrak{L}(S) = \prod_{i=1}^{n} \left[S(\ell_i^+) - S(r_i^-) \right]. \tag{5.2}$$

The nonparametric maximum likelihood estimator (NPMLE) of $S(t)$ has been developed by Peto (1973) and Turnbull (1976). Define the sets $\mathcal{L} = \{\ell_i,\ i = 1, \ldots, n\}$ and $\mathcal{R} = \{r_i,\ i = 1, \ldots, n\}$, and the intervals

$$\mathcal{I} = \{(q_1, p_1], \ldots, (q_m, p_m]\}, \tag{5.3}$$

where $q_j \in \mathcal{L}$, $p_j \in \mathcal{R}$, and no other elements from $\mathcal{L}$ and $\mathcal{R}$ are contained in the intervals. Turnbull proved that in this set of intervals $\mathcal{I}$ is where the NPMLE concentrates its mass. That is, he proved that any function maximizing (5.2) is constant between intervals $[q_j, p_j]$ and decreasing within them.

Define the weights for each interval in $\mathcal{I}$, $w_j = P(q_j < T \leq p_j) = S(q_j) - S(p_j)$ for $j = 1, \ldots, m$. If $\alpha_{ij} = I\big((q_j, p_j] \subseteq (\ell_i, r_i]\big)$, then the likelihood (5.2) can be rewritten by

$$\mathfrak{L}(w_1, \ldots, w_m) = \prod_{i=1}^{n} \left(\sum_{j=1}^{m} \alpha_{ij} w_j, \right) \tag{5.4}$$

and thus maximizing (5.2), a non-finite dimensional problem reduces to the finite-dimensional problem of maximizing (5.4), subject to $w_j \leq 0$ and $\sum_{j=1}^{m} w_j = 1$. The NPMLE for $S(t)$ is then given by

$$\widehat{S}(t) = \begin{cases} 1 & \text{if} \quad t \leq q_1 \\ 1 - (\widehat{w}_1 + \cdots + \widehat{w}_k) & \text{if} \quad p_k \leq t \leq q_{k+1},\ 1 \leq k \leq m-1 \\ 0 & \text{if} \quad t \geq p_m, \end{cases}$$

and is not specified within $(q_k, p_k]$, $k = 1 \leq k \leq m$.

While Peto uses the Newton-Raphson algorithm to solve this maximization problem, Turnbull proposes the self-consistency algorithm, based on the simultaneous solution of the following self-consistent equations:

$$\widehat{w}_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\alpha_{ij}}{\sum_{l=1}^{m} \alpha_{il} \widehat{w}_l},\ 1 \leq j \leq m.$$

Given the somehow slow convergence of the self-consistency algorithm, more efficient algorithms have appeared, such as the Iterative Convex Minorant (ICM) algorithm or the EM-Iterative Convex Minorant (EM-ICM) algorithm. Details on these algorithms can be found in the cited reviews at the beginning of this chapter.

5.1.3 Parametric maximum likelihood estimation

We now consider now the problem of regression modelling for interval-censored data via parametric models. That is, assume that T_i follows a parametric model with survival function $S(t;\boldsymbol{\theta},\boldsymbol{Z})$, with $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_p)'$ denoting the unknown parameters of the parametric model, and $\boldsymbol{Z}=(Z_1,\ldots,Z_p)$ a vector of known covariates.

Most common survival models can be expressed as a log-linear model:

$$\ln T=\mu+\boldsymbol{\beta}'\boldsymbol{Z}+\sigma W$$

where W stands for the error term distribution. In this case, the vector of parameters corresponds to $\boldsymbol{\theta}=(\mu,\boldsymbol{\beta},\sigma)$. Common choices for T are the Weibull, the log-logistic and the log-normal model, in which cases the error term W follows the extreme value, the logistic and the normal distribution, respectively.

Given a sample of n individuals with data $\{(\ell_i,r_i],\boldsymbol{Z}_i;i=1,\ldots,n\}$, where $\boldsymbol{Z}_i$ is the vector of covariates of the i^{th} individual, estimates for $\boldsymbol{\theta}$ can be obtained by maximizing the likelihood function

$$\mathfrak{L}(\boldsymbol{\theta})=\prod_{i=1}^{n}\left[S(\ell_i;\boldsymbol{\theta},\boldsymbol{Z}_i)-S(r_i;\boldsymbol{\theta},\boldsymbol{Z}_i)\right]\underset{\text{def.}}{=}\prod_{i=1}^{n}\mathfrak{L}_i(\boldsymbol{\theta}),$$

under the assumption that $\ell_i<r_i$ for all $i=1,\ldots,n$. That is, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ is obtained as the solution of the score equation

$$U(\boldsymbol{\theta})=\sum_{i=1}^{n}\frac{\partial\mathfrak{L}_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}=0.$$

Under some regularity conditions, $\widehat{\boldsymbol{\theta}}$ is consistent and unique, and asymptotically normal with mean $\boldsymbol{\theta}$ and covariance matrix $I^{-1}(\boldsymbol{\theta})$, where $I(\boldsymbol{\theta})=\sum_{i=1}^{n}\partial^2\mathfrak{L}_i(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$.

5.1.4 Competing risks analysis when data is interval-censored

Now we address the problem of competing risks, described extensively in Chapter 2, when data is interval-censored. Literature in this topic is scarce. Recall that a competing risk model is specified through the joint distribution of the time to failure, T, and the cause of failure C, which takes values in the finite set $\{1,\ldots,J\}$. The joint distribution for (T,C) is completely described either by means of the cause-specific hazard functions,

$$\lambda_j(t)=\lim_{\Delta t\to 0}\frac{P\big(T<t+\Delta t,C=j|T\geq t\big)}{\Delta t},$$

representing the rate of occurrence of the j^{th} failure, or by the cumulative incidence function for type j failure,

$$F_j(t)=P(T\leq t,C=j)$$

corresponding to the probability of a subject failing from cause j in the presence of all the competing risks, $j = 1, \ldots, J$. The survival function of time T, $S(t) = P(T > t)$ satisfies that $S(t) = 1 - \sum_{j=1}^{J} F_j(t)$. In our setting, T is interval-censored, that is, there exists L and R such that $T \in (L, R]$ with probability one.

Hudgens *et al.* (2001), in the line of Turnbull's estimator, derived the NPMLE of the cumulative incidence function for each cause of failure when data is subject to interval censoring and truncation. For teh sake of simplicity, we ignore truncation in this exposition. Consider a sample of n observed individuals

$$\{(\ell_i, r_i, \delta_i, c_i), i = 1, \ldots, n\},$$

where $T_i \in (\ell_i, r_i]$, $\delta_i = 1$ if the failure type is known, 0 otherwise, and $c_i \in \{1, \ldots, J\}$ is the type of failure (if known). The likelihood function in this setting is proportional to

$$L = \prod_{i=1}^{n} [F_{c_i}(r_i^+) - F_{c_i}(\ell_i^-)]^{\delta_i} \big[\sum_{j=1}^{J} F_j(r_i^+) - F_j(\ell_i^-) \big]^{1-\delta_i}. \tag{5.5}$$

Let $N_j = \{i : \delta_i = 0 \text{ or } c_i = j\}$ be the index set for observations with type j failure, or having unknown failure type. Define the sets $\mathcal{L}_j = \{\ell_i : i \in N_j\}$ and $\mathcal{R}_j = \{r_i : i \in N_j\}$, and the intervals

$$\mathcal{I}_j = \{(q_{j1}, p_{j1}], \ldots, (q_{jm_j}, p_{jm_j}]\}, \tag{5.6}$$

where $q_{jk} \in \mathcal{L}_j$, $p_{jk} \in \mathcal{R}_j$, and no other elements from $\mathcal{L}_j$ and $\mathcal{R}_j$ are contained in the intervals. The authors provide two lemmas to show that the NPMLE of $F_j(t)$ is constant outside $\mathcal{I}_j$ and increasing in some or all of the intervals.

Define now the weights for each interval in $\mathcal{I}_j$, $w_{jk} = F_j(p_{jk}^+) - F_j(q_{jk}^-)$ for $k = 1, \ldots, m_j$. Let α_{ijk} be an indicator variable equal to one if $[q_{jk}, p_{jk}] \subseteq (\ell_i, r_i]$, and $i \in N_j$, zero otherwise. The likelihood (5.5) can be rewritten by

$$L(\boldsymbol{w_1}, \ldots, \boldsymbol{w_J}) = \prod_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{m_j} \alpha_{ijk} w_{jk}, \tag{5.7}$$

where $\boldsymbol{w_j} = (w_{j1}, \ldots, w_{jm_j})'$, $j = 1, \ldots, J$. Then, maximizing (5.5) reduces to a finite-dimensional problem of maximizing (5.7), subject to $w_{jk} \leq 0$ for all j, k and $\sum_{j=1}^{J} \sum_{k=1}^{m_j} w_{jk} = 1$. Once $(\widehat{\boldsymbol{w}}_{\boldsymbol{1}}, \ldots, \widehat{\boldsymbol{w}}_{\boldsymbol{J}})$ are obtained, the NPMLE of $F_j(t)$ is given by

$$\widehat{F}_j(t) = \begin{cases} 0 & \text{if} \quad t < q_{j1} \\ \widehat{w}_{j1} + \cdots + \widehat{w}_{jk} & \text{if} \quad p_{jk} < t < q_{jk+1}, 1 \leq k \leq m_j - 1 \\ \widehat{w}_{j1} + \cdots + \widehat{w}_{jm_j} & \text{if} \quad t > p_{m_j}, \end{cases}$$

For $t \in [q_{jk}, p_{jk}]$, $\widehat{F}_j(t)$ is undefined if $\widehat{w}_{jk} > 0$ and $[q_{jk}, p_{jk}] \subseteq \mathcal{I}_j$ and equals $\widehat{w}_{j1} + \cdots + \widehat{w}_{jk-1}$ otherwise. If $p_{jm_j} = \infty$ and $w_{jm_j} > 0$, $F_j(t)$ is undefined for $t > q_{jm_j}$. Hudgens propose an EM algorithm to solve the maximization problem.

Special mention deserves the problem of competing risks in the presence of current status data,

object of an active research in recent years. Current status data, also known as interval-censored data case 1, arises when T is only known to be larger or smaller than an observed monitoring time, C. In this case, the study subject is observed only once producing either a left ($T \in (0, C]$) or a right-censored observation ($T \in (C, \infty)$). Jewell *et al.* (2003), Maathuis (2006) and Groeneboom *et al.* (2008a,b) develop the inference of this particular problem.

When it comes to regression modelling, inferences based on cause-specific hazards can be derived using regression methods to account for interval-censored data (some of which are reported in Section 5.1.3). On the other hand, and up to our knowledge, the problem of extending Fine and Gray's model when data are interval-censored has not been addressed in scientific publications.

5.1.4.1 Illustration: the Spanish Bladder Cancer Study

In the competing risks analysis of the time to the first event (see Chapter 2) we computed the cumulative incidence functions for recurrence, progression of disease (including deaths from bladder cancer) and deaths due to other causes. In that analysis, we ignored the presence of interval censoring, but in fact, recurrences and progressions of the tumour were detected between consecutive visits of the patient, and therefore, the exact time of recurrence or progression was known to lie within a censoring interval. It was decided then to use midpoint imputation: it was assumed that the exact time of recurrence or progression was the midpoint of the observed interval.

In this section we want to apply the methodology of Hudgens to estimate the cumulative incidence functions taking into account that: (i) type 1 (recurrence) is interval-censored, and (ii) type 2 (progression or death due to bladder cancer) is interval-censored or exactly observed (death). We compare the results with the nonparametric estimator obtained by midpoint imputation. We must check that the estimated cumulative incidence function for death due to other causes (type 3), which is NOT interval-censored, and thus no imputation is necessary, coincides coincides for both methods. We use the SAS macro implemented by the authors and available in `http://www.bios.unc.edu/~mhudgens/cr_npmle_8.sas`.

Figure 5.1 contains the estimates for the cumulative incidence functions for recurrence and progression (and death due to bladder cancer) obtained by assuming midpoint imputation (black solid line) or accounting for interval censoring (red dashed line). Only small differences are observed between both nonparametric estimators of the cumulative incidence functions, which justify the use of the midpoint imputation in this example for simplicity reasons. However, as we will show in next chapters, this strategy of ignoring interval censoring can lead to biased results in other situations, in particular, when estimating the correlation between two survival times.

5.2 Bivariate interval-censored data

5.2.1 Notation and likelihood function

Let T_1 and T_2 be two survival times observed in the same individual. We denote their bivariate distribution function by $F(s,t)$, their joint survival function by $S(s,t)$, and their corresponding

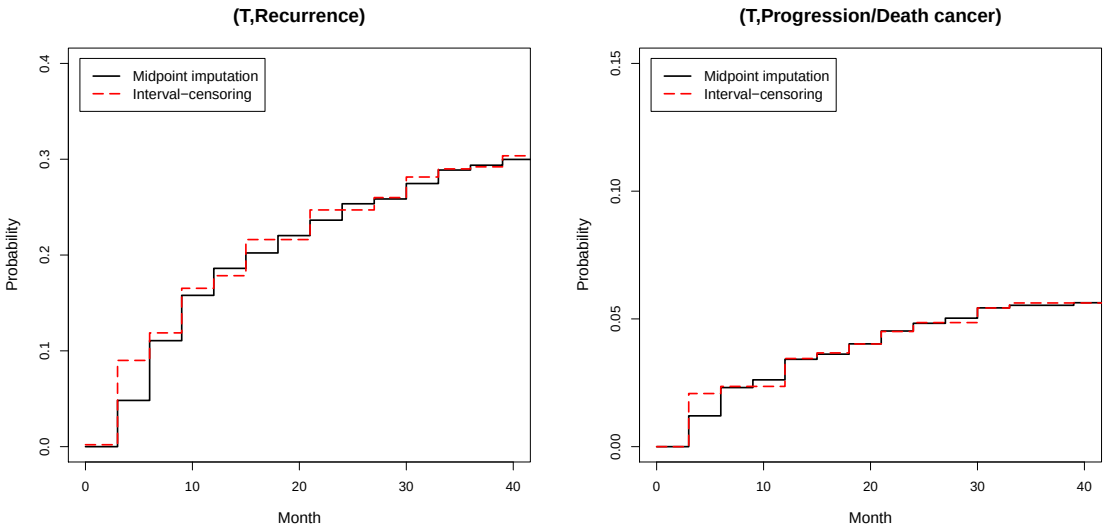


Figure 5.1: Competing risks analysis with interval-censored data: analysis of the time to the first event occurring, either recurrence (left), or progression/death due to bladder cancer (right).

marginal distribution and survival functions by $F_1(s)$, $F_2(t)$, $S_1(s)$ and $S_2(t)$, as defined in Section 4.1.

We assume that both times of interest are interval-censored, that is, there exist L_1, R_1 and L_2, R_2 random variables such as $T_1 \in (L_1, R_1]$ and $T_2 \in (L_2, R_2]$ with probability one, and $L_1 \leq R_1$ and $L_2 \leq R_2$ almost surely.

For n independent individuals, let (T_{1i}, T_{2i}) their bivariate vector of times. Then, the observed bivariate interval-censored data consists of the rectangles B_i

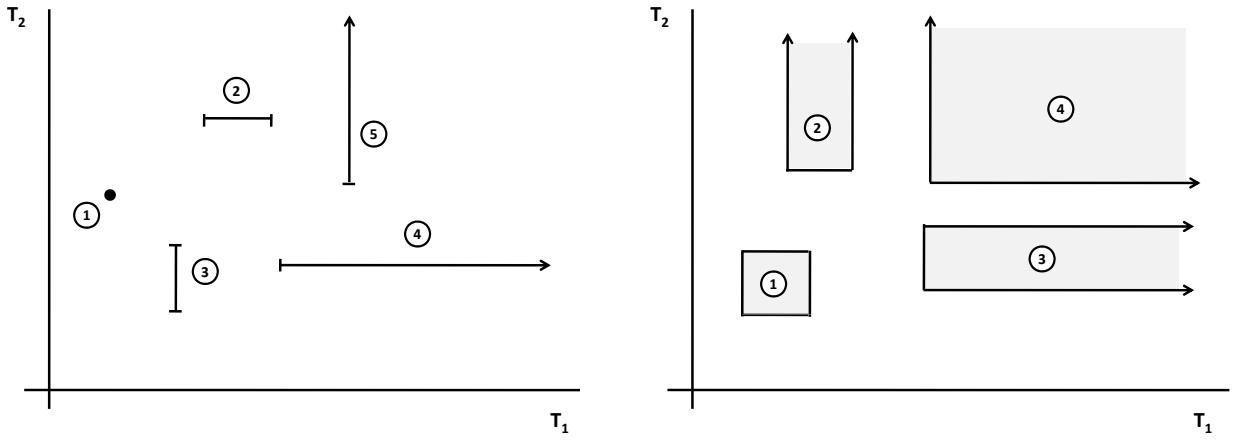
$$\{B_i = (\ell_{1i}, r_{1i}] \times (\ell_{2i}, r_{2i}], i = 1, \ldots, n\}$$

such that $T_{1i} \in (\ell_{1i}, r_{1i}]$ and $T_{2i} \in (\ell_{2i}, r_{2i}]$, and $(\ell_{1i}, r_{1i}, \ell_{2i}, r_{2i})$ are realizations of the vector (L_1, R_1, L_2, R_2). For every individual $i = 1, \ldots, n$, and for each $k = 1, 2$, T_{ki} is either exactly observed ($\ell_{ki} = r_{ki}$), interval-censored ($\ell_{ki} < r_{ki} < \infty$) or right-censored ($r_{ki} = \infty$). Therefore, B_i is a region of the plane $T_1 \times T_2$ which can be a point, a line-segment or a rectangle.

For instance, in Figure 5.2(a), we have plotted five different types of observed individuals, corresponding to the case when either T_{1i} or/and T_{2i} are exactly observed. Then, the regions B_i for each case are:

$$B_i = (\ell_{1i}, r_{1i}] \times (\ell_{2i}, r_{2i}] = \begin{cases} \{l_{1i}\} \times \{l_{2i}\} & \text{for} \quad i = 1 \\ (l_{1i}, r_{1i}] \times \{l_{2i}\} & \text{for} \quad i = 2 \\ \{l_{1i}\} \times (l_{2i}, r_{2i}] & \text{for} \quad i = 3 \\ (l_{1i}, \infty) \times \{r_{2i}\} & \text{for} \quad i = 4 \\ \{l_{1i}\} \times (l_{2i}, \infty) & \text{for} \quad i = 5 \end{cases}$$

On the other hand, Figure 5.2(b) contains 4 types of observed individuals, corresponding to the



(a) T_1 or/and T_2 observed exactly.

(b) T_1 nor T_2 observed exactly.

Figure 5.2: Examples of bivariate interval-censored data

cases where nor T_{1i} neither T_{2i} are exactly observed. The regions B_i for each case are:

$$B_i = (\ell_{1i}, r_{1i}] \times (\ell_{2i}, r_{2i}] = \begin{cases} (l_{1i}, r_{1i}] \times (l_{2i}, r_{2i}] & \text{for} \quad i = 1 \\ (l_{1i}, r_{1i}] \times (l_{2i}, \infty) & \text{for} \quad i = 2 \\ (l_{1i}, \infty) \times (l_{2i}, r_{2i}] & \text{for} \quad i = 3 \\ (l_{1i}, \infty) \times (l_{2i}, \infty) & \text{for} \quad i = 4 \end{cases}$$

We assume noninformative censoring of (L_1, R_1) on T_1 and (L_2, R_2) on T_2, but (L_1, R_1) and (L_2, R_2) could be dependent. Under this assumption, the likelihood function of the observed data is proportional to

$$L(F, \boldsymbol{B}) = \prod_{i=1}^{n} F(B_i) = \prod_{i=1}^{n} \Big[F(r_{1i}, r_{2i}) - F(r_{1i}, \ell_{2i}) - F(\ell_{1i}, r_{2i}) + F(\ell_{1i}, \ell_{2i}) \Big], \tag{5.8}$$

where $\boldsymbol{B} = (B_1, \ldots, B_n)$.

5.2.2 Nonparametric estimation of $F(s,t)$

The approach to obtain the NPMLE of the joint distribution $F(s,t)$ generalizes the NPLME for the univariate case (Betensky and Finkelstein, 1999). By studying the likelihood function (5.8), it can be shown that the NPMLE has to be discrete, and that it concentrates its mass on the observed rectangles or on intersections of such rectangles (mimicking the role of Turnbull's intervals defined in (5.3)). Let

$$\mathcal{I}_b = \big\{ \bar{B}_j = (q_{1j}, p_{1j}] \times (q_{2j}, p_{2j}], j = 1, \ldots, m \big\}$$

be the set of disjoint rectangles defining the possible support of the NPMLE of F.

The likelihood function is independent of the behavior of F within these regions, that is, its probability mass can be distributed arbitrarily within the rectangles. Define then the weights within

each rectangle by $w_j = F(\bar{B}_j) = P\big((T_1, T_2) \in (q_{1j}, p_{1j}] \times (q_{2j}, p_{2j}]\big)$ for $j = 1, \ldots, m$. Given $\alpha_{ij} = I\big(\bar{B}_j \subseteq (L_{1i}, R_{1i}] \times (L_{2i}, R_{2i}]\big)$, we can rewrite the likelihood function by

$$L(w_1, \ldots, w_m) = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} w_j.$$

We obtain the NPMLE by maximizing this function over the unknown quantities w_j, subject to $w_j \geq 0$ and $\sum_{j=1}^{m} w_j = 1$. This maximization can be achieved for instance by using the self-consistent algorithm for the univariate case (5.1.2). The greatest difficulty for bivariate interval-censored data is posed by determining the regions $\mathcal{I}_b$. In order to characterize $\bar{B}_j$ we encounter a computational challenge since the number of iterations to decide whether or not a rectangle is a subset of another rectangle can be very large. Betensky and Finkelstein (1999) propose an iterative algorithm which searches directly for all rectangles in $\mathcal{I}_b$, but it is very time-consuming. Other proposed algorithms by Bogaerts and Lesaffre (2004) and Maathuis (2005), among others, try to reduce the dimension of the problem in order to reduce computation time .

5.2.3 Estimation of the correlation structure

5.2.3.1 Direct estimation of association measures

The dependency between two completely observed times of interest can be described through measures of association such as Spearman's rank correlation ρ_S, or Kendall's coefficient of concordance, $\tau_{\mathcal{K}}$, as defined in Section 4.1.2.

When data is censored, the computation of these measures of association is not straightforward. In the case of right-censored data, pairs of observations may not be comparable (comparability issues were discussed in Section 4.2), and ranks between observations cannot be computed. Oakes (1986) extends $\tau_{\mathcal{K}}$ by assigning a score equal to 0 to those non-comparable pairs. Following this idea, the proposal of Betensky and Finkelstein (1999) extends $\tau_{\mathcal{K}}$ in the case of bivariate interval-censored data, by analysing the comparability condition between pairs of observed rectangles in the plane.

In general, it is not possible to compare pairs of bivariate interval-censored data because the rectangles of observation may overlap, or even some non-overlapping intervals cannot be ordered. More details of this bivariate comparability problem will be discussed in the framework of semi-competing risks in chapter 6, where the same phenomenon arises. The idea of Betensky and Finkelstein (1999) is to use multiple imputation to fill in interval as much as possible: by first estimating the joint survival function $S(s,t)$ parametrically or nonparametrically (following any of the procedures described above), impute the most refined failure time following this joint distribution to the whole interval, and then compute $\tau_{\mathcal{K}}$ and its variance. This procedure is repeated M times, and the final estimate is obtained by averaging the obtained M measures.

Bogaerts and Lesaffre (2008a) propose an alternative based on smoothing techniques to estimate the joint density function. Then, estimating $\tau_{\mathcal{K}}$ by means of the expression

$$\tau_{\mathcal{K}} = 4 \int \int S(u,v) f(u,v) du dv - 1,$$

is straightforward. They also provide local measures of association for bivariate interval-censored data using the same techniques.

5.2.3.2 The copula approach

Another approach to study the association structure between two times of interest is via copula models. These models have been described with some detail in Section 4.1.3, their advantage being that the marginal distributions of the bivariate times (T_1, T_2) and the association structure are modelled separately. This property is very convenient in the present approach.

Indeed, assume that the joint survival function $S(s,t)$ follows a general copula model

$$S(s,t) = P(T_1 > s, T_2 > t) = C_\alpha\big(S_1(s), S_2(t)\big),$$

where $C_\alpha(u,v)$ is a bivariate copula function defined in $[0,1]\times[0,1]$, continuous and non-decreasing, such that $C_\alpha(u,0) = C_\alpha(0,v) = 0$, and $C_\alpha(u,1) = u$, $C_\alpha(1,v) = v$. We consider copula functions ruled by a a single dependence parameter α, that is, the association structure can be summarized by a single parameter for all (s,t).

Under this model, Kendall's $\tau_{\mathcal{K}}$ is

$$\tau_{\mathcal{K}} = 4\int_0^1\int_0^1 C_\alpha(u,v)dC_\alpha(u,v) - 1.$$

Since for copula models the marginal distributions and the association parameter are modelled separately, to obtain $\tau_{\mathcal{K}}$ we only need to estimate α.

Sun *et al.* (2006) considered a two-stage estimation procedure to estimate the association parameter α. They based their method in the log-likelihood function obtained from (5.8), but expressed in terms of the joint survival function

$$\ell(S, \boldsymbol{B}) = \ell(S_1, S_2, \alpha, \boldsymbol{B}) = \sum_{i=1}^{n} \log\big[S(\ell_{1i}, \ell_{2i}) - S(r_{1i}, \ell_{2i}) - S(\ell_{1i}, r_{2i}) + S(r_{1i}, r_{2i})\big].$$

The first stage consists in obtaining estimates $\widehat{S}_1(t)$ and $\widehat{S}_2(t)$ of S_1 and S_2, respectively. They propose the use of the nonparametric methods described in Section 5.1.2, based on the univariate interval-censored data samples, $\{(\ell_{1i}, r_{1i}), i = 1,\ldots,n\}$ and $\{(\ell_{2i}, r_{2i}), i = 1,\ldots,n\}$, respectively.

The second stage consists in estimating α by maximizing the pseudo log-likelihood given by $\ell(\alpha, \widehat{S}_1(\cdot), \widehat{S}_2(\cdot))$, that is, $\widehat{\alpha}$ is obtained as a solution of the pseudo score equation

$$\frac{\partial \ell}{\partial \alpha}(\alpha, \widehat{S}_1(\cdot), \widehat{S}_2(\cdot)) = 0.$$

A numerical maximization method, such as Newton-Raphson, is proposed to solve this equation.

Bogaerts and Lesaffre (2008b) extend this two-stage procedure allowing for covariates, both in the marginal distributions, by means of an accelerated failure time model, and in the dependence parameter α.

5.2.4 Final comments

Specific methods for bivariate interval-censored data, as the described above, could be extended to interval-censored semi-competing risks data. However, some of these methods do not treat adequately dependent censoring, and are not straightforward applicable. For instance, the methods based on the copula approach are not valid because we cannot reproduce the two-stage process explained above: since we cannot empirically estimate $S_1(\cdot)$, we need to first estimate α, inverting the two-stage estimation process.

In line with the methods aiming at estimating directly an association measure, in the next chapter we proceed to determine the concordance status (measure of dependence) between pairs of individuals by comparing their rectangles of observation. In next chapter, we proceed by taking advantage of the need for an assumed model in the semi-competing risks framework, and show how to estimate the dependence structure when dependent censoring is present.

CHAPTER 6

Interval-Censored Semi-Competing Risks Data

In Chapter 4 we presented the semi-competing risks problem for right-censored data. Now we consider the situation where the time to the intermediate event is interval-censored. This situation is quite common in longitudinal medical studies: in the Spanish Bladder Cancer Study, for instance, interval censoring arises because recurrence or progression of the tumour are detected between consecutive visits of the patient. Until Chapter 4, interval censoring was ignored by imputing the midpoint of the observed interval to T_1.

Existing methods accounting for bivariate interval-censored data, reviewed in Chapter 5, are not appropriate in this setting, because they do not take into account the dependent censoring induced by the terminating event on the intermediate event. In this chapter, we propose a new methodology to deal with interval-censored semi-competing risks data.

In the following Section 6.1 we introduce the notation and the assumed Clayton copula model for the joint law of the times of interest. Next, we make a brief outline of the estimating strategy in Section 6.2, which is based in two points: (i) the estimation of the dependence parameter of the copula model, and (ii) the estimation of the marginal law of the intermediate event. In Sections 6.3 and 6.4, we deal with estimation of the dependence structure, and in Section 6.5, with the estimation of the survival functions involved. The chapter ends up with the iterative algorithm which performs the proposed methodology in Section 6.6.

6.1 Notation and model

6.1.1 Interval-censored semi-competing risks data

Consider the semi-competing risks framework for (T_1, T_2) introduced in Chapter 4, where T_1 represents the time until an intermediate event and T_2 represents the time to a terminating event.

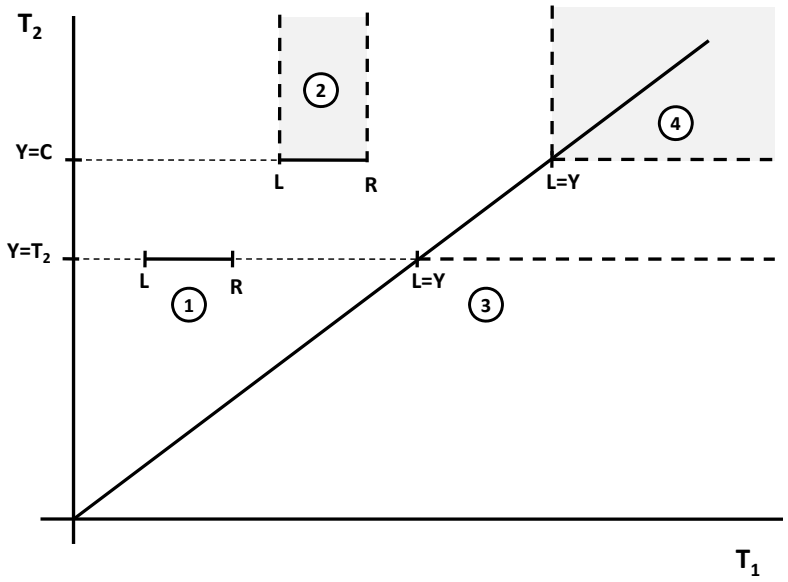


Figure 6.1: Different situations of observed interval-censored semi-competing risks data.

This is a particular case of a bivariate survival problem where the occurrence of the terminating event, if first, prevents the observation of the intermediate event. The support of (T_1, T_2), denoted by $\text{supp}(T_1, T_2)$, is included in the upper wedge of the plane defined by

$$\mathcal{D}_1 = \{\, (s,t) \,|\, 0 \leq s \leq t \leq \infty \,\}.$$

In our setup we encounter three different censoring mechanisms acting concurrently:

(a) **Right-censoring:** The first mechanism occurs because data is subject to a non-informative right-censoring, C, due, for instance, to the end of the study. Hence, T_1 and T_2 are right-censored by C, which is assumed independent of both T_1 and T_2.

(b) **Dependent censoring:** Because T_1 and T_2 are possibly related, T_2 induces a dependent right-censoring on T_1.

(c) **Interval censoring:** T_1 is interval-censored, that is, there exist two random variables $L \geq 0$ and $R > 0$, with $L < R$ almost surely, such that $P(T_1 \in (L, R]) = 1$. Note here that exact observations for T_1 are not allowed. We use the convention that $R = \infty$ represents a right-censored observation when $T_1 > T_2$.

We assume that (L, R) censors non-informatively T_1 in the sense described by Oller *et al.* (2004) (see section 5.1.1). These assumption ensures that $P(T_1 \in (a, b], L = a, R = b)$ is proportional to $P(T_1 \in (a, b])$. We also assume that C and (L, R) are independent.

The observable data consists of the vector $V = (L, R, Y, \delta_1, \delta_2)$ where $Y = \min(T_2, C)$, $\delta_1 = I(R < +\infty)$ and $\delta_2 = I(T_2 \leq C)$. We adopt the usual convention that realizations of random variables are denoted by lower-case letters. The different censoring patterns give raise to four distinct types of observed individuals, as shown in Figure 6.1:

(1) **T_1 is interval-censored and T_2 is exactly observed:**
It corresponds to $\delta_1 = \delta_2 = 1$, that is, $T_1 \in (L, R]$, $L < R \leq Y$ and $Y = T_2$. This situation describes the region of the plane: $(L, R] \times \{Y\}$.

(2) **T_1 is interval-censored and T_2 is right-censored:**
It corresponds to $\delta_1 = 1$ and $\delta_2 = 0$, that is, $T_1 \in (L, R]$, $L < R \leq Y$ and $Y = C < T_2$. This situation describes the region of the plane: $(L, R] \times (Y, \infty)$.

(3) **T_1 is dependently censored by T_2 and T_2 is exactly observed:**
It corresponds to $\delta_1 = 0$ and $\delta_2 = 1$, that is, $T_1 \in (L, \infty)$, $L = Y$ and $Y = T_2$. This situation describes the region of the plane: $(L, \infty) \times \{Y\}$.

(4) **T_1 and T_2 are right-censored:**
Corresponds to $\delta_1 = \delta_2 = 0$, that is, $T_1 \in (L, \infty)$, $L = Y$ and $Y = C < T_2$. This situation describes the region of the plane: $(L, \infty) \times (Y, \infty)$.

For simplicity, we assume that at the end of follow-up (Y) of each individual, the status of his intermediate event (occurred or not) is known. However, the methods proposed in this chapter are also valid when the status of the intermediate event at the end of the follow-up is uncertain.

6.1.2 Model for (T_1, T_2)

The joint law for (T_1, T_2) is specified through a survival copula model. We have chosen a Clayton's archimedean copula in the upper wedge $\mathcal{D}_1$ in order to extend to the interval-censored case the right-censored semi-competing risks method in Section 4.2. That is, given the marginal survival functions $S_1(\cdot)$ and $S_2(\cdot)$ for T_1 and T_2, respectively, the joint survival is given by

$$S(s,t) = P(T_1 > s, T_2 > t) = (S_1(s)^{1-\alpha} + S_2(t)^{1-\alpha} - 1)^{\frac{1}{1-\alpha}}, \tag{6.1}$$

for $0 \leq s \leq t \leq \infty$, where α is the association parameter.

The main purpose of this chapter is to estimate the joint law of (T_1, T_2), that is, the association parameter α, and the marginal laws $S_1(\cdot)$ and $S_2(\cdot)$, based on n independent and identically distributed realizations of the observable data, $\{(a_i, b_i, y_i, \delta_{1i}, \delta_{2i}), i = 1, \ldots, n\}$, under the censoring mechanisms described previously in 6.1.1. We outline the estimation strategy in Section 6.2.

6.2 Outline of the estimating strategy

We extend the two-steps estimation procedure proposed by Fine *et al.* (2001) to the interval-censored framework. Fine's proposal for **right-censored semi-competing risks data** is based on the concordance indicator,
$\Delta_{ij} = I((T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0)$, an unbiased estimate of $\alpha/\alpha + 1$:

$$E[\Delta_{ij}] = \frac{\alpha}{1+\alpha}, \tag{6.2}$$

and consists of the following two steps:

(a) *Estimation of the association parameter α:* An estimate $\widehat{\alpha}$ is obtained as the unique root of the equation $U^R(\alpha) = 0$, where

$$U^R(\alpha) = \binom{n}{2}^{-1} \sum_{i<j} O_{ij}^R \left\{ \Delta_{ij} - \frac{\alpha}{\alpha+1} \right\}, \tag{6.3}$$

is a zero-mean random variable presented in Section 4.2.3. For simplicity, we consider the unweighted estimating equation ($W^R = 1$ in equation (4.13)).

(b) *Estimation of the marginal survival function of T_1:* A consistent estimate for $S_1(\cdot)$ is obtained by plugging-in consistent estimates for $S_2(s)$, $S_T(s)$ and α in

$$S_1(s; S_T(s), S_2(s), \alpha) = \{S_T(s)^{1-\alpha} - S_2(s)^{1-\alpha} + 1\}^{\frac{1}{1-\alpha}}, \tag{6.4}$$

where $S_T(s) = S(s,s) = P(\min(T_1, T_2) > s)$.

The adaptation of the above procedure to interval censoring arises two main difficulties: the non observability of the concordance indicator Δ_{ij}, and the definition of comparable pairs. Our approach also consists of the two previous steps, (a) estimation of α, and and (b) estimation of $S_1(\cdot)$, but within an iterative process:

(a') For the association parameter α we will

  (i) propose a new measure of concordance for interval-censored semi-competing risks data (Section 6.3.1),

  (ii) define a new condition of comparable pairs, O_{ij} (Section 6.3.2), and finally

  (iii) propose two alternative estimating equations $U_1(\alpha)$ and $U_2(\alpha)$ for α, based respectively on

    - correction of the bias induced by the comparable sample (Section 6.4.1), and
    - inverse weighting by the probability of being comparable (Section 6.4.2).

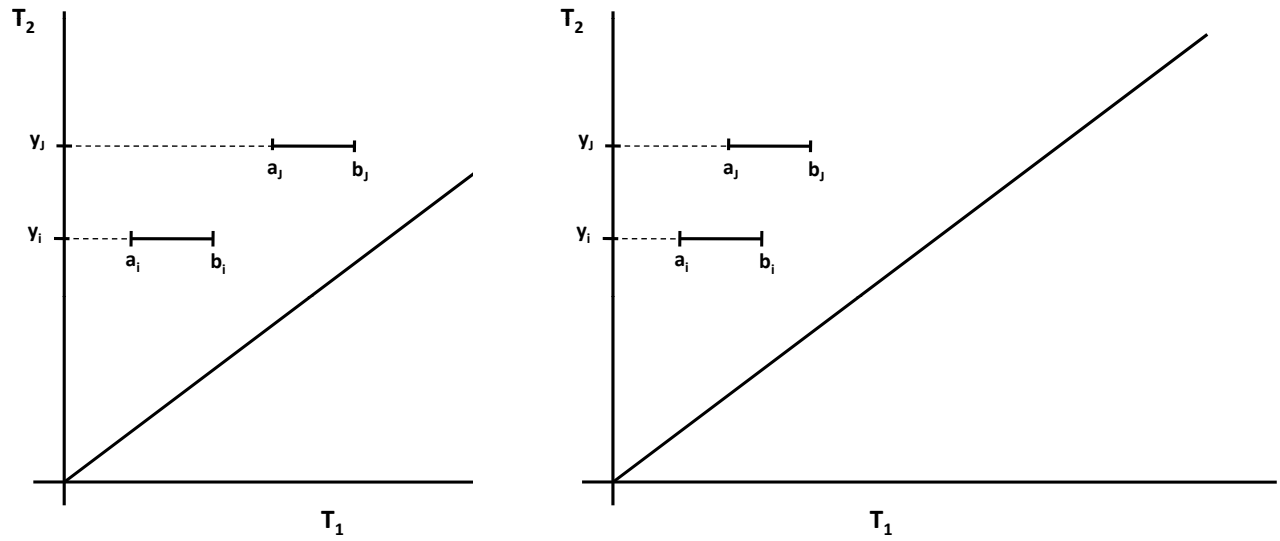
(b') For the estimation of the marginal $S_1(\cdot)$ we will consider the same plug-in estimator as for the right-censoring case, but taking into account the interval-censored nature of data (Section 6.5).

Section 6.6, contains the iterative algorithm to jointly estimate the association parameter α and the marginal S_1. The asymptotic properties of the proposed estimates are developed and assessed in Chapters 7 and 9.

6.3 The expected concordance and the comparable sample

6.3.1 Definition and estimation of the expected concordance

For a pair (i, j) of individuals, when T_{1i} and T_{1j} are interval-censored, the concordance indicator cannot in general be determined. Figure 6.2 shows two examples of observed pairs. Let (T_{1i}, T_{2i})



(a) Δ_{ij} can be determined.

(b) Δ_{ij} cannot be determined.

Figure 6.2: Examples of observed pairs of individuals

and (T_{1j}, T_{2j}) be their corresponding bivariate times, and assume the observed data consists of $T_{1i} \in (a_i, b_i]$, $T_{2i} = y_i$, $T_{1j} \in (a_j, b_j]$ and $T_{2j} = y_j$. In Figure 6.2(a), the pair (i, j) is concordant, because y_i is smaller than y_j, and, though we do not know exactly T_{1i} nor T_{1j}, the disposition of the intervals, without overlapping, ensures that $T_{1i} < T_{1j}$. On the contrary, for the pair in Figure 6.2(b), Δ_{ij} is unknown because, since the intervals overlap, the exact position of T_{1i} and T_{1j} within their corresponding intervals is unknown. In a given data set, many of these pairs with overlapping intervals will be found. If we discard these pairs, much information will be lost. For this reason we propose a new concordance measure, Z_{ij}, which unbiasedly estimates $\alpha/\alpha + 1$ as the concordance indicator does.

Definition 6.1. *Given two individuals i and j, we define the* ***expected concordance*** *Z_{ij} between the pair (i, j) as the expectation of the concordance indicator given their observed data. That is*

$$Z_{ij} = E\big[\Delta_{ij}|\mathcal{H}_{ij}\big] = P\big(\Delta_{ij} = 1|\mathcal{H}_{ij}\big) = P\big((T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0|\mathcal{H}_{ij}\big), \tag{6.5}$$

where $\mathcal{H}_{ij} = \{(a_i, b_i, y_i, \delta_{1i}, \delta_{2i}), (a_j, b_j, y_j, \delta_{1j}, \delta_{2j})\}$, the observed data for the pair (i, j), consists of two realizations of vector $V = (L, R, Y, \delta_1, \delta_2)$. When the intervals including T_{1i} and T_{1j} do not overlap, $Z_{ij} = \Delta_{ij}$, and if they overlap, $0 < Z_{ij} < 1$.

The random variable Z_{ij} depends on α, $S_1(\cdot)$ and $S_2(\cdot)$, as it is shown in Section 6.3.1.

Proposition 6.1. *Z_{ij} is an unbiased estimator of $\alpha/\alpha + 1$.*

Proof. Z_{ij} is a random variable with expectation

$$E\big[Z_{ij}\big] = E\big[E[\Delta_{ij}|\mathcal{H}_{ij}]\big] = E\big[\Delta_{ij}\big] = \frac{\alpha}{\alpha + 1}.$$

□

Proposition 6.2. *Given a bivariate model for (T_1, T_2) in the upper wedge $\mathcal{D}_1$, the expected concordance Z_{ij} for a comparable[1] pair (i, j) is*

$$Z_{ij} = \frac{1}{P(\mathcal{H}_{ij})} \left(\delta_{2i}\delta_{2j}P_1(i,j) + \delta_{2i}(1-\delta_{2j})P_2(i,j) + (1-\delta_{2i})\delta_{2j}P_2(j,i)\right) \tag{6.6}$$

where $\mathcal{H}_{ij} = \{(a_i, b_i, y_i, \delta_{1i}, \delta_{2i}), (a_j, b_j, y_j, \delta_{1j}, \delta_{2j})\}$ is the observed data for the pair (i, j) and

$$P_1(i,j) = K^2 \int_{a_i}^{b_i} dx \int_{a_j}^{b_j} I((x-u)(y_i - y_j) > 0) f(x, y_i) f(u, y_j) du, \tag{6.7}$$

$$P_2(i,j) = K^2 \int_{y_j}^{\infty} dv \int_{a_i}^{b_i} dx \int_{a_j}^{b_j} I((x-u)(y_i - v) > 0) f(x, y_i) f(u, v) du, \tag{6.8}$$

and

$$P(\mathcal{H}_{ij}) = \prod_{\ell=i,j} \left[\delta_{2\ell} K \int_{a_\ell}^{b_\ell} f(x, y_\ell) dx + (1-\delta_{2\ell}) K \int_{a_\ell}^{b_\ell} \int_{y_\ell}^{\infty} f(x, y) dx dy\right],$$

where K is a constant of proportionality derived of the noninformativity condition.

Proof. Given $\mathcal{H}_{ij} = \{(a_i, b_i, y_i, \delta_{1i}, \delta_{2i}), (a_j, b_j, y_j, \delta_{1j}, \delta_{2j})\}$, the observed data for the pair (i, j), the expected concordance Z_{ij} can be explicitly computed when a model is assumed for (T_{1i}, T_{2i}) in the upper wedge $\mathcal{D}_1$, and when the pair is comparable. In next Section 6.3.2 a thoroughly description of comparable pairs is given, which requires, at least, that one of the two individuals is not right-censored neither for T_1 ($\delta_{1i} + \delta_{1j} \geq 1$) nor for T_2 ($\delta_{2i} + \delta_{2j} \geq 1$). We develop the expectation in (6.5) by

$$\begin{aligned} Z_{ij} &= E\left[\Delta_{ij} | \mathcal{H}_{ij}\right] = \frac{P\left(\Delta_{ij} = 1, \mathcal{H}_{ij}\right)}{P\left(\mathcal{H}_{ij}\right)} \\ &= \frac{1}{P(\mathcal{H}_{ij})} \left(\delta_{2i}\delta_{2j}P_1(i,j) + \delta_{2i}(1-\delta_{2j})P_2(i,j) + (1-\delta_{2i})\delta_{2j}P_2(j,i)\right) \end{aligned}$$

where

$$\begin{aligned} P_1(i,j) &= P(\Delta_{ij} = 1, \mathcal{H}_{ij}, \delta_{2i} = 1, \delta_{2j} = 1) \\ P_2(i,j) &= P(\Delta_{ij} = 1, \mathcal{H}_{ij}, \delta_{2i} = 1, \delta_{2j} = 0). \end{aligned}$$

and $P(\mathcal{H}_{ij})$ is the probability of the observed data. This probability, given that i and j are independent, can be factorized into

$$P(\mathcal{H}_{ij}) = P\Big((L_i, R_i, Y_i, \delta_{1i}, \delta_{2i}) = (a_i, b_i, y_i, \delta_{1i}, \delta_{2i})\Big) P\Big((L_j, R_j, Y_j, \delta_{1j}, \delta_{2j}) = (a_j, b_j, y_j, \delta_{1j}, \delta_{2j})\Big).$$

[1] In next Section 6.3.2 a description of comparable pairs is given.

Now, the subject-specific probability can be expressed by

$$
\begin{aligned}
&P\Big((L_\ell, R_\ell, Y_\ell, \delta_{1\ell}, \delta_{2\ell}) = (a_\ell, b_\ell, y_\ell, \delta_{1\ell}, \delta_{2\ell})\Big) \\
&= \delta_{2\ell} P\big(T_{1\ell} \in (a_\ell, b_\ell], T_{2\ell} = y_\ell, L_\ell = a_\ell, R_\ell = b_\ell\big) \\
&\qquad + (1-\delta_{2\ell}) P\big(T_{1\ell} \in (a_\ell, b_\ell], T_{2\ell} > y_\ell, L_\ell = a_\ell, R_\ell = b_\ell\big) \\
&\underset{(i)}{=} \delta_{2\ell} K P\big(T_{1\ell} \in (a_\ell, b_\ell], T_{2\ell} = y_\ell\big) + (1-\delta_{2\ell}) K P\big(T_{1\ell} \in (a_\ell, b_\ell], T_{2\ell} > y_\ell\big) \\
&= \delta_{2\ell} K \int_{a_\ell}^{b_\ell} f(x, y_\ell) dx + (1-\delta_{2\ell}) K \int_{a_\ell}^{b_\ell} \int_{y_\ell}^{\infty} f(x,y) dx dy,
\end{aligned}
$$

$\ell = i, j$. Equality (i) is justified by the noninformativity condition of (L, R) over T_1, implying that $P(T_1 \in (a, b], L = a, R = b) = KP(T_1 \in (a, b])$ for some constant $K > 0$. Function f is the joint density function under the copula model (6.1), which is well defined in the upper wedge and has the following expression:

$$
f(x,y) = \frac{\partial^2 S(x,y)}{\partial x \partial y} = \alpha f_1(s) f_2(t) \left(S_1(s) S_2(t)\right)^{\alpha} S(s,t)^{2\alpha - 1}. \tag{6.9}
$$

We can develop the probability $P_1(i,j)$ and obtain

$$
\begin{aligned}
&P_1(i,j) = \\
&P\left(\Delta_{ij} = 1, T_{1i} \in (a_i, b_i], T_{1j} \in (a_j, b_j], T_{2i} = y_i, T_{2j} = y_j, L_i = a_i, R_i = b_i, L_j = a_j, R_j = b_j\right) \\
&\underset{(i)}{=} K^2 P\left(\Delta_{ij} = 1, T_{1i} \in (a_i, b_i], T_{1j} \in (a_j, b_j], T_{2i} = y_i, T_{2j} = y_j\right) \\
&= K^2 \int_{a_i}^{b_i} dx \int_{a_j}^{b_j} I((x-u)(y_i - y_j) > 0) f(x, y_i) f(u, y_j) du.
\end{aligned}
$$

Analogously, the expression for $P_2(i,j)$ is given by

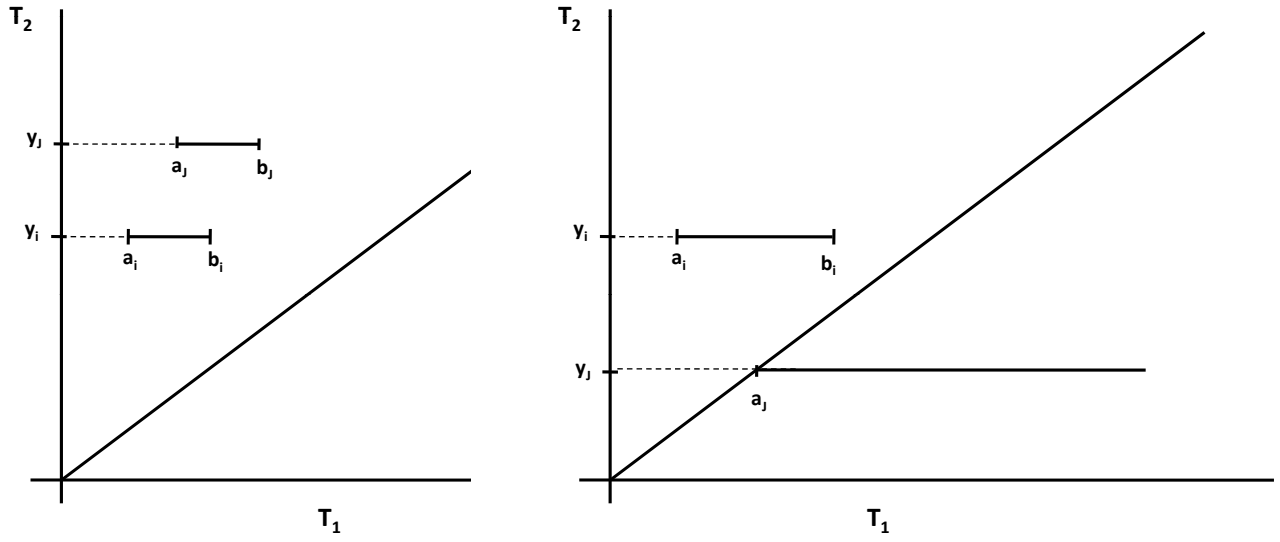
$$
\begin{aligned}
&P_2(i,j) = \\
&P(\Delta_{ij} = 1, T_{1i} \in (a_i, b_i], T_{1j} \in (a_j, b_j], T_{2i} = y_i, T_{2j} \in (y_j, \infty), L_i = a_i, R_i = b_i, L_j = a_j, R_j = b_j) \\
&\underset{(i)}{=} K^2 P(\Delta_{ij} = 1, T_{1i} \in (a_i, b_i], T_{1j} \in (a_j, b_j], T_{2i} = y_i, T_{2j} \in (y_j, \infty)) \\
&= K^2 \int_{y_j}^{\infty} dv \int_{a_i}^{b_i} dx \int_{a_j}^{b_j} I((x-u)(y_i - v) > 0) f(x, y_i) f(u, v) du.
\end{aligned}
$$

□

The constant K^2 resulting from the noninformativity condition does not appear in the final expression (6.6) because it appears in both the numerator and the denominator, and so it cancels out. We postponed to Appendix B.4 the expanded expression for the integrals (6.7) and (6.8).

6.3.2 The comparable sample

A pair (i,j) is comparable if Z_{ij} can be computed based on observed data $\mathcal{H}_{ij}$ and on the assumed underlying model. In other words, the comparable sample determines the pairs contributing in



(a) Z_{ij} can be determined.

(b) Z_{ij} cannot be determined.

Figure 6.3: Determination of Z_{ij}

the estimation of α. When T_1 is interval-censored, these pairs in general do not coincide with the comparable sample for right-censored data, $\mathcal{C}^R = \{i < j | O_{ij}^R = 1\}$. In this section, we provide the necessary conditions for a pair (i, j) to be comparable given their observed data (in Section 6.3.2.1), summarizing 24 situations which are explored in detail in Appendix B.5. We illustrate some of the most important cases, the comparable pairs, in Section 6.3.2.2.

6.3.2.1 Conditions of comparability for the interval censoring setting

Figure 6.3 shows an example of a comparable pair and one of a non-comparable pair. For the pair plotted in Figure 6.2(a), we can be sure that both events occur in the upper wedge $\mathcal{D}_1$, where the assumed model is valid, so we can compute Z_{ij} following expression (6.6). On the other hand, for the pair in Figure 6.2(b), Z_{ij} cannot be estimated. Indeed, suppose for example that both $T_{2i} = y_i$ and $T_{2j} = y_j$ are exactly observed, and $y_i < y_j$. The intermediate event for the i^{th} individual occurs in $\mathcal{D}_1$ within the interval $T_{1i} \in (a_i, b_i]$, while for the j^{th} individual T_{1j} does not belong to $\mathcal{D}_1$ (so $T_{1j} \in (a_j, \infty)$). To compute Z_{ij} we need to integrate out in the overlapping region $(a_j, b_i) \times \{y_j\}$, which is not included in $\mathcal{D}_1$. Therefore, the copula model is not valid and Z_{ij} cannot be computed.

Before giving a general condition of comparability we need to introduce some notation. Given a pair of individuals and their observed data, $\mathcal{H}_{ij} = \{(a_i, b_i, y_i, \delta_{1i}, \delta_{2i}), (a_j, b_j, y_j, \delta_{1j}, \delta_{2j})\}$, consider the following scenarios:

- In the case that only one T_1 time of the pair is interval-censored, $\delta_{1i} + \delta_{1j} = 1$, define the index o, which stands for observed, by

$$o = \begin{cases} i & \text{if } \delta_{1i} = 1 \\ j & \text{if } \delta_{1j} = 1 \end{cases}$$

as the index of the interval-censored individual. Define the index c, which stands for censored, by

$$c = \{i, j\} \backslash o,$$

the index of the other individual, the right-censored one.

- Similarly, in the case that only one T_2 time is exactly observed, $\delta_{2i} + \delta_{2j} = 1$, define the index

$$o' = \begin{cases} i & \text{if } \delta_{2i} = 1 \\ j & \text{if } \delta_{2j} = 1 \end{cases}$$

to indicate the index of the exactly observed pair, and

$$c' = \{i, j\} \backslash o',$$

the index of the right-censored T_2 time.

Proposition 6.3. *A pair of individuals (i, j) is comparable, given their observed data $\mathcal{H}_{ij}$, if the three following conditions are satisfied:*

1. *At least one of T_{1i} or T_{1j} must be interval-censored ($\delta_{1i} + \delta_{1j} \geq 1$), and at least one of T_{2i} or T_{2j} must be exactly observed ($\delta_{2i} + \delta_{2j} \geq 1$).*

2. *When only one T_1 time is interval-censored in $D1$ ($\delta_{1i} + \delta_{1j} = 1$), the intervals must not overlap, that is, $b_o < a_c$, which can also be expressed as*

$$\delta_{1i} b_i + \delta_{1j} b_j < (1 - \delta_{1i}) a_i + (1 - \delta_{1j}) a_j.$$

3. *When only one T_2 time is exactly observed ($\delta_{2i} + \delta_{2j} = 1$), this time must be smaller than the right-censored value, that is $y_{o'} < y_{c'}$, which can also be expressed as*

$$\delta_{2i} y_i + \delta_{2j} y_j < (1 - \delta_{2i}) y_i + (1 - \delta_{2j}) y_j.$$

Proof. The previous three conditions results from the study of the 24 possible dispositions of the pair (i, j) on the plane $T_1 \times T_2$ according to the observed values for the pair. An exhaustive description of all 24 situations is described in Appendix B.5. It results that only six of the 24 initial dispositions correspond to comparable pairs, the pairs satisfying conditions 1 to 3. □

Given a pair (i, j), if the previous conditions hold, the comparable indicator O_{ij} equals 1, in any other case is 0. We define the set of comparable pairs by

$$\mathcal{C}^{IC} = \{i < j | O_{ij} = 1\}.$$

Let $\widetilde{T}_{1ij} = \min(T_{1i}, T_{1j})$ and $\widetilde{T}_{2ij} = \min(T_{2i}, T_{2j})$. Consider also $\widetilde{C}_{ij} = \min(C_i, C_j)$, $\widetilde{L}_{ij} = \min(L_i, L_j)$ and $\widetilde{R}_{ij} = \min(R_i, R_j)$, the minimums among the right censoring variables and the extremes of the intervals, respectively. Given that $L_i < T_{1i} < R_i$ and $L_j < T_{1j} \leq R_j$, it is satisfied

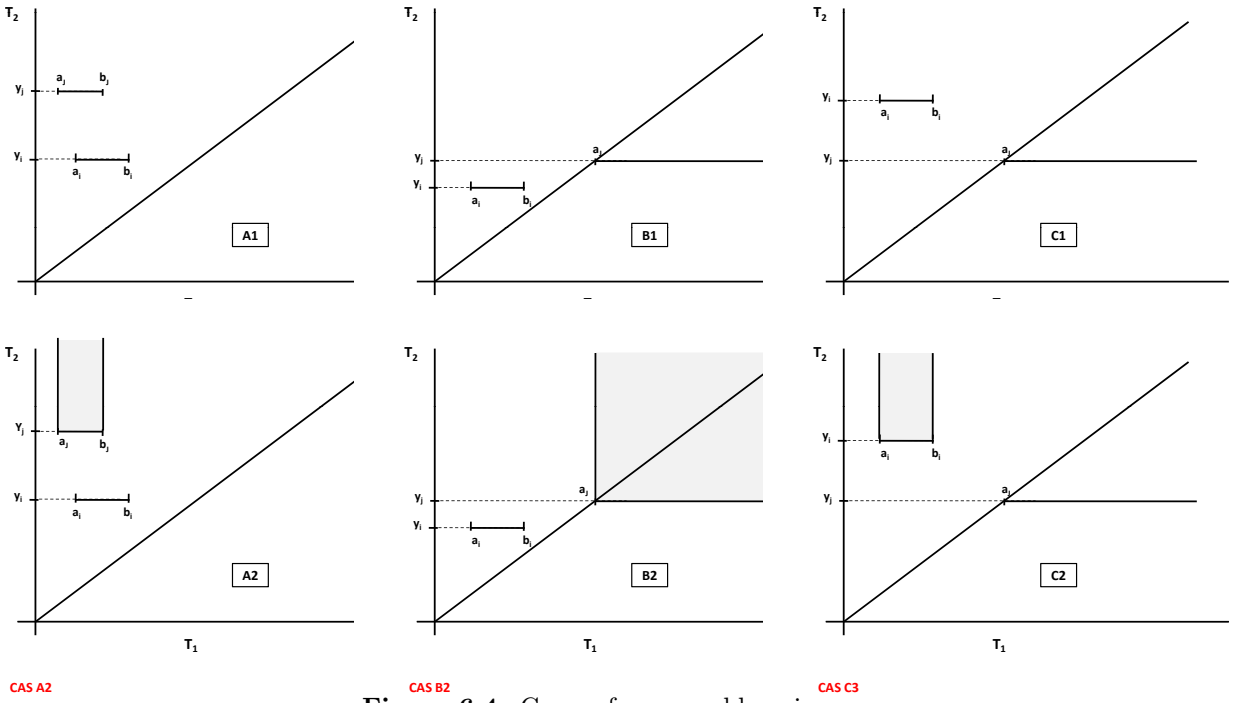


Figure 6.4: Cases of comparable pairs.

that $\widetilde{a}_{ij} < \widetilde{T}_{2ij} \leq \widetilde{b}_{ij}$. An equivalent formulation for the comparability condition is given by the following proposition.

Proposition 6.4. *A pair of individuals (i, j) is comparable, given their observed data $\mathcal{H}_{ij}$, if the following conditions hold:*

(i) $\widetilde{T}_{1ij} < \widetilde{C}_{ij}$,

(ii) $\widetilde{T}_{2ij} < \widetilde{C}_{ij}$, *and*

(iii) $\widetilde{R}_{ij} < \widetilde{T}_{2ij}$.

Proof. The proof of this proposition can be found in Appendix B.6. □

The two first conditions mimic the concept of *orderable* pairs proposed by Oakes (1986), determining those pairs in the whole plain for whom the concordance indicator could be computed regardless of the censoring mechanism. Fine *et al.* (2001) pointed out that in the context of semi-competing risks data, where only the upper wedge is observable, it is trivially satisfied that $\widetilde{T}_{1ij} < \widetilde{T}_{2ij}$. The third condition is necessary when the intermediate event is interval-censored.

6.3.2.2 The comparable pairs

In this section we characterize the 6 situations, among the 24 possible, which correspond to comparable pairs. They are plotted in Figure 6.4.

The first condition in Proposition 6.3 states that $\delta_{2i} + \delta_{2j}$ must be either 1 or 2. The first row in Figure 6.4, cases A1, B1 and C1, correspond to $\delta_{2i} + \delta_{2j} = 2$, that is a situation where both T_2 times are exactly observed: $T_{2i} = y_i$ and $T_{2j} = y_j$.

- **Case A1** $\delta_{1i} + \delta_{1j} = 2$: in this case, $T_{1i} \in (a_i, b_i]$ and $T_{1j} \in (a_j, b_j]$, $b_i \leq y_i$ and $b_j \leq y_j$.
- **Case B1** $\delta_{1i} + \delta_{1j} = 1$ and $y_o < y_c$: in this case, $T_{1o} \in (a_o, b_o]$, $T_{1c} \in (a_c, \infty)$, $b_o \leq y_o < y_c$ and $a_c = y_c$ (in the plot, $o = i$ and $c = j$).
- **Case C1** $\delta_{1i} + \delta_{1j} = 1$ and $y_o > y_c$ and $b_o < a_c$: in this case, $T_{1o} \in (a_o, b_o]$, $T_{1c} \in (a_c, \infty)$, $b_o \leq y_c < y_o$ and $a_c = y_c$ (in the plot, $o = i$ and $c = j$).

Cases B1 and C1 correspond to a situation where only one of the T_1 times is interval-censored, but the intervals do not overlap (condition 2 of Proposition 6.3).

The second row in Figure 6.4, cases A2, B2 and C2, correspond to $\delta_{2i} + \delta_{2j} = 1$, that is, only one T_2 time is exactly observed, and the other is right-censored. Using the notation introduced above, we have $T_{2o'} = y_{o'}$ and $T_{2c'} > y_{c'}$, and, according to the third condition in Proposition 6.3, a comparable case must satisfy $y_{o'} < y_{c'}$. Under this assumption, the comparable cases are:

- **Case A2** $\delta_{1o'} + \delta_{1c'} = 2$: in this case, $T_{1o'} \in (a_{o'}, b_{o'}]$ and $T_{1c'} \in (a_{c'}, b_{c'}]$, $b_{o'} \leq y_{o'}$ and $b_{c'} \leq y_{c'}$ and $y_{o'} < y_{c'} < T_{2c'}$ (in the plot, $o' = i$, $c' = j$).
- **Case B2** $\delta_{1o'} = 1$, $\delta_{1c'} = 0$: in this case, $T_{1o'} \in (a_{o'}, b_{o'}]$, $T_{1c'} \in (a_{c'}, \infty)$, $b_{o'} \leq y_{o'} < y_{c'} < T_{2c'}$ and $a_{c'} = y_{c'}$ (in the plot, $o' = i$ and $c' = j$).
- **Case C2** $\delta_{1o'} = 0$, $\delta_{1c'} = 1$ and $b_{c'} < y_{o'}$: in this case, $T_{1o'} \in (a_{o'}, \infty)$, $T_{1c'} \in (a_{c'}, b_{c'}]$, $b_{c'} \leq y_{o'} < y_{c'} < T_{2c'}$ and $a_{o'} = y_{o'}$ (in the plot $o' = j$, $c' = i$).

6.4 Estimating equations for α

Assume functions $S_1(\cdot)$ and $S_2(\cdot)$ are known. Since the expected concordance Z_{ij} has the same expectation as Δ_{ij}, our first attempt to extend the estimating equation for the right-censoring case,

$$U^R(\alpha) = \binom{n}{2}^{-1} \sum_{i<j} O_{ij}^R \left\{ \Delta_{ij} - \frac{\alpha}{\alpha + 1} \right\}.$$

was to substitute the concordance indicator Δ_{ij} by its expected concordance Z_{ij},

$$U_0(\alpha) = \frac{1}{\binom{n}{2}} \sum_{i<j} O_{ij} \left\{ Z_{ij}(\alpha) - \frac{\alpha}{\alpha + 1} \right\},$$

where only the comparable pairs $\mathcal{C}^{IC} = \{(i, j) \mid O_{ij} = 1\}$ contribute to the summation. However, we will prove that though $U^R(\alpha)$ is an unbiased estimating equation for α, $U_0(\alpha)$ is biased. The reason for this is that the sample of comparable pars $\mathcal{C}^{IC}$ is not a random sample of the set of observed pairs. In fact, in Section 6.4.1 we prove that the summation in $U_0(\alpha)$ has less terms than in $U^R(\alpha)$. We will show that the bias of equation $U_0(\alpha)$ is given by the lost terms and that this yields to a systematic overestimation of α. We propose a new estimating equation by explicitly correcting the bias. On the other hand, in Section 6.4.2, we propose an alternative unbiased estimating equation based on inverse weighting by the probability if being comparable.

6.4.1 Estimation of α by direct estimation of bias

Proposition 6.5. *Any pair (i, j) that is comparable in the interval censoring framework ($O_{ij} = 1$) would also be a comparable pair if T_1 was exactly observed ($O^R_{ij} = 1$). Thus, the set of comparable pairs for interval censoring, $\mathcal{C}^{IC}$, is a subset of the comparable sample for right-censored data $\mathcal{C}^R$,*

$$\mathcal{C}^{IC} \subseteq \mathcal{C}^R.$$

We refer to as excluded pairs those pairs in

$$\mathcal{C}^R \setminus \mathcal{C}^{IC} = \{(i, j) \mid O^R_{ij} = 1, O_{ij} = 0\},$$

that is, those pairs that would be comparable if T_1 was exactly observed but they are not comparable with interval censoring.

Proposition 6.6. *The excluded pairs are characterized by the following three conditions:*

(i) $\delta_{2i} + \delta_{2j} = 2$ *or* $(\delta_{2i} + \delta_{2j} = 1$ *and* $y_{o'} < y_{c'})$*, and*

(ii) $\delta_{1i} + \delta_{1j} = 1, y_o > y_c, a_c = y_c, a_o < a_c, b_o > a_c$*, and*

(iii) $T_{1o} \in (a_o, a_c]$.

Proof of Propositions 6.5 and 6.6. The excluded cases are determined by revision of all possible cases. Indeed, Figure 6.5 shows two examples of excluded pairs. In Figure 6.5(a), a pair with both T_{2i} and T_{2j} exactly observed which would be comparable if T_{1i} was exactly observed ($O^R_{ij} = 1$), but in the presence of interval censoring, Z_{ij} cannot be determined because the intervals overlap outside the region of observation $\mathcal{D}_1$. Figure 6.5(b) shows a similar case when $\delta_{2i} + \delta_{2j} = 1$. □

Proposition 6.7. *The contribution of any excluded pair $(i, j) \in \mathcal{C}^R \setminus \mathcal{C}^{IC}$ to the summation in the estimating equation $U^R(\alpha) = 0$ is $-\frac{\alpha}{\alpha+1}$.*

Proof. All excluded pairs correspond to discordant pairs: indeed, from condition (ii) of the previous theorem, $y_o > y_c$, and $T_{1c} > a_c$. From *(iii)*, $T_{1o} \in (a_o, a_c]$, so $T_{1o} \leq a_c$. Therefore, $T_{1o} < T_{1c}$ and $(T_{1o} - T_{1c})(y_o - y_c) < 0$, and the pair considered is discordant. Hence, $\Delta_{ij} = 0$, and the term in the summation of $U^R(\alpha)$ is

$$\left(\Delta_{ij} - \frac{\alpha}{\alpha + 1}\right) = -\frac{\alpha}{\alpha + 1}.$$

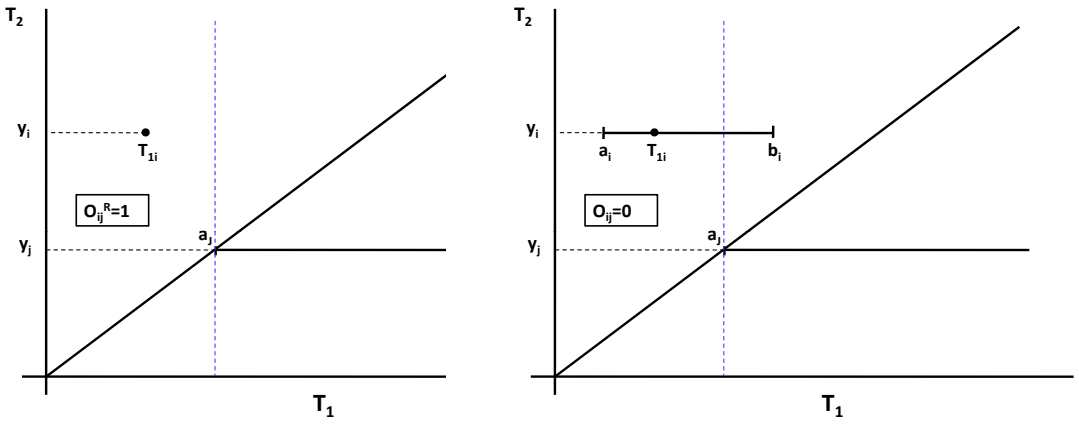
□

Denote by n_e the cardinal of the set of excluded pairs, $\mathcal{C}^R \setminus \mathcal{C}^{IC} = \{(i, j) \mid O^R_{ij} = 1, O_{ij} = 0\}$.

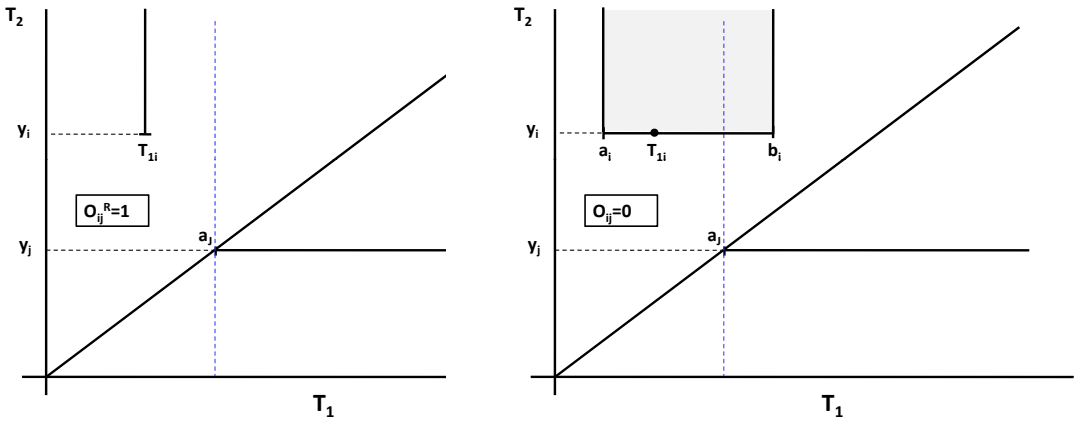
Proposition 6.8. *The bias of the estimating equation $U_0(\alpha)$ is*

$$E[U_0(\alpha)] = p_e \frac{\alpha}{\alpha + 1},$$

where $p_e = n_e / \binom{n}{2}$ is the proportion of excluded pairs among all possible pairs.



(a) Example of excluded pair when $\delta_{2i} + \delta_{2j} = 2$



(b) Example of excluded pair when $\delta_{2i} + \delta_{2j} = 1$.

Figure 6.5: Excluded pairs $(i, j) \in \mathcal{C}^R \setminus \mathcal{C}^{IC}$, satisfying $O_{ij}^R = 1, O_{ij} = 0$.

Proof. We directly compute the expectation of $U_0(\alpha)$. By Proposition 6.5 we can split up the sum in two terms:

$$\begin{aligned}
\mathrm{E}\left[U_0(\alpha)\right] &= \mathrm{E}\left[\binom{n}{2}^{-1}\left\{\sum_{i<j} O_{ij}\left(Z_{ij}(\alpha) - \frac{\alpha}{\alpha+1}\right)\right\}\right] \\
&\underset{\text{Prop. } 6.5}{=} \binom{n}{2}^{-1}\left\{\mathrm{E}\left[\sum_{i<j} O_{ij}^R\left(Z_{ij}(\alpha) - \frac{\alpha}{\alpha+1}\right)\right] - \mathrm{E}\left[\sum_{i<j} O_{ij}^R(1-O_{ij})\left(Z_{ij}(\alpha) - \frac{\alpha}{\alpha+1}\right)\right]\right\} \\
&\underset{\text{Prop. } 6.7}{=} E[U^R(\alpha)] + \frac{n_e}{\binom{n}{2}}\frac{\alpha}{\alpha+1} \underset{\text{Prop. } 4.4}{=} 0 + p_e\frac{\alpha}{\alpha+1}.
\end{aligned}$$

□

The existing bias in equation $U_0(\alpha)$ prevents from obtaining an unbiased estimate for α as the root of the estimating equation. Indeed, a root of $U_0(\alpha)$ would overestimate the real value of α. Recall from Proposition 4.1 that under Clayton's copula model α can be estimated by the number of

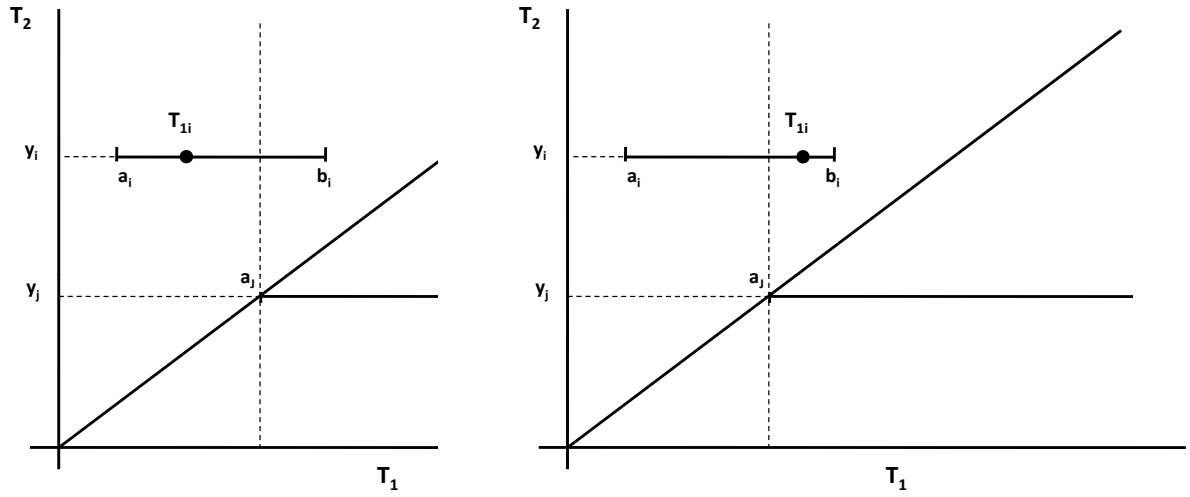


Figure 6.6: Examples of pairs from set D satisfying (i) and (ii) from Proposition 6.5.

concordant pairs divided by the number of discordant pairs. If we systematically exclude discordant pairs, we are removing terms from the denominator of this ratio, and thus increasing the estimation of α.

We propose an alternative estimating equation $U_1(\alpha)$ that explicitly corrects this bias:

$$U_1(\alpha) = \binom{n}{2}^{-1} \sum_{i<j} O_{ij} \left(Z_{ij}(\alpha) - \frac{\alpha}{\alpha+1} \right) - p_e \frac{\alpha}{\alpha+1}. \tag{6.10}$$

$U_1(\alpha)$ is a zero-mean random-variable, and a root of equation $U_1(\alpha) = 0$ provides an unbiased estimate of α, namely $\widehat{\alpha}_1$. An implicit form for such root is given by

$$\widehat{\alpha}_1 = \frac{\sum_{i<j} O_{ij} Z_{ij}(\widehat{\alpha}_1)}{\sum_{i<j} O_{ij} \left(1 - Z_{ij}(\widehat{\alpha}_1)\right) + n_e}. \tag{6.11}$$

As we will discuss in detail in the following section, the exact number of excluded pairs, n_e, cannot be exactly determined. We propose a procedure for estimating n_e.

6.4.1.1 Estimation of n_e

Recall Proposition 6.6, where the excluded pairs were characterized. The number n_e of excluded pairs cannot be exactly determined because of the third condition (iii) in that proposition: the event $T_{1o} \in (a_o, a_c]$ is usually unknown under interval censoring.

We propose to estimate n_e as

$$\widehat{n}_e = \sum_{(i,j)} P\left(T_{1o} \in (a_o, a_c] \middle| \text{(i) and (ii) are satisfied}\right).$$

Define D as the set of pairs (i, j) satisfying (i) and (ii). These pairs are represented in Figure 6.6. Now $\widehat{n}_e$ can be expressed as:

$$\widehat{n}_e = \sum_{(i,j)\in D} \mathbb{1}_a P\big(T_{1o} < a_c | T_{1o} \in (a_o, b_o], T_{2o} = y_o, T_{1c} > a_c, T_{2c} = y_c = a_c, y_o > y_c, b_o > a_c\big)$$
$$+ \sum_{(i,j)\in D} \mathbb{1}_b P\big(T_{1o} < a_c | T_{1o} \in (a_o, b_o], T_{2o} > y_o, T_{1c} > a_c, T_{2c} = y_c = a_c, y_o > y_c, b_o > a_c\big)$$

where $\mathbb{1}_a = I(\delta_{2i} + \delta_{2j} = 2)$ and $\mathbb{1}_b = I(\delta_{2i} + \delta_{2j} = 1)I(y_{o'} < y_{c'})$. By developing the conditional probabilities and simplifying expressions, we finally have:

$$\widehat{n}_e = \sum_{(i,j)\in D} \mathbb{1}_a \frac{P\big(a_o < T_{1o} < a_c, T_{2o} = y_o\big)}{P\big(a_o < T_{1o} < b_o, T_{2o} = y_o\big)} + \mathbb{1}_b \frac{P\big(a_o < T_{1o} < a_c, T_{2o} > y_o\big)}{P\big(a_o < T_{1o} < b_o, T_{2o} > y_o\big)}, \tag{6.12}$$

where $P(a < T_1 < b, T_2 = c) = \int_a^b f(x,c)dx$, $P(a < T_1 < b, T_2 > c) = \int_a^b \int_c^\infty f(x,y)dxdy$ and $f(x,y)$ is the joint density of Clayton's copula model as defined in (6.9).

6.4.2 Estimation of α by inverse probability weighting

The bias induced by the comparable sample can be corrected by inverse probability weighting (IPW), following the ideas in Lakhal *et al.* (2009). We consider the comparable pairs as a sample of the $\binom{n}{2}$ possible pairs of individuals. Since the comparable sample is not obtained at random, the contribution of each comparable pair to the estimation of α is weighted by the inverse probability of being comparable. This type of correction is known in the survey sample framework as Horvitz-Thomson correction (Horvitz and Thompson, 1952).

Our proposed estimating equation for α is

$$U_2(\alpha) = \binom{n}{2}^{-1} \sum_{i<j} \frac{O_{ij}}{w_{ij}} \left\{ Z_{ij}(\alpha) - \frac{\alpha}{\alpha + 1} \right\}, \tag{6.13}$$

with w_{ij} the probability of being comparable.

We propose then an estimator of α obtained as the root of equation $U_2(\alpha) = 0$, namely $\widehat{\alpha}_2$. An implicit form for such root is given by

$$\widehat{\alpha}_2 = \frac{\sum_{i<j} O_{ij} Z_{ij}(\widehat{\alpha}_2)/w_{ij}}{\sum_{i<j} O_{ij} \left(1 - Z_{ij}(\widehat{\alpha}_2)/w_{ij}\right)}. \tag{6.14}$$

This solution requires the computation of the probabilities of being comparable, w_{ij}. Note that for all comparable pairs, we observe that $\widetilde{T}_{1ij} \in (\widetilde{a}_{ij}, \widetilde{b}_{ij}]$, and that $\widetilde{T}_{2ij} = \widetilde{y}_{ij}$, and in addition, we know that $\widetilde{L}_{ij} < \widetilde{T}_{1ij} \leq \widetilde{R}_{ij}$. Taking the conditions of comparability given in Proposition 6.4, a pair's

probability of being comparable is given by:

$$
\begin{aligned}
w_{ij} &= P(O_{ij}=1|\widetilde{T}_{1ij}\in(\widetilde{a}_{ij},\widetilde{b}_{ij}],\widetilde{L}_{ij}<\widetilde{T}_{1ij}\leq\widetilde{R}_{ij},\widetilde{T}_{2ij}=\widetilde{y}_{ij})\\
&= P(\widetilde{T}_{1ij}<\widetilde{C}_{ij},\widetilde{T}_{2ij}<\widetilde{C}_{ij},\widetilde{R}_{ij}<\widetilde{T}_{2ij}|\widetilde{T}_{1ij}\in(\widetilde{a}_{ij},\widetilde{b}_{ij}],\widetilde{L}_{ij}<\widetilde{T}_{1ij}\leq\widetilde{R}_{ij},\widetilde{T}_{2ij}=\widetilde{y}_{ij})\\
&\underset{\widetilde{T}_{1ij}<\widetilde{T}_{2ij}}{=} \frac{P(\widetilde{T}_{2ij}<\widetilde{C}_{ij},\widetilde{T}_{1ij}\in(\widetilde{a}_{ij},\widetilde{b}_{ij}],\widetilde{L}_{ij}<\widetilde{T}_{1ij}\leq\widetilde{R}_{ij},\widetilde{R}_{ij}<\widetilde{T}_{2ij}|\widetilde{T}_{2ij}=\widetilde{y}_{ij})}{P(\widetilde{T}_{1ij}\in(\widetilde{a}_{ij},\widetilde{b}_{ij}],\widetilde{L}_{ij}<\widetilde{T}_{1ij}\leq\widetilde{R}_{ij}|\widetilde{T}_{2ij}=\widetilde{y}_{ij})}\\
&= \frac{\int_{\widetilde{a}_{ij}}^{\widetilde{b}_{ij}} P(\widetilde{y}_{ij}<\widetilde{C}_{ij},\widetilde{L}_{ij}<u,u\leq\widetilde{R}_{ij}<\widetilde{T}_{2ij}|\widetilde{T}_{2ij}=\widetilde{y}_{ij})P(\widetilde{T}_{1ij}=u|\widetilde{T}_{2ij}=\widetilde{y}_{ij})du}{\int_{\widetilde{a}_{ij}}^{\widetilde{b}_{ij}} P(\widetilde{L}_{ij}<u,u\leq\widetilde{R}_{ij}|\widetilde{T}_{2ij}=\widetilde{y}_{ij})P(\widetilde{T}_{1ij}=u|\widetilde{T}_{2ij}=\widetilde{y}_{ij})du}\\
&\underset{C\perp(L,R)}{=} P(\widetilde{y}_{ij}<\widetilde{C}_{ij}|\widetilde{T}_{2ij}=\widetilde{y}_{ij})\frac{\int_{\widetilde{a}_{ij}}^{\widetilde{b}_{ij}} P(\widetilde{L}_{ij}<u,u\leq\widetilde{R}_{ij}<\widetilde{T}_{2ij}|\widetilde{T}_{2ij}=\widetilde{y}_{ij})P(\widetilde{T}_{1ij}=u|\widetilde{T}_{2ij}=\widetilde{y}_{ij})du}{\int_{\widetilde{a}_{ij}}^{\widetilde{b}_{ij}} P(\widetilde{L}_{ij}<u,u\leq\widetilde{R}_{ij}|\widetilde{T}_{2ij}=\widetilde{y}_{ij})P(\widetilde{T}_{1ij}=u|\widetilde{T}_{2ij}=\widetilde{y}_{ij})du}.
\end{aligned}
\tag{6.15}
$$

In the last expression, we can use $S_C(c)$, the survival function of the random variable C, to compute

$$P(\widetilde{y}_{ij}<\widetilde{C}_{ij}|\widetilde{T}_{2ij}=\widetilde{y}_{ij}) = P(\widetilde{y}_{ij}<C_i,\widetilde{y}_{ij}<C_j|\widetilde{T}_{2ij}=\widetilde{y}_{ij}) = [S_C(\widetilde{y}_{ij})]^2.$$

On the other hand, if we observe the last quotient in expression 6.15, we obtain similar integrals in both the numerator and the denominator involving both the distribution of $(\widetilde{L}_{ij},\widetilde{R}_{ij})$ given $\widetilde{T}_{2ij}$, and the distribution of $\widetilde{T}_{1ij}$ given $\widetilde{T}_{2ij}$. We start studying the latter.

6.4.2.1 The distribution of $\widetilde{T}_{1ij}|\widetilde{T}_{2ij}$

We first note that:

$$P(\widetilde{T}_{1ij}=u|\widetilde{T}_{2ij}=\widetilde{y}_{ij}) = \frac{P(\widetilde{T}_{1ij}=u,\widetilde{T}_{2ij}=\widetilde{y}_{ij})}{P(\widetilde{T}_{2ij}=\widetilde{y}_{ij})} = \frac{\tilde{f}(u,\widetilde{y}_{ij})}{\tilde{f}_2(\widetilde{y}_{ij})}.$$

Function $\tilde{f}(s,t)$ is the joint density function of $(\widetilde{T}_{1ij},\widetilde{T}_{2ij})$ and can be recovered from the joint survival function of (T_1,T_2). Indeed, because individuals i and j are assumed independent,

$$
\begin{aligned}
\tilde{S}(s,t) &= P(\widetilde{T}_{1ij}>s,\widetilde{T}_{2ij}>t) = P(T_{1i}>s,T_{1j}>s,T_{2i}>t,T_{2j}>t)\\
&= P(T_{1i}>s,T_{2i}>t)P(T_{1j}>s,T_{2j}>t) = [S(s,t)]^2.
\end{aligned}
$$

Therefore,

$$\tilde{f}(s,t) = 2\left[\frac{\partial^2\tilde{S}(s,t)}{\partial s\partial t} = \frac{\partial S(s,t)}{\partial s}\frac{\partial S(s,t)}{\partial t} + f(s,t)S(s,t)\right].$$

Similarly, function $\tilde{f}_2(t)$ is the density of the variable $\widetilde{T}_{2ij}$, and can be obtained from the corresponding survival function:

$$\tilde{S}_2(t) = P(\widetilde{T}_{2ij}>t) = P(T_{2i}>t,T_{2j}>t) = [S_2(t)]^2 \Longrightarrow \tilde{f}_2(t) = -2S_2(t)f_2(t).$$

Equation (6.15) is reexpressed by

$$w_{ij} = [S_C(\widetilde{y}_{ij})]^2 \frac{\int_{\widetilde{a}_{ij}}^{\widetilde{b}_{ij}} P(\widetilde{L}_{ij} < u, u \leq \widetilde{R}_{ij} < \widetilde{T}_{2ij} | \widetilde{T}_{2ij} = \widetilde{y}_{ij}) \frac{\tilde{f}(u,\widetilde{y}_{ij})}{\tilde{f}_2(\widetilde{y}_{ij})} du}{\int_{\widetilde{a}_{ij}}^{\widetilde{b}_{ij}} P(\widetilde{L}_{ij} < u, u \leq \widetilde{R}_{ij} | \widetilde{T}_{2ij} = \widetilde{y}_{ij}) \frac{\tilde{f}(u,\widetilde{y}_{ij})}{\tilde{f}_2(\widetilde{y}_{ij})} du}$$
$$= [S_C(\widetilde{y}_{ij})]^2 \frac{\int_{\widetilde{a}_{ij}}^{\widetilde{b}_{ij}} P(\widetilde{L}_{ij} < u, u \leq \widetilde{R}_{ij} < \widetilde{T}_{2ij} | \widetilde{T}_{2ij} = \widetilde{y}_{ij}) \tilde{f}(u,\widetilde{y}_{ij}) du}{\int_{\widetilde{a}_{ij}}^{\widetilde{b}_{ij}} P(\widetilde{L}_{ij} < u, u \leq \widetilde{R}_{ij} | \widetilde{T}_{2ij} = \widetilde{y}_{ij}) \tilde{f}(u,\widetilde{y}_{ij}) du}. \tag{6.16}$$

6.4.2.2 The distribution of $(\widetilde{L}_{ij}, \widetilde{R}_{ij})|\widetilde{T}_{2ij}$

Finally, we have to study the probabilities in the numerator and the denominator of expression (6.16) involving the joint distribution of $\widetilde{L}_{ij}$ and $\widetilde{R}_{ij}$) conditional to $\widetilde{T}_{2ij}$, $(\widetilde{L}_{ij}, \widetilde{R}_{ij})|\widetilde{T}_{2ij}$. Indeed, define by $\tilde{G}(l, r|y)$ the joint survival function of $(\widetilde{L}_{ij}, \widetilde{R}_{ij})$ given $\widetilde{T}_{2ij} = y$,

$$\tilde{G}(l, r|y) = P(\widetilde{L}_{ij} > l, \widetilde{R}_{ij} > r | \widetilde{T}_{2ij} = y).$$

Now, we can write

$$P(\widetilde{L}_{ij} < u, u \leq \widetilde{R}_{ij} < \widetilde{T}_{2ij} | \widetilde{T}_{2ij} = \widetilde{y}_{ij}) = \tilde{G}(0, u|\widetilde{y}_{ij}) - \tilde{G}(0, \widetilde{y}_{ij}|\widetilde{y}_{ij}) - \tilde{G}(u, u|\widetilde{y}_{ij}) + \tilde{G}(u, \widetilde{y}_{ij}|\widetilde{y}_{ij})$$
$$P(\widetilde{L}_{ij} < u, u \leq \widetilde{R}_{ij} | \widetilde{T}_{2ij} = \widetilde{y}_{ij}) = \tilde{G}(0, u|\widetilde{y}_{ij}) - \tilde{G}(u, u|\widetilde{y}_{ij})$$

Again, we can obtain $\tilde{G}(l, r|y)$ from the joint distribution of $(L, R)|Y$:

$$\begin{aligned}
\tilde{G}(l, r|y) &= P(\widetilde{L}_{ij} > l, \widetilde{R}_{ij} > r | \widetilde{T}_{2ij} = y) = P(L_i > l, L_j > l, R_i > r, R_j > r | \widetilde{T}_{2ij} = \widetilde{y}_{ij}) \\
&= I(T_{2i} < T_{2j}) \left[P(L_i > l, L_j > l, R_i > r, R_j > r | T_{2i} = \widetilde{y}_{ij}, \widetilde{y}_{ij} < T_{2j})\right] \\
&+ I(T_{2i} > T_{2j}) \left[P(L_i > l, L_j > l, R_i > r, R_j > r | T_{2j} = \widetilde{y}_{ij}, T_{2i} > \widetilde{y}_{ij})\right] \\
&= P(L > l, R > r | T_2 = \widetilde{y}_{ij}) P(L > l, R > r | T_2 > \widetilde{y}_{ij}) \\
&= G_1(l, r|\widetilde{y}_{ij}) G_2(l, r|\widetilde{y}_{ij}),
\end{aligned}$$

where $G_1(l, r|y)$ and $G_2(l, r|y)$ are the joint survival functions of $(L, R)|\{T_2 = y\}$ and $(L, R)|\{T_2 > y\}$, respectively.

Therefore, to estimate w_{ij} we need consistent estimations of $S_C(c)$, $G_1(l, r|y)$ and $G_2(l, r|y)$, but also of $f(s, t)$ and $S(s, t)$, and thus implicitly on $S_1(\cdot)$, $S_2(\cdot)$ and α, to plug-in in the expression of w_{ij}. The dependency of w_{ij} on the latter justifies, the need for an iterative algorithm to jointly estimate α and $S_1(\cdot)$. In the next section, issues concerning the estimation of all the marginal survivals are discussed.

Before, we need to establish the next property, which follows easily from the computation of w_{ij}.

Proposition 6.9. *$U_2(\alpha)$ is an unbiased zero-mean random variable.*

Proof. Recall that we are in a scenario where S_1 and S_2 are assumed to be known. We start directly

computing the expectation of $\binom{n}{2} E\left[U_2(\alpha)\right]$:

$$\begin{aligned}\binom{n}{2} E\left[U_2(\alpha)\right] &= \sum_{i<j} E\left[\frac{O_{ij}}{w_{ij}}\left\{Z_{ij}(\alpha) - \frac{\alpha}{\alpha+1}\right\}\right] \\ &= \sum_{i<j} E\left[E\left[\frac{O_{ij}}{w_{ij}}\left\{Z_{ij}(\alpha) - \frac{\alpha}{\alpha+1}\right\}\Big|\widetilde{T}_{1ij}, \widetilde{T}_{2ij}\right]\right] \\ &= \sum_{i<j} E\left[\frac{1}{w_{ij}} E\left[O_{ij}\left\{Z_{ij}(\alpha) - \frac{\alpha}{\alpha+1}\right\}\Big|\widetilde{T}_{1ij}, \widetilde{T}_{2ij}\right]\right].\end{aligned}$$

Once $\widetilde{T}_{1ij}$ and $\widetilde{T}_{2ij}$ are given, we have seen in the computation of w_{ij} (Equation 6.15) that the comparability indicator O_{ij} only depends on the distribution of the censoring variables $\widetilde{C}_{ij}$, $\widetilde{L}_{ij}$ and $\widetilde{R}_{ij}$, while the expected concordance Z_{ij} depends on the original data $(T_{1i}, T_{2i}, T_{1j}, T_{2j})$. Therefore, O_{ij} and Z_{ij} are conditionally independent:

$$\binom{n}{2} E\left[U_2(\alpha)\right] = \sum_{i<j} E\left[\frac{1}{w_{ij}} E\left[O_{ij}|\widetilde{T}_{1ij}, \widetilde{T}_{2ij}\right] E\left[\left\{Z_{ij}(\alpha) - \frac{\alpha}{\alpha+1}\right\}\Big|\widetilde{T}_{1ij}, \widetilde{T}_{2ij}\right]\right],$$

which equals zero, by Proposition 6.1 (the expectation of the expected concordance is $\alpha/\alpha+1$. □

6.5 Estimation of the marginal survivals

6.5.1 Estimation of $S_C(c)$, $G_1(l, r|y)$ and $G_2(l, r|y)$

In the computation of the subject-specific weights proposed in Section 6.4.2, estimates for $S_C(\cdot)$, the survival function of the censoring variable C, $G_1(l, r|y)$, the joint distribution function of the interval censoring variables (L, R) given $T_2 = y$, and $G_2(l, r|y)$, the joint distribution function of the interval censoring variables (L, R) given $T_2 > y$, are needed.

Function $S_C(c) = P(C > c)$ can be estimated by $\widehat{S}_C(\cdot)$, the Kaplan-Meier estimator based on $\{(y_i, 1 - \delta_{2i}), i = 1, \ldots, n\}$.

For the estimation of $G_1(l, r|y) = P(L \leq l, R \leq r|T_2 = y)$ and $G_2(l, r|y) = P(L \leq l, R \leq r|T_2 > y)$, we restrict to the subsample of $n_k < n$ individuals such that $\delta_{1i} = 1$, that is, the individuals for whom the intermediate event is known to have occurred in the upper wedge $\mathcal{D}_1$, and therefore, the time of the intermediate event is interval-censored between L and R, with $R < \infty$. Denote by

$$\{(a_k, b_k, y_k, \delta_{1k} = 1, \delta_{2k}), k = 1, \ldots, n_k\},$$

the subsample of interval-censored semi-competing risks data of such individuals.

We will obtain stratified estimations of G_1 and G_2 by stratifying the observed $y_1, \ldots, y_{n_k}$ in M groups defined by the corresponding $m - 1$ quartiles (with M=4 or 5 groups, it behaves nicely). Let S_m be the m^{th} stratum, and n_m the cardinal of this group.

Now, to approximate $G_1(l, r|y)$, we obtain the empirical joint survival function within each stratum.

That is, for stratum S_m, consider only the subsample $\{(a_k, b_k), k = 1, \ldots, n_k, y_k \in S_m\}$, and the estimate within this stratum is

$$\widehat{G}_1(l, r|m) = \frac{1}{n_m} \sum_{k=1, y_k \in S_m}^{n_k} I(a_k > l, b_k > r).$$

To approximate $G_2(l, r|y)$, consider the cumulative stratums

$$S_m^* = \bigcup_{k=m}^{M} S_m \qquad m = 1, \ldots, M.$$

The cumulative stratum S_m^* contains all y_k such that $y_k \geq y_m$ for all $y_m \in S_m$. Let n_m^* be the cardinal of S_m^*. Now we obtain the empirical joint survival function within each cumulative stratum. For stratum S_m^*, we restrict to the subsample $\{(a_k, b_k), k = 1, \ldots, n_k, y_k \in S_m^*\}$, and the estimate within this stratum is

$$\widehat{G}_2(l, r|m) = \frac{1}{n_m^*} \sum_{k=1, y_k \in S_m^*}^{n_k} I(a_k > l, b_k > r).$$

6.5.2 Estimation of $S_2(\cdot)$ and $S_T(\cdot)$

We need to consistently estimate $S_2(\cdot)$ and $S_T(\cdot)$ to recover the marginal survival $S_1(\cdot)$ of T_1. On one hand, $S_2(\cdot)$ can be estimated nonparametrically through the Kaplan-Meier estimator or by adjusting a parametric model based on data $\mathfrak{D}_2 = \{(y_i, \delta_{2i}), i = 1, ..., n\}$. On the other hand, to estimate $S_T(\cdot)$ interval-censored data must be accounted for. Indeed, $T = \min(T_1, T_2)$, the time to the first event occurring, is either interval-censored when $T = T_1$, exactly observed when $T = T_2$ or right-censored when $T > C$. Therefore, $S_T(\cdot)$ can be estimated through the nonparametric maximum likelihood estimate (NPMLE) proposed by Turnbull (1976), or through a parametric fit taking into account the intervals of observation, based on data

$$\mathfrak{D}_T = \{(a_i^T, b_i^T, \delta_{Ti} = \delta_{1i} + (1 - \delta_{1i})\delta_{2i}), i = 1, \ldots, n\},$$

where $a_i^T = a_i$ and $b_i^T = b_i$ when $T_i = T_{1i}$; $a_i^T = b_i^T = y_i$ when $T_i = T_{2i}$; and $a_i^T = y_i$, $b_i^T = \infty$ if T is right-censored at y_i.

Nonparametric estimates have been presented in Sections 2.1.3 and 5.1.2 for the Kaplan-Meier and Turnbull's estimates, respectively. Parametric models can be fitted by assuming that T_2 and T follow a log-linear model:

$$\ln T_2 = \mu_2 + \sigma_2 W_2$$
$$\ln T = \mu_T + \sigma_T W_T.$$

where W_2, W_T stand for the error term distributions. Common choices for T_2 and T are the Weibull, the log-logistic and the log-normal model, that correspond to error terms W_2, W_T following the extreme value, the logistic and the normal distributions, respectively.

Estimates of parameters $\boldsymbol{\theta_2} = (\mu_2, \sigma_2)^t$ and $\boldsymbol{\theta_T} = (\mu_T, \sigma_T)^t$ are obtained by maximizing the log-

likelihood of the given samples, that is:

$$\ln \mathcal{L}_2(\boldsymbol{\theta_2}; \mathfrak{D}_2) = \sum_{i=1}^{n} \ln \mathcal{L}_{2i}(\boldsymbol{\theta_2}),$$
$$\ln \mathcal{L}_T(\boldsymbol{\theta_T}; \mathfrak{D}_T) = \sum_{i=1}^{n} \ln \mathcal{L}_{Ti}(\boldsymbol{\theta_T}).$$

where the contribution of the ith subject to the likelihood is given by

$$\mathcal{L}_{2i}(\boldsymbol{\theta_2}) = f_2(y_i)^{\delta_{2i}} S_2(y_i)^{1-\delta_{2i}}$$
$$\mathcal{L}_{Ti}(\boldsymbol{\theta_T}) = \left[S_T(a_i^T; \boldsymbol{\theta_T}) - S_T(b_i^T; \boldsymbol{\theta_T})\right]^{\delta_{Ti}\delta_{1i}} f_T(a_i^T; \boldsymbol{\theta_T})^{\delta_{Ti}\delta_{2i}} S_T(a_i^T; \boldsymbol{\theta_T})^{1-\delta_{Ti}}$$

6.5.3 The plug-in estimation of $S_1(\cdot)$

An estimate for $S_1(\cdot)$ is obtained by plugging-in estimates for $\widehat{S}_2(\cdot)$, $\widehat{S}_T(\cdot)$ and $\widehat{\alpha}$ in equation (6.4), that is,

$$\widehat{S}_1(s; \widehat{S}_T(s), \widehat{S}_2(s), \widehat{\alpha}) = g(\widehat{S}_T(s), \widehat{S}_2(s), \widehat{\alpha}), \tag{6.17}$$

where $g(a, b, c) = \left(a^{1-c} - b^{1-c} + 1\right)^{\frac{1}{1-c}}$. In fact, two estimates of $S_1(\cdot)$ are considered, depending on which estimating method for α is assumed: we can plug-in $\widehat{\alpha}_1$ (resulting from direct estimation of the bias) or $\widehat{\alpha}_2$ (resulting from inverse probability weighting) in (6.17). Both estimates are consistent (see next chapter), deriving thus consistent estimates for $S_1(\cdot)$.

The plug-in estimator has some problems of definition, as stated for the right-censored case (Fine *et al.*, 2001). In finite samples, $\widehat{S}_T(s)$ may be greater than $\widehat{S}_2(s)$, although $S_T(s) \leq S_2(s)$ for all s. In addition, $\widehat{\alpha}$ might be less than one, because no restriction is assumed in the concordance-based method. To address this issue and ensure monotonicity, we restrict inferences to the interval $[0, \tau]$ where

$$\tau \leq \max\{s : \widehat{S}_T(s)^{1-\widehat{\alpha}} - \widehat{S}_2(s)^{1-\widehat{\alpha}} > -1, 0 \leq \widehat{S}_1(u) \leq 1, u \leq s\} \tag{6.18}$$

For $t \leq \tau$, define $\widehat{S}_1^*(t) = \min_{s \leq t}\{\widehat{S}_1(s)\}$, which is monotone.

6.6 Estimation algorithm

6.6.1 Algorithm

The algorithm to jointly estimate α and $S_1(\cdot)$ runs as follows:
INITIAL PHASE:

(i) Obtain $\widehat{S}_2(\cdot)$ and $\widehat{S}_T(\cdot)$, estimates of $S_2(t)$ and $S_T(t)$ respectively.

(ii) Obtain initial estimates of α *and* $S_1(\cdot)$, $\widehat{\alpha}^{(0)}$ *and* $\widehat{S}_1(\cdot)^{(0)}$.

(iii) Determine the comparable sample, $\mathcal{C}^{IC} = \{i < j \mid O_{ij} = 1\}$

ITERATIVE PHASE: repeat until convergence

1. *Compute the expected concordance* $Z_{ij}^{(k-1)} = Z_{ij}(\widehat{\alpha}^{(k-1)}, \widehat{S}_1(\cdot)^{(k-1)}, \widehat{S}_2(\cdot))$.
2. *Obtain* $\widehat{\alpha}^{(k)}$ *as the unique root of the corrected estimating equation, according to*
   (a) *Strategy I: Estimate* n_e *from (6.12),* $\widehat{n}_e$, *and then*

$$\widehat{\alpha}_1^{(k)} = \frac{\sum_{i<j} O_{ij} Z_{ij}^{(k-1)}}{\sum_{i<j} O_{ij} \left(1 - Z_{ij}^{(k-1)}\right) + \widehat{n}_e}$$

   (b) *Strategy II: Compute* $\widehat{w}_{ij}$ *and then*

$$\widehat{\alpha}_2^{(k)} = \frac{\sum_{i<j} \frac{O_{ij}}{\widehat{w}_{ij}} Z_{ij}^{(k-1)}}{\sum_{i<j} \frac{O_{ij}}{\widehat{w}_{ij}} \left(1 - Z_{ij}^{(k-1)}\right)}$$

3. *Update* $S_1(\cdot)$*: for* $t \leq \tau$, *with* τ *defined as in (6.18)*

$$\widehat{S}_1(t)^{(k)} = \min_{s \leq t} \{g(\widehat{S}_T(s), \widehat{S}_2(s), \widehat{\alpha}^{(k)})\}.$$

Further details on the implementation of this algorithm in R are given in Chapter 10.

CHAPTER 7

Asymptotic theory

In this chapter, we derive the asymptotic properties of the estimators presented in the previous chapter. First, in section 7.1.1, we show that the estimating equations given in (6.10) and (6.13) can be seen as U-statistics. We use their properties to show the consistency and the limiting distribution of $n^{1/2}(\widehat{\alpha}-\alpha)$ when $S_1(\cdot)$ and $S_2(\cdot)$ are known (Sections 7.1.2 and 7.1.3) or estimated (Section 7.1.4). In Section 7.2, we study the properties of $\sqrt{n}\left(\widehat{S}_1(t)-S_1(t)\right)$, where $\widehat{S}_1(t)$ is the proposed estimator for the marginal distribution of the intermediate event T_1.

7.1 Inference on the copula parameter α

7.1.1 Estimating equations and U-statistics

Under the assumption that $S_1(\cdot)$ and $S_2(\cdot)$ are known, in the previous chapter we presented two unbiased estimating equations for α. The estimating equation resulting from strategy I presented in Section 6.4.1 was:

$$U_1(\alpha)=\binom{n}{2}^{-1}\sum_{i<j}O_{ij}\left(Z_{ij}(\alpha)-\frac{\alpha}{\alpha+1}\right)-p_e\frac{\alpha}{\alpha+1},$$

where $Z_{ij}=\mathrm{E}[\Delta_{ij}|\mathcal{H}_{ij}]$ is the expected concordance defined in Section 6.3.1, and p_e is the proportion of excluded pairs within all possible pairs with $i<j$ defined in Proposition 6.8. In that proposition, excluded pairs were defined as pairs for which, under interval censoring, the expected concordance Z_{ij} is not well defined based on observed data (the pair (i,j) is not comparable, $O_{ij}=0$), but, if interval censoring was not present, the concordance indicator Δ_{ij} could be determined (the pair (i,j) would be comparable in a right-censoring data scenario, $O_{ij}^R=1$).

On the other hand, the estimating equation resulting from strategy II presented in Section 6.4.2 was:

$$U_2(\alpha) = \binom{n}{2}^{-1} \sum_{i<j} \frac{O_{ij}}{w_{ij}} \left\{ Z_{ij}(\alpha) - \frac{\alpha}{\alpha+1} \right\}, \tag{7.1}$$

where w_{ij} is the probability of being comparable given the observed data $\mathcal{H}_{ij}$,

$$w_{ij} = P(O_{ij} = 1 | \mathcal{H}_{ij}).$$

Proposition 7.1. *The two statistics $U_1(\alpha)$ and $U_2(\alpha)$ are inverse weighting estimators. We can express $U_1(\alpha)$ in a unique summation*

$$U_1(\alpha) = \binom{n}{2}^{-1} \sum_{i<j} \frac{O_{ij}}{v_{ij}} \left\{ Z_{ij}(\alpha) - \frac{\alpha}{\alpha+1} \right\}, \tag{7.2}$$

where

$$v_{ij} = \begin{cases} \frac{n_c}{n_e+n_c} & \text{if } (i,j) \in \mathcal{C}^{IC} \bigcap \{(i,j) \text{ is type C1 or C3}\} \\ 1 & \text{in any other case.} \end{cases}$$

The constant n_c is the cardinal of the set of comparable pairs from type C1 or C3 (see Section 6.3.2.1), and n_e is the cardinal of the excluded pairs.

The statistic $U_2(\alpha)$ is already expressed in a single weighted summation.

Proof. Indeed, recall that the comparable sample $\mathcal{C}^{IC}$ consists of six types of pairs, namely A1, A2, B1, B2, C1 and C3 pairs. The contribution of these pairs to the statistic U_1 is $0 \leq Z_{ij} \leq 1$ for A1 and A2 pairs, 1 for B1 and B2 pairs, and 0 for C1 and C3 pairs. The sum of $O_{ij}\{Z_{ij}(\alpha) - \alpha/\alpha+1\}$ over all (i,j) with $i<j$ can be split into three terms as follows:

$$\begin{aligned} &\sum_{(i,j)\in A1\cup A2} \left(Z_{ij}(\alpha) - \frac{\alpha}{\alpha+1} \right) + \sum_{(i,j)\in B1\cup B2} \left(1 - \frac{\alpha}{\alpha+1} \right) + \sum_{(i,j)\in C1\cup C2} \left(0 - \frac{\alpha}{\alpha+1} \right) \\ &= \sum_{(i,j)\in A1\cup A2} \left(Z_{ij}(\alpha) - \frac{\alpha}{\alpha+1} \right) + n_b \left(1 - \frac{\alpha}{\alpha+1} \right) - n_c \frac{\alpha}{\alpha+1}, \end{aligned}$$

where n_b is the number of B1 and B2 pairs, and n_c the number of C1 and C3 pairs. Thus, from (6.10),

$$\begin{aligned} \binom{n}{2} U_1(\alpha) &= \sum_{(i,j)\in A1\cup A2} \left(Z_{ij}(\alpha) - \frac{\alpha}{\alpha+1} \right) + n_b \left(1 - \frac{\alpha}{\alpha+1} \right) - (n_c + n_e) \frac{\alpha}{\alpha+1} \\ &= \sum_{(i,j)\in A1\cup A2} \left(Z_{ij}(\alpha) - \frac{\alpha}{\alpha+1} \right) + n_b \left(1 - \frac{\alpha}{\alpha+1} \right) - n_c \frac{(n_c+n_e)}{n_c} \frac{\alpha}{\alpha+1}. \end{aligned}$$

We observe in the last term of this equation that the contribution $-\frac{\alpha}{\alpha+1}$ of individuals C1 and C3

is weighted by $\frac{n_c+n_e}{n_c}$, which suggests the definition of the inverse weights as

$$v_{ij} = \begin{cases} 1 & \text{if } (i,j) \in A1 \cup A2 \\ 1 & \text{if } (i,j) \in B1 \cup B2 \\ \frac{n_c}{n_c+n_e} & \text{if } (i,j) \in C1 \cup C3 \end{cases} .$$

we obtain a unique summation where only comparable pairs contribute. □

Theorem 7.1. *Given $S_1(\cdot)$ and $S_2(\cdot)$, the statistics $U_1(\alpha)$ and $U_2(\alpha)$ defined in (7.2) and (7.1), are U-statistics of degree 2.*

Proof. In this proof, we use the properties of U-statistics, which are briefly reviewed in Appendix B.7. For further theoretical background see, for instance Lehman (1999).

Consider n i.i.d realizations of the random vector $X = (L, R, Y, \delta_1, \delta_2)$, that is, the observed interval-censored semi-competing risks data $\{X_i = (L_i, R_i, Y_i, \delta_{1i}, \delta_{2i}), i = 1, .., n\}$. The 2-degree kernel for U_1 and U_2 are given by

$$\phi_1(X_i, X_j) = \frac{O_{ij}}{v_{ij}} \left\{ Z_{ij}(\alpha) - \frac{\alpha}{\alpha+1} \right\},$$

and

$$\phi_2(X_i, X_j) = \frac{O_{ij}}{w_{ij}} \left\{ Z_{ij}(\alpha) - \frac{\alpha}{\alpha+1} \right\}.$$

In addition, the expectation of both U-statistics is zero. In Proposition 6.8 we showed that $E[U_1(\alpha)] = 0$. In proposition 6.9 we showed the same for $U_2(\alpha)$. Therefore, $U_1\big(\alpha\big)$ and $U_2\big(\alpha\big)$ are zero-mean U-statistics. □

Corollary 7.1. *Given $S_1(\cdot)$ and $S_2(\cdot)$, the roots of $U_1(\alpha) = 0$ and $U_2(\alpha) = 0$, $\widetilde{\alpha}_1$ and $\widetilde{\alpha}_2$, provide unbiased estimates of α.*

It can be checked empirically that $\widetilde{\alpha}_1$ and $\widetilde{\alpha}_2$ are unique roots of their respective equations (see Appendix B.8).

In what follows we use general theory of U-statistics to derive the asymptotic properties of $\widetilde{\alpha}_1$ and $\widetilde{\alpha}_2$. Since the structure of U_1 and U_2, and their relationship with $\widetilde{\alpha}_1$ and $\widetilde{\alpha}_2$, is analogous, we will only derive the asymptotic properties for $\widetilde{\alpha}_1$.

7.1.2 Consistency of $\widetilde{\alpha}_1$ when $S_1(\cdot)$ and $S_2(\cdot)$ are known

Assume $S_1(\cdot)$ and $S_2(\cdot)$ are known functions and denote by α_0 the true value of α. The following lemma is needed to prove the consistency of $\widetilde{\alpha}_1$:

Lemma 7.1. *The second derivative of $U_1(\alpha)$ with respect to α, that is,*

$$U_1''(\alpha) = \frac{1}{\binom{n}{2}} \sum_{i<j} \frac{O_{ij}}{v_{ij}} \left(\frac{\partial^2 Z_{ij}}{\partial \alpha^2}(\alpha) + \frac{1}{(\alpha+1)^3} \right),$$

is bounded for all values of α, that is, it exists $C > 0$ such that $|U''(\alpha)| \leq C < \infty$.

Proof. Since $0 < v_{ij} \leq 1$, then $0 \leq \frac{O_{ij}}{v_{ij}} < \infty$. Therefore, $\widetilde{U}''(\alpha)$ is bounded when $\partial^2 Z_{ij}/\partial\alpha^2$ is bounded. Given that Z_{ij} is expressed as sums, products and definite integrals of functions $S(s,t)$ and $H(s,t)$, which have bounded second derivatives, then $\partial^2 Z_{ij}/\partial\alpha^2$ is also bounded. □

Proposition 7.2. *$\widetilde{\alpha}_1$ is a strongly consistent estimator for α, that is $\widetilde{\alpha}_1 \xrightarrow{a.s} \alpha$.*

Proof. Consider the first-order Taylor expansion of function $U_1(\alpha)$ around α_0, and evaluated at the estimator $\widetilde{\alpha}_1$:

$$0 = U_1(\widetilde{\alpha}_1) = U_1(\alpha_0) + U_1'(\alpha_0)(\widetilde{\alpha}_1 - \alpha_0) + R_{U_1,1}(\widetilde{\alpha}_1, \alpha_0), \tag{7.3}$$

where

$$U_1'(\alpha_0) = \frac{\partial U_1}{\partial \alpha}\Big|_{\alpha=\alpha_0} = \frac{1}{\binom{n}{2}} \sum_{i<j} \frac{O_{ij}}{v_{ij}} \left(\frac{\partial Z_{ij}}{\partial \alpha}(\alpha_0) - \frac{1}{(\alpha_0+1)^2} \right).$$

Taylor's remainder is defined by $R_{U_1,1}(\widetilde{\alpha}_1, \alpha_0) = \frac{1}{2} U_1''(\xi)(\widetilde{\alpha}_1 - \alpha_0)^2$, where ξ is a value between $\widetilde{\alpha}_1$ and α_0. From Taylor's theorem, the remainder satisfies

$$R_{U,1}(\widehat{\alpha}, \alpha_0) = o_P(\widetilde{\alpha}_1 - \alpha_0) \tag{7.4}$$

when $U_1''(\alpha)$ is bounded, which is satisfied by Lemma 7.1. Denote by $I_1(\alpha_0) = -U_1'(\alpha_0)$. Because $\widetilde{\alpha}_1$ is the root of $U_1(\widetilde{\alpha}_1) = 0$, from (7.3) and (7.4) we obtain that the random variables $\widetilde{\alpha}_1 - \alpha_0$ and $I_1(\alpha_0)^{-1} U_1(\alpha_0)$ are asymptotically equivalent.

The variable $I_1(\alpha_0)^{-1} U_1(\alpha_0)$ converges to zero almost surely by the strong law of large numbers for U-statistics, which states that

$$U_1(\alpha_0) \xrightarrow{\text{a.s}} 0.$$

Therefore, $\widetilde{\alpha}_1 - \alpha_0$ must also converge almost surely to zero, and so $\widetilde{\alpha}_1$ is strongly consistent of α_0. □

7.1.3 Asymptotical distribution of $\widetilde{\alpha}_1$ when $S_1(\cdot)$ and $S_2(\cdot)$ are known

Proposition 7.3. *Given $S_1(\cdot)$ and $S_2(\cdot)$, the distribution of $\sqrt{n}(\widetilde{\alpha}_1 - \alpha_0)$ is asymptotically normal with mean zero and variance equal to*

$$\Sigma = 4 I_1^{-2} \sigma_1^2,$$

where $\sigma_1^2 = \sigma_1^2(\alpha_0) = Cov\Big[Q_{ij}(\alpha_0), Q_{ir}(\alpha_0)\Big]$, $Q_{ij}(\alpha) = \frac{O_{ij}}{v_{ij}} \Big\{ Z_{ij}(\alpha) - \frac{\alpha}{\alpha+1} \Big\}$, and $I_1 = I_1(\alpha_0) = -U_1'(\alpha_0)$ as defined in the previous section.

The variance Σ can be approximated by $\widetilde{\Sigma} = \widetilde{I}_1^{-2} \widetilde{J}$, where

$$\widetilde{J} = \frac{4}{n(n-1)^2} \sum_{i<j<r} \widetilde{Q}_{ij}\widetilde{Q}_{ir} + \widetilde{Q}_{ij}\widetilde{Q}_{jr} + \widetilde{Q}_{ir}\widetilde{Q}_{jr},$$

obtained replacing α_0 by its estimate $\widetilde{\alpha}_1$, $\widetilde{Q}_{ij} = Q_{ij}(\widetilde{\alpha}_1)$ and $\widetilde{I}_1 = I_1(\widetilde{\alpha}_1)$.

Consequently, confidence intervals at the δ-level are given by

$$\left(\widetilde{\alpha}_1 - z_\delta\sqrt{\frac{\widetilde{\Sigma}}{n}}\,,\ \widetilde{\alpha}_1 + z_\delta\sqrt{\frac{\widetilde{\Sigma}}{n}}\right),$$

where z_δ is the critical value of the standard normal distribution for a level δ.

Proof. Due to the central limit theorem for U-statistics (see Theorem B.1(c)) applied to the U-statistic

$$U_1(\alpha) = \binom{n}{2}^{-1}\sum_{i<j} Q_{ij}(\alpha),$$

we have that $\sqrt{n}U_1(\alpha_0)$ is asymptotically normal with mean zero and variance $4\sigma_1^2(\alpha_0)$, that is

$$\sqrt{n}U_1(\alpha_0) \to \mathcal{N}\Big(0, 4\sigma_1^2(\alpha_0)\Big) \tag{7.5}$$

where $\sigma_1^2(\alpha_0) = \mathrm{Cov}\Big[Q_{ij}(\alpha_0), Q_{ir}(\alpha_0)\Big]$. Hence, it suffices to compute the covariance function between Q_{ij} and Q_{kr} when they share a common index, $k = i$. To simplify notation, we write $Q_{ij} = Q_{ij}(\alpha_0)$ and $\sigma_1^2 = \sigma_1^2(\alpha_0)$.

As shown in the proof of Proposition (7.2), $\widetilde{\alpha}_1 - \alpha_0$ and $I_1(\alpha_0)^{-1}U_1(\alpha_0)$ are asymptotically equivalent. Hence, $\sqrt{n}(\widetilde{\alpha}_1 - \alpha_0)$ and $\sqrt{n}I_1(\alpha_0)^{-1}U_1(\alpha_0)$ are also asymptotically equivalent. From (7.5), the distribution of $\sqrt{n}(\widetilde{\alpha}_1 - \alpha_0)$ is asymptotically normal with mean 0 and variance $\Sigma = 4I_1(\alpha_0)^{-2}\sigma_1^2$, that is

$$\sqrt{n}(\widetilde{\alpha}_1 - \alpha_0) \to \mathcal{N}\Big(0, 4I_1(\alpha_0)^{-2}\sigma_1^2\Big).$$

In order to make inferences on $\widetilde{\alpha}_1$, notice first that, if $S_1(\cdot)$ and $S_2(\cdot)$ were known functions, we could estimate $4\sigma_1^2$ consistently. Indeed, since $E\left[U_1(\alpha_0)\right] = 0$,

$$\begin{aligned}\mathrm{Var}\left[\sqrt{n}U_1(\alpha_0)\right] &= \mathrm{E}\left[nU_1(\alpha_0)^2\right] = \mathrm{E}\left[\frac{4n}{n^2(n-1)^2}\left(\sum_{i<j} Q_{ij}\right)^2\right] \\ &= \mathrm{E}\left[\frac{4}{n(n-1)^2}\sum_{i<j<r} Q_{ij}Q_{ir} + Q_{ij}Q_{jr} + Q_{ir}Q_{jr}\right].\end{aligned}$$

The limiting variance $4\sigma_1^2$ is then equal to $\lim_{n\to\infty}\frac{4}{n(n-1)^2}\mathrm{E}\Big[\sum_{i<j<r} Q_{ij}Q_{ir} + Q_{ij}Q_{jr} + Q_{ir}Q_{jr}\Big]$, and can be approximated by replacing α_0 by its estimate $\widetilde{\alpha}_1$, that is, by expression

$$\widetilde{J} = \frac{4}{n(n-1)^2}\sum_{i<j<r} \widetilde{Q}_{ij}\widetilde{Q}_{ir} + \widetilde{Q}_{ij}\widetilde{Q}_{jr} + \widetilde{Q}_{ir}\widetilde{Q}_{jr}$$

where $\widetilde{Q}_{ij} = Q_{ij}(\widetilde{\alpha}_1)$. Therefore, the variance $\Sigma = 4I_1^{-2}\sigma_1^2$ can be approximated by $\widetilde{\Sigma} = \widetilde{I}_1^{-2}\widetilde{J}$. □

The general situation, however, is that $S_1(\cdot)$ and $S_2(\cdot)$ are unknown functions. What we can obtain, at most, are consistent estimates for them.

7.1.4 Inference on α when $S_1(\cdot)$ and $S_2(\cdot)$ are estimated

Assume we can derive strongly consistent estimates of $S_1(t)$ and $S_2(t)$, uniformly in $t \in [0, \tau]$, and denote them by $\widehat{S}_1(t)$ and $\widehat{S}_2(t)$, respectively. Recall from (6.18) that τ satisfies

$$\tau \leq \max\{s : \widehat{S}_T(s)^{1-\widehat{\alpha}} - \widehat{S}_2(s)^{1-\widehat{\alpha}} > -1, 0 \leq \widehat{S}_1(u) \leq 1, u \leq s\}.$$

Consider now $\widehat{\alpha}_1$, the root of equation $U_1\big(\alpha, \widehat{S}_1(\cdot), \widehat{S}_2(\cdot)\big) = 0$. In this section we derive the asymptotic properties of this estimator.

Proposition 7.4. *The estimate $\widehat{\alpha}_1$, root of equation $U_1\big(\alpha, \widehat{S}_1(\cdot), \widehat{S}_2(\cdot)\big) = 0$, is asymptotically equivalent to $\widetilde{\alpha}_1$.*

Proof. Consider the continuous function $g_{\alpha_0}(u, v) = U_1(\alpha_0, u, v)$. Because we have consistent estimates for $S_1(\cdot)$ and $S_2(\cdot)$, then by Theorem 4.1, $g_{\alpha_0}(\widehat{S}_1(t), \widehat{S}_2(t)) \xrightarrow{P} g_{\alpha_0}(S_1(t), S_2(t))$. Thus $U_1\big(\alpha_0, \widehat{S}_1(\cdot), \widehat{S}_2(\cdot)\big)$ and $U_1\big(\alpha_0, S_1(\cdot), S_2(\cdot)\big)$ are asymptotically equivalent.

In the proof of Proposition 7.2 we have shown that

$$\widetilde{\alpha}_1 - \alpha_0 \quad \text{and} \quad I_1\big(\alpha_0, S_1(\cdot), S_2(\cdot)\big)^{-1} U_1\big(\alpha_0, S_1(\cdot), S_2(\cdot)\big) \tag{7.6}$$

were asymptotically equivalent. Following similar arguments, we can see that

$$\widehat{\alpha}_1 - \alpha_0 \quad \text{and} \quad I_1\big(\alpha_0, \widehat{S}_1(\cdot), \widehat{S}_2(\cdot)\big)^{-1} U_1\big(\alpha_0, \widehat{S}_1(\cdot), \widehat{S}_2(\cdot)\big) \tag{7.7}$$

are asymptotically equivalent. It follows naturally that $\widehat{\alpha}_1$ and $\widetilde{\alpha}_1$ are also asymptotically equivalent. □

Corollary 7.2. *The estimate $\widehat{\alpha}_1$ is strongly consistent and asymptotically normal with the same limiting distribution as $\widetilde{\alpha}_1$:*

$$\sqrt{n}(\widehat{\alpha}_1 - \alpha_0) \to \mathcal{N}\Big(0, 4I_1(\alpha_0)^{-2}\sigma_1^2\big(\alpha_0, S_1(\cdot), S_2(\cdot)\big)\Big).$$

The consistency of $\widehat{\alpha}_1$ is directly inferred by the strong consistency of $\widetilde{\alpha}_1$.

The limiting variance $\sigma_1^2\big(\alpha_0, S_1(\cdot), S_2(\cdot)\big)\Big)$ can be approximated by

$$\widehat{J} = \frac{4}{n(n-1)^2} \sum_{i<j<r} \widehat{Q}_{ij}\widehat{Q}_{ir} + \widehat{Q}_{ij}\widehat{Q}_{jr} + \widehat{Q}_{ir}\widehat{Q}_{jr}$$

where $\widehat{Q}_{ij} = Q_{ij}\big(\widehat{\alpha}_1, \widehat{S}_1(\cdot), \widehat{S}_2(\cdot)\big) = \frac{O_{ij}}{v_{ij}}\left\{Z_{ij}\big(\widehat{\alpha}_1, \widehat{S}_1(\cdot), \widehat{S}_2(\cdot)\big) - \frac{\widehat{\alpha}_1}{\widehat{\alpha}_1+1}\right\}$, and $I_1(\alpha_0, S_1(\cdot), S_2(\cdot))$ can be approximated by $I_1(\widehat{\alpha}_1, \widehat{S}_1(\cdot), \widehat{S}_2(\cdot)) = -U_1'\big(\widehat{\alpha}_1, \widehat{S}_1(\cdot), \widehat{S}_2(\cdot)\big)$.

7.2 Inference on the survival function $S_1(\cdot)$

An estimate for $S_1(\cdot)$ could be obtained by plugging in $\widehat{S}_T(\cdot)$, $\widehat{S}_2(\cdot)$, $\widehat{\alpha}$, consistent estimates for $S_2(\cdot)$, $S_T(\cdot)$ and α respectively, in $g(a, b, c) = \big(a^{1-c} + b^{1-c} - 1\big)^{\frac{1}{1-c}}$. That is, as was defined in

Section 6.5

$$\widehat{S}_1(t) = g(\widehat{S}_T(t), \widehat{S}_2(t), \widehat{\alpha}).$$

Note that we employ the general notation $\widehat{\alpha}$ to refer to $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ indistinctively, the roots of $U_1\big(\alpha, \widehat{S}_1(\cdot), \widehat{S}_2(\cdot)\big)$ and $U_2\big(\alpha, \widehat{S}_1(\cdot), \widehat{S}_2(\cdot)\big)$, respectively.

To study the asymptotic properties of $\widehat{S}_1(t)$ we need to study the asymptotic properties of $\widehat{S}_T(t)$ and $\widehat{S}_2(t)$ (properties of $\widehat{\alpha}$ have been studied in the previous sections).

We begin discussing the asymptotics when a parametric model is fitted for $S_2(t)$ and $S_T(t)$. Maximum likelihood is hence used to estimate the parameters of the models, $\boldsymbol{\theta_2} = (\mu_2, \sigma_2)^t$ and $\boldsymbol{\theta_T} = (\mu_T, \sigma_T)^t$, as defined in Section 6.5.2. Under some regularity conditions, their corresponding estimators, $\widehat{\boldsymbol{\theta}}_{\boldsymbol{2}}$ and $\widehat{\boldsymbol{\theta}}_{\boldsymbol{T}}$, are unique and consistent, and their asymptotic distributions are multivariate normal distributions

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{2}} - \boldsymbol{\theta_2}) \to \mathcal{N}_{p_2}(0, \mathcal{I}_2^{-1}(\boldsymbol{\theta_2})) \qquad \text{and} \qquad \sqrt{n}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{T}} - \boldsymbol{\theta_T}) \to \mathcal{N}_{p_T}(0, \mathcal{I}_T^{-1}(\boldsymbol{\theta_T})),$$

where $\mathcal{I}_2(\boldsymbol{\theta_2})$ and $\mathcal{I}_T(\boldsymbol{\theta_T})$ are their corresponding information matrices, and $p_2 = \dim \boldsymbol{\theta_2}$ and $p_T = \dim \boldsymbol{\theta_T}$. The information matrices can be approximated by their observed information matrices, $I_2(\boldsymbol{\theta_2}) = -\sum_{i=1}^n \frac{\partial^2 \ln \mathcal{L}_{2i}(\boldsymbol{\theta_2})}{\partial \boldsymbol{\theta_2} \boldsymbol{\theta_2}'}$ and $I_T(\boldsymbol{\theta_T}) = -\sum_{i=1}^n \frac{\partial^2 \ln \mathcal{L}_{Ti}(\boldsymbol{\theta_T})}{\partial \boldsymbol{\theta_T} \boldsymbol{\theta_T}'}$, where $\mathcal{L}_{2i}$ and $\mathcal{L}_{Ti}$ are given by

$$\begin{aligned}
\mathcal{L}_{2i}(\boldsymbol{\theta_2}) &= f_2(y_i)^{\delta_{2i}} S_2(y_i)^{1-\delta_{2i}} \\
\mathcal{L}_{Ti}(\boldsymbol{\theta_T}) &= \left[S_T(a_i^T; \boldsymbol{\theta_T}) - S_T(b_i^T; \boldsymbol{\theta_T})\right]^{\delta_{1i}} f_T(a_i^T; \boldsymbol{\theta_T})^{\delta_{Ti}\delta_{2i}} S_T(a_i^T; \boldsymbol{\theta_T})^{1-\delta_{Ti}}.
\end{aligned}$$

The form of these likelihood functions is determined by the type of observed data. Recall that, while T_2 is exactly observed or right-censored, T, the minimum between T_1 and T_2, may be interval-censored (when $T = T_1$), exactly censored (when $T = T_2$) or right censored (when $T = C$).

Due to the invariance property of the maximum likelihood estimators, $S_2(t; \widehat{\boldsymbol{\theta}}_{\boldsymbol{2}})$ is the maximum likelihood estimator of $S_2(t; \boldsymbol{\theta_2})$ for all $t \in [0, \tau]$, and therefore, it is uniformly consistent for all $t \in [0, \tau]$. Moreover, by applying the delta method, we can conclude that the asymptotic distribution for the parametric estimate $S_2(t; \widehat{\boldsymbol{\theta}}_{\boldsymbol{2}})$ is

$$\sqrt{n}\left(S_2(t; \widehat{\boldsymbol{\theta}}_{\boldsymbol{2}}) - S_2(t; \boldsymbol{\theta_2})\right) \to \mathcal{N}\left(0, W_1(\boldsymbol{\theta_2})\right), \tag{7.8}$$

where $W_1(\boldsymbol{\theta_2}) = \left(\frac{\partial S_2(t;\boldsymbol{\theta_2})}{\partial \boldsymbol{\theta_2}}\right)^t \mathcal{I}_2^{-1}(\boldsymbol{\theta_2}) \left(\frac{\partial S_2(t;\boldsymbol{\theta_2})}{\partial \boldsymbol{\theta_2}}\right)$. The same arguments apply to $S_T(t, \widehat{\boldsymbol{\theta}}_{\boldsymbol{T}})$, and its asymptotic distribution is given by

$$\sqrt{n}\left(S_T(t, \widehat{\boldsymbol{\theta}}_{\boldsymbol{T}}) - S_T(t, \boldsymbol{\theta_T})\right) \to \mathcal{N}\left(0, W_2(\boldsymbol{\theta_T})\right), \tag{7.9}$$

where $W_2(\boldsymbol{\theta_T}) = \left(\frac{\partial S_T(t;\boldsymbol{\theta_T})}{\partial \boldsymbol{\theta_T}}\right)^t \mathcal{I}_T^{-1}(\boldsymbol{\theta_2}) \left(\frac{\partial S_T(t;\boldsymbol{\theta_T})}{\partial \boldsymbol{\theta_T}}\right)$.

The consistency of $\widehat{S}_1(t)$ is guaranteed by Theorem 4.1 and the consistency of $\widehat{\alpha}$, $S_2(t; \widehat{\boldsymbol{\theta}}_{\boldsymbol{2}})$ and $S_T(t; \widehat{\boldsymbol{\theta}}_{\boldsymbol{T}})$. With respect to the weak convergence of $\widehat{S}_1(t)$ for $t \in [0, \tau]$, applications of the functional

and finite-dimensional delta methods show that $\sqrt{n}(\widehat{S}_1(t) - S_1(t))$ is asymptotically equivalent to

$$J_x(t) = g_1(S_T(t), S_2(t), \alpha_0)\left[\sqrt{n}\{S_T(t; \widehat{\boldsymbol{\theta}}_{\boldsymbol{T}}) - S_T(t; \boldsymbol{\theta}_{\boldsymbol{T}})\}\right] + g_2(S_T(t), S_2(t), \alpha_0)\left[\sqrt{n}\{S_2(t; \widehat{\boldsymbol{\theta}}_{\boldsymbol{2}}) - S_2(t; \boldsymbol{\theta}_{\boldsymbol{2}})\}\right] + g_3(S_T(t), S_2(t), \alpha_0)\left[\sqrt{n}\{\widehat{\alpha}_k - \alpha_0\}\right]$$

for $t \in [0, \tau]$, where

$$g_1(a,b,c) = \partial g(a,b,c)/\partial a = a^{-c}(a^{1-c} - b^{1-c} + 1)^{c/(1-c)},$$
$$g_2(a,b,c) = \partial g(a,b,c)/\partial b = -b^{-c}(a^{1-c} - b^{1-c} + 1)^{c/(1-c)},$$
$$g_3(a,b,c) = \partial g(a,b,c)/\partial c = g(a,b,c)\left\{\frac{\log(a^{1-c} - b^{1-c} + 1)}{(1-c)^2} + \frac{-a^{1-c}\log(a) + b^{1-c}\log(b)}{(a^{1-c} - b^{1-c} + 1)(1-c)}\right\}.$$

Being a sum of asymptotically normal distributions, we can conclude the asymptotical normality of $\sqrt{n}(\widehat{S}_1(t) - S_1(t))$. The computation of the variance, however, is not straightforward because we cannot use the theory of U-statistics just as in Fine *et al.* (2001) or Lakhal *et al.* (2008), because since data is interval-censored, theory of counting processes does not apply here and we cannot derive a martingale representation of $\sqrt{n}\{S_T(t; \widehat{\boldsymbol{\theta}}_{\boldsymbol{T}}) - S_T(t; \boldsymbol{\theta}_{\boldsymbol{T}})\}$. We suggest the use of the jackknife or bootstrap methods to approximate the variance.

More difficulties are found to make inferences on $S_1(t)$ when $S_2(t)$ and $S_T(t)$ are estimated non-parametrically. In particular, the asymptotical behaviour of the NPMLE (Turnbull's estimate) for $S_T(t)$ is far much complex than for its counterpart for right-censored data. Indeed, the Kaplan-Meier estimate $\widehat{S}_2(t)$ is strongly consistent and asymptotically normal (a similar expression to (7.8) holds). On the contrary, the consistency, the $\sqrt{n}$ convergence and the asymptotic normality of Turnbull's estimate $\widehat{S}_T(t)$ can only be assured under some restrictive situations. We summarize the discussion on the asymptotics of the NPMLE estimator found in Gomez *et al.* (2009) and refer the reader to this tutorial for a complete discussion on the convergence of $\widehat{S}_T(t)$.

In general, the NPMLE estimator is uniformly strongly consistent if the theoretical survival $S_T(t)$ is continuous and its support is contained in the support of the inspection times. In addition, Yu *et al.* (1998) assumed a setting where the inspection times are discrete random variables (for instance, a longitudinal study with a fixed number of scheduled visits) and the survival function $S_T(t)$ is continuous. In this setting, they showed the uniformly consistency of $\widehat{S}_T(t)$ whenever the closure of the subset $\mathcal{A}$ of all the possible values of L and R contained all the points where $\widehat{S}_T(t)$ jumps. Moreover, if $\mathcal{A} = \{a_1, \ldots, a_m\}$ is a finite set, and for all a_i, a_j with $a_i < a_j$ it is satisfied that $0 < S_T(a_i) < S_T(a_j) < 1$, then

$$\sqrt{n}\{\widehat{S}_T(a_1) - S_T(a_1), \ldots, \widehat{S}_T(a_m) - S_T(a_m)\}$$

is asymptotically normal. Moreover, Huang (1999) proved that $\sqrt{n}(\widehat{S}_T(t_0) - S_T(t_0))$ converges to a Gaussian process whenever we have enough exact failure times together with enough censoring intervals (with the right-extreme $R < \infty$), $S_T(t)$ is continuous and $\mathcal{A}$ is finite.

Whenever these restrictions are satisfied, we could obtain similar expression as in (7.9), and thus derive the asymptotic normality of the plug-in estimator for $S_1(t)$ for a given t.

CHAPTER 8

Interval-censored semi-competing risks analysis of the Spanish Bladder Cancer Study

Understanding bladder cancer disease requires knowledge about the following three processes: recurrence of the tumour, progression and death (due or not to the tumour). Recurrence is an intermediate event and its study is hindered by the dependent censoring provoked by progression or death, terminating events of recurrence. Analogously, the observation of progression may be prevented by the occurrence of death. Competing risks is the the usual approach for modelling these two intermediate processes. In Chapter 2 we used competing risks for modelling relapse free survival and progression free survival and obtained some relevant conclusions on the most important risk factors for these two processes. However, the competing risks approach treated both recurrence and progressions as terminating events, ignoring that after the occurrence of each of these events the patient has been followed, hence the course of the disease after recurrence or progression has been ignored.

The semi-competing risk approach provides a new insight into these two processes by making use of the whole history of the patient in order to recover the marginal distribution of recurrence and progression. The results obtained from competing risk and semi-competing risk approaches are complementary, since the first focus on the cumulative incidence of events (the events that are observed in presence of other causes of failure) while the second focus on the marginal distributions of the process (if other causes of failure would not be present). In the Spanish Bladder Cancer/EPICURO Study both recurrence and progression are interval-censored, hence the semi-competing risks methods have to acknowledge this incompleteness. In this chapter we apply the two estimation strategies presented in Chapter 6 to deal with interval-censored semi-competing risk (ICSCR) data.

In Table 8.1 we briefly summarize the most relevant endpoints of bladder cancer, as defined in

Table 8.1: Relevant lifetime variables for bladder cancer.

	Right-censored		**Int.cens./exact data†**	
Endpoint	%	Median	%	Median
T_R	66.8%	79.9 months	33.2%	8.2 months
T_P	92.5%	75.9 months	7.5%	9.1 months
T_{DBC}	93.4%	78.9 months	6.6%	38.8 months
T_{DOC}	78.7%	81.2 months	20.3%	42.7 months

†T_R and T_P are interval-censored, T_{DBC}, T_{DOC} exactly observed.

Chapter 1: T_R, the time to the first recurrence; T_P, the time to the first progression; T_{DBC}, the time to death due to bladder cancer; and T_{DOC}, the time to death due to other not disease-related causes. In the table, we describe for each time, the percentage of right-censored observations, the median time among these, the percentage of interval-censored (in the case of recurrence or progression) or exactly observed data (in the cases of death), and the median time among these. In Part I we ignored the presence of interval censoring by imputing the midpoint of the corresponding observed intervals to T_R and T_P. In this Chapter, we will account for interval-censored data whenever the events are considered intermediate, while we keep the midpoint imputation strategy when these events are terminating. In Section 8.1, we present the setting where recurrence acts as an intermediate event for progression or death. Section 8.2 is devoted to the analysis where progression acts as an intermediate event for death. In Section 8.3 we present a simulated example in which the association between T_R and T_P, for instance, is strong and the amount of dependent censoring is higher than that encountered in the SBC/EPICURO Study. We use this illustration to show how, when strong association and dependent censoring exists, the semi-competing risks methodology enables us to recover the marginal laws of T_R and T_P.

8.1 The recurrence process in bladder cancer: a situation with low association

The recurrence process is right-censored by the occurrence of progression or death. This semi-competing risks situation is depicted in Figure 8.1. Recurrence is the intermediate event thus $T_1 = T_R$ is right-censored by the time until the first between progression or death occurs, namely $T_2 = \min(T_P, T_{DBC}, T_{DOC})$. In Part I we focused the analysis on $T_2|T_R$, assessing how the occurrence of a recurrence modified the risk of a posterior progression (Chapter 3). We are now interested in studying the recurrence process if progression or death had not occurred before; that is, we want to recover T_R and assess its behavior marginally.

Denote by δ_1 the indicator for recurrence (1 if occurs, 0 otherwise), and by δ_2, the corresponding indicator for the occurrence of progression or death. Table 8.2 summarizes the number of events for each type. Observe that only 87 patients (8.7% of the total 995) experience first a recurrence and then progress or die, while for 243 patients only recurrence is recorded (24.4%). On the other hand, there is a 22.5% of dependent censoring caused by 224 patients experiencing progression or

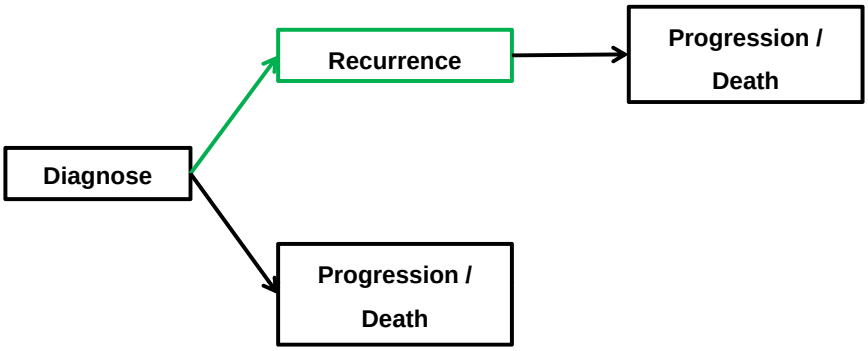


Figure 8.1: Semi-competing risks data for Recurrence and Progression/death.

death ($\delta_2 = 1$), but not recurrence ($\delta_1 = 0$).

Table 8.2: Events of interest (intermediate=recurrence).

	Progression/Death		
Recur.	**No**	**Yes**	**Total**
No	441	224	665
	(44.3%)	(22.5%)	(66.8%)
Yes	243	87	330
	(24.4%)	(8.7%)	(33.2%)
Total	684	311	995
	(68.7%)	(31.3%)	(100.0%)

In the following sections, we obtain estimators for α, the association parameter and $S_1(t)$, the survival function of T_R, for the whole group of superficial bladder cancer cases ($n = 995$), and we also perform stratified analysis for each of the prognostic factors presented in Chapter 1: gender, age, tumour number, tumour size, stage, grade and smoking status. We have used nonparametric estimates for $S_T(t)$ and $S_2(t)$ and we plugged them in Equation (6.4) to obtain an estimator for $S_1(t)$. Since in the SBC/EPICURO Study patients are observed at regular visits due to their medical antecedents caused by the primary tumour, it is reasonable to assume that the inspection process has been made at discrete time points. Under this assumption (see the discussion in Section 7.2) the NPMLE for $S_T(t)$ converges to a Gaussian process at a $\sqrt{n}$ rate, therefore, the asymptotic properties of the ICSCR estimate for $S_1(t)$ are fulfilled.

We start checking whether Clayton's copula can be assumed in the upper wedge of observation. An approximate check of this assumption can be done by imputing the midpoint of the censoring interval and applying the goodness-of-fit test proposed by Fine *et al.* (2001) for right-censored semi-competing risks data. The results are given in Table 8.3 which shows that the fit is plausible except for the female's group. We analyse subsequently the data without incorporating gender.

Table 8.3: Goodness-of-fit tests for the Clayton's model (Fine *et al.*, 2001).

Variable	**Categories**	n	**Z test**	**p-value**
Total group		995	0.679	0.249
Gender	Male	868	1.343	0.090
	Female	127	2.707	0.003
Age (years)	≤ 60	254	0.423	0.336
	61-70	378	0.679	0.249
	>70	363	0.694	0.244
Tumour number	Single	660	0.812	0.208
	Multiple	283	0.504	0.307
Size	< 3cm	854	1.136	0.128
	> 3 cm	141	0.261	0.397
Stage	Ta	828	1.460	0.072
	T1/Tis	167	1.234	0.109
Grade	G1/Benign	424	0.767	0.222
	G2	332	1.148	0.126
	G3	239	1.142	0.127
Smoking status	Non-smoker	155	0.616	0.269
	Smoker	728	0.480	0.316

8.1.1 Estimation of the association parameter α

The two estimation strategies presented in Chapter 6 are labeled Strategy 1 or ICSCR1 (direct estimation and correction of bias) and Strategy 2 or ICSCR2 (inverse-weighting by the probability of being comparable). The estimators for α together with its standard error for the two methods in the total group and within each category are given in Table 8.4.

We observe, for all cases, small differences between Strategy 1 and Strategy 2, and, in particular, we do not observe a systematic trend in both estimates such as one being consistently larger than the other. The estimation and variability of α show little association between recurrence and progression, with values near the unity. However, a 95% confidence interval shows that for age greater than 70 years, Stage Ta, Grade 2 and non-smokers the association is greater than 1.

Both strategies require the specification of initial values for α and $S_1(t)$. The estimate from mid-point imputation is usually chosen, provided it is greater than one. When this condition is not fulfilled, we can use other initial values but we checked that our method is robust event if an arbitrary initial value (such as $\alpha = 10$) is chosen. In addition, initial values taking midpoint, left or right imputation results as a starting point resulted in similar estimates.

8.1.2 Estimation of the time to recurrence

Aiming to analyse the marginal distribution of recurrence, $1 - S_1(t)$, and to compare it to the incidence for recurrence, we present plots in Figures 8.2a to 8.2d for those covariates exhibiting more evidence of association, namely: age, stage, grade and smoking status. Since Strategies 1 and

Table 8.4: Estimates for α when recurrence is an intermediate event (ICSCR analysis).

			Strategy 1			**Strategy 2**		
Variable	**Categories**	n	$\widehat{\alpha}_1$	$\widehat{\mathbf{SD}}_1$	$\mathbf{IC}_1\mathbf{95\%}$	$\widehat{\alpha}_2$	$\widehat{\mathbf{SD}}_2$	$\mathbf{IC}_2\mathbf{95\%}$
Total group		995	1.271	0.152	(0.507, 2.035)	1.190	0.151	(0.428, 1.951)
Age (years)	<60	254	1.219	0.485	(0.268, 2.169)	1.261	0.487	(0.307, 2.216)
	60-70	378	1.056	0.212	(0.640, 1.471)	0.966	0.031	(0.904, 1.027)
	>70 years	363	1.518	0.243	(1.042 ,1.993)	1.513	0.237	(1.048, 1.978)
Tumour number	Single	660	1.241	0.202	(0.846, 1.637)	1.176	0.196	(0.792, 1.561)
	Multiple	283	1.497	0.294	(0.920, 2.074)	1.526	0.285	(0.968, 2.084)
Tumour size	≤3cm	854	1.334	0.172	(0.997, 1.671)	1.214	0.172	(0.877, 1.551)
	>3cm	141	1.094	0.389	(0.331, 1.856)	1.127	0.366	(0.410, 1.844)
Stage	Ta	828	1.562	0.200	(1.170, 1.954)	1.483	0.192	(1.105, 1.860)
	T1/Tis	167	0.452	0.234	(-0.006, 0.910)	0.889	0.028	(0.835, 0.943)
Grade	G1/Benign	424	1.327	0.281	(0.776, 1.878)	1.276	0.301	(0.686, 1.866)
	G2	332	1.742	0.353	(1.050, 2.434)	1.561	0.320	(0.934, 2.187)
	G3	239	0.844	0.370	(0.118, 1.570)	0.781	0.031	(0.721, 0.841)
Smoking status	Non-smoker	155	2.536	0.955	(0.664, 4.409)	2.653	0.953	(0.785, 4.521)
	Smoker	728	1.189	0.164	(0.867, 1.511)	1.168	0.157	(0.859, 1.477)

$\widehat{\mathbf{SD}}$: estimated standard deviations for $\widehat{\alpha}$.

2 provide similar results and the simulation study in next Chapter 9 has shown a better behaviour for Strategy 1, only results for this approach are shown.

The fact that the association between T_R and $T_2 = \min(T_P, T_{DBC}, T_{DOC})$ is mild or even non-existent and the relatively small number of terminating events (8.7%) which furthermore tend to occur later than recurrences, causes a very similar behaviour between the cumulative incidence functions and the marginal distributions for recurrence. In such a situation the benefit of a semi-competing risks methodology to recover the marginal distribution of the intermediate event is not substantial.

Nevertheless, we can observe a modest trend of the semi-competing risks methodology to recover the marginal distribution of the events that the competing event did not allowed to observe. For instance, in the case of age, the competing risks analysis showed that older patients (more than 70 years-old) had less recurrences as a first event than younger patients (Chapter 2), of course induced by the competing terminating event (progression or death). The marginal distribution recovered by the ICSCR analysis shows a higher incidence of recurrences in this group than the CR estimate, and shows a similar rate as for younger patients. At this marginal scale, age is not related with the recurrence process. In the case of Stage, differences between Ta and T1/Tis tumours are detected both by competing and semi-competing risks analysis, though the probabilities of the marginal distribution, as expected, are higher. In the case of grade of the tumour, the probability of recurrence in G3 tumours is pushed up to become similar to G1/Benign tumours, while in the CR setting this category showed a probability of recurring as a first event behind other grades. Finally, differences between non-smokers and smokers are emphasized at the marginal scale: in this case, the competing event prevents observing the true differences between these two groups.

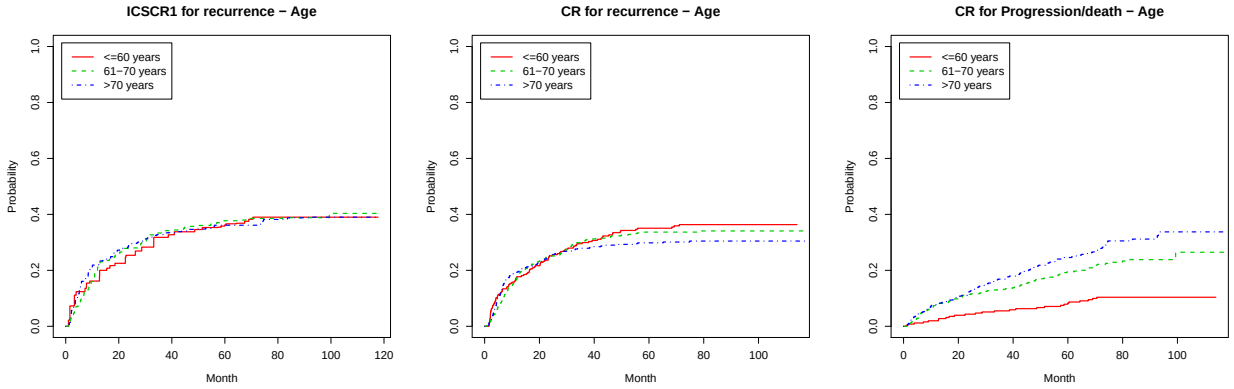


Figure 8.2a: ICSCR vs CR for Age (intermediate=recurrence)

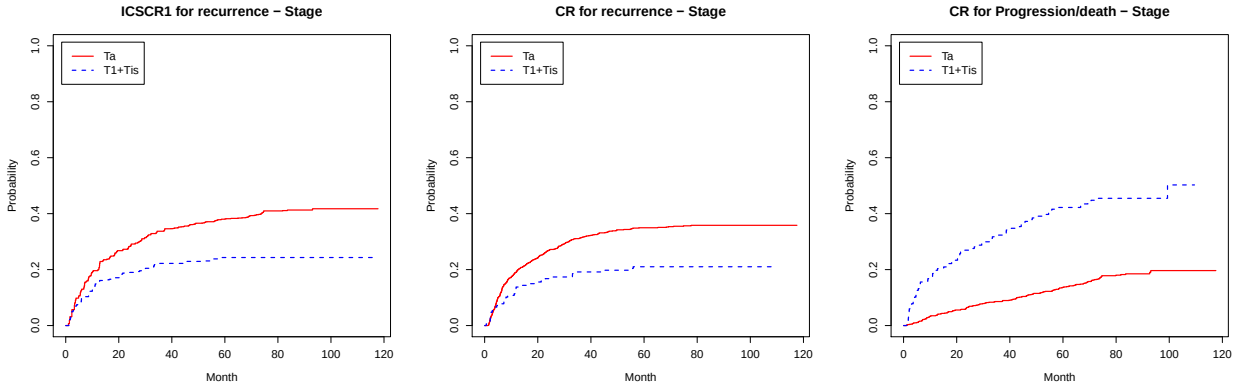


Figure 8.2b: ICSCR vs CR for Tumour stage (intermediate=recurrence)

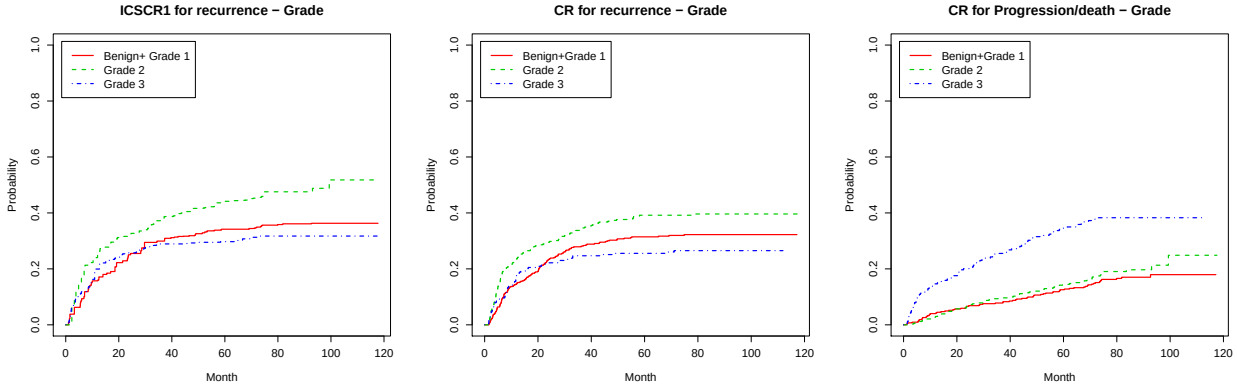


Figure 8.2c: ICSCR vs CR for Grade (intermediate=recurrence)

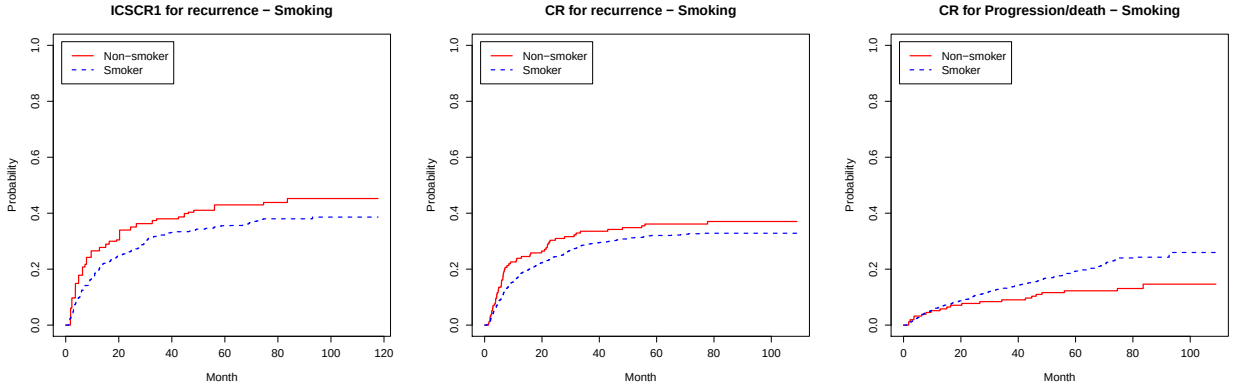


Figure 8.2d: ICSCR vs CR for Smoking status (intermediate=recurrence)

Besides the above considerations, the results of the ICSCR analysis for the Spanish Bladder Cancer/EPICURO study show no substantial gain on information on the recurrence process, not more information that it could be extracted from a competing risks analysis. The reason why competing risks and semi-competing risks analysis are similar is because of the small amount of dependent censoring present in the bladder cancer data.

8.2 The progression process in bladder cancer: a situation with moderate association

Now we are interested in the progression of disease, taking into account that death due to other causes acts as a competing event (see Section 2.3). Given that for some patients, death for not disease-related causes can occur after a progression, we want to include this information to recover the marginal distribution of the time describing Progression Free Survival. Figure 8.3 describes this semi-c
by $T_1 = \mathrm{m}$

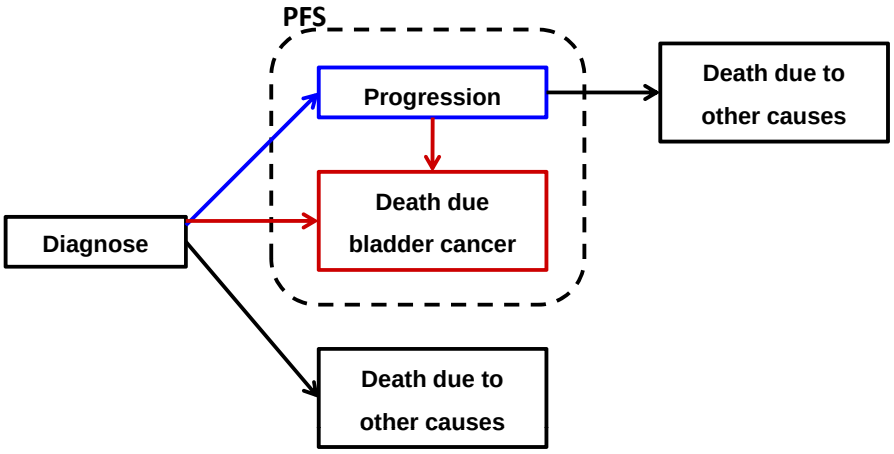


Figure 8.3: Semi-competing risks data for Progression and Death due to Other Causes.

the indicator for progression (1 if occurs, 0 otherwise), and by δ_2, the indicator for death due to other causes. The different combinations of these indicators for the total group (n=995) are given in Table 8.5. Observe that the percentage of dependent censoring is around 20%, that both events of interest are observed in only 1.2% of the patients, while 8.6% patients experience progression but not death due to other causes. Given that both events are only observed in 1.2% of patients, the estimation of the joint survival function of (T_1, T_2) is hopeless and all the results must be taken with caution. The exploration of Clayton's copula fitting yields a valid modelling only for multiplicity and stage.

8.2.1 Estimation of the association parameter α

The estimates for α for the two strategies of ICSCR are given in Table 8.6. More association is observed between progression and deaths due to other causes than between recurrence and

Table 8.5: Events of interest (intermediate=progression).

	Death OC		
Progression	**No**	**Yes**	**Total**
No	697 (70.1%)	200 (20.1%)	897 (90.2%)
Yes	86 (8.6%)	12 (1.2%)	98 (9.8%)
Total	783 (78.7%)	212 (21.3%)	995 (100.0%)

PFS/Death. Nevertheless, this is a moderate association: around 1.7 for the total group, slightly higher for Ta tumours, which would represent, on a complete plane, a correlation of 0.42. In addition, the estimates for the standard deviation are high, because in this setting, there are not enough events to properly estimate α.

Table 8.6: Estimates for α when progression is an intermediate event (ICSCR analysis).

			Strategy 1			**Strategy 2**		
Variable	**Categories**	n	$\widehat{\alpha}_1$	$\widehat{\mathrm{SD}}_1$	$\mathrm{IC}_1 95\%$	$\widehat{\alpha}_2$	$\widehat{\mathrm{SD}}_2$	$\mathrm{IC}_2 95\%$
Total group		995	1.731	0.537	(0.295, 3.167)	1.703	0.548	(0.252, 3.154)
Tumour number	Single	660	1.831	0.784	(0.294, 3.368)	1.771	0.773	(0.255, 3.287)
	Multiple	283	1.999	0.950	(0.136, 3.862)	2.135	1.016	(0.144, 4.126)
Stage	Ta	828	2.139	0.864	(0.445, 3.833)	2.167	0.923	(0.359, 3.976)
	T1/Tis	167	1.051	0.512	(0.047, 2.055)	0.905	0.035	(0.836, 0.974)

$\widehat{\mathbf{SD}}$: estimated standard deviations for $\widehat{\alpha}$.

8.2.2 Estimation of the time to progression

The analysis of time to progression both marginally and as the cumulative incidence is explored and plotted in Figures 8.4a and 8.4b. The probability of progressing to bladder tumour computed by CR and ICSCR before 12, 24 and 60 months is presented in Table 8.7.

Observe that, while 5% of individuals with multiple tumours progress before 12 months, about 14% do so among individuals with T1/Tis. Furthermore, while only 6% progress in 5 years among Stage Ta, about 28% of the individuals do so in 5 years if their stage is T1/Tis.

For stage T1/Tis tumours, the probability of progression has almost doubled from 12 to 60 months, while for multiple tumours, the probability at 60 months has almost tripled the probability at 12 months.

Though the marginal probabilities for 12 and 24 months do not differ significantly from the cumulative incidence curves for progression estimated at the same time points, differences are encountered

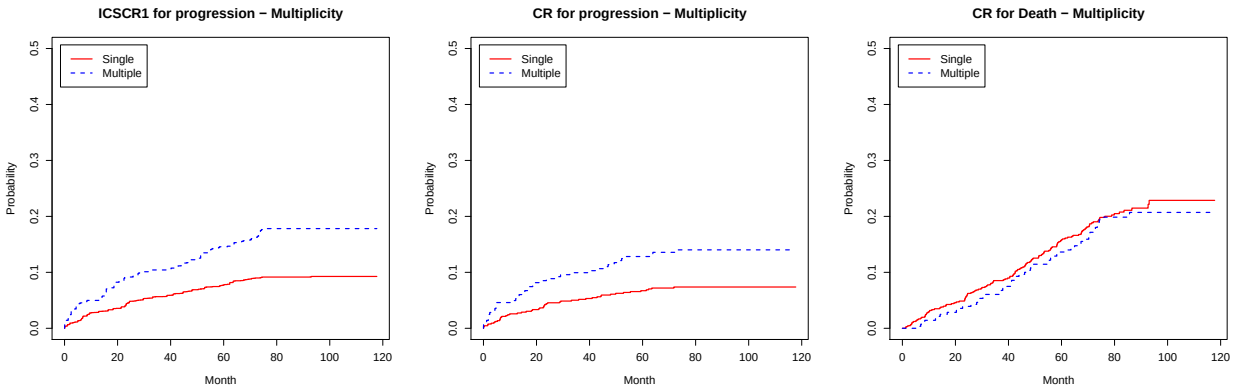


Figure 8.4a: ICSCR vs CR for Tumour number (intermediate=progression)

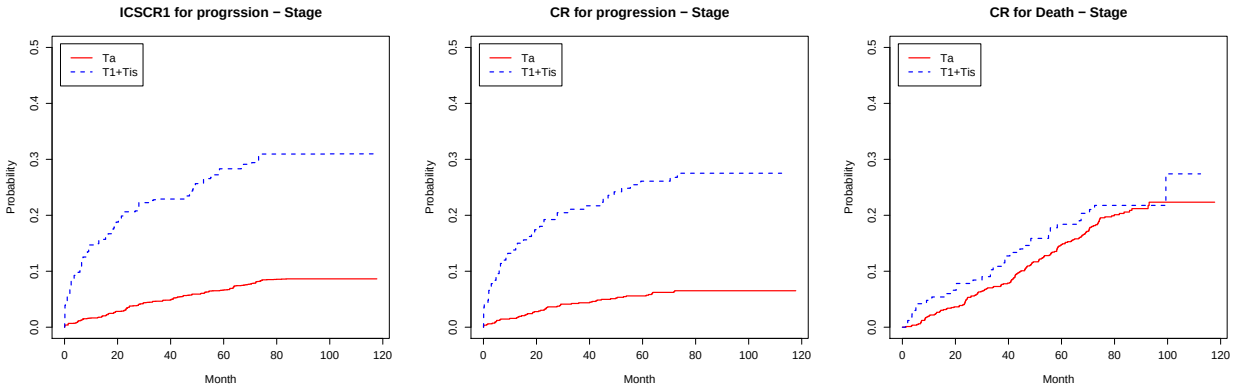


Figure 8.4b: ICSCR vs CR for Tumour stage (intermediate=progression)

at 5 years. Regarding the plotted curves, indeed larger differences exist for larger times. Marginal curves are pushed up with respect to cumulative incidence functions due to the correction of the dependent censoring performed by the ICSCR analysis.

Again in this setting the two times of interest exhibit small dependent censoring and low association, thus no substantial information is gained from a semi-competing risks analysis. However, we have observed that differences exists with regard to a competing risks analysis. In our situation, marginal curves and cumulative incidence functions are similar, but they can be radically different (see next section). Indeed, they estimate distinct quantities: via ICSCR, the distribution of T_1 is estimated, while a CR analysis provides an estimate of the joint distribution of $(T, C = \text{Prog})$, where $T = \min(T_1, T_2)$. A common error in competing risks is to interpret this distribution marginally.

8.3 Illustration: strongly associated simulated data

We turn back to the example presented in Chapter 4, Section 4.3, in which bivariate data was generated following Clayton's copula model with a strong dependency between T_1 and T_2 ($\alpha = 4$). These data were generated assuming significant differences between a Treatment variable at the marginal scale. We discussed there how the semi-competing risks methodology permitted to recover the marginal effect of the intermediate event which was hidden by the presence of the competing terminating event. A considerable amount of dependent censoring (around 50%) is present in this illustration, contrary to the bladder cancer situation, were a small amount of dependent censoring was present. On these data, censoring intervals have been generated for those T_1 events observed

Table 8.7: Estimates of the probability of progression at 12, 24 and 60 months.

Variable	Months	CR	ICSCR1	ICSCR2
Single tumours	12	0.026	0.028	0.028
	24	0.044	0.045	0.045
	60	0.067	0.078	0.077
Multiple tumours	12	0.050	0.050	0.050
	24	0.089	0.092	0.092
	60	0.128	0.146	0.148
Stage Ta	12	0.016	0.017	0.017
	24	0.036	0.035	0.035
	60	0.056	0.066	0.067
Stage T1/Tis	12	0.138	0.149	0.148
	24	0.192	0.206	0.204
	60	0.261	0.283	0.274

in the upper wedge $\mathcal{D}_1$: therefore, we can apply the ICSCR strategies 1 and 2 developed in Chapter 6.

Table 8.8 contains the estimates of α according to Strategy 1 (ICSCR1) and Strategy 2 (ICSCR2). While Strategy 1 obtains accurate estimates of α, specially in treatment group A, Strategy 2 performs poorly in treatment group B (a value of 6.707 is obtained for a true parameter of $\alpha = 4$). Figure 8.5 shows the estimated marginal distributions $S_1(t)$ at each treatment arm and for each Strategy. Again, non-parametric estimates for $S_T(t)$ and $S_2(t)$ were used in the plug-in estimate to obtain $S_1(t)$. For treatment B, Strategy 2 (blue dotted line) is surpassed by Strategy 1 (blue solid line), as reflected by the worse estimate of α. Despite the small differences, we observe that in both treatment arms and for both strategies, the shape of the real marginal distribution of T_1 is recovered.

Table 8.8: Estimates for $\alpha = 4$ for the simulated set (ICSCR analysis).

			Strategy 1		**Strategy 2**	
Variable	**Categories**	n	$\widehat{\alpha}_1$	$\widehat{\mathbf{SD}}_1$	$\widehat{\alpha}_2$	$\widehat{\mathbf{SD}}_2$
Treatment	A	250	4.064	0.745	4.465	0.857
	B	250	4.451	0.909	6.707	1.861

$\widehat{\mathbf{SD}}$: estimated standard deviations for $\widehat{\alpha}$.

Some questions remain concerning the impact of interval-censored data: except for Strategy 1 in treatment A, midpoint imputation estimates show less bias than ICSCR strategies, but ICSCR strategies present more variability so in fact no differences exists between them. Moreover, in this setting with a considerable amount of dependent censoring, Strategy 2 provided biased estimates of α, while in the bladder cancer case, with small percentages of dependent censoring, both strategies performed similar.

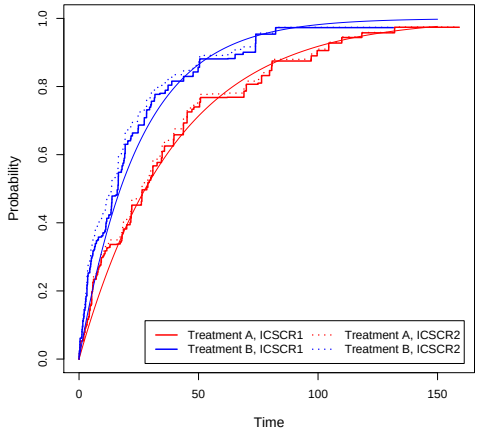


Figure 8.5: Interval-censored semi-competing risks analysis: Estimated distribution function vs real distribution function for T_1.

A simulation study is performed in Chapter 9 to compare both approaches. We need to further explore different scenarios of dependent censoring as well as different association levels, and their impact on both estimation strategies.

CHAPTER 9

Simulation Study

In this chapter we explore the performance of the methodologies proposed for interval-censored semi-competing risks data (ICSCR) in Chapter 6 by means of a simulation study. The goal of a semi-competing risks analysis that assumes a bivariate Clayton's copula model is to estimate both the association parameter α and the marginal survival function, $S_1(t)$, of the intermediate event. The estimators of α and $S_1(t)$ will be obtained and compared under the following three strategies:

1. Midpoint imputation (Midpoint),
2. Strategy 1, direct estimation of bias (ICSCR1), and
3. Strategy 2, inverse probability weighting (ICSCR2).

The midpoint imputation approach is the simplest strategy for avoiding the problem of interval censoring. If $T_1 \in (L, R]$, we assign to T_1 the value of the midpoint of the interval, $T_1 = (L + R)/2$, and we perform the semi-competing risks analysis for right-censored data proposed by Fine *et al.* (2001) and presented in Chapter 4. The second and third approaches are our proposals for analysing ICSCR data, and have been developed in Chapter 6. The difference between these two approaches rely in the estimation of the copula parameter. For the first approach, the proposed estimating equation explicitly corrects the bias induced by the comparable sample, while the second approach employs inverse weighting techniques to account for such bias. The estimation of the marginal distribution of T_1 is taken similarly using a plug-in estimate based on the assumed structure of the joint survival function, and for both methods, an iterative algorithm is required.

We first present the 48 different simulation settings and the evaluation criteria used for the comparison of the methods in Sections 9.1 and 9.2, respectively. The simulation results are summarized in Section 9.3, and the corresponding tables are presented in Appendix C.

9.1 Simulation scenarios and data generation

9.1.1 Parameters defining simulation scenarios

The simulation scenarios are defined by the following parameters. A total of 48 scenarios are considered, and $B = 1000$ data sets were generated for each scenario.

- **Sample size,** n
  Number of observations per generated data set, $n \in \{200, 500\}$
- **Copula parameter,** α
  Value of the dependence parameter $\alpha \in \{3, 5\}$. Under Clayton's copula model and following expressions (4.8), the chosen parameters represent values for Kendall's tau of $\tau_{\mathcal{K}} \in \{0.50, 0.67\}$ and Spearman's rho of $\rho_S \in \{0.65, 0.81\}$ corresponding to moderate and strong association, respectively.
- **Average length of intervals** $(L, R]$**,** w
  On one hand, we have narrow intervals referring to a scenario where, on average, the generated intervals have length around 6 units of time. On the other hand, we refer to wide intervals when the average length of the intervals is within 12 to 15 units of time.
- **Percentage of dependent censoring,** p
  The dependent censoring appears when $T_1 > T_2$, thus this percentage is obtained from $p = P(T_1 > T_2) \in \{0.25, 0.50, 0.75\}$, corresponding to moderate, high and heavy dependent censoring.
- **Distribution of the marginals** T_1 **and** T_2
  We assume that both marginals have the same distribution, which can be:
  - Exponential: $T_1 \sim \text{Exp}(\lambda_1)$ and $T_2 \sim \text{Exp}(\lambda_2)$.
    The following parametrization is taken: the survival function is $S_k(t) = e^{-\lambda_k t}$ for $k = 1, 2$. We assume that $\text{E}[T_2] = 1/\lambda_2 = 40$, and
  - Weibull: $T_1 \sim \text{Weibull}(\rho_1, \mu_1)$ and $T_2 \sim \text{Weibull}(\rho_2, \mu_2)$.
    The following parametrization is adopted: the survival function is $S_k(t) = e^{-\mu_k t^{\rho_k}}$ for $k = 1, 2$. We assume $\rho_1 = \rho_2 = 1/2$ and $E[T_2] = 40$.

A summary of the simulation parameters and the simulation settings is given in Table 9.1.

9.1.1.1 Determination of the parameters of the marginal distributions

The parameters for the distribution of T_1 depend on the percentage of dependent censoring p, the copula parameter α and the parameters for T_2. Indeed, the probability of dependent censoring is obtained from

$$p = P(T_1 > T_2) = \int_0^\infty dt \left[\int_t^\infty f(s,t) ds \right] \underset{(6.9)}{=} \int_0^\infty dt \left[\int_t^\infty f_1(s) f_2(t) (S_1(s) S_2(t))^{-\alpha} S(s,t)^{2\alpha - 1} ds \right]. \tag{9.1}$$

Table 9.1: Simulation parameters.

Parameter	**Values**	**Settings**
n	200, 500	2
α	3, 5	2
w	Narrow,Wide	2
p	0.25, 0.50, 0.75	3
Margins	Exponential, Weibull	2
	Total settings	**48**

By performing the following change of variables:

$$\begin{array}{lll} v = S_2(t) & du = -f_2(t)dt & t = 0 \Rightarrow v = 1, \quad t = \infty \Rightarrow v = 0 \\ u = S_1(s) & du = -f_1(s)ds & s = t \Rightarrow u = S_1(t) = S_1(S_2^{-1}(v)), \quad s = \infty \Rightarrow u = 0, \end{array}$$

we can rewrite (9.1) by

$$p = P(T_1 > T_2) = \int_0^1 dv \left[\int_{S_1(S_2^{-1}(v))}^1 (uv)^{-\alpha} C(u,v)^{2\alpha-1} du \right]. \tag{9.2}$$

In the case of the Exponential distribution,

$$S_1(S_2^{-1}(v)) = S_1\left(\log\left\{ \frac{1}{v^{1/\lambda_2}} \right\} \right) = v^{\frac{\lambda_1}{\lambda_2}}.$$

With this parametrization, $E[T_2] = 40 = 1/\lambda_2$ and thus $\lambda_2 = 0.025$. Therefore, the parameter λ_1 must satisfy

$$p = \int_0^1 dv \left[\int_{v^{\lambda_1/0.025}}^1 (uv)^{-\alpha} C(u,v)^{2\alpha-1} du \right]. \tag{9.3}$$

By varying p and α we obtain λ_1 for each simulation setting.

In the case of the Weibull distribution,

$$S_1(S_2^{-1}(v)) = S_1\left(\left[\frac{1}{\mu_2} \log \frac{1}{v} \right]^{1/\rho_2} \right) = \exp\left\{ -\mu_1 \left[\frac{1}{\mu_2} \log \frac{1}{v} \right]^{\rho_1/\rho_2} \right\}.$$

With the assumed parametrization,

$$E[T_2] = \frac{1}{\mu_2^{1/\rho_2}} \Gamma\left(\frac{1}{\rho_2} + 1 \right),$$

where $\Gamma(t)$ is the gamma function, equal to $(t-1)!$ if t is an integer. Because we assume $\mathrm{E}[T_2] = 40$

and $\rho_2 = 1/2$, then we have

$$\mu_2^2 = \frac{\Gamma(3)}{\mathrm{E}[T_2]} = 0.05 \Rightarrow \mu_2 = 0.2236.$$

To obtain μ_1, and given that $\rho_1 = \rho_2 = 1/2$, we must solve the equation

$$p = \int_0^1 dv \left[\int_{u_0(v)}^1 \left\{ (uv)^{-\alpha} C(u,v)^{2\alpha-1} du \right\} \right], \tag{9.4}$$

where $u_0(v) = \exp\left\{ -\mu_1 \left[\frac{1}{\mu_2} \log \frac{1}{v} \right] \right\} = v^{\frac{\mu_1}{\mu_2}}$.

9.1.1.2 Percentage of independent censoring

In all simulation scenarios we have considered an independent censoring variable C following a uniform distribution, $U[0, C_m]$. To guarantee a fixed percentage of 20% of independent censoring in all scenarios, we take $C_m = 200$. Indeed, if $f_C(c) = 1/(C_m)$ is the density function of C, then C_m is obtained resolving the following equation:

$$0.20 = P(C < T_2) = \int_0^\infty dy \int_0^y \frac{1}{C_m} f_2(y) du = \frac{1}{C_m} \int_0^\infty y f_2(y) dy = \frac{1}{C_m} E[T_2].$$

In all the scenarios considered, we have fixed $E[T_2] = 40$ (both for Exponential and Weibull distributions). Therefore,

$$C_m = \frac{40}{0.2} = 200,$$

and $C \sim U[0, 200]$.

9.1.2 Generation of data sets

For each of the B data sets of a particular scenario, n data vectors of the form $(L, R, Y, \delta_1, \delta_2)$ are generated as follows:

1. **Generation of (T_1, T_2)**
   A bivariate sample is drawn from a Clayton's copula model in all the plane following the Inverse Probability Method (Trivedi and Zimmer, 2007). We draw v_1 and v_2 from a uniform random variable in $(0, 1)$, $U(0, 1)$. We set $u_1 = v_1$ and

   $$u_2 = \frac{\partial C_\alpha(u_1, u_2)}{\partial u_1} = \left(v_1^{1-\alpha} \left(v_2^{(1-\alpha)/\alpha} - 1 \right) + 1 \right)^{1/(1-\alpha)}.$$

   Then, we obtain $T_1 = S_1^{-1}(u_1)$ y $T_2 = S_2^{-1}(u_2)$, where the inverse survivals depend on the chosen parametric model (Exponential or Weibull).

2. **Generation of Y, δ_1 and δ_2**
   We generate C following $U[0, 200]$. Calculate then $Y = \min(T_2, C)$, $\delta_1 = I(T_1 \leq Y)$, and $\delta_2 = I(T_2 \leq C)$.

3. **Generation of (L, R)**

   The censoring intervals were simulated by reproducing a follow-up study where a number of visits were scheduled for each individual and, in order to introduce randomness in the process, a probability of attending the scheduled visit was also considered. See Gomez *et al.* (2009) for a detailed description of the approach. In this simulation study we considered that visits were scheduled for each individual at times $k = 1, 2, \ldots, 200$. Denote by q the probability of a patient to attend the visit, and V_k, $k = 1, \ldots, 200$ the corresponding indicator variable for attendance to visit k, following a Bernoulli distribution of parameter q.

   Then, R is defined as the first visit after T_1, that is, the first visit where the intermediate event can be detected, and L is the previous visit. Formally,

$$L = \begin{cases} \max\{t_k | t_k < T_1, V_k = 1\} & \text{if } \delta_1 = 1 \\ Y & \text{if } \delta_1 = 0 \end{cases}$$

$$R = \begin{cases} \min(Y, \min\{t_k | t_k \geq T_1, V_k = 1\}) & \text{if } \delta_1 = 1 \\ \infty & \text{if } \delta_1 = 0 \end{cases}$$

   To obtain narrow intervals, a probability of $q = 0.225$ is chosen, and to get wide intervals, we need a probability of $q = 0.085$.

The code for the Exponential case is shown in Appendix D.4.8; the code for the Weibull case is similar and can be obtained upon request.

In order to visualize the sort of data that we are dealing in this simulation study, we provide in Figure 9.1 four of the simulated bivariate distribution (T_1, T_2) generated from a Clayton's copula model with Exponential and Weibull distributions and for $\alpha = 3$ and $\alpha = 5$. Fixed α, the Weibull setting presents more dispersion than its Exponential counterpart.

9.2 Evaluation criteria

9.2.1 Estimation of the association parameter $\boldsymbol{\alpha}$

As mentioned at the beginning of the chapter for each simulation setting, B data sets are generated, and for each of them, the association parameter is estimated through three different methods. The estimates of α are denoted by $\widehat{\alpha}_m^b$, when midpoint imputation and right-censoring semi-competing risks analysis is used; $\widehat{\alpha}_1^b$, corresponding to strategy 1 for ICSCR, and $\widehat{\alpha}_2^b$, corresponding to strategy 2 for ICSCR. Thus a vector of parameter estimates is obtained,

$$\widehat{\boldsymbol{\alpha}}^b = (\widehat{\alpha}_m^b, \widehat{\alpha}_1^b, \widehat{\alpha}_2^b)^t, \quad b = 1, .., B.$$

Let

$$\widehat{\boldsymbol{V}}^b = \left(\widehat{V}_m^b, \widehat{V}_1^b, \widehat{V}_2^b\right)^t,$$

denote the estimators for the variance of $\widehat{\boldsymbol{\alpha}}^b$, obtained as follows. For midpoint imputation, the variance is estimated according to the asymptotic behavior of $\widehat{\alpha}_m^b$, as explained in Chapter 4.

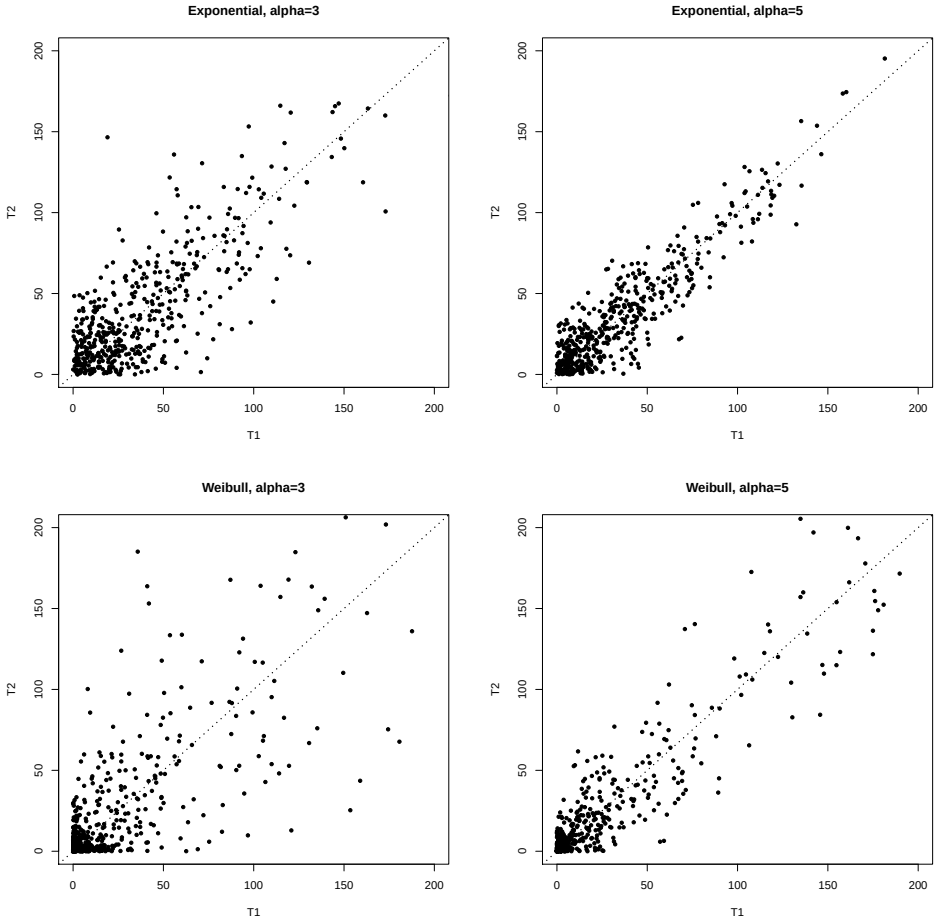


Figure 9.1: Bivariate data generated following Clayton's copula, with Exponential (top) and Weibull margins (bottom)

In the case of $\widehat{V}_1^b$ and $\widehat{V}_2^b$, though the asymptotic properties were also developed in Chapter 7, their computation is time-consuming and thus intractable in this simulation setting. We opted for obtaining the jackknife estimator of the variance, $V_k = \Sigma/n$ (Proposition 7,3),

$$\widehat{V}_k^b = \frac{n-1}{n} \sum_{i=1}^{n} \left(\widehat{\alpha}_k^{b(-i)} - \widehat{\alpha}_k^{b(\cdot)} \right)^2$$

where $\widehat{\alpha}_k^{b(-i)}$ is the estimator of α resulting from the data set obtained by removing the i^{th} individual, and

$$\widehat{\alpha}_k^{b(\cdot)} = \frac{\sum_{i=1}^{n} \widehat{\alpha}_k^{b(-i)}}{n} \qquad k = 1, 2.$$

Based on these estimations and given the true value of the association parameter (in vectorial form) $\boldsymbol{\alpha}_0 = (\alpha_0, \alpha_0, \alpha_0)^t$, known in every setting, we calculate the mean, the bias, the average variance, the mean square error (MSE) and the coverage probability:

$$\overline{\widehat{\boldsymbol{\alpha}}} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\boldsymbol{\alpha}}^b$$

$$\widehat{\text{Bias}}(\widehat{\boldsymbol{\alpha}}) = \overline{\widehat{\boldsymbol{\alpha}}} - \boldsymbol{\alpha}_0$$
$$\widehat{\text{AveVar}}(\widehat{\boldsymbol{\alpha}}) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\boldsymbol{V}}^b$$
$$\widehat{\text{MSE}}(\widehat{\boldsymbol{\alpha}}) = \widehat{\text{AveVar}}(\widehat{\boldsymbol{\alpha}}) + \widehat{\text{Bias}}(\widehat{\boldsymbol{\alpha}})^2$$
$$\text{Cover95} = \frac{1}{B}\sum I\Big(\alpha_0 \in \text{IC95}(\widehat{\boldsymbol{\alpha}}^b)\Big),$$

where $\text{IC95}(\widehat{\boldsymbol{\alpha}}^b)$ corresponds to the 95% confidence interval for $\widehat{\boldsymbol{\alpha}}^b$ ($z_{\alpha/2} = 1.96$),

$$\left(\widehat{\boldsymbol{\alpha}}^b - 1.96\sqrt{\widehat{\boldsymbol{V}}^b}, \widehat{\boldsymbol{\alpha}}^b + 1.96\sqrt{\widehat{\boldsymbol{V}}^b}\right).$$

9.2.2 Estimation of the marginal survival $S_1(\cdot)$

When it comes to the estimation of $S_1(\cdot)$, for each strategy of estimation of α, we obtain the corresponding estimations of the marginal. Therefore, for each setting and each data set generated, we ob tain,

$$\widehat{\boldsymbol{S}}_{\boldsymbol{1b}}(\boldsymbol{t}) = \Big(\widehat{S}^b_{1,m}(t), \widehat{S}^b_{1,1}(t), \widehat{S}^b_{1,2}(t)\Big)^t$$

for t in their corresponding support. For the purpose of comparison, we estimate $\widehat{\boldsymbol{S}}_{\boldsymbol{1b}}(\boldsymbol{t})$ at five percentiles of the theoretical distribution (10%,30%,50%,70%,90%). These theoretical percentiles z_p, for p in $\{0.9, 0.7, 0.5, 0.3, 0.1\}$ are computed according to the generating distribution for T_1, that is, for $T_1 \sim \text{Exp}(\lambda_1)$,

$$z_p = -\frac{\log(1-p)}{\lambda_1},$$

and for $T_1 \sim \text{Weibull}(\rho_1, \mu_1)$,

$$z_p = \left\{-\frac{\log(1-p)}{\mu_1}\right\}^{1/\rho_1}.$$

So we obtain $\widehat{\boldsymbol{S}}_{\boldsymbol{1b}}(\boldsymbol{z_p})$ for each data set, and calculate their mean, bias, average variance (from the jackknife estimators at each percentile) and mean square error.

9.3 Simulation results

In the following, we present the simulation results for the 24 scenarios considered. In Section 9.3.1 we discuss the relevant findings regarding estimation of the association parameter α, while Section 9.3.2 is devoted to the estimation of $S_1(t)$.

9.3.1 Results for $\boldsymbol{\alpha}$

9.3.1.1 Exponential marginal distributions

In this Section, we present the results of the 12 simulated settings corresponding to Exponential marginal distributions. We will compare the performance of the three available methodologies,

Table 9.2: Estimation of α: comparison of bias between ICSCR1, ICSCR2 and Midpoint (Exponential margins).

			$\alpha = 3$			$\alpha = 5$		
Width	**p†**	n	**ICSCR1**	**ICSCR2**	**Midpoint**	**ICSCR1**	**ICSCR2**	**Midpoint**
Narrow	25%	200	0.021	0.167	0.133	0.031	0.358	0.154
intervals‡		500	0.005	0.068	0.107	0.021	0.162	0.146
	50%	200	0.006	0.319	0.063	0.066	0.773	0.113
		500	0.004	0.164	0.071	0.011	0.356	0.061
	75%	200	0.061	0.624	0.109	0.112	1.319	0.123
		500	0.008	0.297	0.039	0.047	0.675	0.050
Wide	25%	200	−0.056	0.195	0.253	−0.156	0.449	0.130
intervals‡		500	−0.082	−0.032	0.220	−0.173	−0.022	0.112
	50%	200	−0.059	0.452	0.102	−0.058	1.200	−0.094
		500	−0.064	0.098	0.106	−0.142	0.261	−0.150
	75%	200	0.021	1.116	0.114	0.049	2.614	−0.092
		500	−0.039	0.365	0.042	−0.037	0.857	−0.151

†p: Percentage of dependent censoring, $100 \times P(T_1 > T_2)$.

‡Narrow intervals: average width 6 time units.

Wide intervals: average width 12 to 15 time units.

Strategy 1 (ICSCR1), Strategy 2 (ICSCR2) and midpoint imputation (Midpoint), for estimating the association parameter α in terms of bias, relative bias, mean square error and confidence interval coverage probability.

Bias and relative bias:

Table 9.2 provides the bias of the three approaches of the association parameter estimate. In terms of absolute bias, Strategy 1 outperforms the other two approaches in all the settings considered. Imputation methods are known to be biased (Sun, 2006), therefore Strategy 1 targets better the true value of the parameter than midpoint imputation. It also performs systematically better than Strategy 2, which, on the contrary, only provides better results than midpoint imputation in those scenarios with low dependent censoring (25%). As expected, the three strategies benefits of larger sample sizes, lower percentages of censoring and narrow intervals. Not so obvious is the fact that the three methodologies perform worst when $\alpha = 5$ which corresponds to a strong association between T_1 and T_2. We will discuss this later in Subsection 9.4.

Similar results and conclusions are obtained in terms of relative bias. The relative bias is defined by $\widehat{\text{Bias}}(\widehat{\alpha})/\alpha_0$, where α_0 is the true value for α. The relative bias permits to compare the bias of different parameter values on the same scale. These results are highlighted in Figure 9.2, where we plot the relative bias for the three strategies, distinguishing by different α and selecting those scenarios with $n = 500$. Again, Strategy 1 clearly outperforms midpoint imputation and Strategy 2 in most of the settings. Its behavior is robust in terms of percentage of dependent censoring and association parameter, and it is only seen to be affected for the length of the censoring intervals. With narrow intervals the observed relative bias of Strategy 1 is fairly non-existent, while for wider intervals Strategy 1 tends to underestimate a little the true association parameter. For low censoring (25%), Strategy 2 outperforms midpoint imputation but, as censoring increases, the performance

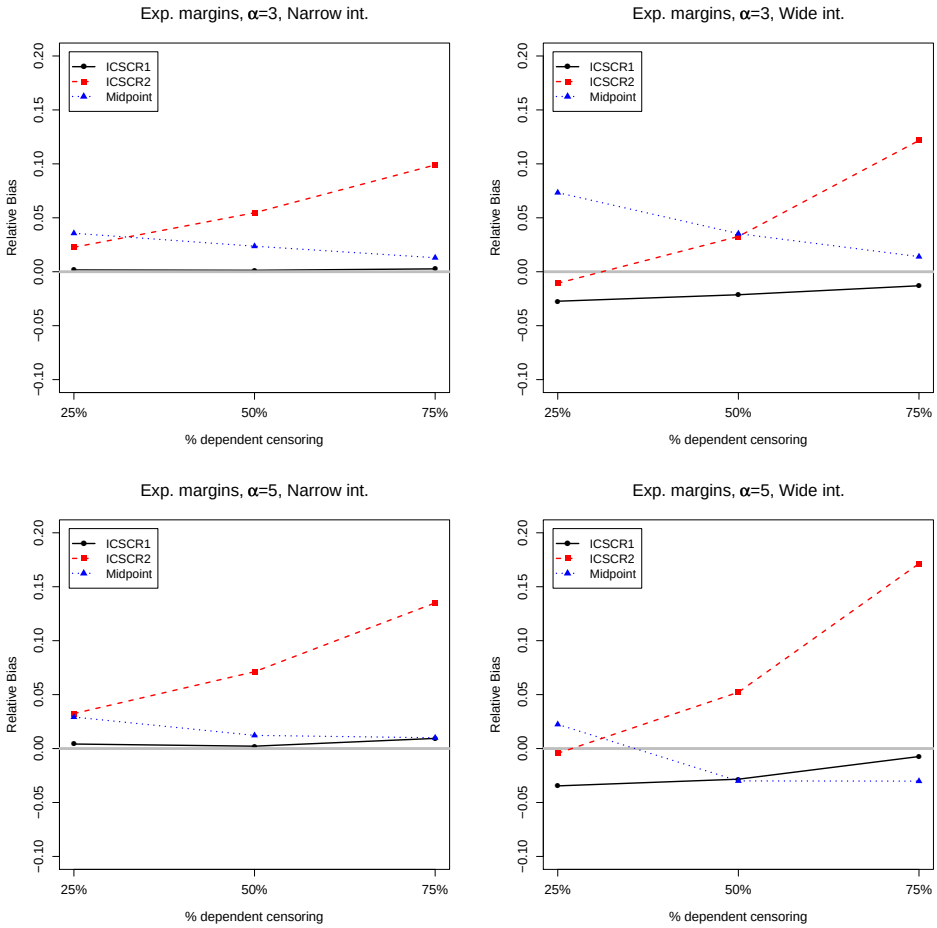


Figure 9.2: Estimation of α: comparison of relative bias between ICSCR1, ICSCR2 and Midpoint (Exponential margins, sample size $n = 500$)

of Strategy 2 get worse providing strongly biased results when the censoring percentage is 75% for both, narrow and wide intervals and both values of α. On the contrary, midpoint imputation seems to benefit of larger censoring percentages.

MSE and coverage probability:

The simulation results for Strategy 1 and Strategy 2 are summarized in Table 9.3, where we provide the mean, the bias, the mean square error (MSE) and the confidence interval coverage probability. Though we have already discussed the behaviour of the three approaches in terms of bias, we keep the mean and bias values in this table. However, now we focus on the variability of the estimates by evaluating the MSE and the coverage probability of the three strategies across the different simulation settings. The performance in terms of mean square error (MSE) can be visualized in Figure 9.3 for $n = 500$. Note that the scale of the plots for $\alpha = 3$ and for $\alpha = 5$ are different.

Midpoint imputation have not been included in the comparison because such a comparison would be unfair: imputation techniques ignore the variability contained in the interval of time and thus underestimate the real variability of the estimator.

As it was expected, the variability of both estimates is affected by (1) the sample size, with larger samples giving more precise estimates, (2) the percentage of censoring, where higher percentages results in more variability, and (3) the length of the censoring intervals, with wider intervals larger

MSE. As it was the case in terms of bias, the association parameter also affects variability, with more precise estimates for $\alpha = 3$ than for $\alpha = 5$. Strategy 1 always outperform Strategy 2 in terms of MSE but the advantage is small for low and moderate censoring percentages, specially with $n = 500$, where both strategies perform similarly. As already mentioned, both strategies give more variable results with heavy censoring (75%), but Strategy 2 is clearly more affected for this than Strategy 1.

The results for the coverage probabilities are less robust. Coverage probability is highly affected by the variability of results in some settings, which result in wider intervals. For instance, consider the results for $\alpha = 3$ and narrow intervals: the bias is almost inexistent for all 6 cases. Those settings with $n = 200$ have higher coverage probability than $n = 500$, because for equal bias, the $n = 200$ settings obtain wider CI intervals, increasing the chances to contain the true parameter. Moreover, within these settings with $n = 200$, the coverage probability increases as the dependent censoring

Table 9.3: ICSCR estimation of α for a model with Exponential marginals, $T_1 \sim \text{Exp}(\lambda_1)$, $T_2 \sim \text{Exp}(\lambda_2)$.

			ICSCR1				**ICSCR2**			
Int.Width	**p†**	***n***	**Mean**	**Bias**	**MSE**	**Cov95‡**	**Mean**	**Bias**	**MSE**	**Cov95‡**
					$\boldsymbol{\alpha = 3}$					
Narrow	25%	200	3.021	0.021	0.178	0.954	3.167	0.167	0.187	0.941
intervals		500	3.005	0.005	0.067	0.957	3.068	0.068	0.063	0.950
	50%	200	3.006	0.006	0.261	0.960	3.319	0.319	0.366	0.950
		500	3.004	0.004	0.097	0.940	3.164	0.164	0.118	0.951
	75%	200	3.061	0.061	0.522	0.963	3.624	0.624	1.014	0.956
		500	3.008	0.008	0.174	0.947	3.297	0.297	0.265	0.952
Wide	25%	200	2.944	−0.056	0.199	0.928	3.195	0.195	0.245	0.941
intervals		500	2.918	−0.082	0.079	0.957	2.968	−0.032	0.079	0.946
	50%	200	2.941	−0.059	0.331	0.972	3.452	0.452	0.635	0.966
		500	2.936	−0.064	0.122	0.942	3.098	0.098	0.151	0.985
	75%	200	3.021	0.021	0.790	0.969	4.116	1.116	2.698	0.970
		500	2.961	−0.039	0.245	0.933	3.365	0.365	0.459	0.988
					$\boldsymbol{\alpha = 5}$					
Narrow	25%	200	5.031	0.031	0.546	0.973	5.358	0.358	0.630	0.956
intervals		500	5.021	0.021	0.205	0.943	5.162	0.162	0.207	0.936
	50%	200	5.066	0.066	0.822	0.958	5.773	0.773	1.512	0.939
		500	5.011	0.011	0.292	0.949	5.356	0.356	0.417	0.946
	75%	200	5.112	0.112	1.599	0.969	6.319	1.319	3.957	0.960
		500	5.047	0.047	0.534	0.933	5.675	0.675	1.053	0.924
Wide	25%	200	4.844	−0.156	0.658	0.952	5.449	0.449	0.936	0.953
intervals		500	4.827	−0.173	0.261	0.925	4.978	−0.022	0.270	0.949
	50%	200	4.942	−0.058	1.137	0.964	6.200	1.200	3.317	0.946
		500	4.858	−0.142	0.403	0.912	5.261	0.261	0.601	0.980
	75%	200	5.049	0.049	2.823	0.978	7.614	2.614	14.641	0.978
		500	4.963	−0.037	0.830	0.895	5.857	0.857	2.082	0.975

†p: Percentage of dependent censoring, $100 \times P(T_1 > T_2)$.

‡Cov95: coverage probability, $P(\alpha \in CI)$.

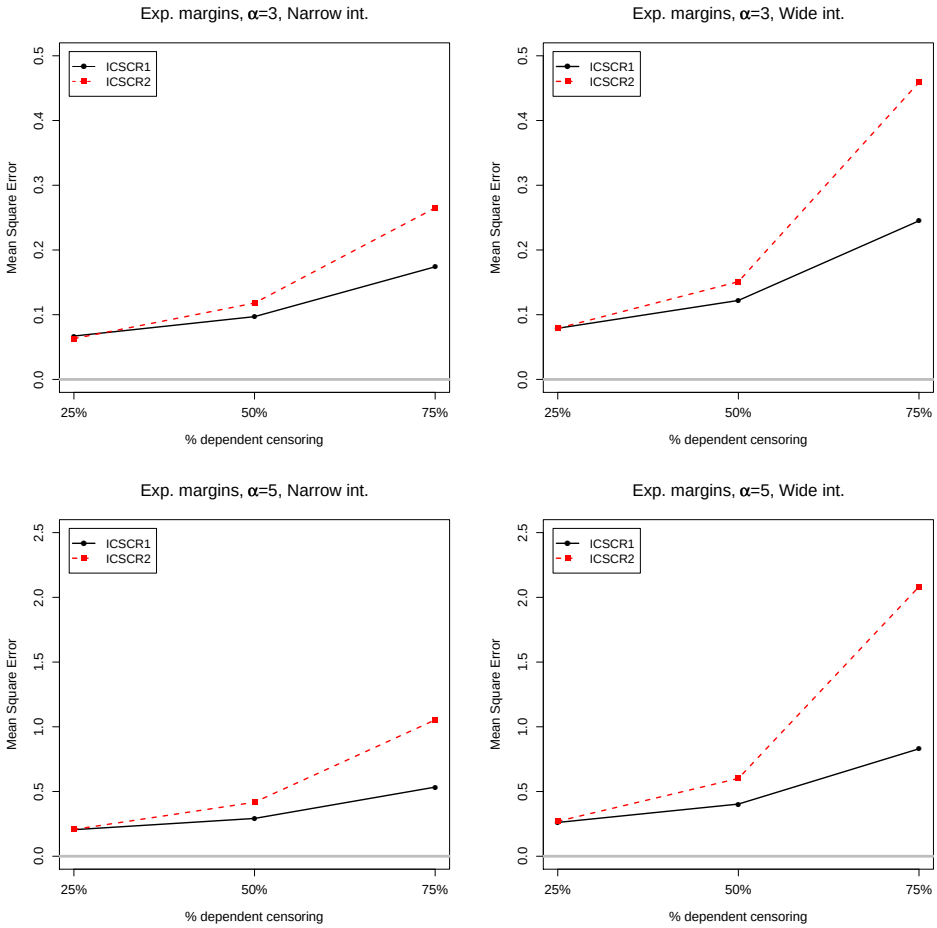


Figure 9.3: Mean Square Error for the ICSCR estimation of α (sample size $n = 500$, Exponential marginals.)

increases, while an inverse relationship is observed in $n = 500$ scenarios.

In general, in those settings for which results have low precision, the coverage probability is over-estimated. This phenomena is also observed for Strategy 2, where several settings (for 50% and 75%) obtain larger coverage probabilities than Strategy 1, which outperforms the others in terms of bias and variability.

9.3.1.2 Weibull Margins

Now we turn to the discussion and comparison of the three methods, Strategy 1 (ICSCR1), Strategy 2 (ICSCR2) and midpoint imputation (Midpoint), to estimate the association in the 12 settings generated with Weibull marginals.

Bias and Relative bias

Table 9.4 contains the absolute bias for the estimation of the association parameter for the three strategies. Strategy 1 clearly outperforms Strategy 2 and midpoint imputation. Following a similar trend as for exponential margins, the three strategies benefit again of larger sample sizes, narrow intervals and lower percentages of censoring. The three strategies, in addition, obtain higher bias for $\alpha = 5$. What most strikes on these results, however, is that the bias obtained is, in magnitude,

Table 9.4: Estimation of α: comparison of bias between ICSCR1, ICSCR2 and Midpoint (Weibull margins).

			$\alpha = 3$			$\alpha = 5$		
Width	**p†**	n	**ICSCR1**	**ICSCR2**	**Midpoint**	**ICSCR1**	**ICSCR2**	**Midpoint**
Narrow	25%	200	0.182	0.955	0.798	0.245	1.617	1.174
intervals‡		500	0.160	0.839	0.748	0.246	1.412	1.179
	50%	200	0.147	1.133	0.636	0.325	2.401	1.126
		500	0.139	0.917	0.631	0.268	1.794	1.054
	75%	200	0.231	1.785	0.725	0.476	4.222	1.347
		500	0.156	1.211	0.614	0.371	2.832	1.225
Wide	25%	200	0.142	1.404	1.414	−0.031	2.419	1.969
intervals‡		500	0.113	1.043	1.328	−0.052	1.641	1.938
	50%	200	0.164	1.771	1.102	0.245	4.004	1.587
		500	0.129	1.165	1.083	0.167	2.325	1.511
	75%	200	0.277	3.210	1.147	0.617	8.752	1.749
		500	0.198	1.765	1.028	0.502	4.535	1.649

†p: Percentage of dependent censoring, $100 \times P(T_1 > T_2)$.

‡Narrow intervals: average width 6 time units.

Wide intervals: average width 12 to 15 time units.

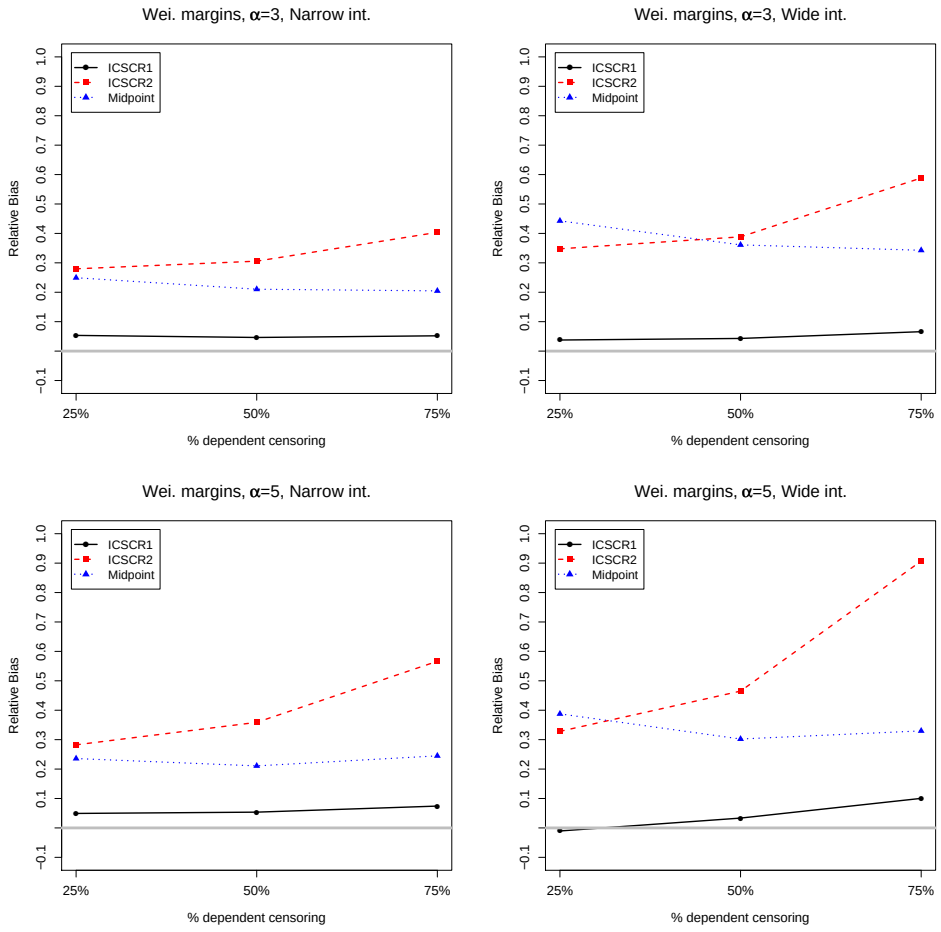


Figure 9.4: Estimation of α: comparison of relative bias between ICSCR1, ICSCR2 and Midpoint (Weibull margins, sample size $n = 500$)

Table 9.5: ICSCR estimation of α for a model with Weibull marginals, $T_1 \sim \text{Weibull}(\mu_1, \rho_1)$, $T_2 \sim \text{Weibull}(\mu_2, \rho_2)$.

			ICSCR1				**ICSCR2**			
Int.Width	**p†**	n	**Mean**	**Bias**	**MSE**	**Cov95‡**	**Mean**	**Bias**	**MSE**	**Cov95‡**
					$\boldsymbol{\alpha = 3}$					
Narrow	25%	200	3.182	0.182	0.263	0.655	3.955	0.955	1.176	0.570
intervals		500	3.160	0.160	0.110	0.938	3.839	0.839	0.800	0.206
	50%	200	3.147	0.147	0.338	0.838	4.133	1.133	1.803	0.749
		500	3.139	0.139	0.136	0.938	3.917	0.917	1.020	0.391
	75%	200	3.231	0.231	0.813	0.934	4.785	1.785	4.847	0.871
		500	3.156	0.156	0.264	0.963	4.211	1.211	1.904	0.614
Wide	25%	200	3.142	0.142	0.300	0.750	4.404	1.404	2.384	0.426
intervals		500	3.113	0.113	0.112	0.981	4.043	1.043	1.243	0.205
	50%	200	3.164	0.164	0.462	0.907	4.771	1.771	4.260	0.692
		500	3.129	0.129	0.168	0.960	4.165	1.165	1.714	0.523
	75%	200	3.277	0.277	1.470	0.966	6.210	3.210	16.957	0.895
		500	3.198	0.198	0.415	0.956	4.765	1.765	4.250	0.747
					$\boldsymbol{\alpha = 5}$					
Narrow	25%	200	5.245	0.245	0.753	0.936	6.617	1.617	3.489	0.638
intervals		500	5.246	0.246	0.315	0.939	6.412	1.412	2.316	0.271
	50%	200	5.325	0.325	1.162	0.967	7.401	2.401	7.840	0.692
		500	5.268	0.268	0.439	0.937	6.794	1.794	3.882	0.377
	75%	200	5.476	0.476	2.933	0.989	9.222	4.222	26.190	0.821
		500	5.371	0.371	0.976	0.934	7.832	2.832	10.047	0.478
Wide	25%	200	4.969	−0.031	0.836	0.923	7.419	2.419	7.341	0.517
intervals		500	4.948	−0.052	0.290	0.980	6.641	1.641	3.250	0.427
	50%	200	5.245	0.245	1.647	0.976	9.004	4.004	21.614	0.646
		500	5.167	0.167	0.544	0.946	7.325	2.325	7.007	0.596
	75%	200	5.617	0.617	7.087	0.992	13.752	8.752	145.794	0.857
		500	5.502	0.502	1.839	0.874	9.535	4.535	28.017	0.720

†p: Percentage of dependent censoring, $100 \times P(T_1 > T_2)$.

‡Cov95: coverage probability, $P(\alpha \in CI)$.

greater than the bias obtained when marginal Exponentials where used to generate the data. This fact may be reflecting the greater dispersion of the Weibull setting (recall Figure 9.1).

In Figure 9.4, we compare the three strategies in terms of relative bias for $n = 500$. The performance of Strategy 2 and Midpoint imputation is not comparable, in any setting, to Strategy 1 even when dependent censoring is low (in which case, Strategy 2 outperforms midpoint imputation).

MSE and coverage probability

The simulation results for Strategy 1 and Strategy 2 are summarized in Table 9.5, where we provide the mean, the bias, the mean square error (MSE) and the confidence interval coverage probability. We discuss the variability of the estimates by studying the MSE and the coverage probability of the ICSCR strategies (midpoint imputation is not comparable in terms of variability). The performance

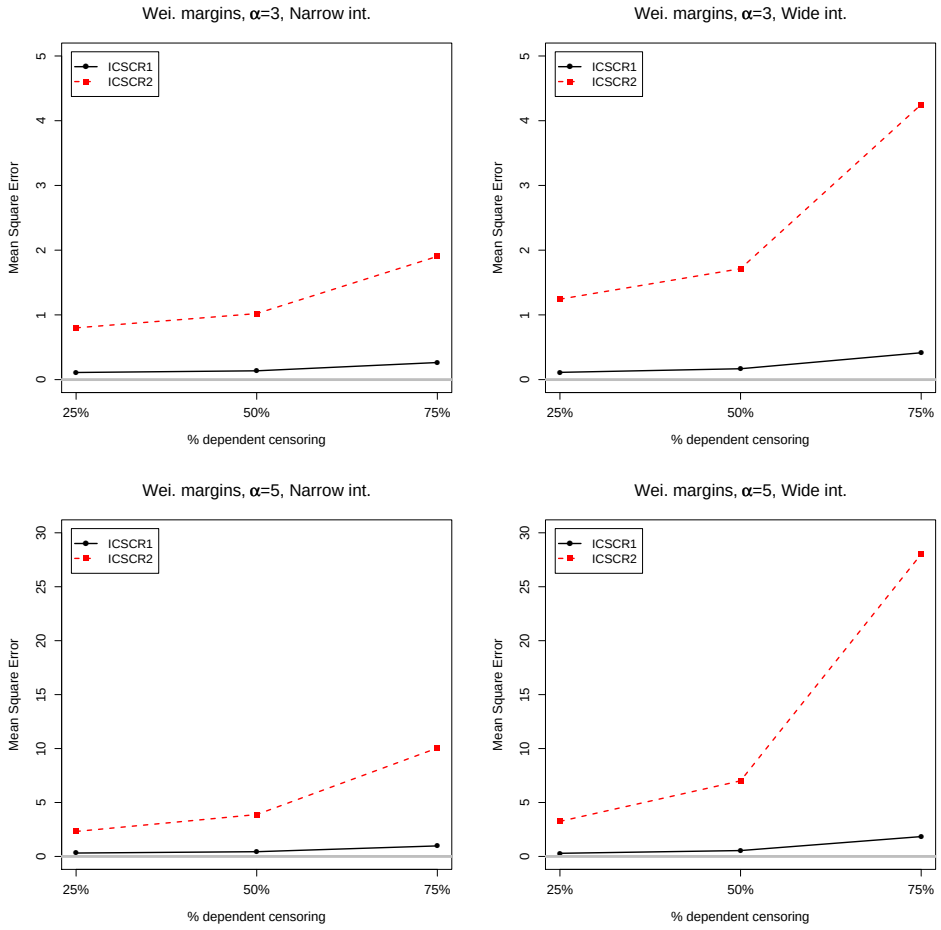


Figure 9.5: Mean Square Error for the ICSCR estimation of α (sample size $n = 500$, Weibull marginals.)

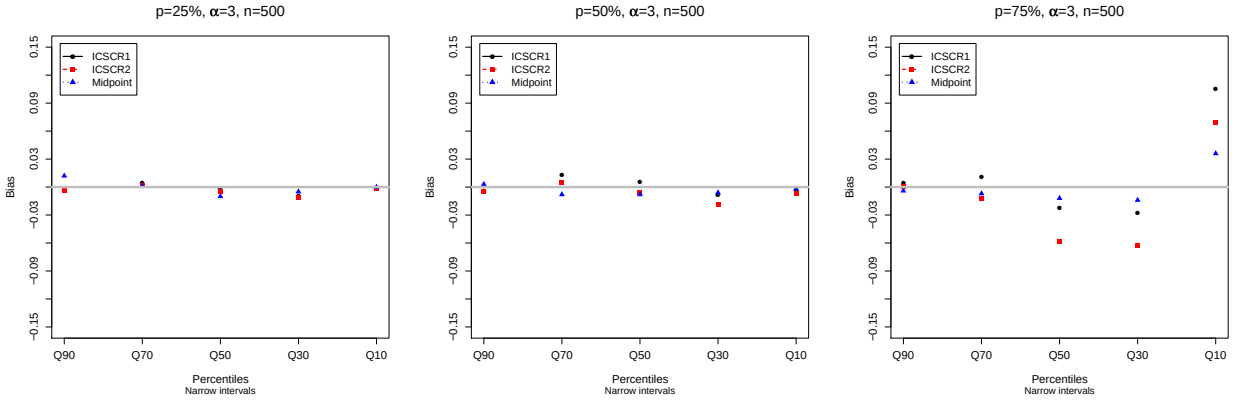
of the MSE can be visualized in Figure 9.5 for $n = 500$. Note that the scale of the plots for $\alpha = 3$ and for $\alpha = 5$ are different.

Again, both strategies exhibit an increase of variability along with (1) smaller samples, (2) larger dependent censoring settings, (3) wider censoring intervals, and (4) the magnitude of α. In terms of MSE, Strategy 1 outperforms Strategy in any setting.

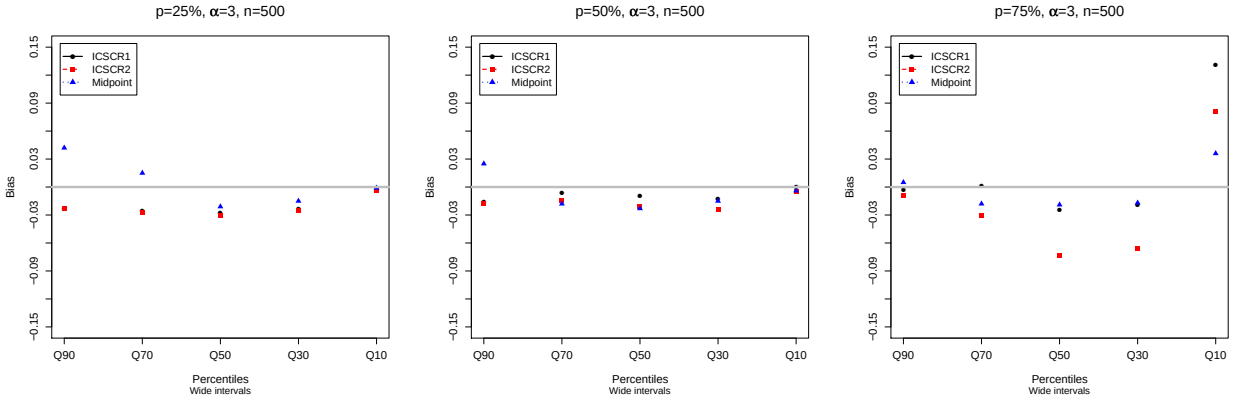
It is difficult to make conclusive statements in terms of coverage probabilities, though, because they are highly affected by the large amount of variability. For Strategy 1, and $\alpha = 3$, higher coverage probabilities are obtained for larger sample sizes, as it would be expected. However, within smaller sample sizes ($n = 200$), the coverage probability increases with dependent censoring. On the other hand, Strategy 1 outperforms in general Strategy 2 in terms of coverage probability, but results from the latter are seldom comparable due to the severe bias in the estimation of α.

9.3.2 Results for $S_1(\cdot)$

In order to explore the performance of the three methodologies, Strategy 1 (ICSCR1), Strategy 2 (ICSCR2) and midpoint imputation (Midpoint), for estimating the marginal distribution of the intermediate event, we obtain the pointwise estimate of the survival function at the 90%, 70%,



(a) Narrow Intervals, $\alpha = 3$



(b) Wide Intervals, $\alpha = 3$

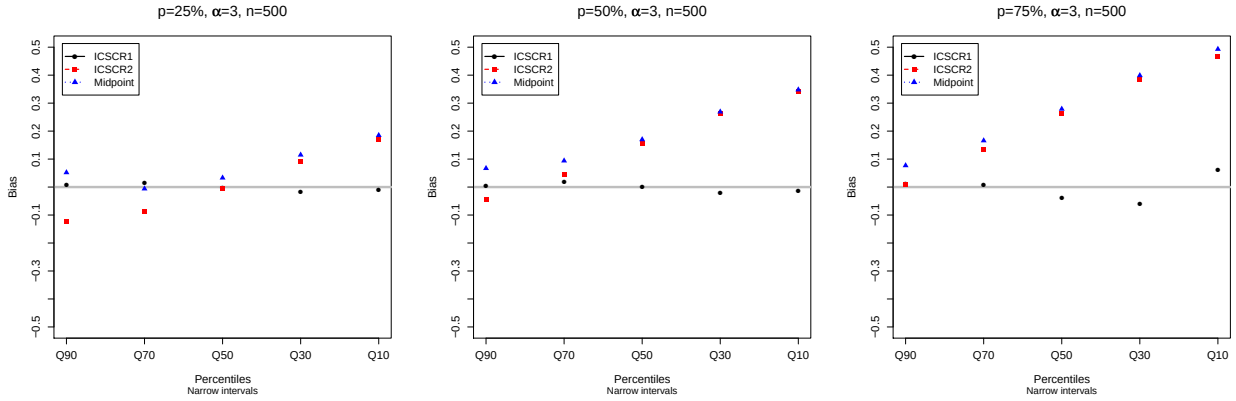
Figure 9.6: Bias of $S_1(t)$ estimates: comparison of ICSC1, ICSCR2 and Midpoint ($\alpha = 3$, Exponential margins)

50%, 30% and 10% percentiles of the theoretical marginal distribution of T1, that we denote by Q90, Q70, Q50, Q30 and Q10.

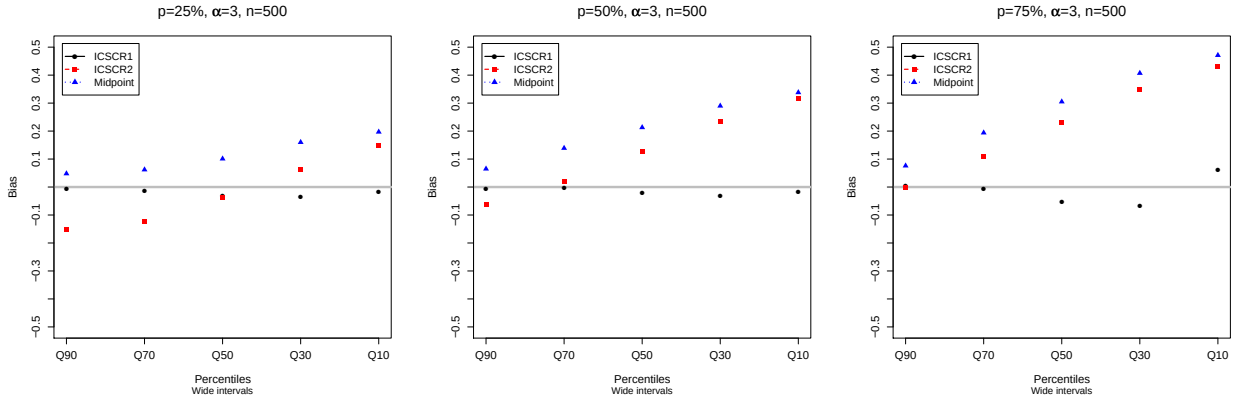
In Figure 9.6 we provide the biases of the survival estimates for Exponential marginal distributions and association parameter $\alpha = 3$. In Figure 9.7 we plot the corresponding biases for the case of Weibull marginal distributions and $\alpha = 3$. The results for $\alpha = 5$ can be found in Appendix C, Section C.1, Figures C.1 and C.2 for exponential and Weibull distributions, respectively. Tables summarizing the mean, bias and MSE for all the percentiles can also be found in Appendix C.1.

Naturally, the accuracy of the marginal distribution estimate is directly related to the accuracy of the association parameter estimate. The results for Exponential marginal distributions (Figure 9.6) are much more accurate than for Weibull distribution (Figure 9.7). Note that different scales are considered in both Figures.

More specifically, for Exponential marginal distributions we observe (Figure 9.6) that the three approaches provide very good results for all percentiles of the marginal distributions when the censoring percentage (p) is low or moderate (p=25% or 50%) and for both, narrow and wide intervals: the bias in absolute terms is less than 0.03. For heavy censoring (p=75%), Strategy 1 and Midpoint imputation still provide very accurate results except in the tail of the distribution



(a) Narrow Intervals, $\alpha = 3$



(b) Wide Intervals, $\alpha = 3$

Figure 9.7: Bias of $S_1(t)$ estimates: comparison of ICSC1, ICSCR2 and Midpoint ($\alpha = 3$, Weibull margins)

where less information is available. In this setting (p=75%), Strategy 2 estimates for α were a bit biased (bias around 0.3) and, consequently, the estimates of the marginal distribution are also a bit biased (absolute bias at Q50 and Q30 around 0.06).

For Weibull marginal distributions (Figure 9.7) we observe how Strategy 1 systematically gives less biased estimates than Strategy 2 and midpoint imputation for all the percentiles. Again, this is related with the accuracy of the three approaches for estimating the association parameters with Weibull marginal distributions. In this setting, both Midpoint imputation and Strategy 2 provide strongly biased estimates for α and, consequently, the estimates of the marginal distribution of the intermediate event is strongly biased. In this setting, only Strategy 1 provides reliable results.

9.4 Discussion

In this chapter we have explored the performance of three methodologies for interval-censored semi-competing risks data: midpoint imputation and the two new approaches proposed in this thesis, Strategy 1 and Strategy 2, described in chapter 6. The simulation scenarios distinguish between Exponential and Weibull distributions, different sample size, dependent censoring, width of intervals and levels of association.

Strategy 1 outperforms both Midpoint imputation and Strategy 2 in most simulated settings. It provides accurate estimates for both, the association parameter α and for the marginal distribution of the intermediate event. In terms of bias the estimates are robust, hardly affected by the percentage of dependent censoring or the length of the censoring intervals. The two factors affect the variability of the estimates. Strategy 1 performs similarly well under Exponential marginal distributions and Weibull marginal distributions.

Contrary to what we had expected, Strategy 2 is not performing accurately in many settings. It performs well for low dependent censoring (25%). For wide intervals and Exponential margins, it performs slightly better than Strategy 1 in terms of bias and identically in terms of MSE. However, as the censoring increases, the bias and variability increases. In the case of Weibull margins, a situation with higher dispersion in the data, some bias is observed even with low dependent censoring. The reason for the bad behavior of Strategy 2 and thus, its weakness, is that this strategy requires estimation of $G_1(l, r|y)$, the joint distribution function of the interval censoring variables (L, R) given $T_2 = y$, and $G_2(l, r|y)$, the joint distribution function of the interval censoring variables (L, R) given $T_2 > y$. We have proposed a strategy to empirically estimate G_1 and G_2 (Subsection 6.5.1) but accurate estimates of these two functions would require large sample sizes. In situations with considerable dependent censoring and wide intervals the available sample size for estimating G_1 and G_2 is insufficient. Future work is to explore alternative strategies to estimate, perhaps parametrically, G_1 and G_2 more accurately.

A similar reason is behind the less accurate results when $\alpha = 5$. This parameter represents a strong association (under Clayton's copula model in all the plane, Kendall's tau of $\tau_{\mathcal{K}} = 0.67$ or Spearman's $\rho_S = 0.81$). Therefore, bivariate points (T_1, T_2) lie close to the diagonal of the plane $T_1 \times T_2$, forcing the right endpoints of the censored interval, R, to be close, if not equal, to Y, the observed value of T_2. Again, less 'pure' intervals (with $R < Y$) are available, and functions G_1 and G_2 are not enough accurately estimated.

CHAPTER 10

Software contributions

In this Chapter, we enumerate the software functions that have been implemented in the statistical software R (R Development Core Team, 2009) to deal with the methods described in this thesis. Firstly, we explain the code needed to obtain a nomogram and a calibration plot in the context of competing risks (Section 10.1). Then we will present the code developed to obtain the predictive process of the multi-state model for progression of bladder cancer in Section 10.2. Section 10.3 contains the programmed functions to perform the semi-competing risks analysis for right-censored data proposed by Fine *et al.* (2001). Finally, in Section 10.4, we explain how to implement the analysis for interval-censored semi-competing risks data and we describe the main functions to run the analysis.

10.1 Competing risks

10.1.1 A nomogram for competing risks

To build a nomogram for competing risks in `R`, we have to first load the packages `Hmisc` (Frank E Harrell Jr and with contributions from many other users., 2009) and `Design` (Harrell Jr., Frank E., 2009), in order to use their specific functions to build a nomogram (see Harrell *et al.* (1996) for further details). We will take advantage of these functions and their already defined data structures to build our competing risks nomogram.

We will follow the steps to build the nomogram in Figure 2.8, in the context of a competing risks model for PFS. Let `Tev2` and `Cev2` two vectors containing, respectively, the time to the first observed event between PFS and DOC, and the cause of failure (`Cev2`=1 for PFS, `Cev2`=2 for DOC or `Cev2`=0 for a censored observation). We first obtain Fine and Gray's model for the time to progression using the `crr` function of the `cmprsk` package (see D.1). To include covariates on

this function, these need to be coded into dummy variables and stored into a matrix structure. Let `cov1` be such a matrix for covariates considered in the model: gender, age, multiplicity, size of the tumour, stage and grade and smoking status.

```
  fine1<-crr(Tev2,Cev2,cov1=cov1,failcode=1)
```

Now we extract from the model the necessary information to construct the nomogram: the coefficients of the model, the times at which the baseline sub-hazard is estimated and the size of the jumps of this estimation:

```
beta<-fine1$coef
  time<-c(0,fine1$uftime,max(Tev2[is.na(Tev2)==F])) #failure times
  jump<-fine1$bfitj #jumps (baseline sub-hazards)
```

Finally, we compute the cumulative subhazard function, the corresponding subsurvival and the linear predictors of each individual of the data set:

```
  bascum<-c(0,cumsum(jump),cumsum(jump)[length(cumsum(jump))])
  bassurv<-exp(-bascum)
  lin.pred<-cov1%*%beta
```

Now we need the structure necessary to apply function `nomogram`: it needs a `cph` object, obtained by applying the `cph` function, a `Design` variation of the original function `coxph` included in the `survival` package:

```
#covariates to make nicer axis in nomogram --> not dummies
  f.sex<-factor(female,label=c('Male','Female'))
  f.age<-factor(agec,label=c('<=60y','61-70y','>70y'))
  f.mult<-factor(mult,label=c('Single','Multiple'))
  f.size<-factor(size3cm,label=c('<3cm','>=3cm'))
  f.stage<-factor(T1Tis,label=c('Ta','T1Tis'))
  f.grade<-factor(grade,label=c('Benign/G1','G2','G3'))
  f.smok<-factor(smoking,label=c('Non-smoker','Smoker'))

  #structure of cph object to fill-in with FGH estimates
  ddist<-datadist(f.sex,f.age,f.mult,f.size,f.stage,f.grade,f.smok)
  cox1<-cph(Surv(Tev2,Cev2==1)~f.sex+f.age+f.mult+f.size+f.stage+f.grade+
  f.smok,method="breslow",surv=T)
  fine2<-Newlabels(cox1,c(f.sex="Gender",f.age="Age(year)",f.mult="Number",
  f.size="Size",f.stage="Stage", f.grade="Grade",f.smok="Smoking status"))
```

Now we replace the coefficients, linear predictors and predictions of the baseline survival function of the `cph` model by the values of Fine and Gray's model:

```
fine2$coefficients<-beta
  fine2$linear.predictors<-lin.pred
  fine2$surv<-bassurv
  fine2$center<-0
```

Finally, we can construct a nomogram that takes into account the presence of competing risks:

```
#construction of nomogram
  ddist <-datadist(f.sex,f.age,f.mult,f.size,f.stage,f.grade,f.smok)
  options(datadist='ddist')
  surv2<-Survival(fine2)
  FP.60<-function(lp) 1-surv2(60,lp)
  at.surv<-c(.01,.05,seq(.1,.9,by=.1),.95,.98,.99,.999)
  nom<-nomogram(fine2,lp=F,conf.int=F,fun=list(FP.60),funlabel=
     c("Probability of PFS before 5 years"),fun.at=list(at.surv),
     intercept=0, force.label=T)
  title(main="Nomogram for Progression Free Survival ",cex.main=1.4)
```

10.1.2 A calibration plot for competing risks

Function `getCalibrateCIF` (the code is given in Appendix D.1.3) implements the calibration plot for predictions of the probability of an event when competing risks are present. In particular, Fine and Gray's model, estimated by the `crr` function, is included. The function also provides the calibration plot for the corresponding Cox model obtained ignoring competing risks (Section 2.3.4, option (a)).

First, we need to estimate the models to be compared. Following the previous example on progression free survival:

```
cox1<-cph(Surv(Tev2,Cev2==1)~female+age+mult+size3cm+T1Tis+G2+G3+smoking,
surv=T,x=T,y=T,data=TOTAL1)
fine1<-crr(Tev2,Cev2,cov1=cov1,failcode=1).
```

Observe the parameters required in the calling of the `cph` function (`surv=T,x=T,y=T`). `TOTAL1` is the data frame of size n containing the variables of interest, and `covar` is the matrix containing the dummy variables representing the covariates in the model.

To obtain hence the calibration plot for the predicted probability of progression at time `u=60` months, which has been modelled by `cox1` and `fine1` (Figure 2.9(b)) the following call is made:

```
getCalibrateCIF(cox.obj=cox1,fine.obj=fine1,g=5,Srv=Srv,dades=TOTAL,
covar=cov1,which.cause=1,u=60, B=100,pl=TRUE,conf.int=0.95,unit='Month').
```

For each model (say generically `M`), the function's program proceeds as follow:

1) Obtain predictions from model `M` at time `u` for all individuals in `dades` using function `getEstimates` (see Appendix D.1.3.1).

2) Compute the apparent calibration:

   - Divide the sample in `dades` into `g` groups according to the predicted values in point 1) and compute the empirical cumulative incidence function (`Obs`) and the mean predicted values within each group (`Pred`). This can be achieved using function `groupCIF` (see Appendix D.1.3.2).
   - Compute the differences `D=Obs-Pred`.

3) Internal validation of the calibration curves (Boostrap): take `B` samples with replacement form the original data `dades`.

   - For each bootstrap sample `dades.b`, fit model `M`. Coefficients for the covariates are hence re-estimated in new model `M.b`.
   - Execute 2) with the bootstrap sample `dades.b` and model `M.b` to obtain `Obs.b`, `Pred.b` and the differences `D.b`.
   - Execute 2) with the original sample `dades` and model `M.b` to obtain the differences `D.b`*.
   - Estimate the possible overoptimistic bias, `bias.b=D.b`*`-D.b`.

4) Obtain bootstrap confidence intervals at a significance level `conf.int` from the percentiles of `Obs.b`, and the estimated bias from the average of `bias.b`, `Mbias` .

5) Plot the apparent calibration points (`Pred` vs `Obs`) with their corresponding confidence intervals.

6) Plot the calibration points corrected by overoptimistic bias (`Pred` vs `Obs+MBias`).

Some parameters of the getCalibrateCIF function have not been mentioned above: `Srv` and `which.cause`. `Srv` is a matrix containing the vector of times and the vector of causes:

```
Srv<-cbind(Tev2,Cev2).
```

On the other hand, `which.cause` specifies the code of the cause of interest in the vector of causes in order to properly fit the successive bootstrap models `M.b`. Other parameters can be added to modify the plots.

10.2 Multi-state models

10.2.1 The predictive process

We have implemented functions `Pi1cr.0` and `Pi1cr.1` (see the code in Appendix D.2.1) corresponding to the predictive processes defined in expressions (3.11) and (3.12). These processes are defined as the predicted probability if progression at time u given the history of past events at the moment t. In order to compute these probabilities, Models 1 to 4 detailed in equations (3.7) to (3.8) must have been fitted, and estimates of the cumulative hazards for all transitions in the desired values of the covariates and in a specified vector of times have to be available (see pages 59 and 60).

For instance, let `new1` be a data.frame containing the different risk profiles in the bladder cancer data:

```
new<-expand.grid(female=0:1,age=c(60,70),mult=0:1,size3cm=0:1,T1Tis=0:1,
  G2=0:1,G3=0:1,smoking=0:1)
new1<-new[new$G2+new$G3<2,].
```

Let `coxP` be the `coxph` object corresponding to Model 2 (3.7). Then, the cumulative hazard function of the transition from Recurrence to Progression, assessed at the distinct risk profiles can be obtained by:

```
new2<-data.frame(new1,Rec=1)
haRP=survfit(coxP,type='br',newdata=new2)
timeRP=c(0,haRP$time,Inf)
HRP0=rbind(rep(0,ncol(haRP$surv)),-log(haRP$surv),
                          -log(haRP$surv[nrow(haRP$surv),]))
```

The cumulative hazard function of the transition from Diagnose to Progression is obtained by:

```
new3<-data.frame(new1,Rec=0)
haP=survfit(coxP,type="breslow",newdata=new3)
timeP=c(0,haP$time,Inf)
HP0=rbind(rep(0,ncol(haP$surv)),-log(haP$surv),
                          -log(haP$surv[nrow(haP$surv),]))
```

The rest of transitions are obtained similarly. Then, the predictive process can be estimated calling functions `Pi1cr.0` and `Pi1cr.1`. For instance, to obtain the residual cumulative incidence of progression after 24 months from diagnosis (Figure 3.4) we execute the following:

```
t<-24
u<-seq(t,96,1)
PPT2<-matrix(nrow=nrow(new1),ncol=length(u))
PPT1<-matrix(nrow=nrow(new1),ncol=length(u))
for (j in 1:length(u))
{
  for(i in 1:nrow(new1))
  {
      PPT1[i,j]<-Pi1cr.1(t,u[j],time,H2[,i],HRP[,i])
      PPT0[i,j]<-Pi1cr.0(t,u[j],time,H1[,i],HP[,i],HR[,i],H2[,i],HRP[,i])
  }
}
```

10.3 Semi-competing risks for right-censored data

We have implemented in `R` the proposal of Fine *et al.* (2001) to deal with semi-competing risks data. The main functions are `corSCR` to estimate the association parameter α of Clayton's copula model, and `margSCR` to obtain the plug-in estimator of the marginal survival function $S_1(t)$. The code for these procedures together with some internal functions for their execution can be found in Appendices D.3.1 and D.3.2 respectively. In the following we explain their use to obtain a semi-competing risks analysis.

10.3.1 Data preparation

In a situation of semi-competing risks data we need to define in advance the role of the variables involved. For instance, consider the bladder cancer situation where progression or death prevents the observation of recurrences (Section 8.1). Let `Tev1` be the minimum between the time to the first event or the censoring time. Let `C1` be the indicator of a recurrence, that is, `C1=1` if the first event is a recurrence, 0 otherwise. Let `Tev2` the minimum between the time to progression or death and the censoring time, and `C2` its correspondent indicator[1]. Define `Cz` to the indicator of the first event observed, no matter if a recurrence, progression or death. We define the semi-competing risks data structure as

```
dSCR<-list(X=Tev1,d1=C1,Y=Tev2,d2=C2,dz=Cz)
```

Functions `corSCR` and `margSCR` will search for the elements `X`,`Y`, `d1`,`d2` and `dz` of such a list.

10.3.2 Estimation of α and $S_1(t)$

Estimates of α are hence obtained by the following call:

```
alp<-corSCR(dSCR,-1,-1)

Returning: >alp
  $con.index [1] 1.098853
  $se [1] 0.1178480
  $a [1] -1
  $b [1] -1
  $I [1] 0.0378401.
```

The function returns a list with the estimate (`alp$con.index`), its standard error (`alp$se`), the parameters a and b employed for the weight function $W_{a,b}$ (equation (4.14)), and a parameter used in the computation of the standard error (`alp$I`) which may be needed in the computation of the standard errors for the marginals. When a=b=-1, the weight function corresponds to $W_{\infty,\infty}$.

Estimates for the marginal survival $S_1(t)$ are given in the `$S1.e` element at the times `$tim` of the list provided by the call

```
hS1<-marg.SCR(dSCR,alp).
```

10.4 Interval-censored semi-competing risks data

The estimation algorithm in Section 6.6 is implemented in function `algICSCR` (see appendix D.4.1) with the following calling:

```
algICSCR<-function(dSCRi,e.alp.m, sh1.0,sc1.0,p0=TRUE,p1=TRUE,survC=NULL,
      datLR=NULL,strat=NULL, breaks=NULL,time=time){..}.
```

[1] In this case, `Tev1` and `Tev2` would correspond to midpoint imputation values, since we are dealing with right-censored data.

The parameters to call this function are the following:

- `dSCRi`: An ICSCR data object.
- `e.alp.m, sh1.0, sc1.0`: Initial values for α and $S_1(t)$.
- `p0`: Logical value to indicate whether estimates for Strategy 1 are returned.
- `p1`: Logical value to indicate whether estimates for Strategy 2 are returned.
- `survC`: Suvfit object containing the survival function of the censoring time C. Only required for Strategy 2.
- `datLR, strat, breaks`: Data to estimate the distribution of $(L, R)|Y$. Only required for strategy 2.
- `time`: Vector with time points where $S_1(t)$ will be estimated.

In the following section, we expose how to initialize the parameters for the algorithm. In Section 10.4.2 we discuss the functioning of `algICSCR`.

10.4.1 Initialize parameters

Data preparation

We need to pre-specify the role of the variables involved in the ICSCR situation. For instance, consider again the bladder cancer situation where progression or death prevents the observation of recurrences (Section 8.1). Let `L1` and `R1` be the variables containing the limits of the censoring intervals. Let `C1` be the indicator of a recurrence, that is, `C1=1` if the first event is a recurrence (in such case, `R1`$< \infty$), 0 otherwise. Let `Tev2` the minimum between the time to progression or death and the censoring time, and `C2` its correspondent indicator. We define the interval-censored semi-competing risks data structure as

```
dSCR<-list(L=L1,R=R1,d1=C1,Y=Tev2,d2=C2)
```

All the functions implemented will search for the elements of this list.

Estimation of $S_C(t)$, $G_1(l, r|y)$ and $G_2(l, r|y)$

To obtain the estimates for the survival distributions of C, $(L, R)|Y = y$ and $(L, R)|Y > y$ (functions $S_C(t)$, $G_1(l, r|y)$ and $G_2(l, r|y)$ as defined in Section 6.5.1), the package `survival` must be loaded. We fit the survival function and implement a small function to access the information at any moment during the algorithm:

```
survC<-survfit(Surv(T2,1-C2)~1)
sC<-function(x,survC){
  ind<-findInterval(x,survC$time)
  return(survC$surv[ind])
}
```

When it comes to G_1 and G_2 we have take the subsample of individuals experiencing a recurrence within an interval (C1=1). Matrix datLR is defined by selecting vectors L1, R1 and Y1 within this subsample. Then, we stratify Y1 in *gr* groups according the corresponding percentiles (which are stored in vector breaks). Vector strat labels each individual to its corresponding stratum.

```
L1<-L[d1==1]
R1<-R[d1==1]
Y1<-Y[d1==1]
datLR<-cbind(L1,R1,Y1)
nk<-nrow(datLR)
gr<-if(nk/n>=0.5) 5 else 4 #sstratum must contain enough data
breaks<-quantile(Y1,probs=seq(1/gr,1,1/gr))
strat<-findInterval(Y1,breaks,rightmost.closed=T) +1
```

The following functions permit us to obtain the empirical bivariate survival function of (L, R) for any given set of observed values contained in (L, R): this way, the stratified distributions of $(L, R)|Y$ and $(L, R)|Y > y$ can be obtained just varying the set datLR.

```
empSurv<-function(l,r,datLR){
  return(mean((datLR[,1]>l)*(datLR[,2]>r)))
}
jLR<-function(l,r,datLR){
  return(mapply(empSurv,l,r,MoreArgs=list(datLR=datLR)))
}
GLR<-function(l,r,datLR1,datLR2){
  return(mapply(jLR,l,r,MoreArgs=list(datLR=datLR1))*
           mapply(jLR,l,r,MoreArgs=list(datLR=datLR2)))
}
```

The algICSCR function call the GLR function t compute the inverse weights of Strategy 2.

Estimation of $S_2(t)$ and $S_T(t)$

Parametric and nonparametric estimates for $S_2(t)$ and $S_T(t)$ can be used to feed the estimating algorithm. Parametric models can be adjusted by means of the survreg function of package survival, which can deal both with right-censored(T_2) and interval-censored data (T). Assume we choose a Weibull fit. Let S2.wei and ST.wei be the survreg functions for each function. Since these estimates remain invariant during the iterative process, we define the following functions to obtain their survival whenever required during the algorithm:

```
S2<-function(t){ return(exp(-sc2*t^sh2))}
Sz<-function(t){ return(exp(-scT*t^shT))}.
```

Caution is needed with the parametrization chosen because estimates from the survreg function are given at the log-linear scale. For the parametrization of the Weibull chosen, the parameters sc2,sh2,scT,shT are obtained by:

```
sh2<-1/S2.wei$scale
sc2<-exp(-S2.wei$coef/S2.wei$scale)
```

```
shT<-1/ST.wei$scale
scT<-exp(-ST.wei$coef/ST.wei$scale)
```

When nonparametric estimates are used, the `survfit` function can be employed to estimate $S_2(t)$. However, to obtain Turnbull's estimate for $S_T(t)$ we have used the `PGM` function of the `Icens` package (Gentleman and ., 2009). Let `hatS2` and `hatST` two lists containing the times and values at which they are estimated. Similarly to the parametric case, we need to define functions `S2` and `Sz`:

```
S2<-function(t){
  ind<-findInterval(t,hatS2$time)
  return(hatS2$S2[ind])
}
Sz<-function(t){
  ind<-findInterval(t,hatSz$time)
  return(hatSz$Sz[ind]).
}
```

Initial values for α and $S_1(t)$

Initial estimates for α and $S_1(\cdot)$ are needed to start the iterative phase. Our choice is to impute to T_1 the midpoint of the interval and then perform a semi-competing risks analysis for right-censored data. The resulting estimates for α and $S_1(\cdot)$ are taken as the initial steps of our algorithm. Let `u0` and `v0` be the results from `corSCR` and `margSCR`, respectively.

In order to run the estimating algorithm, it is convenient to smooth the step function given in `u0` as well as the successive estimations obtained at each iteration step. Indeed, if we choose a parametric fit, only the parameters need to be updated at each step. Moreover, it is a technical solution to avoid problems of definition with the integrals employed to compute expected concordance and weight probabilities, as well as to easily compute the density function $f_1(t)$ which is needed at every call of the joint density $f(s,t)$ (6.9). the provided code assumes that Weibull fits were adopted.

Therefore, in the call of function `algICSCR`, the initial estimate for α would be `e.alp.m=u0$con.index`, while `sh1.0` and `sc1.0` are the parameters of the Weibull fit obtained from `v0`.

10.4.2 Estimation algorithm

Internally, function `algICSCR` execute the following steps:

1) Obtain the comparable sample O_{ij}, and identify the distinct types of observed individuals (defined in detail in Appendix B.5) by means of functions `compIC` and `compIC.1` (Appendix D.4.2).

2) At each iteration step compute:

   - For Strategy 1, the number of excluded pairs n_e (Equation 6.12) with function `nD` (Appendix D.4.3).

- For Strategy 2, the probability for a pair to be comparable w_{ij} (Equation 6.15) with function `pes0.5n` (Appendix D.4.4).
- Solve the corresponding estimating equation $U_1(\alpha)$ or $U_2(\alpha)$, computing Z_{ij} with function `f.Zij` (Appendix D.4.5).
- Update S_1 by the plug-in estimator implemented in function `iS1` (Appendix D.4.6)

3) Repeat until convergence.

4) Compute the jackknife variance of the estimates.

10.4.3 Comments on the estimation algorithm

The iterative phase is repeated until convergence , when the difference in estimated consecutive α is smaller than a tolerance value, or whether a maximum number of iterations is reached.

The several definite integrals that must be computed during the iterative algorithm have been implemented with the `integrate` function included in the `R` software, which performs numerical integration by means of the adaptive quadrature of functions. Nevertheless, this function is time consuming, and during the simulation study, an approximation using Simpson's composite rule has been implemented. In the simulation study, specially for Strategy 2, it was mandatory to use this approximation to reduce computation time. In addition, when $S_2(t)$ and $S_T(t)$ are estimated non-parametrically, the integrate function may cause if we do not take care to select the grid of points among those with $S_2(t)$ and $S_T(t)$ being strictly greater than zero.

CHAPTER 11

Discussion and future research

This thesis has been developed with two main goals in mind: the first, of an applied nature, was to model the course of bladder cancer by means of survival analysis techniques; the second, to propose a methodological contribution to the problem of semi-competing risks data when interval censoring is present. In the following sections we summarize the results of this work with respect to these two main goals, while suggesting limitations as well as other possible approaches to the problem.

11.1 Modelling the evolution of bladder cancer

In the first part of the PhD manuscript, we have described and analysed the Spanish Bladder Cancer Study by means of survival analysis techniques, emphasizing the use of competing risks and multi-state models, which, although well developed from a methodological point of view, they are not routinely used. We have focused on two survival endpoints of interest: Event Free Survival (EFS), defined by the time from diagnosis until some disease-related event occurs, and Progression Free Survival (PFS), defined by the time from diagnosis until progression or death due to bladder cancer occurs. These endpoints are frequently analysed in the literature of bladder cancer and in other types of cancer.

First, competing risk was used in order to appropriately accommodate the existence of non disease-related deaths in the study. The presence of this competing event could prevent the complete observation of both endpoints of interest, EFS and PFS. Actually, since EFS includes recurrences that usually occur in the first years of follow-up, this process is reasonably well observed. Instead, progressions take a longer time to occur and, consequently, PFS is more affected by non disease-related deaths. We assessed and confirmed in the Spanish Bladder Cancer/EPICURO Study the effect of the agreed prognostic factors for EFS (multiplicity, tumor size and grade) and PFS (age, multiplicity, stage and grade). In addition, gender turned out to be also a risk factor for EFS. This

effect of gender is specific of the SBC/EPICURO study and raises the question of why this higher risk of recurrence among Spanish women.

Competing risks methods were also used to analyse the time to the first observed event, distinguishing among recurrence, progression of disease (including deaths due to disease) and deaths due to other causes. The clinical motivation for this approach was the characterization of those patients who may experience a progression as a first detected event. These patients are probably suffering from a more aggressive disease and they might benefit of a more strict follow-up protocol. A part from the most important prognostic factors of progression (stage T1/Tis and Grade 3 tumours), we also found differences in gender: females resulted to have higher risk of progression as a first event than males.

Multi-state modelling permitted us to obtain a complete picture of the evolution of bladder cancer, because distinct events, intermediate and terminating, are explicitly linked into the multi-state model. We explore the effect of having or not a recurrence during the follow-up on the rate of progression, finding that experiencing at least one recurrence increased significantly the risk of progression of disease. Additionally, in order to assess the effect of this recurrence on the probability of progression (at a cumulative scale), we computed the predictive process of progression, thus obtaining a dynamical model which permits to update the predicted probability for a patient, at a certain time point t, given the path of the disease the patient has followed until that particular moment. These dynamic predictions may prove useful for clinicians in the management of the patients with bladder cancer.

During the multi-state modelling we acknowledged the presence of distinct types of event: intermediate and terminating. For instance, recurrence was intermediate event for progression and death. After the occurrence of a progression or death, an individual is no longer at risk of recurrence, despite still being under observation. As such, in the modelling of the time until recurrence, terminating events cannot be simply treated as independent censoring. In the search of methods to account for this induced dependent censoring, we came across with semi-competing risks.

The semi-competing risks problem describes a situation where two events compete but one can be observed after the other. Therefore, for a portion of the observed individuals, information on the time of the two events is available. This extra amount of information permits to explore the marginal distribution of the time to the intermediate event, T_1, by taking advantage of the modelling of the joint distribution of T_1 and the time to the terminating event, T_2. This approach has been explored in depth in the second part of the thesis, and it has given raise to an extension of the semi-competing risks problem in order to account for interval-censoring.

The modelling approaches followed to describe the course of bladder cancer, mainly competing risks and multi-state models, present some limitations: (i) interval censoring was ignored, (ii) only the first of the recurrences was modelled, ignoring that several were possible, and (iii) more flexible models could have been used.

Firstly, the problem of interval censoring has been simplified in order to apply techniques for right-censored data. The time to recurrence or progression was assumed to be exactly observed by imputing the midpoint of the interval where they where known to lie in. However, there exist methods to deal with interval-censoring both in competing risks (the work of Hudgens *et al.* (2001)

has been presented in Chapter 5) and in multi-state models (Commenges and Gégout-Petit, 2007, Foucher *et al.*, 2010, Joly *et al.*, 2002). It would be interesting then to study the impact of the presence of interval-censoring data and compare the performance of right-censored (plus midpoint imputation) and interval-censored specific methods.

Secondly, in order to describe all aspects of the evolution of bladder cancer, the distinct recurrences and progressions must be modelled. In the SBC/EPICURO Study, out of 330 patients with at least one recurrence, 179 (54.2%) had a single recurrence, 62 (18.8%) had 2 recurrences, and 89 (27%) had three or more recurrences. Recurrences were solely explored to assess if the impact of the number of recurrences on the risk of progression was significant. However, it would be of interest to characterize those patients who tend to experience more recurrences than others. An extensive literature on this issue exists, which was reviewed to some extent in my Thesis proposal project at the beginning of our research in 2006. Different methodologies can be used depending on whether the main goal is to study marginal and conditional distributions, or the degree of association among the distinct failure times. Regression models conditioning on previous recurrences, marginal models undertaking a working assumption on the dependence of the distinct recurrent events, frailty models introducing between-subject heterogeneity by a random effect, and multivariate parametric or copula models are some of the approaches for this type of data (Cook and Lawless, 2002, Lawless, 2003). In addition, the problem of dependent censoring in the recurrent process induced by a terminating event has been described (Cook *et al.*, 2009, Ye *et al.*, 2007).

Last but not least, we have been pointed out during the referral process of this work that more flexible nonparametric regression models could have been used to fit cause-specific hazards models in Chapter 2 and the transition intensities in Chapter 3. A good overview on these models can be found in Martinussen and Scheike (2006). Among others, it has been suggested the use of additive hazards models (Aalen, 1993), or the use of smoothing techniques such as P-splines (Eilers and Marx, 1996) to introduce flexibility into the Cox model. These models, free of the assumptions that can invalidate results from parametric and semi-parametric models, are an attractive alternative. Additive hazards models have been explored at the beginning of our research on competing risks (Porta *et al.*, 2007), but since we focused the methodological research in the semi-competing risks problem, we did not further explore these models. However, we believe them to be useful to explore how the effects of the covariates diminish with time, which is a more realistic framework for long follow-up studies such as the SBC/EPICURO Study.

11.2 Interval-censored semi-competing risks data

We have proposed a methodology to deal with semi-competing risks data when T_1 is interval-censored. Assuming a Clayton's copula model to fit the dependency between T_1 and T_2, we have developed an algorithm to jointly estimate the copula parameter and the marginal distribution of T_1. Firstly, a new measure of concordance between two pairs of individuals given their observed interval-censored data has been defined. The concordance is computed among these pairs for which an ordering scheme can be defined, the so-called comparable sample. The properties of the concordance sample lead us to define two unbiased estimating equations for α: in Strategy 1, the bias induced by the comparable sample is explicitly corrected; Strategy 2 employs inverse-weighting

techniques to correct such bias. Finally, $S_1(t)$ is approximated using a plug-in estimator based on the Clayton's copula parametric form.

Both the illustrations in Chapter 8 as well as the simulation study in Chapter 9 showed that Strategy 1 provides accurate estimates for α, while Strategy 2 is more sensible to the amount of information present in our data, performing poorly in scenarios where much association as well as a high percentage of dependent censoring are present.

We believe that Strategy 1 provides an appropriate tool to account for interval censoring in this setting. It behaves better than midpoint imputation, which would be the easiest way to reduce the problem of interval censoring to right-censored data. Furthermore, it provides unbiased estimates of the copula parameter.

One of the reasons for the poor performance of Strategy 2 is that it involves the estimation of the joint distribution of L and R, the extremes of the censoring intervals. In settings where not enough 'pure intervals' (those with $R < Y < \infty$) are available, this joint distribution is estimated with large variability, and this fact has an important impact on the inverse-weighted estimator of α. Moreover, Strategy 2 is more sensible to inaccuracies on the estimates of $S_T(t)$, $S_2(t)$ and on the numerical integration method employed during the computation.

On the other hand, results of the simulation study were limited due to the fact that it resulted highly time-consuming. To complete a simulation setting (1000 iterations) with sample size equal to 500, Strategy 1 took an average of 2.5 days, while Strategy 2 could last around 5 to 6 days. Several computers have undertaken the simulations, some of them being servers, but they were not specially empowered for such computations (the most powerful being an Intel® Core$^T M$ 2 Quad Processor with 3-25 of RAM memory). The 'neck of the bottle' of the algorithm are the two-by-two comparisons between individuals, which greatly increases as sample size gets larger. With the simulation process being so cumbersome, this prevented us to explore more scenarios.

We applied the proposed methodologies for interval-censored semi-competing risks data to the SBC/EPICURO study since it was the study that motivated this methodology. Firstly, we explored the recurrence process as an intermediate event for progression or death. Secondly, we analysed progressions acting as a non-terminating event and death due to other causes as terminating event. However, the results should be taken with caution or just as an illustrative real example since the two basic elements required for a study to enjoy the benefits of a semi-competing risks analysis, say the dependence between the intermediate and the terminating event and a certain amount of dependent censoring, were not present in the SBC/EPICURO study.

11.3 Future work

Several lines of research are now open to keep on working on the issues proposed in this thesis: firstly, to complete the analysis of the course of bladder cancer; secondly, to consolidate and extend the methodology of interval-censored semi-competing risks data; thirdly, the software implementation and transferral of the developed methods.

A simulation study is needed to assess the impact between methods accounting for interval censoring and standard methods for right censoring in the framework of competing risks and multi-state

models. The SBC/EPICURO data might potentially be re-analyzed taking into account interval censoring. On the other hand, the modelling will be completed by the analysis of recurrent events. The use of more flexible models, if suitable, shall be considered to gain insight into the bladder cancer data. Another point of interest in our research is on the prediction of the bladder cancer events, and in the tools to validate a predictive model for this disease, such as the calibration plot. Some tools are available for several regression models (Harrell *et al.*, 1996), but in the context of competing risks and multi-state models tools for the validation of a predictive model, like the calibration plot we have proposed in this thesis, need to be adapted.

When it comes to future work related to the proposed methodology for interval-censored semi-competing risks data, further simulation settings need to be explored. Larger sample sizes, smaller association parameters (to cover more realistic scenarios, such as weak association), and varying amounts of 'pure' intervals (to confirm the dependency of Strategy 2 on this quantity) would allow for more robust conclusions on the methods proposed. In addition, more insight into the choice of the generating marginals (Exponential or Weibulls) as well as on the choice of the estimating procedure for $S_T(t)$ and $S_2(t)$ (only parametric fits were used in the simulation study) would complete the results.

From a theoretical point of view, there are several aspects of the proposed methodology for ICSCR data which could be extended to a more general setting: (i) T_2 could also be interval-censored, (ii) a less restrictive copula model could be assumed, and (iii) regression modelling is needed.

(i) We have covered the frequent situation where the exact time of the terminating event could be easily obtained, for instance in the case of death. However, situations arise where T_2 could be exactly observed or interval-censored. We have ignored this fact in the SBC/EPICURO Study when analysing progression as a terminating event for recurrence (see Chapter 8). The interval-censored semi-competing risks methods could be extended with a moderate effort to bivariate interval censoring by redefining the comparable sample O_{ij} (Section 6.3.2) and the expected concordance (Section 6.3.1) in order to include two 'pure' interval comparisons.

(ii) On the other hand, extensions to other copula models are not straightforward and deeper work is needed. An important milestone of our methodology is the relationship between the expectation of the concordance indicator and the copula parameter α,

$$E[\Delta_{ij}] = \frac{\alpha}{\alpha + 1},$$

maintained by the new concordance measure Z_{ij}. In the case of an archimedean copula (Definition 4.2), the previous expectation, for right-censored semi-competing risks data equals (Lakhal *et al.*, 2008)

$$\mathrm{E}\left[\Delta_{ij} | (\widetilde{T}_{1ij}, \widetilde{T}_{2ij})\right] = \frac{\theta_\alpha\big[S(\widetilde{T}_{1ij}, \widetilde{T}_{2ij})\big]}{\theta_\alpha\big[S(\widetilde{T}_{1ij}, \widetilde{T}_{2ij})\big] + 1},$$

which depends on the copula parameter α and also on the observed data through the joint survival $S(s,t)$. The impact of this property in the extension of the ICSCR methods with a more general copula must be carefully assessed.

(iii) Finally, in Chapter 8 we performed stratified analysis for distinct levels of a covariate, but no inferential tests specific for ICSCR data when covariates are present were performed. It would be useful to generalize to this situation tests for goodness-of-fit of Clayton's model (such as the one referred in Fine *et al.* (2001) and used in this work for right-censored data), or for testing constancy of association across strata of a covariate (such as the ones proposed by Ghosh (2006) for the right-censored case). In order to deal with regression modelling for interval-censored semi-competing risks data we would need first to go into the semi-competing risks problem in depth, a problem that, besides few references (Hsieh *et al.*, 2008, Peng and Fine, 2007) is far beyond from being solved.

Future contributions arising from the modelling of bladder cancer includes the implementation of an `R` package containing the prediction tools developed for competing risks (nomograms and calibration tools) and multi-state models (predictive process). In addition, a paper concerning the multi-state modelling of the course of bladder cancer based on the SBC/EPICURO data is in preparation.

As future research, we want to improve the efficiency of the algorithm, with the target in mind of building an `R` package to make these ICSCR functions available. One option could be to employ the combined features of `R` with `Fortran` or `C++` implementations. We have to assess these tools, of course, in order to transfer this aspect of our work to the statistical community as soon as possible.

Bibliography

Aalen, O. O. (1993). Further results on the non-parametric linear regression model in survival analysis. *Statistics in Medicine*, **12**(17), 1569–1588.

Aalen, O. O. (2010). Understanding disease processes. *Statistics in Medicine*, **29**(11), 1159–1160.

Aalen, O. O., Borgan, O., and Fekjaer, H. (2001). Covariate adjustment of event histories estimated from markov chains: the additive approach. *Biometrics*, **57**(4), 993–1001.

Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, **11**(2), 91–115.

Andersen, P. K., Abildstrom, S. Z., and Rosthøj, S. (2002). Competing risks as a multi-state model. *Statistical Methods in Medical Research*, **11**(2), 203–215.

Babjuk, M., Oosterlinck, W., Sylvester, R., Kaasinen, E., Böhle, A., Palou-Redorta, J., and of Urology (EAU), E. A. (2008). Eau guidelines on non-muscle-invasive urothelial carcinoma of the bladder. *European Urology*, **54**(2), 303–314.

Babjuk, M., Oosterlinck, W., Sylvester, R., Kaasinen, E., Böhle, A., and Palou, J. (2009). *Guidelines on TaT1 (Non-muscle invasive) Bladder Cancer*. European Association of Urology. `http://www.uroweb.org/fileadmin/tx_eauguidelines/2009/Full/TaT1_BC.pdf`.

Betensky, R. A. and Finkelstein, D. M. (1999). An extension of kendall's coefficient of concordance to bivariate interval censored data. *Statistics in Medicine*, **18**(22), 3101–3109.

Bogaerts, K. and Lesaffre, E. (2004). A new, fast algorithm to find the regions of possible support for bivariate interval censored data. *Journal of Computational and Graphical statistics*, **2**, 330–340.

Bogaerts, K. and Lesaffre, E. (2008a). Estimating local and global measures of association for bivariate interval censored data with a smooth estimate of the density. *Statistics in Medicine*, **27**(28), 5941–5955.

Bogaerts, K. and Lesaffre, E. (2008b). Modeling the association of bivariate interval-censored data using the copula approach. *Statistics in Medicine*, **27**(30), 6379–6392.

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Duxbury Press.

Chen, B., Yi, G. Y., and Cook, R. J. (2010). Analysis of interval-censored disease progression data via multi-state models under a nonignorable inspection process. *Statistics in Medicine*, **29**(11), 1175–1189.

Chun, F. K.-H., Karakiewicz, P. I., Briganti, A., Gallina, A., Kattan, M. W., Montorsi, F., Huland, H., and Graefen, M. (2006). Prostate cancer nomograms: an update. *European Urology*, **50**(5), 914–26; discussion 926.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**(1), 141–151.

Commenges, D. and Gégout-Petit, A. (2007). Likelihood for generally coarsened observations from multistate or counting process models. *Scandinavian Journal of Statistics*, **34**, 432–450.

Cook, R. J. and Lawless, J. F. (2002). Analysis of repeated events. *Statistical Methods in Medical Research*, **11**(2), 141–166.

Cook, R. J., Zeng, L., and Lee, K.-A. (2008). A multistate model for bivariate interval-censored failure time data. *Biometrics*.

Cook, R. J., Lawless, J. F., Lakhal-Chaieb, L., and Lee, K.-A. (2009). Robust estimation of mean functions and treatment effects for recurrent events under event-dependent censoring and termination: Application to skeletal complications in cancer metastatic to bone. *Journal of the American Statistical Association*, **104**(485), 60–75.

Coviello, V. and Boggess, M. (2004). Cumulative incidence estimation in the presence of competing risks. *Stata Journal*, **4**(2), 103–112(10).

Cox, D. and Oakes, D. (1984). *Analysis of survival data*. London: Chapman-Hall.

Day, R., Bryant, J., and Lefkopoulou, M. (1997). Adaptation of bivariate frailty models for prediction, with application to biological markers as prognostic indicators. *Biometrika*, **84**(1), 45–56.

de Wreede, L. C., Fiocco, M., and Putter, H. (2010). The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. *Computer Methods and Programs in Biomedicine*, **In Press, Corrected Proof**, –.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, **11**(2), 89–121.

Fine, J. and Gray, R. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, **94**(446), 496–509.

Fine, J. P. (2001). Regression modeling of competing crude failure probabilities. *Biostatistics*, **2**(1), 85–97.

Fine, J. P., Jiang, H., and Chappell, R. (2001). On semi-competing risks data. *Biometrika*, **88**(4), 907–919.

Foucher, Y., Giral, M., Soulillou, J., and Daures, J. (2010). A flexible semi-Markov model for interval-censored data and goodness-of-fit testing. *Statistical Methods in Medical Research*, **19**(2), 127–145.

Frank E Harrell Jr and with contributions from many other users. (2009). *Hmisc: Harrell Miscellaneous*. R package version 3.7-0.

García-Closas, M., Malats, N., Silverman, D., Dosemeci, M., Kogevinas, M., Hein, D. W., Tardón, A., Serra, C., Carrato, A., García-Closas, R., Lloreta, J., Castaño-Vinyals, G., Yeager, M., Welch, R., Chanock, S., Chatterjee, N., Wacholder, S., Samanic, C., Torà, M., Fernández, F., Real, F. X., and Rothman, N. (2005). Nat2 slow acetylation, gstm1 null genotype, and risk of bladder cancer: results from the spanish bladder cancer study and meta-analyses. *The Lancet*, **366**(9486), 649 – 659.

Gentleman, R. and ., A. V. (2009). *Icens: NPMLE for Censored and Truncated Data*. R package version 1.16.0.

Georges, P., Lamy, A.-G., Nicolas, E., Quibel, G., and Roncalli, T. (2001). Multivariate survival modelling: a unified approach with copulas. Technical report, Groupe de Recherche Opérationnelle Crédit Lyonnais.

Geskus, R. B. (2010). Cause-specific cumulative incidence estimation and the fine and gray model under both left truncation and right censoring. *Biometrics*.

Ghosh, D. (2006). Semi-parametric inferences for association with semi-competing risks data. *Statistics in Medicine*, **25**(12), 2059–2070.

Gómez, G., Calle, M., and Oller, R. (2004). Frequentist and bayesian approaches for interval-censored data. *Statistical Papers*, **45**(2), 139–173.

Gomez, G., Calle, M., Oller, R., and Langohr, K. (2009). Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modelling*, **9**(4), 259–297. *In press*.

Gray, R. (2004). *The cmprsk package*. The Comprehensive R Archive network. http://cran.r-project.org/src/contrib/Descriptions/cmprsk.html.

Gray, R. J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, **16**(3), 1141–1154.

Groeneboom, P., Maathuis, M. H., and Wellner, J. A. (2008a). Current status data with competing risks: consistency and rates of convergence of the mle. *The Annals of Statistics*, **36**(3), 1031–1063.

Groeneboom, P., Maathuis, M. H., and Wellner, J. A. (2008b). Current status data with competing risks: limiting distribution of the mle. *The Annals of Statistics*, **36**(3), 1064–1089.

Guey, L. T., García-Closas, M., Murta-Nascimento, C., Lloreta, J., Palencia, L., Kogevinas, M., Rothman, N., Vellalta, G., Calle, M. L., Marenne, G., Tardón, A., Carrato, A., García-Closas, R., Serra, C., Silverman, D. T., Chanock, S., Real, F. X., Malats, N., and for the EPICURO/Spanish Bladder Cancer Study investigators (2009). Genetic susceptibility to distinct bladder cancer subphenotypes. *European Urology*.

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**(4), 361–387.

Harrell Jr., Frank E. (2009). *Design: Design Package*. R package version 2.3-0.

Hernandez, S., Lopez-Knowles, E., Lloreta, J., Kogevinas, M., Amoros, A., Tardon, A., Carrato, A., Serra, C., Malats, N., and Real, F. X. (2006). Prospective Study of FGFR3 Mutations As a Prognostic Factor in Nonmuscle Invasive Urothelial Bladder Carcinomas. *Journal of Clinical Oncology*, **24**(22), 3664–3671.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**(260), 663–685.

Hougaard, P. (1999). Multi-state models: a review. *Lifetime Data Analysis*, **5**(3), 239–264.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Statistics for Biology and Health. Springer-Verlag.

Hsieh, J.-J., Wang, W., and Ding, A. A. (2008). Regression analysis based on semicompeting risks data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(1), 3–20.

Huang, J. (1999). Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statistica Sinica*, **9**, 501–519.

Hudgens, M. G., Satten, G. A., and Longini, I. M. (2001). Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation. *Biometrics*, **57**(1), 74–80.

Hui-Min, W., Ming-Fang, Y., and Chen, T. H.-H. (2004). Sas macro program for non-homogeneous markov process in modeling multi-state disease progression. *Comput Methods Programs Biomed*, **75**(2), 95–105.

Iasonos, A., Schrag, D., Raj, G. V., and Panageas, K. S. (2008). How to build and interpret a nomogram for cancer prognosis. *Journal of Clinical Oncology*, **26**(8), 1364–1370.

Jackson, C. (2009). *msm: Multi-state Markov and hidden Markov models in continuous time*. R package version 0.9.5.

Jewell, N. P., Van der Laan, M., and Henneman, T. (2003). Nonparametric estimation from current status data with competing risks. *Biometrika*, **90**(1), 183–197.

Jiang, H., Chappell, R., and Fine, J. P. (2003). Estimating the distribution of nonterminal event time in the presence of mortality or informative dropout. *Controlled Clinical Trials*, **24**(2), 135–146.

Jiang, H., Fine, J. P., Kosorok, M. R., and Chappel, R. (2005a). Pseudo self-consistent estimation of a copula model with informative censoring. *Scandinavian Journal of Statistics*, **32**(1), 1–20.

Jiang, H., Fine, J. P., and Chappell, R. (2005b). Semiparametric analysis of survival data with left truncation and dependent right censoring. *Biometrics*, **61**(2), 567–575.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Monographs on statistics ans Applied Probability. Chapman-Hall.

Joly, P., Commenges, D., Helmer, C., and Letenneur, L. (2002). A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics*, **3**(3), 433–443.

Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.

Karakiewicz, P. I., Shariat, S. F., Palapattu, G. S., Perrotte, P., Lotan, Y., Rogers, C. G., Amiel, G. E., Vazina, A., Gupta, A., Bastian, P. J., Sagalowsky, A. I., Schoenberg, M., and Lerner, S. P. (2006). Precystectomy nomogram for prediction of advanced bladder cancer stage. *European Urology*, **50**(6), 1254–60; discussion 1261–2.

Kattan, M. W., Heller, G., and Brennan, M. F. (2003a). A competing-risks nomogram for sarcoma-specific death following local recurrence. *Statistics in Medicine*, **22**(22), 3515–3525.

Kattan, M. W., Zelefsky, M. J., Kupelian, P. A., Cho, D., Scardino, P. T., Fuks, Z., and Leibel, S. A. (2003b). Pretreatment nomogram that predicts 5-year probability of metastasis following three-dimensional conformal radiation therapy for localized prostate cancer. *Journal of Clinical Oncology*, **21**(24), 4568–4571.

Keiding, N., Klein, J. P., and Horowitz, M. M. (2001). Multi-state models and outcome prediction in bone marrow transplantation. *Statistics in Medicine*, **20**(12), 1871–1885.

Klein, J. and Moeschberger, M. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag New York.

Klein, J. P. (2006). Modelling competing risks in cancer studies. *Statistics in Medicine*, **25**(6), 1015–1034.

Klein, J. P. and Andersen, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*, **61**(1), 223–229.

Klein, J. P. and Shu, Y. (2002). Multi-state models for bone marrow transplantation studies. *Statistical Methods in Medical Research*, **11**(2), 117–139.

Klein, J. P., Keiding, N., and Copelan, E. A. (1994). Plotting summary predictions in multistate survival models: probabilities of relapse and death in remission for bone marrow transplantation patients. *Statistics in Medicine*, **13**(24), 2315–2332.

Klein, J. P., Rizzo, J. D., Zhang, M. J., and Keiding, N. (2001a). Statistical methods for the analysis and presentation of the results of bone marrow transplants. part 2: Regression modeling. *Bone Marrow Transplant*, **28**(11), 1001–1011.

Klein, J. P., Rizzo, J. D., Zhang, M. J., and Keiding, N. (2001b). Statistical methods for the analysis and presentation of the results of bone marrow transplants. part i: unadjusted analysis. *Bone Marrow Transplant*, **28**(10), 909–915.

Kneib, T. and Hennerfeind, A. (2008). Bayesian semi parametric multi-state models. *Statistical Modeling*, **8**(2), 169–198.

Lakhal, L., Rivest, L.-P., and Abdous, B. (2008). Estimating survival and association in a semi-competing risks model. *Biometrics*, **64**, 180–188.

Lakhal, L., Rivest, L.-P., and Beaudoin, D. (2009). Ipcw estimator for kendall's tau under bivariate censoring. *The International Journal of Biostatistics*, **5**(1), 8.

Lan, L. and Datta, S. (2010). Non-parametric estimation of state occupation, entry and exit times with multistate current status data. *Statistical Methods in Medical Research*, **19**(2), 147–165.

Lawless, J. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.

Lehman, E. (1999). *Elements of large-sample theory*. Springer-Verlag, New York.

Lesaffre, E., Komárek, A., and Declerk, D. (2005). An overview of methods for interval-censored data with an emphasis on applications in dentistry. *Statistical Methods in Medical Research*, **14**(6), 539–552.

Lin, D. Y. and Ying, Z. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika*, **80**(3), 573–581.

Lotan, Y., Svatek, R. S., and Malats, N. (2008). Screening for bladder cancer: a perspective. *World Journal of Urology*, **26**(1), 13–18.

Maathuis, M. (2006). *Nonparametric estimation for current status data and competing risks.* Ph.D. thesis, University of Washington.

Maathuis, M. H. (2005). Reduction algorithm for the npmle for the distribution function of bivariate interval censored data. *Journal of Computational and Graphical Statistics*, pages 352–362.

Martinussen, T. and Scheike, T. H. (2006). *Dynamic regression models for survival data*. Statistics for Biology and Health. Springer, New York.

Meira-Machado, L., Cadarso-Suárez, C., and de Uña-Álvarez, J. (2007). tdc.msm: An r library for the analysis of multi-state survival data. *Computer Methods and Programs in Biomedicine,*, **86**(2), 131–140.

Meira-Machado, L., de Uña-Alvarez, J., Cadarso-Suárez, C., and Andersen, P. K. (2009). Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research*, **18**(2), 195–222.

Murta-Nascimento, C., Silverman, D. T., Kogevinas, M., García-Closas, M., Rothman, N., Tardón, A., García-Closas, R., Serra, C., Carrato, A., Villanueva, C., Dosemeci, M., Real, F. X., and Malats, N. (2007). Risk of bladder cancer associated with family history of cancer: do low-penetrance polymorphisms account for the increase in risk? *Cancer Epidemiol Biomarkers Prev*, **16**(8), 1595–1600.

Nelsen, R. B. (2006). *An introduction to copulas*. Springer.

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **44**(3), 414–422.

Oakes, D. (1986). Semiparametric inference in a model for association in bivariate survival data. *Biometrika*, **73**(2), 353–361.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, **84**(406), 487–493.

Oller, R., Gomez, G., and Calle, M. L. (2004). Interval censoring: model characterizations for the validity of the simplified likelihood. *Canadian Journal Of Statistics-Revue Canadienne De Statistique*, **32**(3), 315–326.

Oller, R., Gómez, G., and Calle, M. (2007). Interval censoring: identifiability and the constant-sum property. *Biometrika*, **94**(1), 61–70.

Peng, L. and Fine, J. P. (2006). Nonparametric estimation with left-truncated semicompeting risks data. *Biometrika*, **93**(2), 367–383.

Peng, L. and Fine, J. P. (2007). Regression modeling of semicompeting risks data. *Biometrics*, **63**(1), 96–108.

Pepe, M. S. and Mori, M. (1993). Kaplan-meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine*, **12**(8), 737–751.

Peto, R. (1973). Experimental survival curves for interval-censored data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **22**(1), 86–91.

Pintilie, M. (2006). *Competing risks: a practical perspective*. Wiley.

Porta, N., Calle, M., and Gómez, L. (2007). Competing risks methods. Technical Report 2007/14, Department of Statistics and Operations Research, Univeristat Politècnica de Catalunya.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., J., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, **34**(4), 541–554.

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, **26**(11), 2389–2430.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rosthøj, S., Andersen, P. K., and Abildstrom, S. Z. (2004). Sas macros for estimation of the cumulative incidence functions based on a cox regression model for competing risks survival data. *Computer Methods and Programs in Biomedicine*, **74**(1), 69–75.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**(3), 581–592.

Samanic, C., Kogevinas, M., Dosemeci, M., Malats, N., Real, F. X., Garcia-Closas, M., Serra, C., Carrato, A., García-Closas, R., Sala, M., Lloreta, J., Tardón, A., Rothman, N., and Silverman, D. T. (2006). Smoking and bladder cancer in spain: Effects of tobacco type, timing, environmental tobacco smoke, and gender. *Cancer Epidemiology Biomarkers & Prevention*, **15**(7), 1348–1354.

Scheike, T. H. and Zhang, M.-J. (2008). Flexible competing risks regression modeling and goodness-of-fit. *Lifetime Data Analysis*, **14**(4), 464–483.

Scheike, T. H., Zhang, M.-J., and Gerds, T. A. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, **95**(1), 205–220.

Shabsigh, A. and Bochner, B. H. (2006). Use of nomograms as predictive tools in bladder cancer. *World Journal of Urology*, **24**(5), 489–498.

Shih, J. H. (1998). A goodness-of-fit test for association in a bivariate survival model. *Biometrika*, **85**(1), 189–200.

Sun, J. (2006). *The Statistical Analysis of Interval-censored Failure Time Data*. Statistics for Biology and Health. Springer.

Sun, L., Wang, L., and Sun, J. (2006). Estimation of the association for bivariate interval-censored failure time data. *Scandinavian Journal of Statistics*, **33**(4), 637–649.

Sweeting, M. J., Farewell, V. T., and Angelis, D. D. (2010). Multi-state markov models for disease progression in the presence of informative examination times: an application to hepatitis c. *Statistics in Medicine*, **29**(11), 1161–1174.

Therneau, T. and Grambsch, P. (2000). *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer.

Therneau, T. and original R port by Thomas Lumley (2009). *survival: Survival analysis, including penalised likelihood*. R package version 2.35-4.

Trivedi, P. K. and Zimmer, D. M. (2007). Copula modeling: an introduction for practitioners. *Foundations and Trends in Econometrics*, **1**(1), 1–111.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences of the USA*, **72**(1), 20–22.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **38**(3), 290–295.

van der Aa, M., Dönmez, M., Eijkemans, M. J., van der Kwast, T. H., Zwarthoff, e. C., and Steyerberg, E. W. (2009). Clinical and pathological prognostic factors for recurrence, progression and mortality in non-muscle invasive bladder cancer: a meta-analysis. *Current Urology*, **3**, 113–123.

Wang, W. (2003). Estimating the association parameter for copula models under dependent censoring. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**(1), 257–273.

Wangler, M. and Beyersmann, J. (2009). *changeLOS: Change in LOS*. R package version 2.0.9-2.

Wolbers, M., Koller, M. T., Witteman, J. C. M., and Steyerberg, E. W. (2009). Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology*, **20**(4), 555–561.

Ye, Y., Kalbfleisch, J. D., and Schaubel, D. E. (2007). Semiparametric analysis of correlated recurrent and terminal events. *Biometrics*, **63**(1), 78–87.

Yu, Q., Schick, A., Li, L., and Wong, G. Y. (1998). Asymtotic properties of the gmle with case 2 interval-censored data. *Statistics & Probability Letters*, **37**, 223–228.

Zhang, Z. and Sun, J. (2010). Interval censoring. *Statistical Methods in Medical Research*, **19**(1), 53–70.

Zheng, M. and Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, **82**(1), 127–138.

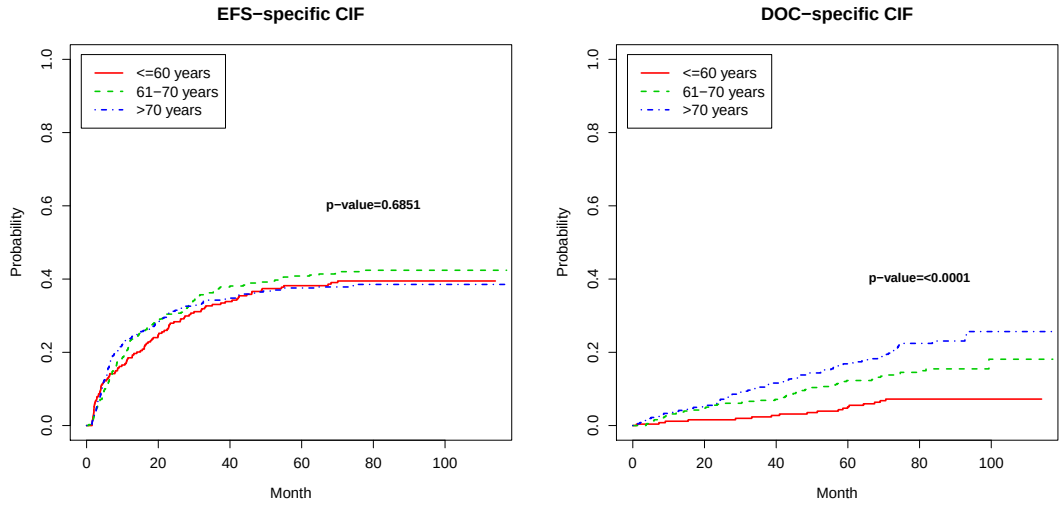
Appendix

APPENDIX A

The Spanish Bladder Cancer Study

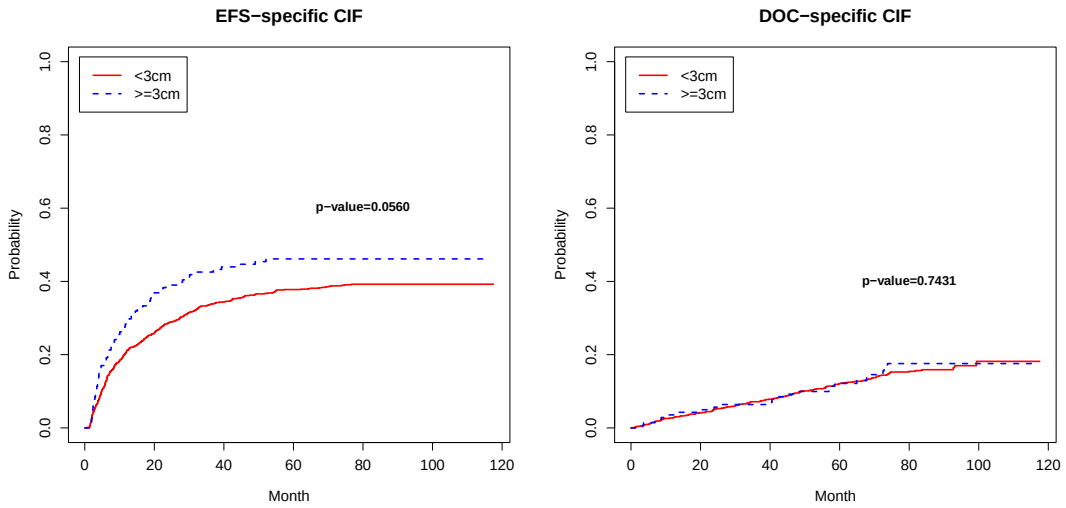
A.1 Cumulative incidence functions for (T_1, C_1)

Cumulative incidence functions for each type of event are empirically estimated from observed data. The time T_1 correspond to the minimum between T_{EFS}, the time to any disease-related events, and T_{DOC}, time to death due to other non disease-related causes. We present the curves distinguishing between stratums of age, tumour size and stage of the tumour.

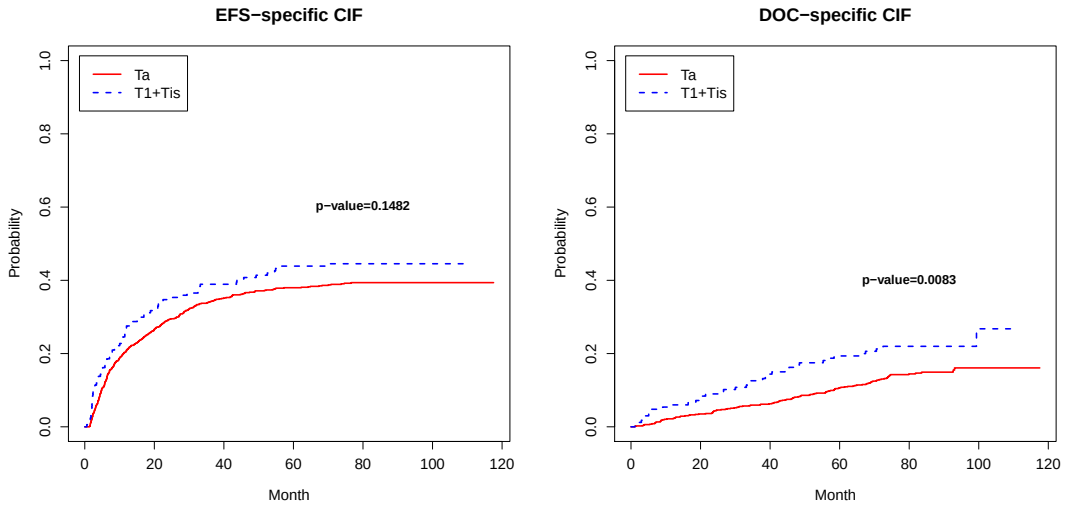


(a) Age: $\leq$ 60 vs 61-70 vs >70 years old

Figure A.1: Cumulative incidence functions for EFS and DOC in the analysis of (T_1, C_1) across age



(a) Tumour size: <3cm vs $\geq$ 3cm



(b) Stage of the tumour: Ta vs T1/Tis

Figure A.2: Cumulative incidence functions for EFS and DOC in the analysis of (T_2, C_2) across the tumour's features: size, stage and grade

A.2 Cumulative incidence functions for (T_2, C_2)

Cumulative incidence functions for each type of event are empirically estimated from observed data. The time T_2 correspond to the minimum between T_{PFS}, the time to progression or death due to bladder cancer, and T_{DOC}, time to death due to other non disease-related causes. We present the curves distinguishing between stratums of gender, tumour size and smoking.

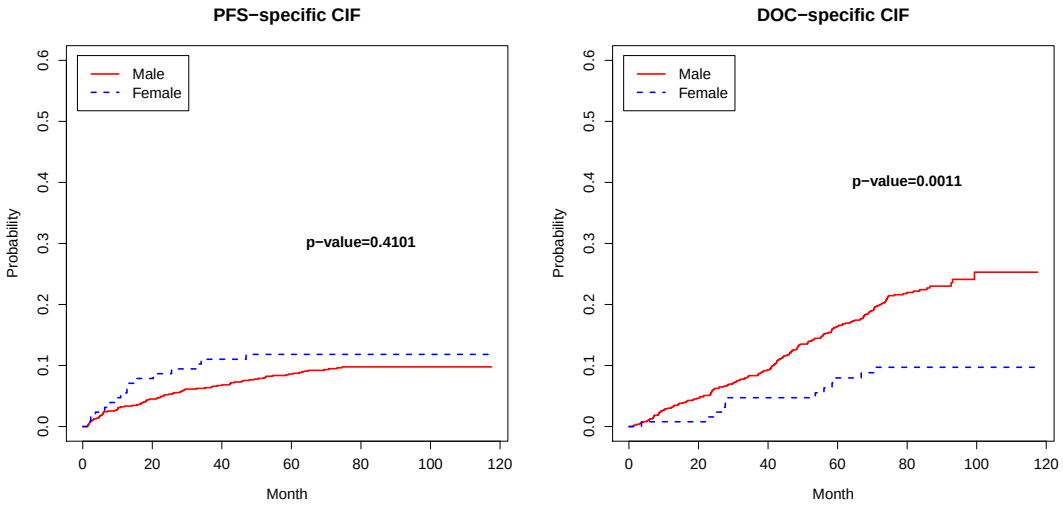


Figure A.3a: Cumulative incidence functions for (T_2, C_2) by Gender

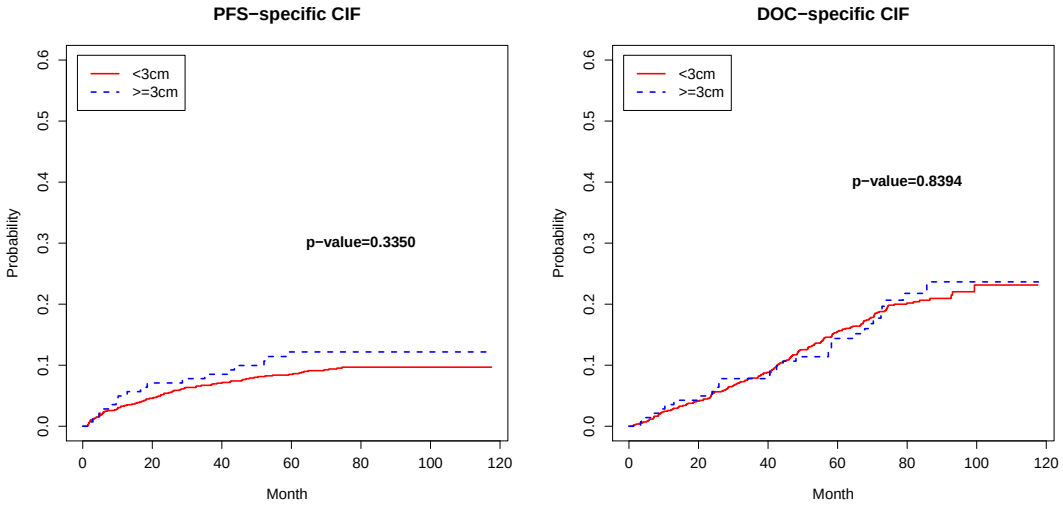


Figure A.3b: Cumulative incidence functions for (T_2, C_2) by Tumour size

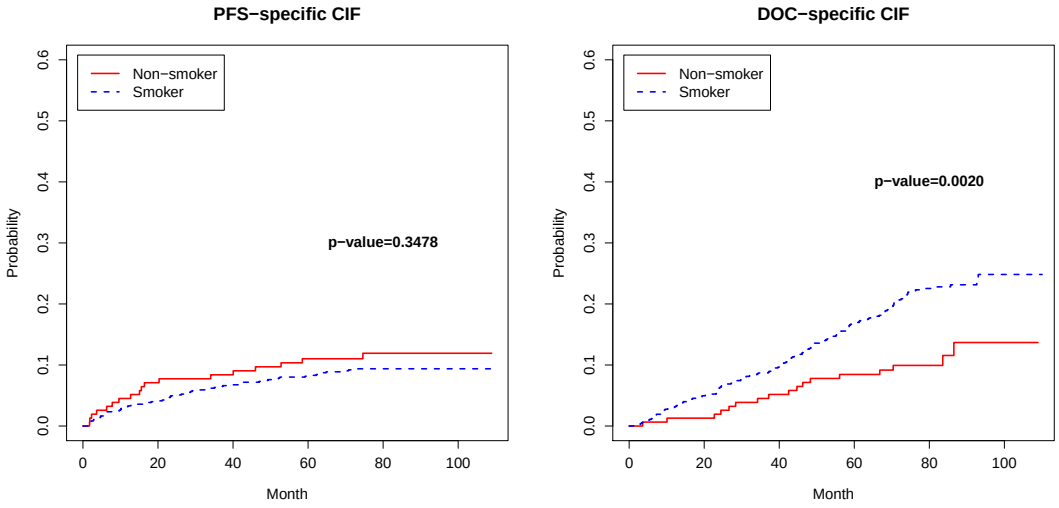


Figure A.3c: Cumulative incidence functions for (T_2, C_2) by Smoking status

A.3 Cumulative incidence functions for (T_1, C_1^*)

Cumulative incidence functions for each type of event are empirically estimated from observed data. The time T_1 correspond to the minimum between T_R, the time to recurrence, T_{PFS}, the time to progression or death due to bladder cancer, and T_{DOC}, time to death due to other non disease-related causes. It corresponds to the minimum between T_{EFS} and T_{DOC}. We present the curves distinguishing between stratums of gender, age, smoking status, tumour number, tumour size, stage and grade.

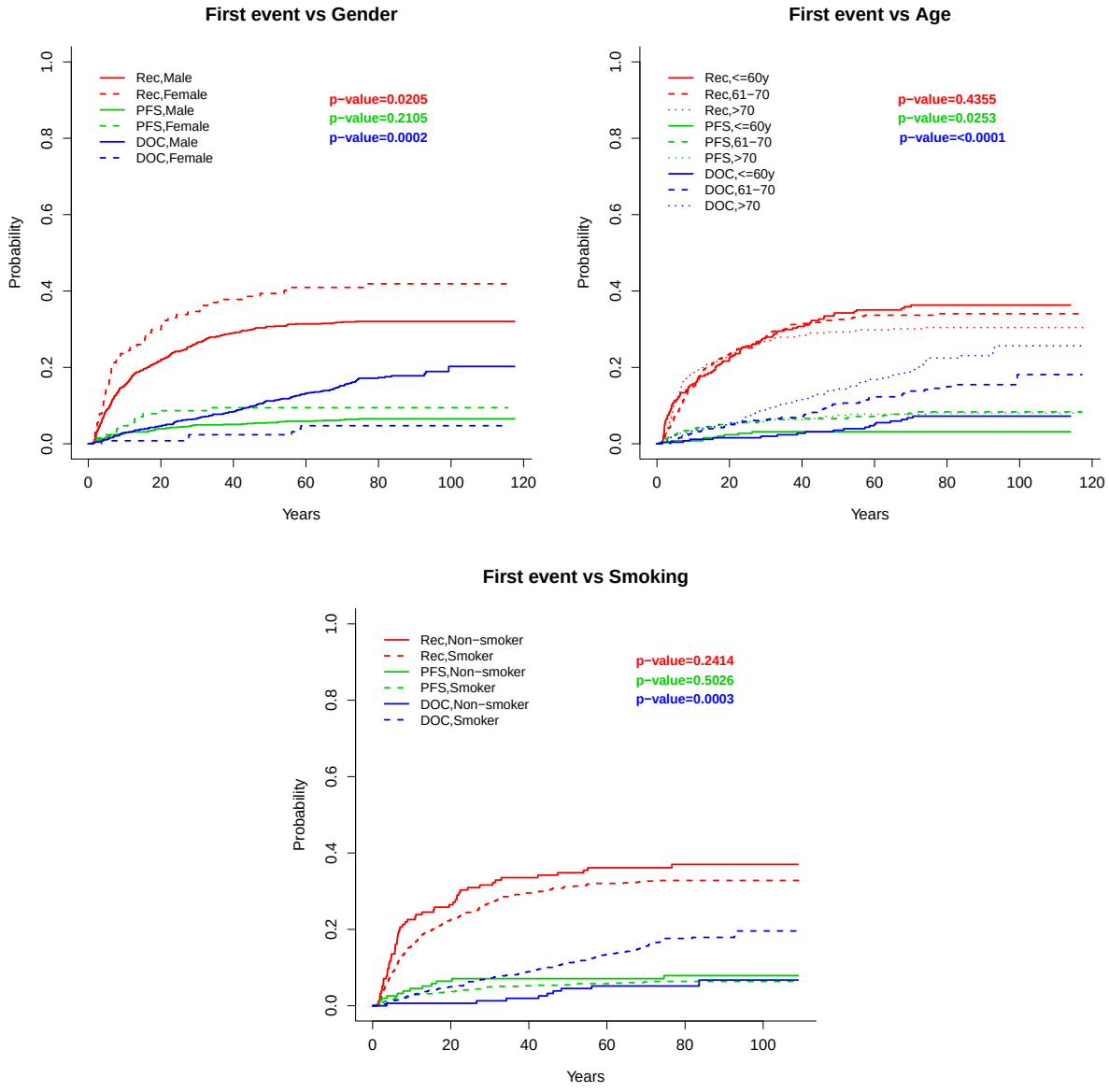


Figure A.4: Cumulative incidence functions for (T_1, C_1^*) across the individual's features

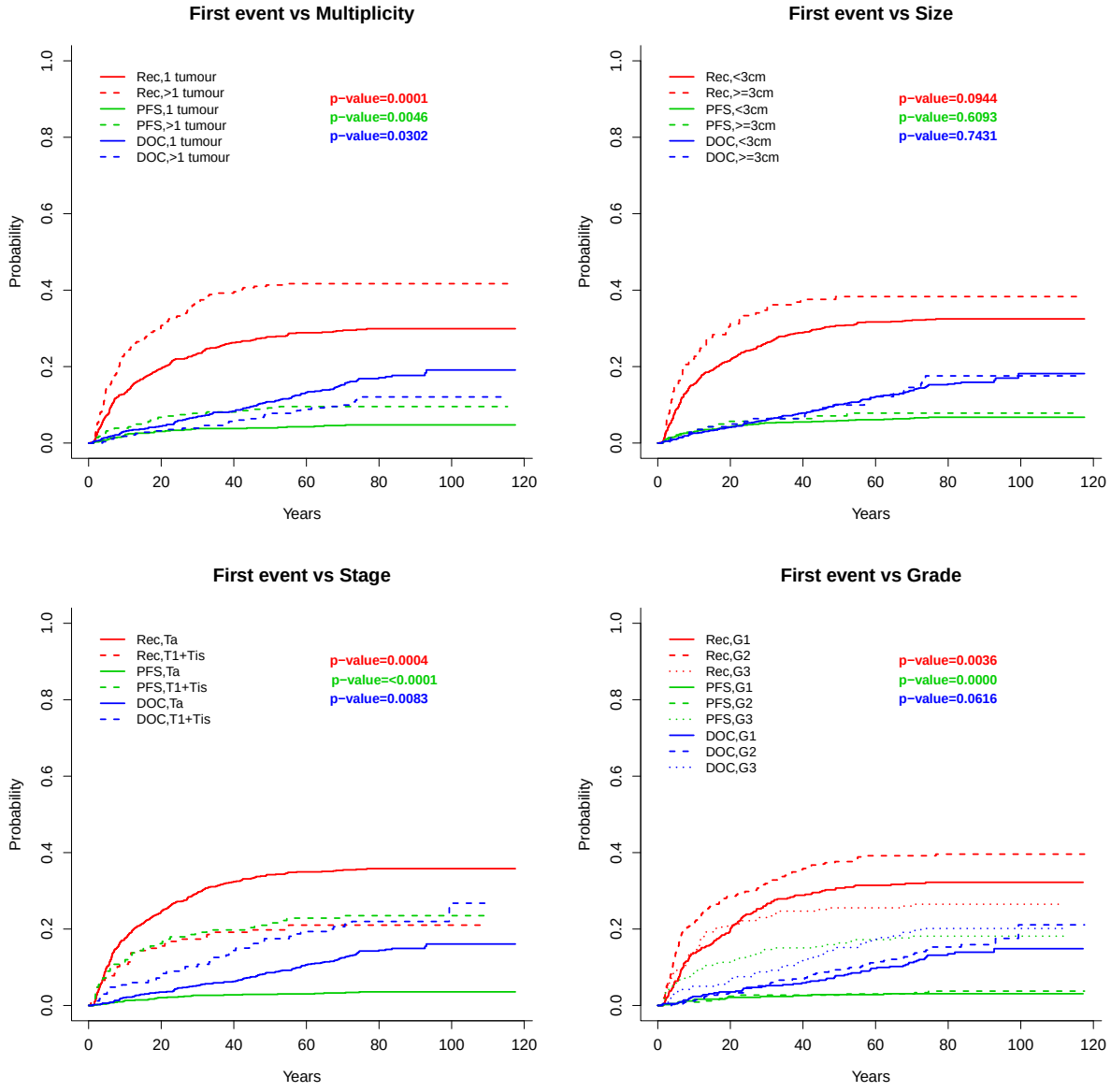


Figure A.5: Cumulative incidence functions for (T_1, C_1^*) across the tumour's features

A.4 Members of the participating centres

Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra (Barcelona) (coordinating center): M. Kogevinas, N. Malats, F.X. Real, M. Sala, G. Castaño, M. Torà, D. Puente, C. Villanueva, C. Murta-Nascimento, J. Fortuny, E. López, S. Hernández, R. Jaramillo, F Fernandez, A. Amorós, G Vellalta, L Palencia, A Alfaro, G Carretero. Hospital del Mar, Universitat Autònoma de Barcelona (Barcelona): J. Lloreta, S. Serrano, L. Ferrer, A. Gelabert, J. Carles, O. Bielsa, K. Villadiego. Hospital Germans Trias i Pujol (Badalona, Barcelona): L. Cecchini, J.M. Saladié,

L. Ibarz. Hospital de Sant Boi (Sant Boi, Barcelona): M. Céspedes. Centre Hospitalari Parc Taulí (Sabadell, Barcelona): C. Serra, D. García, J. Pujadas, R. Hernando, A. Cabezuelo, C. Abad, A. Prera, J. Prat. Centre Hospitalari i Cardiològic (Manresa, Barcelona): M. Domènech, J. Badal, J. Malet. Hospital Universitario (La Laguna, Tenerife): R. García-Closas, J. Rodríguez de Vera, A.I. Martín. Hospital La Candelaria (Santa Cruz, Tenerife): J. Taño, F. Cáceres. Hospital General Universitario de Elche, Universidad Miguel Hernández (Elche, Alicante): A. Carrato, F. García-López, M. Ull, A. Teruel, E. Andrada, A. Bustos, A. Castillejo, J.L. Soto. Universidad de Oviedo (Oviedo, Asturias): A. Tardón. Hospital San Agustín (Avilés, Asturias): J.L. Guate, J.M. Lanzas, J. Velasco. Hospital Central Covadonga (Oviedo, Asturias): J.M. Fernández, J.J. Rodríguez, A. Herrero. Hospital Central General (Oviedo, Asturias): R. Abascal, C. Manzano, T. Miralles. Hospital de Cabueñes (Gijón, Asturias): M. Rivas, M. Arguelles. Hospital de Jove (Gijón, Asturias): M. Díaz, J. Sánchez, O. González. Hospital de Cruz Roja (Gijón, Asturias): A. Mateos, V. Frade. Hospital Alvarez-Buylla (Mieres, Asturias): P. Muntañola, C. Pravia. Hospital Jarrio (Coaña, Asturias): A.M. Huescar, F. Huergo. Hospital Carmen y Severo Ochoa (Cangas, Asturias): J. Mosquera.

APPENDIX B

Theoretical Aspects

B.1 Expressions of the cross-sectional ratio $\theta(s,t)$

B.1.1 Equivalence of (4.4) and (4.5)

Consider the random variable $T_2|T_1$, its hazard function given by

$$\lambda_2(t|T_1=s) = \frac{f_{T_2|T_1=s}(t)}{S_{T_2|T_1=s}(t)} = \frac{\frac{f(s,t)}{f_1(s)}}{\frac{P[T_2>t,T_1=s]}{f_1(s)}} = \frac{f(s,t)}{\int_t^\infty f(s,v)dv}, \tag{B.1}$$

where $f_{T_2|T_1=s}(t)$ and $S_{T_2|T_1=s}(t)$ are the density and survival functions of the random variable $T_2|T_1$, $f(s,t)$ is the joint density function of (T_1,T_2) and $f_1(s)$ is the density function of T_1.

The survival and hazard functions of the random variable $T_2|\{T_1>s\}$ are given by

$$\begin{aligned} S_{T_2|\{T_1>s\}}(t) &= P[T_2>t|T_1>s] = \frac{P[T_2>t,T_1>s]}{P[T_1>s]} = \frac{S(s,t)}{S_1(s)} \\ \lambda_2(t|T_1>s) &= \lambda_{T_2|\{T_1>s\}}(t) = \frac{d}{dt}\left[-\log S_{T_2|\{T_1>s\}}(t)\right] = \frac{\int_s^\infty f(u,t)du}{S(s,t)}. \end{aligned} \tag{B.2}$$

Then, from (4.5), substituting the numerator and denominator by the previous expressions (B.1) and (B.2), we have:

$$\theta(s,t) = \frac{\lambda_2(t|T_1=s)}{\lambda_2(t|T_1>s)} = \frac{\frac{f(s,t)}{\int_t^\infty f(s,v)dv}}{\frac{\int_s^\infty f(u,t)du}{S(s,t)}} = \frac{f(s,t)S(s,t)}{\int_s^\infty f(u,t)du\int_t^\infty f(s,v)dv}$$

B.1.2 Proof of Proposition 4.1:

Proof. To show expression (4.6),

$$\theta(s,t) = \frac{P((T_{1i}-T_{1j})(T_{2i}-T_{2j})>0|\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t)}{P((T_{1i}-T_{1j})(T_{2i}-T_{2j})<0|\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t)},$$

we develop the conditional probabilities:

$$\begin{aligned}
p_c &= P((T_{1i}-T_{1j})(T_{2i}-T_{2j})>0|\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t)\\
&= P(T_{1i}>T_{1j},T_{2i}>T_{2j}|\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t)+P(T_{1i}>T_{1j},T_{2i}>T_{2j}|\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t)\\
&\underset{(1)}{=} 2P(T_{1i}>T_{1j},T_{2i}>T_{2j}|\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t)\\
&= \frac{2P(T_{1i}>T_{1j},T_{2i}>T_{2j},T_{1j}=s,T_{2j}=t)}{P(\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t)}\\
&\underset{(1)}{=} \frac{2P(T_{1i}>s,T_{2i}>t)P(T_{1j}=s,T_{2j}=t)}{P(\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t)} = \frac{2S(s,t)f(s,t)}{P(\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t)}
\end{aligned}$$

$$\begin{aligned}
p_{nc} &= P((T_{1i}-T_{1j})(T_{2i}-T_{2j})<0|\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t)\\
&= P(T_{1i}>T_{1j},T_{2i}<T_{2j}|\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t)+P(T_{1i}<T_{1j},T_{2i}>T_{2j}|\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t)\\
&\underset{(1)}{=} 2P(T_{1i}>T_{1j},T_{2i}<T_{2j}|\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t)\\
&= \frac{2P(T_{1i}>T_{1j},T_{2i}<T_{2j},T_{1j}=s,T_{2i}=t)}{P(\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t)} \underset{(1)}{=} \frac{2P(T_{1i}>s,T_{2i}=t)P(T_{1j}=s,T_{2j}>t)}{P(\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t)}\\
&= \frac{2\int_s^\infty f(u,t)du\int_t^\infty f(s,v)dv}{P(\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t)} = \frac{2(-\partial_2 S(s,t))(-\partial_1 S(s,t))}{P(\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t)},
\end{aligned}$$

where (1) holds for independence of the par (i,j) and the symmetry of the problem. Substituting these expressions in the right side of equation (4.6), we recover the predictive hazard ratio given in (4.5).

Note that p_c is the probability for a pair of being concordant, and p_{nc} is the probability for a pair being discordant, and $p_{nc} = 1-p_c$. The previous expression can be rewritten by $\theta(s,t) = p_c/1-p_c$, and, ailing p_c,

$$p_c = \frac{\theta(s,t)}{\theta(s,t)+1}.$$

We then develop τ^*:

$$\begin{aligned}
\tau^*(s,t) &= E[\text{sgn}(T_{1i}-T_{1j})(T_{2i}-T_{2j})|\widetilde{T}_{1ij}=s,\widetilde{T}_{2ij}=t] = p_c - p_{nc} = 2p_c - 1\\
&= 2\frac{\theta(s,t)}{\theta(s,t)+1} - 1 = \frac{\theta(s,t)-1}{\theta(s,t)+1}
\end{aligned}$$

□

B.2 Equivalence of Clayton's copula model

Day *et al.* (1997) showed that model

$$S(s,t) = P(T_1 > s, T_2 > t) = \left(S_1(s)^{1-\alpha} + S_2(t)^{1-\alpha} - 1\right)^{1/(1-\alpha)}. \tag{B.3}$$

is equivalent to a model where the predictive hazard ratio is constant and equal to the copula association parameter, that is

$$\theta(s,t) = \frac{\lambda_2(t|T_1 = s)}{\lambda_2(t|T_1 > s)} = \alpha \qquad \forall (s,t) \in \mathcal{D}_1, \tag{B.4}$$

where $\lambda_2(t|A)$ is the hazard function of T_2 given event A occurs.

Here are some hints on the proof: to see that model (B.3), implies model (B.4), first note that under the copula model, the joint density function $f(s,t)$ of (T_1, T_2) is given by

$$f(s,t) = \frac{\partial^2 S(s,t)}{\partial s \partial t} = \alpha D^{\frac{2\alpha-1}{1-\alpha}} \frac{f_1(s) f_2(t)}{S_1(s)^\alpha S_2(t)^\alpha},$$

where $D = \{S_1(s)^{1-\alpha} + S_2(t)^{1-\alpha} - 1\}$ and the marginal density functions are given by $f_1(s) = -dS_1(s)/ds$ and $f_2(t) = -dS_2/dt$.

Now consider the expression (4.5) of the cross-ratio function, and express the joint survival and density functions in terms of the copula model:

$$\theta(s,t) = \frac{f(s,t)S(s,t)}{\int_t^\infty f(s,v)dv \int_s^\infty f(u,t)du} = \frac{\left[\alpha D^{\frac{2\alpha-1}{1-\alpha}} f_1(s) f_2(t) / \left(S_1(s)^\alpha S_2(t)^\alpha\right)\right] \left[D^{\frac{1}{1-\alpha}}\right]}{\left[D^{\frac{\alpha}{1-\alpha}} (-f_1(s))/S_1(s)^\alpha\right] \left[D^{\frac{\alpha}{1-\alpha}} (-f_2(t))/S_2(t)^\alpha\right]} = \alpha$$

for $(s,t) \in \mathcal{D}_1$. The proof that the inverse (B.4) $\Rightarrow$ (B.3) is valid can be found in Day *et al.* (1997).

B.3 The expectation of the concordance indicator

Proof of Proposition 4.3.

$$\begin{aligned}
&\mathrm{E}\Big[\Delta_{ij}\Big] = P\Big(\Delta_{ij}=1\Big) = P\Big((T_{1i}-T_{1j})(T_{2i}-T_{2j})>0\Big) = P\Big(T_{1i}>T_{1j},T_{2i}>T_{2j}\Big)\\
&+P\Big(T_{1i}<T_{1j},T_{2i}<T_{2j}\Big) = 2P\Big(T_{1i}>T_{1j},T_{2i}>T_{2j}\Big)\\
&=2\int_0^\infty\int_0^\infty P(T_{1i}>x,T_{2i}>y)f(x,y)dxdy = 2\int_0^\infty\int_0^\infty S(x,y)dS(x,y)\\
&\underset{(4.11)}{=} 2\int_0^\infty\int_0^\infty C_\alpha(S_1(x),S_2(y))dS(x,y) = (1)
\end{aligned}$$

Now, given the change of variable $u=S_1(x), v=S_2(y)$ and the following functions,

$$\begin{aligned}
\frac{\partial C_\alpha(u,v)}{\partial u} &= C_\alpha(u,v)^\alpha u^{-\alpha}\\
\frac{\partial C_\alpha(u,v)}{\partial v} &= C_\alpha(u,v)^\alpha v^{-\alpha}\\
dC_\alpha(u,v) &= \frac{\partial^2 C_\alpha(u,v)}{\partial u\partial v} = \alpha C_\alpha(u,v)^{2\alpha-1}u^{-\alpha}v^{-\alpha}
\end{aligned}$$

we can solve integral (1), which is equivalent to

$$\begin{aligned}
&2\int_0^1\int_0^1 C_\alpha(u,v)\Big[\alpha C_\alpha(u,v)^{2\alpha-1}u^{-\alpha}v^{-\alpha}dudv\Big] = \frac{2\alpha}{\alpha+1}\int_0^1 u^{-\alpha}\Big[C_\alpha(u,1)^{\alpha+1}-C_\alpha(u,0)^{\alpha+1}\Big]du\\
&\underset{C_\alpha \text{ is copula}}{=} \frac{2\alpha}{\alpha+1}\int_0^1 udu = \frac{\alpha}{\alpha+1}
\end{aligned}$$

B.4 The expected concordance

Let (T_{1i}, T_{2i}) and (T_{1j}, T_{2j}) be the bivariate times of two independent individuals (i, j). Consider the expectation of the concordance indicator for two individuals, given their observed data $\mathcal{H}_{ij} = \{(a_i, b_i, y_i, \delta_{1i}, \delta_{2i}), (a_j, b_j, y_j, \delta_{1j}, \delta_{2j})\}$,

$$Z_{ij} = E\left[\Delta_{ij}|\mathcal{H}_{ij}\right] = P\left(\Delta_{ij} = 1|\mathcal{H}_{ij}\right) = \frac{P\left(\Delta_{ij} = 1, \mathcal{H}_{ij}\right)}{P\left(\mathcal{H}_{ij}\right)}.$$

The previous expression is developed obtaining

$$Z_{ij} = \frac{1}{P(\mathcal{H}_{ij})}\left(\delta_{2i}\delta_{2j}P_1(i,j) + \delta_{2i}(1-\delta_{2j})P_2(i,j) + (1-\delta_{2i})\delta_{2j}P_2(j,i)\right).$$

Under Clayton's copula model for the survival joint function, the following functions are well defined in the upper wedge:

$$\begin{aligned} H(s,t) &= \frac{\partial S(s,t)}{\partial t} = -f_2(t)S_2(t)^{-\alpha}S(s,t)^{\alpha} \\ f(x,y) &= \frac{\partial^2 S(x,y)}{\partial x \partial y} = \alpha f_1(s)f_2(t)\left(S_1(s)S_2(t)\right)^{\alpha} S(s,t)^{2\alpha-1}. \end{aligned}$$

Define $\widetilde{a}_{ij} = \min(a_i, a_j)$, $a^*_{ij} = \max(a_i, a_j)$, $\widetilde{b}_{ij} = \min(b_i, b_j)$ and $b^*_{ij} = \max(b_i, b_j)$. The expression for $P_1(i,j)$ is given by

$$\begin{aligned} P_1(i,j) &= P(\Delta_{ij} = 1, \mathcal{H}_{ij}) = P(\Delta_{ij} = 1, T_{1i} \in (a_i, b_i], T_{1j} \in (a_j, b_j], T_{2i} = y_i, T_{2j} = y_j) \\ &= \int_{a_i}^{b_i} dx \int_{a_j}^{b_j} I((x-u)(y_i - y_j) > 0) f(x, y_i) f(u, y_j) du \\ &= I\Big(\widetilde{b}_{ij} < a^*_{ij}\Big)\left[\int_{a_i}^{b_i} dx \int_{a_j}^{b_j} I((a_i - b_j)(y_i - y_j) > 0) f(x, y_i) f(u, y_j)\right] \\ &+ I\Big(\widetilde{b}_{ij} > a^*_{ij}\Big)\Bigg[\\ &\quad I(y_i < y_j)\left\{I(a_i < a_j)\int_{a_i}^{a_j} dx \int_{a_j}^{b_j} f(x, y_i) f(u, y_j) du + \int_{a^*_{ij}}^{\widetilde{b}_{ij}} dx \int_{x}^{b_j} f(x, y_i) f(u, y_j) du\right\} \\ &+ \quad I(y_i > y_j)\left\{I(b_i > b_j)\int_{b_j}^{b_i} dx \int_{a_j}^{b_j} f(x, y_i) f(u, y_j) du + \int_{a^*_{ij}}^{\widetilde{b}_{ij}} dx \int_{a_j}^{x} f(x, y_i) f(u, y_j) du\right\}\Bigg] \\ &\underset{(1)}{=} I\Big(\widetilde{b}_{ij} < a^*_{ij}\Big)\left[I\big((a_i - a_j)(y_i - y_j) > 0\big)\big(H(b_i, y_i) - H(a_i, y_i)\big)\big(H(b_j, y_j) - H(a_j, y_j)\big)\right] \\ &+ I\Big(\widetilde{b}_{ij} > a^*_{ij}\Big)\Bigg[I(y_i < y_j)\Big\{I(a_i < a_j)\big(H(a_j, y_i) - H(a_i, y_i)\big)\big(H(b_j, y_j) - H(a_j, y_j)\big) \\ &\qquad + \int_{a^*_{ij}}^{\widetilde{b}_{ij}} \big(H(b_j, y_j) - H(x, y_j)\big) f(x, y_i) dx\Big\} \\ &\quad + I(y_i > y_j)\Big\{I(b_i > b_j)\big(H(b_i, y_i) - H(b_j, y_i)\big)\big(H(b_j, y_j) - H(a_j, y_j)\big) \end{aligned}$$

$$
\begin{aligned}
&\quad + \int_{a_{ij}^*}^{\widetilde{b}_{ij}} \big(H(x,y_j) - H(a_j,y_j)\big) f(x,y_i)dx \Big\}\Big] \\
&= I\Big(\widetilde{b}_{ij} < a_{ij}^*\Big)\Big[I((a_i-a_j)(y_i-y_j)>0)P(\mathcal{H}_{ij})\Big] \\
&+ I\Big(\widetilde{b}_{ij} > a_{ij}^*\Big)\Big[I(y_i<y_j)\Big\{I(a_i<a_j)\big(H(a_j,y_i)-H(a_i,y_i)\big)\big(H(r_j,y_j)-H(a_j,y_j)\big) \\
&\qquad + H(b_j,y_j)\big(H(\widetilde{b}_{ij},y_i)-H(a_{ij}^*,y_i)\big) - \int_{a_{ij}^*}^{\widetilde{b}_{ij}} H(x,y_j)f(x,y_i)dx\Big\} \\
&\quad + I(y_i>y_j)\Big\{I(b_i>b_j)\big(H(b_i,y_i)-H(b_j,y_i)\big)\big(H(b_j,y_j)-H(a_j,y_j)\big) \\
&\qquad + \int_{a_{ij}^*}^{\widetilde{b}_{ij}} H(x,y_j)f(x,y_i)dx - H(a_j,y_j)\big(H(\widetilde{b}_{ij},y_i)-H(a_{ij}^*,y_i)\big)\Big\}\Big],
\end{aligned}
$$

where (1) is justified because

$$
\int_a^b \frac{\partial^2 S(x,y)}{\partial x \partial y}dx = \frac{\partial S(x,y)}{\partial y}\Big|_{x=a}^{x=b} = H(b,y) - H(a,y).
$$

Notice, in the expression, that $\widetilde{b}_{ij} < a_{ij}^*$ indicates that the observed intervals in T_1 do not overlap. These scenarios is particularly simple, because $Z_{ij} = \Delta_{ij}$: the concordance indicator can be defined.

The expression for $P_2(i,j)$ is given by

$$
\begin{aligned}
&P_2(i,j) = P(\Delta_{ij}=1,\mathcal{H}_{ij}) = P(\Delta_{ij}=1, T_{1i}\in(a_i,b_i], T_{1j}\in(a_j,b_j], T_{2i}=y_i, T_{2j}\in(y_j,\infty)) \\
&= \int_{y_j}^{\infty} dv \int_{a_i}^{b_i} dx \int_{a_j}^{b_j} I((x-u)(y_i-v)>0) f(x,y_i)f(u,v)du \\
&= I(y_i<y_j)\Big[I\Big(\widetilde{b}_{ij} < a_{ij}^*\Big)\Big\{\int_{y_j}^{\infty} dv \int_{a_i}^{b_i} dx \int_{a_j}^{b_j} I(a_i<a_j) f(x,y_i)f(u,v)du\Big\} \\
&+ I\Big(\widetilde{b}_{ij} > a_{ij}^*\Big)\Big\{\int_{y_j}^{\infty}\Big(I(a_i<a_j)\int_{a_i}^{a_j} dx \int_{a_j}^{b_j} f(x,y_i)f(u,v)du \\
&\qquad + \int_{a_{ij}^*}^{\widetilde{b}_{ij}} dx \int_{x}^{b_j} f(x,y_i)f(u,v)du\Big)dv\Big\}\Big] \\
&= I(y_i<y_j)\Big[I\Big(\widetilde{b}_{ij} < a_{ij}^*\Big)\Big\{I(a_i<a_j)\int_{y_j}^{\infty} dv \int_{a_i}^{b_i} dx \int_{a_j}^{b_j} f(x,y_i)f(u,v)du\Big\} \\
&+ I\Big(\widetilde{b}_{ij} > a_{ij}^*\Big)\Big\{I(a_i<a_j)\Big(H(a_j,y_i)-H(a_i,y_i)\Big)\left(\int_{y_j}^{\infty}\int_{a_j}^{b_j} f(u,v)dudv\right) \\
&\quad + \int_{a_{ij}^*}^{\widetilde{b}_{ij}}\left(\int_{y_j}^{\infty}\int_{x}^{b_j} f(u,v)dudv\right) f(x,y_i)dx\Big\}\Big] \\
&= I(y_i<y_j)\Big[I\Big(\widetilde{b}_{ij} < a_{ij}^*\Big)\Big\{I(a_i<a_j)P(\mathcal{H}_{ij})\Big\} \\
&\qquad + I\Big(\widetilde{b}_{ij} > a_{ij}^*\Big)\Big\{I(a_i<a_j)\Big(H(a_j,y_i)-H(a_i,y_i)\Big)\left(S(a_j,y_j)-S(b_j,y_j)\right)
\end{aligned}
$$

$$+\int_{a_{ij}^*}^{\widetilde{b}_{ij}} \left(S(x,y_j) - S(b_j,y_j)\right) f(x,y_i)dx\Bigg\}\Bigg]$$

$$= I(y_i < y_j)\Bigg[I\Big(\widetilde{b}_{ij} < a_{ij}^*\Big) I(a_i < a_j) P(\mathcal{H}_{ij})$$

$$+ I\Big(\widetilde{b}_{ij} > a_{ij}^*\Big)\Bigg\{I(a_i < a_j)\Big(H(a_j,y_i) - H(a_i,y_i)\Big)\left(S(a_j,y_j) - S(b_j,y_j)\right)$$

$$+\int_{a_{ij}^*}^{\widetilde{b}_{ij}} \left(S(x,y_j) - S(b_j,y_j)\right) f(x,y_i)dx\Bigg\}\Bigg]$$

$P_2(i,j)$ can only be estimated for comparable pairs, and in these scenario, it is necessary that, if $\delta_{2i} = 1$ and $\delta_{2j} = 0$, then $y_i < y_j$.

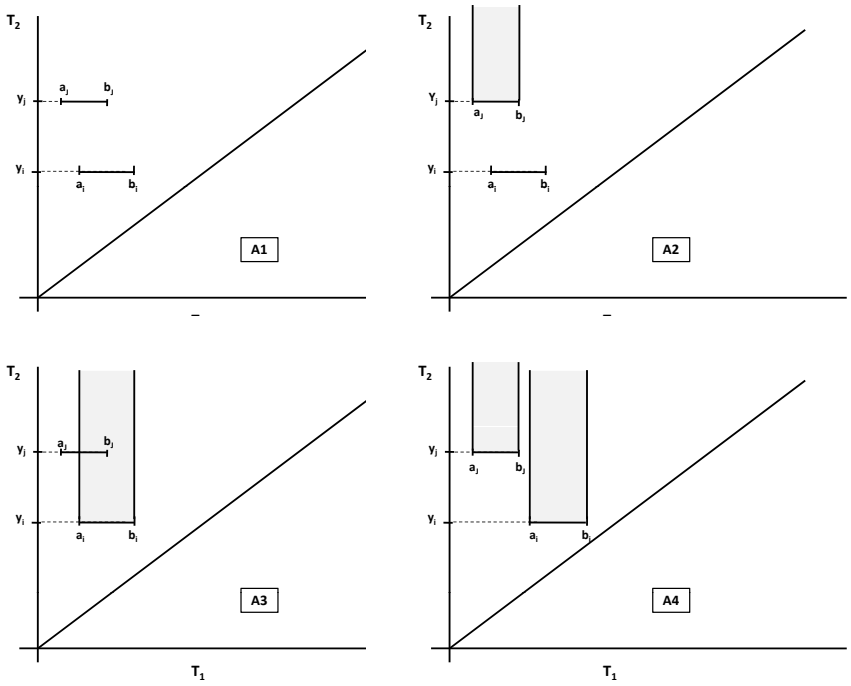


Figure B.1: Case A: Comparable and not comparable pairs when $\delta_{1i} + \delta_{1j} = 2$.

B.5 The comparable sample

B.5.1 Case A: when $\delta_{1i} = 1$, $\delta_{1j} = 1$

We face to fourth different scenarios, shown in Figure B.1:

A.1 $\delta_{2i} + \delta_{2j} = 2 \rightarrow T_{2i} = y_j$, $T_{2j} = y_j \rightarrow$ the pair (i, j) is comparable.

A.2 $\delta_{2i} = 1$, $\delta_{2j} = 0$ and $y_i < y_j \rightarrow T_{2i} = y_i$, $y_j < T_{2j} \rightarrow$ the pair (i, j) is comparable.

A.3 $\delta_{2i} = 0$, $\delta_{2j} = 1$ and $y_i < y_j \rightarrow y_i < T_{2i}$, $y_j = T_{2j} \rightarrow$ the pair (i, j) is NOT comparable.

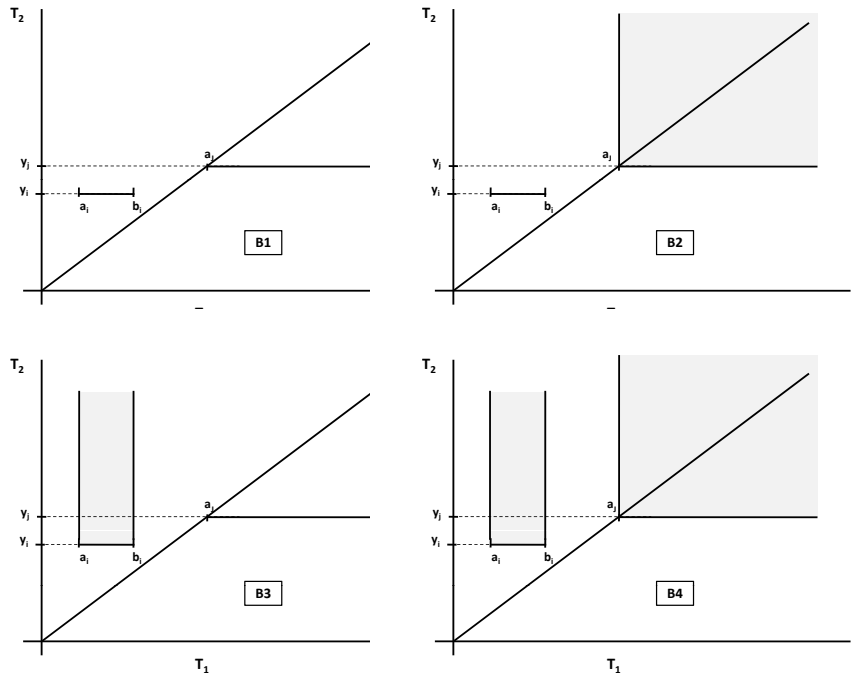
A.4 $\delta_{2i} = 0$, $\delta_{2j} = 0 \rightarrow y_i < T_{2i}$, $y_j < T_{2j} \rightarrow$ the pair (i, j) is NOT comparable.

Notice that the previous scenarios are valid independently of the relative position of the intervals, given that $\delta_{1i} = \delta_{1j} = 1$. They can overlap or not, but this does not change the comparability status. In addition, cases A.2 and A.3 are also valid for the symmetric situation, that is (i) $\delta_{2i} = 0$, $\delta_{2j} = 1$ and $y_i > y_j$ and (ii) $\delta_{2i} = 1$, $\delta_{2j} = 0$ and $y_i > y_j$.

B.5.2 Case B: $\delta_{1i} = 1$, $\delta_{1j} = 0$, and $y_i < y_j$

Assume now that $\delta_{1i} = 1$, $\delta_{1j} = 0$ and consider the setting where $y_i < y_j$, shown in Figure B.2:

B.1 $\delta_{2i} + \delta_{2j} = 2 \rightarrow y_i = T_{2i}$, $y_j = T_{2j}$, and $T_{1i} < T_{1j} \rightarrow$ the pair (i, j) is comparable.



CAS B3 CAS B4

Figure B.2: Case B: Comparable and not comparable pairs when $\delta_{1i} = 1$, $\delta_{1j} = 0$, $y_i < y_j$.

B.2 $\delta_{2i} = 1$, $\delta_{2j} = 0 \rightarrow y_i = T_{2i}$, $y_j < T_{2j}$, $T_{1i} < T_{1j} \rightarrow$ the pair (i, j) is comparable.

B.3 $\delta_{2i} = 0$, $\delta_{2j} = 1 \rightarrow y_i < T_{2i}$, $y_j = T_{2j} \rightarrow$ the pair (i, j) is NOT comparable.

B.4 $\delta_{2i} = 0$, $\delta_{2j} = 0 \rightarrow y_i < T_{2i}$, $y_j < T_{2j} \rightarrow$ the pair (i, j) is NOT comparable.

The previous settings are also valid for their symmetric counterparts, that is, when $\delta_{1i} = 0$, $\delta_{1j} = 1$ and $y_j < y_i$.

B.5.3 Cases C, D and E: When $\delta_{1i} = 1$, $\delta_{1j} = 0$, and $y_i > y_j$

Consider now that $\delta_{1i} = 1$, $\delta_{1j} = 0$ and $y_i > y_j$. In this setting, we must distinguish between three cases: (C) $b_i \leq a_j$, (D) $a_i < a_j$ and $b_i > a_j$, and (E) $a_i > a_j$. Case (C) is shown in Figure B.3:

C.1 $\delta_{2i} + \delta_{2j} = 2 \rightarrow y_i = T_{2i}$, $y_j = T_{2j}$, and $T_{1i} < T_{1j} \rightarrow$ the pair (i, j) is comparable.

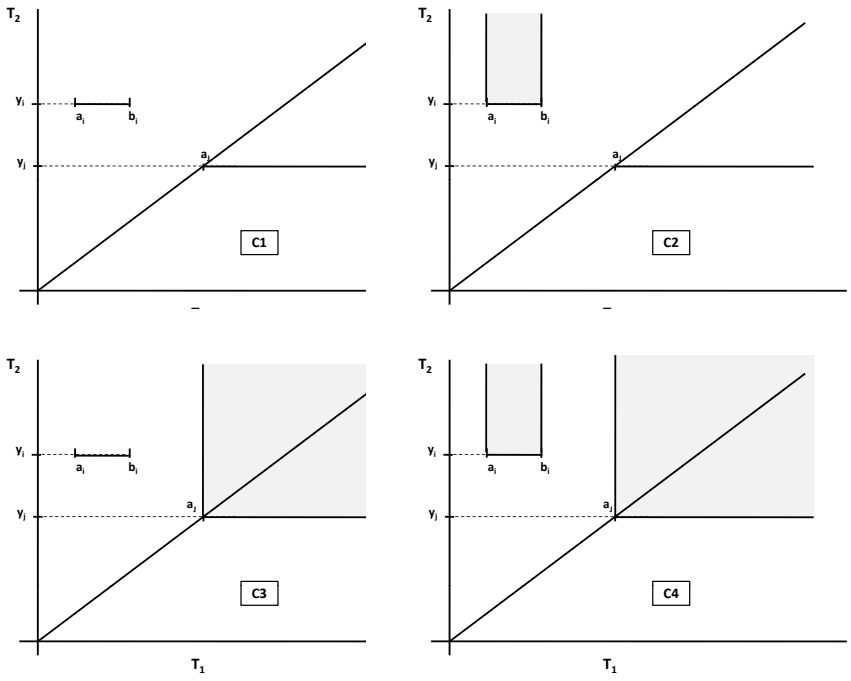
C.2 $\delta_{2i} = 0$, $\delta_{2j} = 1 \rightarrow y_i < T_{2i}$, $y_j = T_{2j}$, $y_i > y_j \rightarrow$ the pair (i, j) is comparable.

C.3 $\delta_{2i} = 1$, $\delta_{2j} = 0 \rightarrow y_i = T_{2i}$, $y_j < T_{2j}$, $y_i > y_j \rightarrow$ the pair (i, j) is NOT comparable.

C.4 $\delta_{2i} = 0$, $\delta_{2j} = 0 \rightarrow y_i < T_{2i}$, $y_j < T_{2j} \rightarrow$ the pair (i, j) is NOT comparable.

Figures B.4 and B.5 show the second (D) and third (E) cases, which can never give raise to a comparable pair. Indeed, in all the scenarios, overlapped area is not contained in $\mathcal{D}_1$, and since Clayton's copula model cannot be valid outside, Z_{ij} cannot be computed.

As before, the comparable distribution among cases C,D and E is also valid for their symmetric counterparts, that is, when $\delta_{1i} = 0$, $\delta_{1j} = 1$, $y_j > y_i$ and (C) $b_j < a_i$, (D) $a_j < a_i$, $a_i < b_j$, and (E) $a_j > a_i$.



CAS C2 CAS C4

Figure B.3: Case C: Comparable and not comparable pairs when $\delta_{1i} = 1$, $\delta_{1j} = 0$, $y_i > y_j$ and $b_i < a_j$.

B.5.4 Case F: $\delta_{1i} = 0$, $\delta_{1j} = 0$

This setting, which we refer to as case F, always produces non comparable pairs, no matter the values of δ_{2i} and δ_{2j}: no T_1 event is observed in the upper wedge $\mathcal{D}_1$ and the underlying bivariate model outside the region $\mathcal{D}_1$ is unknown. Figure B.6 contain the complete description of cases.

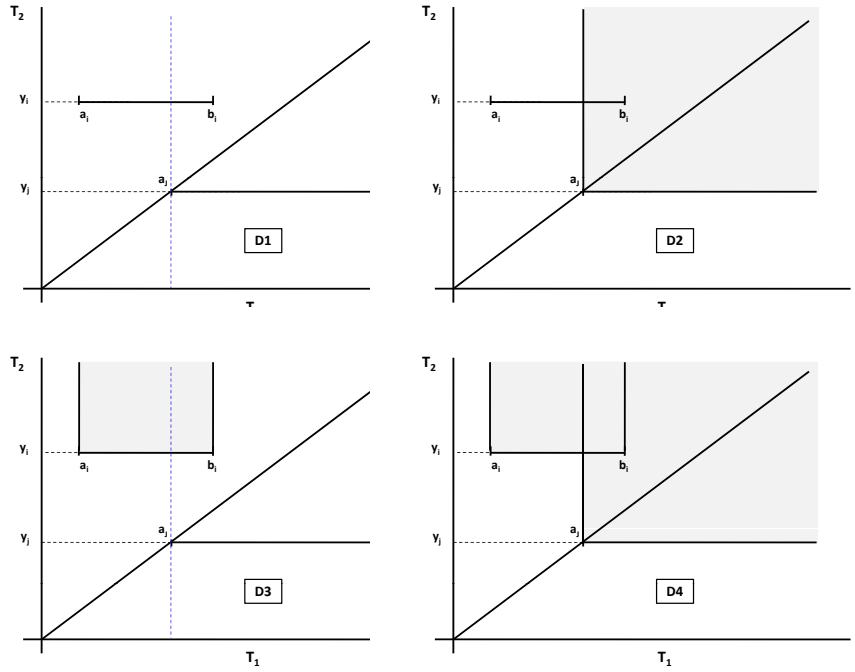
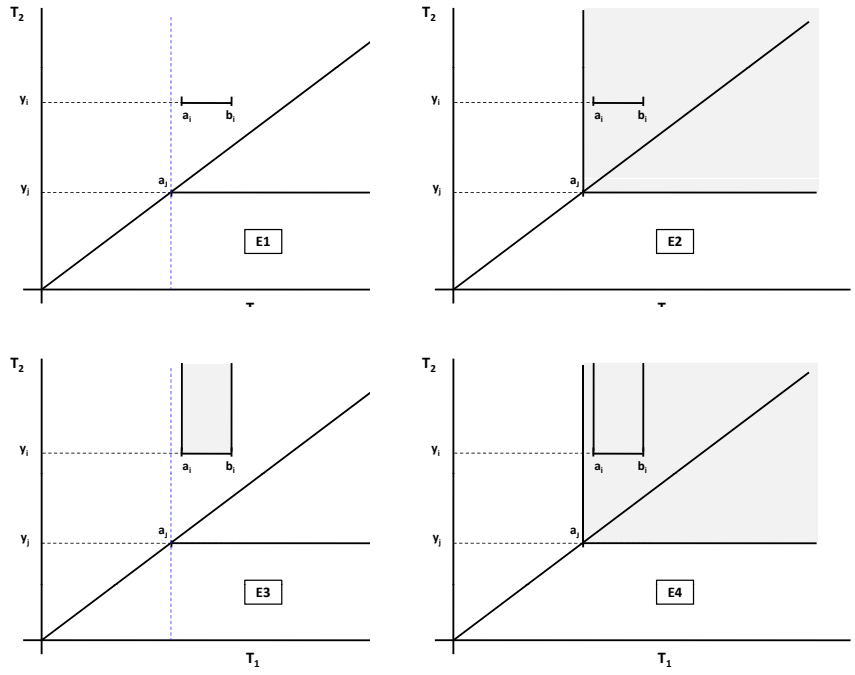


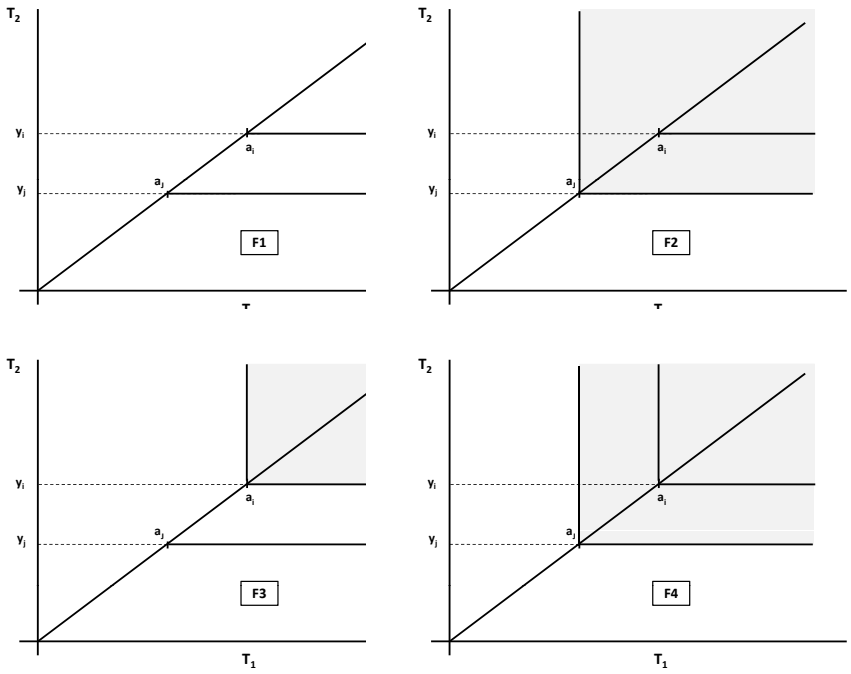
Figure I

CAS D3 CAS D4



CAS F3 CAS F4

Figure B.5: Case E: Non comparable pairs when $\delta_{1i} = 1$, $\delta_{1j} = 0$, $y_i > y_j$ and $a_i > a_j$.



CAS G3 CAS G4

Figure B.6: Case F: Non comparable pairs when $\delta_{1i} = 0$ and $\delta_{1j} = 0$.

B.6 Equivalence on the conditions of comparability

In this Section we prove Proposition 6.4. To do so, we must see that the conditions in Proposition 6.3 (Conditions 1) and the conditions in Proposition 6.4 (Conditions 2) are equivalent.

B.6.1 Conditions 1 $\Longrightarrow$ Conditions 2:

Let $\mathcal{H}_{ij} = \{(a_i, b_i, \delta_{1i}, y_i, \delta_{2i}), (a_j, b_j, \delta_{1j}, y_j, \delta_{2j})\}$ be the observed data for the pair (i, j). We assume that Conditions 1 hold:

1. $\delta_{1i} + \delta_{1j} \geq 1$ and $\delta_{2i} + \delta_{2j} \geq 1$.
2. If $\delta_{1i} + \delta_{1j} = 1$, $\delta_{1i} b_i + \delta_{1j} b_j < (1 - \delta_{1i}) a_i + (1 - \delta_{1j}) a_j$.
3. If $\delta_{2i} + \delta_{2j} = 1$, $\delta_{2i} y_i + \delta_{2j} y_j < (1 - \delta_{2i}) y_i + (1 - \delta_{2j}) y_j$.

To show $\widetilde{T}_{1ij} < \widetilde{C}_{ij}$, from Condition 1, $\delta_{1i} + \delta_{1j} \geq 1$, so we have $\delta_{1i} + \delta_{1j} = 2$ or $\delta_{1i} + \delta_{1j} = 1$:

- If $\delta_{1i} + \delta_{1j} = 2$ and, for instance, $\widetilde{T}_{1ij} = T_{1i}$:

$$\left.\begin{array}{l} \delta 1i = 1 \Rightarrow \widetilde{T}_{1ij} = T_{1i} \leq \min(T_{2i}, C_i) \leq C_i \\ \widetilde{T}_{1ij} = T_{1i} < T_{1j} \underset{\delta 1j=1}{\leq} \min(T_{2j}, C_j) \leq C_j \end{array}\right\} \Rightarrow \widetilde{T}_{1ij} < \widetilde{C}_{ij}$$

  The case $\widetilde{T}_{1ij} = T_{1j}$ is obtained similarly.

- If $\delta_{1i} + \delta_{1j} = 1$, and, for instance, $\delta_{1i} = 1$, then from Condition 2,

$$\left.\begin{array}{l} \delta 1i = 1 \Rightarrow T_{1i} \leq \min(T_{2i}, C_i) \leq C_i \\ \delta_{1i} = 1 \Rightarrow T_{1i} \leq b_i \underset{Cond2}{\leq} a_j = y_j \underset{\delta_{1j}=0}{=} \min(T_{2j}, C_j) \leq C_j \\ \delta_{1j} = 0 \Rightarrow T_{1j} > \min(T_{2j}, C_j) \end{array}\right\} \Rightarrow \widetilde{T}_{1ij} = T_{1i} < \widetilde{C}_{ij}$$

  The case $\delta_{1j} = 1$ is obtained similarly.

To show $\widetilde{T}_{2ij} < \widetilde{C}_{ij}$, from Condition 1, $\delta_{2i} + \delta_{2j} \geq 1$, so we have $\delta_{2i} + \delta_{2j} = 2$ or $\delta_{2i} + \delta_{2j} = 1$:

- If $\delta_{2i} + \delta_{2j} = 2$ and, for instance, $\widetilde{T}_{2ij} = T_{2i}$:

$$\left.\begin{array}{l} \delta 2i = 1 \Rightarrow \widetilde{T}_{2ij} = T_{2i} \leq C_i \\ \widetilde{T}_{2ij} = T_{2i} < T_{2j} \underset{\delta 2j=1}{\leq} C_j \end{array}\right\} \Rightarrow \widetilde{T}_{2ij} < \widetilde{C}_{ij}$$

  The case $\widetilde{T}_{2ij} = T_{2j}$ is obtained similarly.

- If $\delta_{2i} + \delta_{2j} = 1$, and, for instance, $\delta_{2i} = 1$, then from Condition 3,

$$\left.\begin{array}{l} \delta 2i = 1 \Rightarrow T_{2i} \leq C_i \\ T_{2i} = y_i \underset{Cond3}{\leq} y_j \underset{\delta_{2j}=0}{=} C_j < T_{2j} \end{array}\right\} \Rightarrow \widetilde{T}_{2ij} = T_{2i} < \widetilde{C}_{ij}$$

  The case $\delta_{2j} = 1$ is obtained similarly.

Finally, to obtain $\widetilde{R}_{ij} < \widetilde{T}_{2ij}$,

- If $\delta_{1i} + \delta_{1j} = 2$ and $\delta_{2i} + \delta_{2j} = 2$, and assume $b_i < b_j$:

$$\left.\begin{array}{l} R_i = b_i \leq y_i = T_{2i} \\ R_i = b_i < b_j = R_j \leq T_{2j} = y_j \end{array}\right\} \Rightarrow \widetilde{R}_{ij} < \widetilde{T}_{2ij}$$

  The case $b_i > b_j$ is done similarly.

- If $\delta_{2i} + \delta_{2j} = 2$ and $\delta_{1i} + \delta_{1j} = 1$, for instance $\delta_{1i} = 1$:

$$\left.\begin{array}{l} \widetilde{R}_{ij} = \min(R_i, R_j = \infty) = R_i \\ R_i = b_i \underset{Cod2}{\leq} y_j = T_{2j} \\ R_i = b_i \leq T_{2i} = y_i \end{array}\right\} \Rightarrow \widetilde{R}_{ij} < \widetilde{T}_{2ij}$$

  The case δ_{1j} is done similarly.

- If $\delta_{1i} + \delta_{1j} = 2$ and $\delta_{2i} + \delta_{2j} = 1$, and assume $\delta_{2i} = 1$ and $b_i < b_j$:

$$\left.\begin{array}{l} \delta_{2i} = 1 \Rightarrow T_{2i} = y_i \underset{Cond3}{\leq} y_j = C_j < T_{2j} \\ \widetilde{R}_{ij} = R_i = b_i \leq T_{2i} = y_i \end{array}\right\} \Rightarrow \widetilde{R}_{ij} < \widetilde{T}_{2ij}$$

If $\delta_{2i} = 1$ and $b_i > b_j$:

$$\left.\begin{aligned} &\delta_{2i} = 1 \Rightarrow T_{2i} = y_i \underset{Cond3}{\leq} y_j = C_j < T_{2j} \\ &\widetilde{R}_{ij} = R_j = b_j < b_i \leq T_{2i} = y_i \end{aligned}\right\} \Rightarrow \widetilde{R}_{ij} < \widetilde{T}_{2ij}$$

The case $\delta_{2j} = 1$ is done similarly.

- If $\delta_{1i} + \delta_{1j} = 1$ and $\delta_{2i} + \delta_{2j} = 1$, and assume $\delta_{1i} = \delta_{2i} = 1$:

$$\left.\begin{aligned} &\delta_{2i} = 1 \Rightarrow T_{2i} = y_i \underset{Cond3}{\leq} y_j = C_j < T_{2j} \\ &\delta_{1i} = 1 \Rightarrow \widetilde{b}_{ij} = R_i = b_i \underset{Cond2}{\leq} y_j = C_j < T_{2j} \\ &\widetilde{b}_{ij} = R_i = b_i \leq y_i = T_{2i} \end{aligned}\right\} \Rightarrow \widetilde{R}_{ij} < \widetilde{T}_{2ij}$$

The case $\delta_{1j} = \delta_{2j} = 1$ is done similarly.

If $\delta_{1i} = 1$ and $\delta_{2j} = 1$,

$$\left.\begin{aligned} &\delta_{2j} = 1 \Rightarrow T_{2j} = y_j \underset{Cond3}{\leq} y_i = C_i < T_{2i} \\ &\delta_{1i} = 1 \Rightarrow \widetilde{b}_{ij} = R_i = b_i \underset{Cond2}{\leq} y_j = C_j < T_{2j} \end{aligned}\right\} \Rightarrow \widetilde{R}_{ij} < \widetilde{T}_{2ij}$$

The case $\delta_{1j} = 1$ and $\delta_{2i} = 1$ is done similarly.

B.6.2 Conditions 2 $\Longrightarrow$ Conditions 1

Assume now the following:

(i) $\widetilde{T}_{1ij} < \widetilde{C}_{ij}$,

(ii) $\widetilde{T}_{2ij} < \widetilde{C}_{ij}$, and

(iii) $\widetilde{R}_{ij} < \widetilde{T}_{2ij}$.

To show that $\delta_{1i} + \delta_{1j} \geq 1$ and $\delta_{2i} + \delta_{2j} \geq 1$, note that from (i) and (ii):

$$\widetilde{T}_{1ij} < \widetilde{C}_{ij} \Rightarrow \widetilde{T}_{1ij} < C_i, \widetilde{T}_{1ij} < C_j$$
$$\widetilde{T}_{2ij} < \widetilde{C}_{ij} \Rightarrow \widetilde{T}_{2ij} < C_i, \widetilde{T}_{1ij} < C_j$$

In addition, in the semi-competing risks setting, $\widetilde{T}_{1ij} < \widetilde{T}_{2ij}$, so $\widetilde{T}_{1ij} < T_{2i}$, $\widetilde{T}_{1ij} < T_{2j}$. Now, if

$$\widetilde{T}_{1ij} = T_{1i} \Rightarrow T_{1i} < C_i, T_{1i} < T_{2i} \Rightarrow \delta_{1i} = 1 \Rightarrow \delta_{1i} + \delta_{1j} \geq 1$$
$$\widetilde{T}_{2ij} = T_{2i} \Rightarrow T_{2i} < C_i \Rightarrow \delta_{2i} = 1 \Rightarrow \delta_{2i} + \delta_{2j} \geq 1$$

To show that, if $\delta_{1i} + \delta_{1j} = 1$, $\delta_{1i} b_i + \delta_{1j} b_j < (1 - \delta_{1i}) a_i + (1 - \delta_{1j}) a_j$, assume that $\delta_{1i} = 1$:

$$T_{1i} \in (a_i, b_i], b_i \leq y_i$$
$$T_{1j} \in (a_j, \infty), a_j = y_j$$

- If $\delta_{2i} + \delta_{2j} = 2$,
$$b_i = \widetilde{b}_{ij} \underset{(iii)}{<} \widetilde{T}_{2ij} < T_{2j} = y_j = a_j.$$

- If $\delta_{2i} = 1$,
$$b_i = \widetilde{b}_{ij} \underset{(iii)}{<} \widetilde{T}_{2ij} < \widetilde{C}_{ij} < C_j \underset{\delta_{1j}=0}{=} a_j.$$

- If $\delta_{2j} = 1$,
$$b_i = \widetilde{b}_{ij} \underset{(iii)}{<} \widetilde{T}_{2ij} < T_{2j} = y_j = a_j.$$

The case $\delta_{1j} = 1$ is done similarly.

Finally, to show that, if $\delta_{2i} + \delta_{2j} = 1$, $\delta_{2i} y_i + \delta_{2j} y_j < (1 - \delta_{2i}) y_i + (1 - \delta_{2j}) y_j$, assume that $\delta_{2i} = 1$:

$$\delta_{2i} = 1 \Rightarrow T_{2i} \leq C_i, T_{2i} = y_i$$
$$\delta_{2j} = 0 \Rightarrow T_{2j} > C_j, y_j = C_j$$
$$\widetilde{T}_{2ij} \underset{(ii)}{<} \widetilde{C}_{ij} < C_j.$$

Therefore, $\widetilde{T}_{2ij} = T_{2i} = y_i = \delta_{2i} y_i + \delta_{2j} y_j = y_i < C_j = Y_j$.

The case $\delta_{2j} = 1$ is done similarly.

B.7 U-statistics

Suppose that $X_1, \ldots, X_n$ are i.i.d. with cumulative density function F, where F is completely unspecified and restricted only to general conditions such as continuity or existence of moments. Denote by $\mathcal{F}$ the nonparametric family of such functions. The parameter $\theta = \theta(F)$ to be estimated is a real-valued function defined over $\mathcal{F}$. $\theta(F)$ is an estimable parameter within $\mathcal{F}$ if, for some integer r and a real-valued measurable function $\phi(x_1, \ldots, x_r)$ of r arguments,

$$\mathrm{E}_F[\phi(X_1, \ldots, X_r)] = \theta(F) \quad \forall F \in \mathcal{F}, \tag{B.5}$$

when $X_1, \ldots, X_r$ are i.i.d with distribution F. That is, there exists an unbiased estimator of $\theta(F)$ based on r i.i.d. random variables distributed according to F. The smallest integer r with this property is called the degree of $\theta(P)$. In the following, we focus on the case for $r = 2$, though the results are valid for any finite r.

Without loss of generality, we can assume function ϕ to be symmetric, because if ϕ was not symmetric, one can build a symmetric function based on ϕ and satisfying (B.5) (Lehman (1999), chapter 6). Now, for a real-valued measurable function $\phi(x_1, x_2)$, and a sample $X_1, \ldots, X_n$ with $n \geq 2$ from distribution F, the U-statistic with kernel ϕ is defined as

$$U_n = \frac{1}{\binom{n}{2}} \sum_{C_{n,2}} \phi(X_{i_1}, X_{i_2}),$$

where the summation is over the set $C_{n,2}$ of all $\binom{n}{2}$ combinations of m integers, $i_1 < i_2$ chosen from $(1, 2, \ldots, n)$. Clearly, $\phi(X_{i_1}, X_{i_2})$ is an unbiased estimator of $\theta(F)$ for any 2-tuple $1 \leq i_1 < i_2 \leq n$, and therefore, U_n is also an unbiased estimator of $\theta(F)$. In fact, it is the only symmetric estimator unbiased for all F for which $\theta(F)$ exists, and it can be shown to have smaller variance than any other such unbiased estimator. The following theorem summarizes the main asymptotical properties of U-statistics.

Theorem B.1. *A U-statistic U_n with a kernel ϕ of degree $r = 2$, satisfies:*

(a) The variance of the U-statistic U_n is given by

$$Var[U_n] = \sum_{i=1}^{2} \binom{2}{i} \binom{n-2}{2-i} \sigma_i^2 \Big/ \binom{n}{2}$$

where $\sigma_1^2 = Cov\big[\phi(X_1, X_2), \phi(X_1, X_2')\big]$ and $\sigma_2^2 = Var\big[\phi(X_1, X_2)\big]$, and X_1, X_2, X_2' are i.i.d according to F.

(b) If $\sigma_1^2 > 0$ and $\sigma_2^2 < \infty$ for all $i = 1, \ldots, n$, then

$$Var[\sqrt{n}U_n] \underset{n \to \infty}{\to} 2^2 \sigma_1^2$$

(c) The central limit theorem for U-statistics states that, if $0 < \sigma_1^2 < \infty$*, then as* $n \to \infty$*,*

$$\sqrt{n}(U_n - \theta) \underset{\mathcal{D}}{\to} \mathcal{N}(0, 2^2\sigma_1^2).$$

(d) If, in addition, $\sigma_2^2 < \infty$*, then also*

$$\frac{U_n - \theta}{\sqrt{Var[U]}} \underset{\mathcal{D}}{\to} \mathcal{N}(0, 1)$$

(e) The strong law of large numbers for U-statistics states that

$$\frac{1}{\binom{n}{2}} \sum_{C_{n,2}} \Big\{ \phi(X_{i_1}, X_{i_2}) - E\big[\phi(X_{i_1}, X_{i_2})\big] \Big\} \underset{a.s.}{\to} 0$$

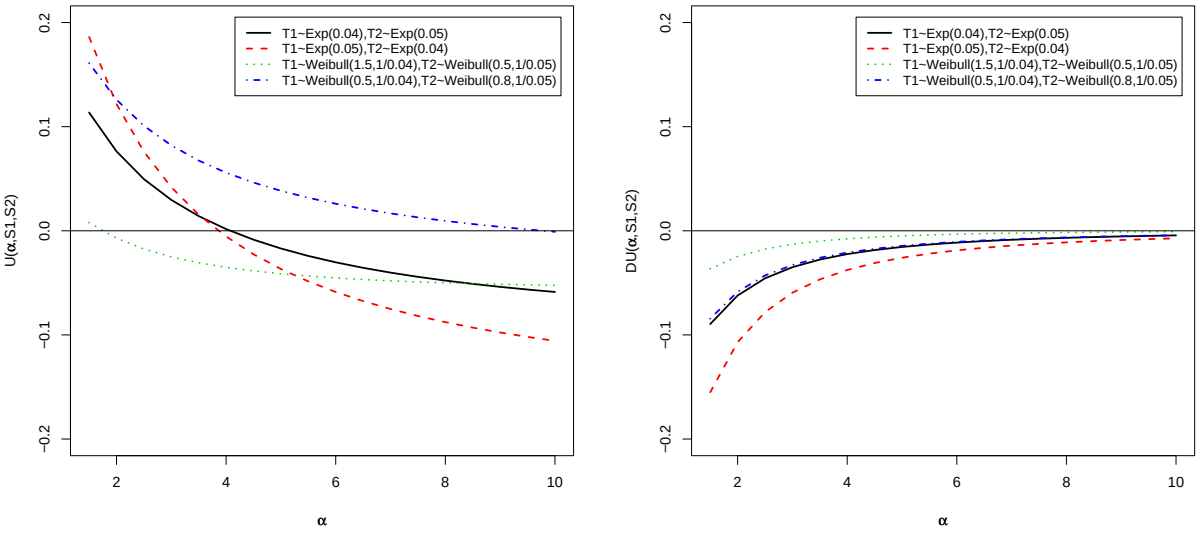
The previous results are defined for the case where $X_1, \ldots, X_n$ are i.i.d random variables, but they equally apply when the X's are i.i.d. random vectors.

B.8 Unicity of $U(\alpha) = 0$

When $S_1(\cdot)$ and $S_2(\cdot)$ are known, a estimates for α are obtained by the root of equations

$$U_1\big(\alpha, S_1(\cdot), S_2(\cdot)\big) = 0$$
$$U_2\big(\alpha, S_1(\cdot), S_2(\cdot)\big) = 0,$$

depending on the strategy to correct bias selected. The root of each equation is unique: for $S_1(\cdot)$ and $S_2(\cdot)$ known, $U_k\big(\alpha, S_1(\cdot), S_2(\cdot)\big)$ is a strictly decreasing function of α, $k = 1, 2$. This can be seen graphically: Figure B.7 contains (a) the function $U_k\big(\alpha, S_1(\cdot), S_2(\cdot)\big)$ for different choices of $S_1(\cdot)$ and $S_2(\cdot)$, where the monotonicity is clear, as well as (b) a numerical approximation of $\partial U_k\big(\alpha, S_1(\cdot), S_2(\cdot)\big)/\partial\alpha$, which results negative for all the values of α considered.



(a) $U\big(\alpha, S_1(\cdot), S_2(\cdot)\big)$ (b) $\partial U\big(\alpha, S_1(\cdot), S_2(\cdot)\big)/\partial\alpha$

Figure B.7: Behavior of function $U\big(\alpha, S_1(\cdot), S_2(\cdot)\big)$ for known $S_1(\cdot)$ and $S_2(\cdot)$.

APPENDIX C

Tables of simulation results

C.1 Estimation of the marginal survival function $S_1(t)$

This section contains 8 tables summarizing the results of the simulation study described in Chapter 9 regarding the estimation of the marginal $S_1(t)$. For each table, we present the mean estimated probabilities at the 0.90, 0.70, 0.50, 0.30 and 0.10 percentiles, together with their bias and mean square error. Results obtained for generated samples from Exponential and Weibull marginal distributions are given in Tables C.1a to C.1d and Tables C.2a to C.2d, respectively. Results for Strategy 1 and Strategy 2 are shown. In Figures C.1 and C.2, comparison of the bias obtained for Strategy 1, Strategy 2 and midpoint imputation for $\alpha = 5$ and Exponential and Weibull marginals is made.

Table C.1a: Estimation of $S_1(t)$, for a model with $T_1 \sim \text{Exp}(\lambda_1)$, $T_2 \sim \text{Exp}(\lambda_2)$ and $\alpha = 3$. Narrow intervals.

Narrow intervals†, $\alpha = 3$								
			ICSCR1			**ICSCR2**		
n	**p ‡**	$S_1(t)$	**Mean**	**Bias**	**MSE**	**Mean**	**Bias**	**MSE**
200	25%	0.9	0.901	0.001	0.000	0.900	0.000	0.000
		0.7	0.708	0.008	0.018	0.704	0.004	0.002
		0.5	0.501	0.001	0.030	0.495	−0.005	0.004
		0.3	0.294	−0.006	0.025	0.289	−0.011	0.012
		0.1	0.100	0.000	0.002	0.099	−0.001	0.000
	50%	0.9	0.898	−0.002	0.002	0.896	−0.004	0.002
		0.7	0.715	0.015	0.077	0.701	0.001	0.005
		0.5	0.510	0.010	0.167	0.487	−0.013	0.044
		0.3	0.295	−0.005	0.155	0.275	−0.025	0.083
		0.1	0.102	0.002	0.047	0.092	−0.008	0.008
	75%	0.9	0.905	0.005	0.011	0.897	−0.003	0.001
		0.7	0.713	0.013	0.306	0.672	−0.028	0.216
		0.5	0.485	−0.015	0.856	0.424	−0.076	1.151
		0.3	0.293	−0.007	1.102	0.232	−0.068	1.231
		0.1	0.225	0.125	2.739	0.165	0.065	1.357
500	25%	0.9	0.897	−0.003	0.001	0.896	−0.004	0.001
		0.7	0.704	0.004	0.006	0.702	0.002	0.000
		0.5	0.498	−0.002	0.012	0.495	−0.005	0.003
		0.3	0.291	−0.009	0.015	0.289	−0.011	0.012
		0.1	0.099	−0.001	0.001	0.098	−0.002	0.000
	50%	0.9	0.896	−0.004	0.002	0.895	−0.005	0.002
		0.7	0.713	0.013	0.037	0.705	0.005	0.004
		0.5	0.506	0.006	0.064	0.494	−0.006	0.015
		0.3	0.292	−0.008	0.059	0.281	−0.019	0.043
		0.1	0.097	−0.003	0.012	0.093	−0.007	0.006
	75%	0.9	0.905	0.005	0.005	0.901	0.001	0.000
		0.7	0.711	0.011	0.125	0.688	−0.012	0.065
		0.5	0.478	−0.022	0.375	0.442	−0.058	0.599
		0.3	0.272	−0.028	0.443	0.237	−0.063	0.633
		0.1	0.205	0.105	1.500	0.169	0.069	0.787

†Narrow intervals: average width 6 time units.

‡**p**: Percentage of dependent censoring, $100 \times P(T_1 > T_2)$.

Table C.1b: Estimation of $S_1(t)$, for a model with $T_1 \sim \text{Exp}(\lambda_1)$, $T_2 \sim \text{Exp}(\lambda_2)$ and $\alpha = 3$. Wide intervals.

Wide intervals†, $\alpha = 3$								
			ICSCR1			**ICSCR2**		
n	**p ‡**	$S_1(t)$	**Mean**	**Bias**	**MSE**	**Mean**	**Bias**	**MSE**
200	25%	0.9	0.882	−0.018	0.032	0.881	−0.019	0.037
		0.7	0.678	−0.022	0.060	0.672	−0.028	0.079
		0.5	0.475	−0.025	0.092	0.466	−0.034	0.117
		0.3	0.279	−0.021	0.065	0.272	−0.028	0.078
		0.1	0.097	−0.003	0.004	0.095	−0.005	0.002
	50%	0.9	0.886	−0.014	0.023	0.881	−0.019	0.035
		0.7	0.696	−0.004	0.072	0.674	−0.026	0.078
		0.5	0.495	−0.005	0.194	0.462	−0.038	0.185
		0.3	0.291	−0.009	0.202	0.261	−0.039	0.177
		0.1	0.105	0.005	0.075	0.091	−0.009	0.012
	75%	0.9	0.898	−0.002	0.015	0.883	−0.017	0.032
		0.7	0.703	0.003	0.420	0.632	−0.068	0.723
		0.5	0.483	−0.017	1.178	0.385	−0.115	1.932
		0.3	0.302	0.002	1.608	0.206	−0.094	1.544
		0.1	0.247	0.147	3.962	0.150	0.050	1.102
500	25%	0.9	0.877	−0.023	0.051	0.877	−0.023	0.053
		0.7	0.674	−0.026	0.074	0.673	−0.027	0.076
		0.5	0.472	−0.028	0.092	0.470	−0.030	0.092
		0.3	0.277	−0.023	0.062	0.275	−0.025	0.062
		0.1	0.096	−0.004	0.002	0.096	−0.004	0.002
	50%	0.9	0.884	−0.016	0.027	0.882	−0.018	0.032
		0.7	0.694	−0.006	0.031	0.686	−0.014	0.024
		0.5	0.491	−0.009	0.080	0.479	−0.021	0.069
		0.3	0.287	−0.013	0.083	0.276	−0.024	0.075
		0.1	0.100	0.000	0.017	0.095	−0.005	0.004
	75%	0.9	0.897	−0.003	0.006	0.891	−0.009	0.008
		0.7	0.701	0.001	0.159	0.669	−0.031	0.241
		0.5	0.475	−0.025	0.514	0.427	−0.073	1.142
		0.3	0.281	−0.019	0.590	0.234	−0.066	1.054
		0.1	0.231	0.131	2.352	0.181	0.081	1.503

†Narrow intervals: average width 6 time units.

‡**p**: Percentage of dependent censoring, $100 \times P(T_1 > T_2)$.

Table C.1c: Estimation of $S_1(t)$, for a model with $T_1 \sim \text{Exp}(\lambda_1)$, $T_2 \sim \text{Exp}(\lambda_2)$ and $\alpha = 5$. Narrow intervals.

Narrow intervals†, $\alpha = 5$								
			ICSCR1			**ICSCR2**		
n	**p ‡**	$S_1(t)$	**Mean**	**Bias**	**MSE**	**Mean**	**Bias**	**MSE**
200	25%	0.9	0.903	0.003	0.002	0.902	0.002	0.000
		0.7	0.699	−0.001	0.019	0.693	−0.007	0.005
		0.5	0.484	−0.016	0.046	0.478	−0.022	0.050
		0.3	0.285	−0.015	0.027	0.283	−0.017	0.031
		0.1	0.104	0.004	0.002	0.103	0.003	0.001
	50%	0.9	0.905	0.005	0.006	0.901	0.001	0.000
		0.7	0.702	0.002	0.059	0.685	−0.015	0.027
		0.5	0.487	−0.013	0.093	0.469	−0.031	0.100
		0.3	0.287	−0.013	0.056	0.275	−0.025	0.063
		0.1	0.116	0.016	0.042	0.109	0.009	0.008
	75%	0.9	0.913	0.013	0.027	0.903	0.003	0.001
		0.7	0.694	−0.006	0.222	0.655	−0.045	0.248
		0.5	0.459	−0.041	0.479	0.419	−0.081	0.708
		0.3	0.280	−0.020	0.337	0.244	−0.056	0.356
		0.1	0.244	0.144	2.405	0.204	0.104	1.135
500	25%	0.9	0.902	0.002	0.001	0.901	0.001	0.000
		0.7	0.695	−0.005	0.010	0.692	−0.008	0.006
		0.5	0.481	−0.019	0.044	0.478	−0.022	0.047
		0.3	0.285	−0.015	0.025	0.283	−0.017	0.028
		0.1	0.104	0.004	0.002	0.103	0.003	0.001
	50%	0.9	0.903	0.003	0.002	0.901	0.001	0.000
		0.7	0.699	−0.001	0.023	0.690	−0.010	0.012
		0.5	0.483	−0.017	0.058	0.473	−0.027	0.076
		0.3	0.284	−0.016	0.038	0.278	−0.022	0.050
		0.1	0.110	0.010	0.015	0.107	0.007	0.005
	75%	0.9	0.911	0.011	0.015	0.905	0.005	0.003
		0.7	0.690	−0.010	0.096	0.667	−0.033	0.134
		0.5	0.454	−0.046	0.323	0.430	−0.070	0.519
		0.3	0.273	−0.027	0.170	0.251	−0.049	0.252
		0.1	0.246	0.146	2.253	0.221	0.121	1.496

†Narrow intervals: average width 6 time units.

‡**p**: Percentage of dependent censoring, $100 \times P(T_1 > T_2)$.

Table C.1d: Estimation of $S_1(t)$, for a model with $T_1 \sim \text{Exp}(\lambda_1)$, $T_2 \sim \text{Exp}(\lambda_2)$ and $\alpha = 5$. Wide intervals.

Wide intervals†, $\alpha = 5$								
			ICSCR1			**ICSCR2**		
n	**p ‡**	$S_1(t)$	**Mean**	**Bias**	**MSE**	**Mean**	**Bias**	**MSE**
200	25%	0.9	0.881	−0.019	0.038	0.878	−0.022	0.050
		0.7	0.666	−0.034	0.136	0.656	−0.044	0.192
		0.5	0.458	−0.042	0.200	0.449	−0.051	0.260
		0.3	0.271	−0.029	0.088	0.267	−0.033	0.109
		0.1	0.102	0.002	0.001	0.101	0.001	0.000
	50%	0.9	0.889	−0.011	0.017	0.881	−0.019	0.035
		0.7	0.678	−0.022	0.120	0.652	−0.048	0.234
		0.5	0.469	−0.031	0.186	0.444	−0.056	0.319
		0.3	0.279	−0.021	0.095	0.262	−0.038	0.146
		0.1	0.118	0.018	0.057	0.106	0.006	0.005
	75%	0.9	0.903	0.003	0.021	0.882	−0.018	0.034
		0.7	0.679	−0.021	0.362	0.617	−0.083	0.741
		0.5	0.451	−0.049	0.681	0.391	−0.109	1.224
		0.3	0.281	−0.019	0.498	0.226	−0.074	0.593
		0.1	0.252	0.152	2.855	0.190	0.090	0.862
500	25%	0.9	0.878	−0.022	0.050	0.878	−0.022	0.046
		0.7	0.656	−0.044	0.192	0.660	−0.040	0.164
		0.5	0.449	−0.051	0.260	0.453	−0.047	0.222
		0.3	0.267	−0.033	0.109	0.270	−0.030	0.093
		0.1	0.101	0.001	0.000	0.101	0.001	0.000
	50%	0.9	0.881	−0.019	0.035	0.884	−0.016	0.025
		0.7	0.652	−0.048	0.234	0.665	−0.035	0.129
		0.5	0.444	−0.056	0.319	0.454	−0.046	0.213
		0.3	0.262	−0.038	0.146	0.269	−0.031	0.097
		0.1	0.106	0.006	0.005	0.108	0.008	0.007
	75%	0.9	0.882	−0.018	0.034	0.892	−0.008	0.007
		0.7	0.617	−0.083	0.741	0.645	−0.055	0.370
		0.5	0.391	−0.109	1.224	0.415	−0.085	0.789
		0.3	0.226	−0.074	0.593	0.245	−0.055	0.348
		0.1	0.190	0.090	0.862	0.221	0.121	1.543

†Narrow intervals: average width 6 time units.

‡**p**: Percentage of dependent censoring, $100 \times P(T_1 > T_2)$.

Table C.2a: Estimation of $S_1(t)$, for a model with $T_1 \sim \text{Weibull}(\mu_1, \rho_1)$, $T_2 \sim \text{Weibull}(\mu_2, \rho_2)$ and $\alpha = 3$. Narrow intervals.

Narrow intervals†, $\alpha = 3$								
			ICSCR1			**ICSCR2**		
n	**p ‡**	$S_1(t)$	**Mean**	**Bias**	**MSE**	**Mean**	**Bias**	**MSE**
200	25%	0.9	0.913	0.013	0.000	0.782	−0.118	0.014
		0.7	0.719	0.019	0.001	0.617	−0.083	0.007
		0.5	0.502	0.002	0.000	0.496	−0.004	0.000
		0.3	0.285	−0.015	0.000	0.393	0.093	0.009
		0.1	0.091	−0.009	0.000	0.272	0.172	0.029
	50%	0.9	0.907	0.007	0.000	0.857	−0.043	0.002
		0.7	0.720	0.020	0.001	0.744	0.044	0.002
		0.5	0.503	0.003	0.002	0.650	0.150	0.023
		0.3	0.282	−0.018	0.002	0.559	0.259	0.067
		0.1	0.089	−0.011	0.000	0.436	0.336	0.113
	75%	0.9	0.912	0.012	0.000	0.908	0.008	0.000
		0.7	0.708	0.008	0.004	0.824	0.124	0.016
		0.5	0.466	−0.034	0.010	0.747	0.247	0.062
		0.3	0.254	−0.046	0.011	0.667	0.367	0.136
		0.1	0.188	0.088	0.017	0.549	0.449	0.203
500	25%	0.9	0.908	0.008	0.000	0.778	−0.122	0.015
		0.7	0.714	0.014	0.000	0.614	−0.086	0.007
		0.5	0.498	−0.002	0.000	0.496	−0.004	0.000
		0.3	0.283	−0.017	0.000	0.391	0.091	0.008
		0.1	0.090	−0.010	0.000	0.271	0.171	0.029
	50%	0.9	0.904	0.004	0.000	0.856	−0.044	0.002
		0.7	0.717	0.017	0.001	0.746	0.046	0.002
		0.5	0.500	0.000	0.001	0.656	0.156	0.024
		0.3	0.279	−0.021	0.001	0.564	0.264	0.070
		0.1	0.086	−0.014	0.000	0.442	0.342	0.117
	75%	0.9	0.910	0.010	0.000	0.910	0.010	0.000
		0.7	0.708	0.008	0.002	0.833	0.133	0.018
		0.5	0.461	−0.039	0.005	0.763	0.263	0.069
		0.3	0.240	−0.060	0.007	0.684	0.384	0.148
		0.1	0.163	0.063	0.007	0.568	0.468	0.220

†Narrow intervals: average width 6 time units.

‡**p**: Percentage of dependent censoring, $100 \times P(T_1 > T_2)$.

Table C.2b: Estimation of $S_1(t)$, for a model with $T_1 \sim \text{Weibull}(\mu_1, \rho_1)$, $T_2 \sim \text{Weibull}(\mu_2, \rho_2)$ and $\alpha = 3$. Wide intervals.

Wide intervals†, $\alpha = 3$								
			ICSCR1			**ICSCR2**		
n	**p ‡**	$S_1(t)$	**Mean**	**Bias**	**MSE**	**Mean**	**Bias**	**MSE**
200	25%	0.9	0.898	−0.002	0.000	0.749	−0.151	0.023
		0.7	0.690	−0.010	0.000	0.579	−0.121	0.015
		0.5	0.473	−0.027	0.001	0.460	−0.040	0.002
		0.3	0.265	−0.035	0.001	0.362	0.062	0.004
		0.1	0.084	−0.016	0.000	0.249	0.149	0.022
	50%	0.9	0.897	−0.003	0.000	0.836	−0.064	0.004
		0.7	0.700	0.000	0.001	0.712	0.012	0.000
		0.5	0.483	−0.017	0.002	0.614	0.114	0.013
		0.3	0.271	−0.029	0.002	0.523	0.223	0.050
		0.1	0.087	−0.013	0.000	0.405	0.305	0.093
	75%	0.9	0.905	0.005	0.000	0.888	−0.012	0.000
		0.7	0.692	−0.008	0.006	0.788	0.088	0.008
		0.5	0.452	−0.048	0.014	0.701	0.201	0.041
		0.3	0.247	−0.053	0.015	0.616	0.316	0.101
		0.1	0.186	0.086	0.019	0.498	0.398	0.160
500	25%	0.9	0.893	−0.007	0.000	0.748	−0.152	0.023
		0.7	0.685	−0.015	0.000	0.578	−0.122	0.015
		0.5	0.469	−0.031	0.001	0.463	−0.037	0.001
		0.3	0.263	−0.037	0.001	0.363	0.063	0.004
		0.1	0.083	−0.017	0.000	0.250	0.150	0.022
	50%	0.9	0.893	−0.007	0.000	0.838	−0.062	0.004
		0.7	0.696	−0.004	0.000	0.720	0.020	0.000
		0.5	0.480	−0.020	0.001	0.627	0.127	0.016
		0.3	0.268	−0.032	0.002	0.535	0.235	0.055
		0.1	0.084	−0.016	0.000	0.417	0.317	0.100
	75%	0.9	0.903	0.003	0.000	0.898	−0.002	0.000
		0.7	0.692	−0.008	0.002	0.810	0.110	0.012
		0.5	0.446	−0.054	0.008	0.731	0.231	0.054
		0.3	0.233	−0.067	0.008	0.648	0.348	0.122
		0.1	0.161	0.061	0.007	0.530	0.430	0.187

†Narrow intervals: average width 6 time units.

‡**p**: Percentage of dependent censoring, $100 \times P(T_1 > T_2)$.

Table C.2c: Estimation of $S_1(t)$, for a model with $T_1 \sim \text{Weibull}(\mu_1, \rho_1)$, $T_2 \sim \text{Weibull}(\mu_2, \rho_2)$ and $\alpha = 5$. Narrow intervals.

Narrow intervals†, $\alpha = 5$								
			ICSCR1			**ICSCR2**		
n	**p ‡**	$S_1(t)$	**Mean**	**Bias**	**MSE**	**Mean**	**Bias**	**MSE**
200	25%	0.9	0.909	0.009	0.000	0.804	−0.096	0.009
		0.7	0.698	−0.002	0.000	0.654	−0.046	0.002
		0.5	0.475	−0.025	0.001	0.546	0.046	0.002
		0.3	0.278	−0.022	0.001	0.453	0.153	0.023
		0.1	0.098	−0.002	0.000	0.338	0.238	0.057
	50%	0.9	0.910	0.010	0.000	0.850	−0.050	0.002
		0.7	0.702	0.002	0.001	0.728	0.028	0.001
		0.5	0.482	−0.018	0.001	0.633	0.133	0.018
		0.3	0.281	−0.019	0.001	0.546	0.246	0.061
		0.1	0.106	0.006	0.000	0.434	0.334	0.112
	75%	0.9	0.915	0.015	0.000	0.882	−0.018	0.000
		0.7	0.691	−0.009	0.003	0.771	0.071	0.005
		0.5	0.451	−0.049	0.006	0.682	0.182	0.033
		0.3	0.263	−0.037	0.004	0.598	0.298	0.089
		0.1	0.219	0.119	0.017	0.486	0.386	0.149
500	25%	0.9	0.905	0.005	0.000	0.801	−0.099	0.010
		0.7	0.693	−0.007	0.000	0.651	−0.049	0.002
		0.5	0.474	−0.026	0.001	0.546	0.046	0.002
		0.3	0.278	−0.022	0.001	0.451	0.151	0.023
		0.1	0.098	−0.002	0.000	0.338	0.238	0.057
	50%	0.9	0.906	0.006	0.000	0.851	−0.049	0.002
		0.7	0.697	−0.003	0.000	0.729	0.029	0.001
		0.5	0.477	−0.023	0.001	0.637	0.137	0.019
		0.3	0.278	−0.022	0.001	0.548	0.248	0.062
		0.1	0.102	0.002	0.000	0.436	0.336	0.113
	75%	0.9	0.914	0.014	0.000	0.887	−0.013	0.000
		0.7	0.687	−0.013	0.001	0.780	0.080	0.007
		0.5	0.446	−0.054	0.004	0.693	0.193	0.037
		0.3	0.257	−0.043	0.003	0.607	0.307	0.094
		0.1	0.218	0.118	0.015	0.494	0.394	0.156

†Narrow intervals: average width 6 time units.

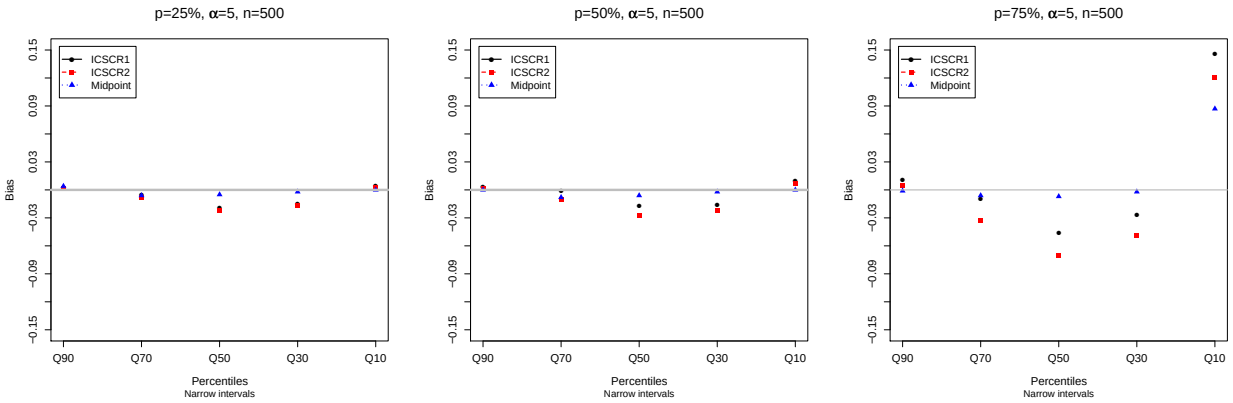
‡**p**: Percentage of dependent censoring, $100 \times P(T_1 > T_2)$.

Table C.2d: Estimation of $S_1(t)$, for a model with $T_1 \sim \text{Weibull}(\mu_1, \rho_1)$, $T_2 \sim \text{Weibull}(\mu_2, \rho_2)$ and $\alpha = 5$. Wide intervals.

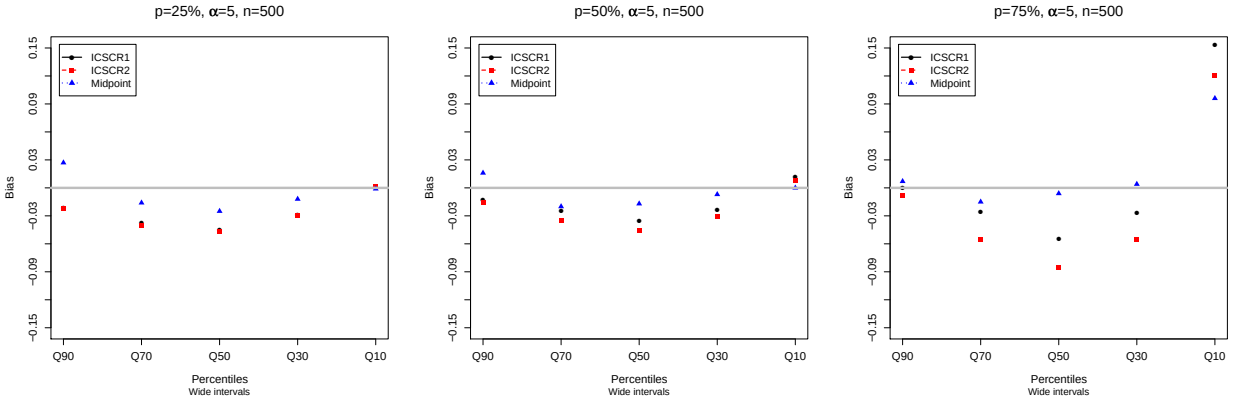
Wide intervals†, $\alpha = 5$								
			ICSCR1			**ICSCR2**		
n	**p ‡**	$S_1(t)$	**Mean**	**Bias**	**MSE**	**Mean**	**Bias**	**MSE**
200	25%	0.9	0.889	−0.011	0.000	0.767	−0.133	0.018
		0.7	0.664	−0.036	0.002	0.613	−0.087	0.008
		0.5	0.447	−0.053	0.003	0.509	0.009	0.000
		0.3	0.260	−0.040	0.002	0.421	0.121	0.015
		0.1	0.092	−0.008	0.000	0.314	0.214	0.046
	50%	0.9	0.895	−0.005	0.000	0.821	−0.079	0.006
		0.7	0.678	−0.022	0.001	0.692	−0.008	0.000
		0.5	0.462	−0.038	0.002	0.598	0.098	0.010
		0.3	0.270	−0.030	0.001	0.515	0.215	0.046
		0.1	0.104	0.004	0.000	0.410	0.310	0.096
	75%	0.9	0.905	0.005	0.000	0.852	−0.048	0.002
		0.7	0.671	−0.029	0.005	0.734	0.034	0.001
		0.5	0.436	−0.064	0.008	0.645	0.145	0.021
		0.3	0.254	−0.046	0.006	0.564	0.264	0.070
		0.1	0.210	0.110	0.016	0.459	0.359	0.129
500	25%	0.9	0.884	−0.016	0.000	0.766	−0.134	0.018
		0.7	0.660	−0.040	0.002	0.614	−0.086	0.007
		0.5	0.446	−0.054	0.003	0.511	0.011	0.000
		0.3	0.260	−0.040	0.002	0.421	0.121	0.015
		0.1	0.092	−0.008	0.000	0.315	0.215	0.046
	50%	0.9	0.891	−0.009	0.000	0.826	−0.074	0.005
		0.7	0.673	−0.027	0.001	0.699	−0.001	0.000
		0.5	0.457	−0.043	0.002	0.607	0.107	0.011
		0.3	0.267	−0.033	0.001	0.521	0.221	0.049
		0.1	0.100	0.000	0.000	0.415	0.315	0.099
	75%	0.9	0.903	0.003	0.000	0.865	−0.035	0.001
		0.7	0.666	−0.034	0.003	0.749	0.049	0.003
		0.5	0.429	−0.071	0.007	0.661	0.161	0.026
		0.3	0.246	−0.054	0.004	0.577	0.277	0.077
		0.1	0.207	0.107	0.013	0.469	0.369	0.137

†Narrow intervals: average width 6 time units.

‡**p**: Percentage of dependent censoring, $100 \times P(T_1 > T_2)$.

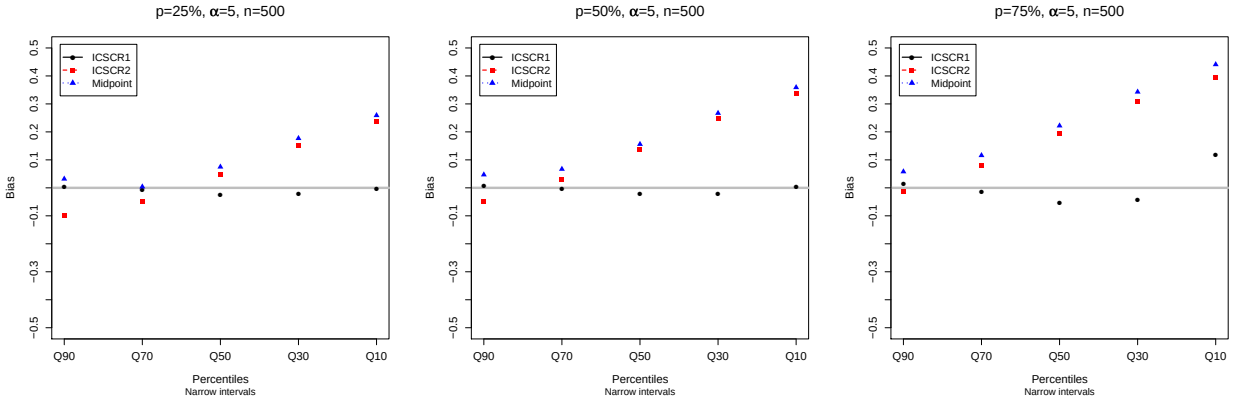


(a) Narrow Intervals, $\alpha = 5$

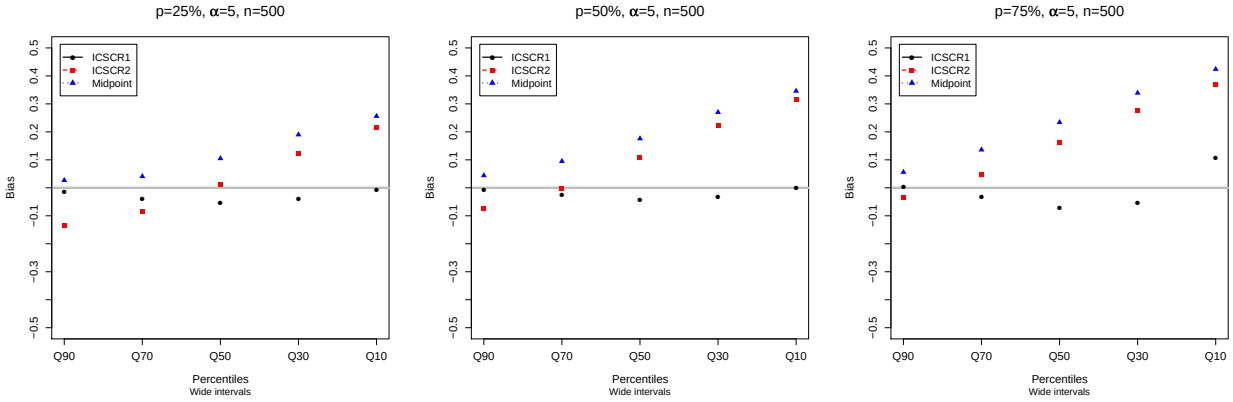


(b) Wide Intervals, $\alpha = 5$

Figure C.1: Bias of $S_1(t)$ estimates: comparison of ICSC1, ICSCR2 and Midpoint ($\alpha = 5$, Exponential margins)



(a) Narrow Intervals, $\alpha = 5$



(b) Wide Intervals, $\alpha = 5$

Figure C.2: Bias of $S_1(t)$ estimates: comparison of ICSC1, ICSCR2 and Midpoint ($\alpha = 5$, Weibull margins)

APPENDIX D

R Programmes

D.1 Competing risks analysis with R

D.1.1 Nonparametric and regression modelling with R

In the following, we describe the necessary code to implement a competing risks analysis in R (R Development Core Team, 2009). Two packages are needed: `survival` (Therneau and original R port by Thomas Lumley, 2009) and `cmprsk` (Gray, 2004). The former is included by default with the software, but `cmprsk` needs first to be downloaded from R's web site. Both packages must be loaded at the beginning of the session.

Assume we have a data frame containing at least two columns `T1` and `C1`, being, respectively, the vector with observed times for each individual, and the vector of failing causes. The length n of these vectors correspond to the sample size we dispose of. We will assume that the $n \times p$ matrix `Z` is the matrix of the p covariates for all n individuals to be included in the regression models. The `C1` vector equals 0 when individuals are censored at their observed time, or takes value j among the distinct possible causes of failure. For this illustration, assume there are only two causes of failure, and therefore, `C1` takes values in $\{0, 1, 2\}$. Data may have this appearance:

```
   T1 C1 cens gender age stage
70.6  2    1      0  47     1
12.8  1    1      1  44     1
42.6  2    1      0  74     0
91.0  0    0      0  64     0
44.3  2    1      0  74     1
71.7  0    0      0  60     0
```

Matrix `Z` of covariates would consist of columns `gender, age` and `stage`.

D.1.1.1 Nonparametric estimation

Two simple functions can be implemented to obtain, at any given time `t`, the number of individuals at risk of failing for any cause and the number of individuals failing from each cause:

```
risk<-function(t=0,vT){
    val<-sum((vT>=t),na.rm=T)
    return(val)
    }
```

```
fail<-function(t=0,vT,vC,c=1){
    val<-sum((vT==t)*(vC==c),na.rm=T)
    return(val)
    }.
```

The `risk` function provides, at any time `t`, the number of individuals at risk, based on the information given by the vector of times `vT`. The `fail` function provides the specific number of failures at time `t` from cause `c`, based on the information given by the vector of times `vT` and the vector of causes `vC`.

Now we apply functions `risk` and `fail` to each element of vector `ts` containing the ordered and unique values of `T1`, in order to obtain vectors of the same length containing the number of individuals at risk (`ni`), and the number of individuals failing from each cause (`d1` and `d2`):

```
ts<-c(0,unique(sort(T1)))
ni<-sapply(ts,risk,vT=T1)
d1<-sapply(ts,fail,vT=T1,vC=C1,c=1)
d2<-sapply(time,fail,vT=T1,vC=C1,c=2)
```

Now, estimates for the cause-specific hazards at any observed time are easily obtained by:

```
lam1<-d1/ni
lam2<-d2/ni.
```

We obtain that, for instance, at time `ts`= 1.4, there are 993 individuals at risk (`ni`), one failure due to cause 1 (`d1`), and one due to cause (d2), and so there are two failures for any cause (`d`). The cause-specific hazard at this point for both causes of failure is 0.001 (`lam1` and `lam2`):

```
       ts  ni d1 d2 d  lam1  lam2
 [1,] 0.0 995  0  0 0 0.000 0.000
 [2,] 1.0 995  0  0 0 0.000 0.000
 [3,] 1.2 994  0  1 1 0.000 0.001
 [4,] 1.4 993  1  1 2 0.001 0.001
 [5,] 1.5 991  1  0 1 0.001 0.000
 [6,] 1.6 990  1  0 1 0.001 0.000
 [7,] 1.8 989  2  0 2 0.002 0.000
 [8,] 1.9 987  1  0 1 0.001 0.000
 [9,] 2.0 986  1  1 2 0.001 0.001
[10,] 2.1 984  2  0 2 0.002 0.000
```

The Kaplan-Meier estimate of the survival function for `T1`, without taking into account distinct causes of failure, is obtained by the `survfit` function. We need to define a censoring indicator for any of the two events: `cens=1` when `cause=1` or `2`, `0` otherwise.

```
cens<-as.integer(C1!=0)
sur<-survfit(Surv(T1,cens)~1)
S<-c(1,sur$surv)
```

The cumulative incidence functions can be obtained using the cuminc function from the cmprsk package, where we must specify, at least, the vector of times T1, the vector of event types C1 and the category of C1 which corresponds to right-censored obervations:

```
cif<-cuminc(T1,C1,cencode=0).
```

From this object cif we can extract the cumulative incidence function from each cause, cif1 and cif2:

```
> cif
Estimates and Variances: $est
            20         40         60        80       100
1 1 0.04940253 0.07370225 0.09016217 0.1003712 0.1003712 1 2
0.04135428 0.08691458 0.15380532 0.2037241 0.2317216
$var
               20           40           60           80           100
1 1 4.739837e-05 6.900797e-05 8.323041e-05 9.293715e-05 9.293715e-05
1 2 4.002899e-05 8.030162e-05 1.328502e-04 1.743529e-04 3.160064e-04
```

Function plot.cuminc directly plots both cumulative incidence functions. Also different curves for each stratum of a categorical covariate can be graphically assessed by using the option group of the cuminc function:

```
plot(cuminc(T1,C1,group=gender,cencode=0)).
```

Plots of the cumulative incidence function in Chapter 2 were obtained from this function.

D.1.1.2 Regression modelling

To adjust Cox proportional hazards model for each cause-specific hazard, we use function coxph of the package survival. We adjust two models, one for each type of event:

```
cox1<-coxph(Surv(T1,C1==1)~gender+age+stage,data=data)
cox2<-coxph(Surv(T1,C1==2)~gender+age+stage,data=data)
```

To fit Fine and Gray's model, we use function crr of the package cmprsk:

```
Z<-cbind(data$gender,data$age,data$stage)
fine1<-crr(T1,C1,cov1=Z,failcode=1)
fine2<-crr(T1,C1,cov1=Z,failcode=2),
```

where we need to introduce specifically the matrix with the covariates Z, and adjust again one model for each cumulative incidence function.

Fine and Gray's model can be estimated using coxph as explained in Geskus (2010). First, data must be preprocessed in order to obtain time-dependent weights. This could be achieved by function crprep, available from the author at

```
source('crprep.R')
data.tr1<-crprep("T1", "C1", data=data, Tstart=0, Z=Z, riskcode=1,censcode = 0)
fine1rev<-coxph(Surv(Tstart,Tstop,status==1)~gender+age+stage,data=data.tr1,
    weight=weight.cens,subset=failcode==1)
data.tr2<-crprep("T1", "C1", data=data, Tstart=0, Z=Z, riskcode=2,censcode = 0)
fine1rev<-coxph(Surv(Tstart,Tstop,status==2)~gender+age+stage,data=data.tr1,
    weight=weight.cens,subset=failcode==2)
```

D.1.1.3 Prediction

Now we want to predict the probability of failure due to cause 1 before time u=60 for a 65 years old male with tumour in stage T1. To use the first approach based on the fitted Cox models, we need to obtain the cumulative hazard functions from the two models. We first obtain the predicted survival function for each model for this individual by means of the `survest` function:

```
cov<-c(0,65,1)
ha1=survfit(cox1,type='br',newdata=cov)
ha2=survfit(cox2,type='br',newdata=cov)
```

Note that the obtained estimation is not a proper survival function, because it is obtained from the cause-specific hazards, that is, $S_j(t) = \exp\{-\int_0^t \lambda_j(u)du\}$. Now we obtain the cause-specific cumulative hazard functions $\Lambda_j(t)$:

```
H10=c(0,-log(ha1$surv))
H20=c(0,-log(ha2$surv))
```

We obtain the estimates for these cumulative hazards in all observed time points:

```
time1=c(0,ha1$time)
time2=c(0,ha2$time)
time<-unique(sort(c(time1,time2)))
k1<-findInterval(time,time1)
k2<-findInterval(time,time2)
H1<-H10[k1]
H2<-H20[k2].
```

To obtain the cumulative incidence function for type 1 we need the cause-specific hazard for type 1, λ_1, as well as overall survival function $S_(t) = P(T1 > t)$:

```
lambda1=c(0,diff(H1))
ST<-exp(-(H1+H2))
```

We use the following function to implement expression 2.8:

```
Cif<-function(t,time,S,lam){
    fi=findInterval(t,time)
    cif=sum(S[1:fi]*lam[1:fi])
    return(cif)
}
```

Now we can obtain the predicted value for this individual at time 60:

```
Cif(60,time,ST,lambda1)
```

To obtain predictions from the Fine and Gray model is more direct using function `predict.crr`:

```
cif1.f<-predict(fine1,cov1=cov)
aux<-findInterval(60,cif1.f[,1])
cif1.f[aux,2]
```

D.1.2 Technical details on the construction of a nomogram

We illustrate with an example how the transformation between the predicted probability and the scoring system is obtained for a nomogram. Indeed, suppose the following Fine and Gray's model for the subhazard of the event of interest and only two binary covariates, x_1 and x_2:

$$\gamma_1(t|x_1, x_2) = \gamma_{1,0}(t) \exp\{\beta_1 x_1 + \beta_2 x_2\},$$

or, equivalently, its cumulative subhazard:

$$\Gamma_1(t|x_1, x_2) = \Gamma_{1,0}(t) \exp\{\beta_1 x_1 + \beta_2 x_2\}.$$

From expression (2.9), we have:

$$1 - F_1(t) = \exp\{-\Gamma_{1,0}(t)\}^{\exp\{\beta_1 x_1 + \beta_2 x_2\}}.$$

If we apply logarithms twice, we obtain the linear relationship:

$$\log(-\log(1 - F_1(t)) = \log(\Gamma_{1,0}(t)) + \beta_1 x_1 + \beta_2 x_2.$$

Now, let $\beta_{\max}$ be the greatest among the estimated effects of the model: $\beta_{\max} = \max\{\beta_1, \beta_2\}$. The points assigned at each variable axis are then

$$\text{Points}(x_i) = 100 \times \frac{\beta_i}{\beta_{\max}} x_i \qquad i = 1, 2.$$

The most significant variable is assigned 100 points, and for the rest of variables, the points assigned correspond to the relative importance of the variable as compared with the most significant variable. The Total Points are defined by

$$\text{TPoints} = 100 \left[\frac{\beta_1}{\beta_{\max}} x_1 + \frac{\beta_2}{\beta_{\max}} x_2 \right].$$

If $a = \log(\Gamma_{1,0}(t))$, the relationship between predicted probabilities and score is given by:

$$F_1(t) = 1 - \exp\{-\exp(a + k * \text{TPoints})\}.$$

D.1.3 Function **getCalibrateCIF**

```
getCalibrateCIF<-function(cox.obj,fine.obj,g=6,Srv,dades,covar,which.cause=1,u, B=20,
    pl=FALSE,conf.int = 0.95, unit='Month',xlab, ylab, main, cex.subtitle = 0.7,...){

    pred<-getEstimates(cox.obj,fine.obj,u,covar)
    pred.group.cox<-groupCIF(pred$getS1,Srv,g=g,which.cause=which.cause,u=u) #Pj.c
    Pr.c<-pred.group.cox$est[,1]
    O.c<-pred.group.cox$est[,4]
    dif.c<-Pr.c-O.c
    pred.group.fine<-groupCIF(pred$getF1,Srv,g=g,which.cause=which.cause,u=u) #Pj.f
    Pr.f<-pred.group.fine$est[,1]
    O.f<-pred.group.fine$est[,4]
    dif.f<-Pr.f-O.f

    CIF.c<-matrix(ncol=g,nrow=B)
    CIF.f<-matrix(ncol=g,nrow=B)
    D.c<-matrix(ncol=g,nrow=B)
    D.f<-matrix(ncol=g,nrow=B)
    D.O.c<-matrix(ncol=g,nrow=B)
    D.O.f<-matrix(ncol=g,nrow=B)
    Bias.c<-matrix(ncol=g,nrow=B)
    Bias.f<-matrix(ncol=g,nrow=B)
    for(b in 1:B){
        set.seed(345/b)
        sampB<-sample(1:nrow(Srv),nrow(Srv),replace=T)
        coxb<-cph(cox.obj$terms,data=dades[sampB,],surv=T,x=T,y=T)
        fineb<-crr(Srv[sampB,1],Srv[sampB,2],cov1=covar[sampB,],failcode=which.cause)
        # partition on bootstrap sample
        prB<-getEstimates(coxb,fineb,u,covar[sampB,])
        estB.c<-groupCIF(prB$getS1,Srv[sampB,],g,which.cause=which.cause,u=u)
        estB.f<-groupCIF(prB$getF1,Srv[sampB,],g,which.cause=which.cause,u=u)
        CIF.c[b,]<-estB.c$est[,4]
        D.c[b,]<-estB.c$est[,1]-CIF.c[b,]
        CIF.f[b,]<-estB.f$est[,4]
        D.f[b,]<-estB.f$est[,1]-CIF.f[b,]

        # partition over the original sample
        prB.O<-getEstimates(coxb,fineb,u,covar)
        estB.O.c<-groupCIF(prB.O$getS1,Srv,g,which.cause=which.cause,u=u)
        estB.O.f<-groupCIF(prB.O$getF1,Srv,g,which.cause=which.cause,u=u)
        D.O.c[b,]<-estB.O.c$est[,1]-estB.O.c$est[,4]
        D.O.f[b,]<-estB.O.f$est[,1]-estB.O.f$est[,4]
        Bias.c[b,]<-D.c[b,]-D.O.c[b,]
        Bias.f[b,]<-D.f[b,]-D.O.f[b,]
    }
    mBias.c<-colMeans(Bias.c)
    mBias.f<-colMeans(Bias.f)

    if (pl) {
      if (missing(xlab))
         xlab <- paste("Predicted ", format(u), "-", unit, " probability", sep = "")
        if (missing(ylab))
```

```
            ylab <- paste("Observed ", format(u), "-", unit, " probability", sep = "")
        if (missing(main))
            main <- paste("Cumulative incidence function for cause ",
            which.cause, sep = "")

        if (conf.int) {
        alp<-(1-conf.int)/2
        low.c<-apply(CIF.c,2,quantile,probs=c(alp))
        hi.c<-apply(CIF.c,2,quantile,probs=c(1-alp))
        low.f<-apply(CIF.f,2,quantile,probs=c(alp))
        hi.f<-apply(CIF.f,2,quantile,probs=c(1-alp))

        xmin<-min(Pr.c,Pr.f,low.c,low.f)
        xmax<-max(Pr.c,Pr.f,hi.c,hi.f)
            errbar.CIF(Pr.c, O.c, hi.c, low.c,type='b',pch=19,xlab = xlab, ylab = ylab,
            ylim=c(xmin,xmax),xlim=c(xmin,xmax),...)
        title(main=main)
            points(Pr.c,O.c+mBias.c, pch=4,lwd=2)
            errbar.CIF(Pr.f, O.f, hi.f, low.f, add = TRUE,col=2,pch=15,type='b')
        points(Pr.f,O.f+mBias.f, pch=4,col=2,lwd=2)

      } else {
        xmin<-min(Pr.c,O.c,O.c+mBias.c,Pr.f,O.f,O.f+mBias.f)
        xmax<-max(Pr.c,O.c,O.c+mBias.c,Pr.f,O.f,O.f+mBias.f)
            plot(Pr.c, O.c, xlab = xlab, ylab = ylab, main=main, type = "b",
            ylim=c(xmin,xmax),xlim=c(xmin,xmax), pch=19,...)
            points(Pr.c,O.c+mBias.c, pch=4,lwd=2)
            lines(Pr.f,O.f,type='b',pch=15,col=2)
        points(Pr.f,O.f+mBias.f, pch=4,col=2,lwd=2)

        }

        if (!is.logical(cex.subtitle)) {
        nn<-nrow(Srv)
        events<-sum(Srv[,2]==which.cause)
            mm <- round(nn/g)
            title(sub = paste("n=", nn, " d=", events,
                ", avg. ", mm, " patients per group", sep = ""),
                adj = 0, cex = cex.subtitle)
        }
      abline(0,1,lty=2)
      legend('bottomright',legend=c('Cox ignoring CR','Fine & Gray model'),lty=1,
      col=c(1,2),pch=c(19,15),inset=0.025,cex=0.95)
    }
}
```

D.1.3.1 Function `getEstimates`

```
getEstimates<-function(cox.obj,fine.obj,u,newdata,...){
    getS1<-1-survest(cox.obj,newdata=newdata,times=u,what='survival',se.fit=FALSE)$surv
    ind<-findInterval(u,fine.obj$uftime)
    getF1<-predict(fine.obj,cov1=as.matrix(newdata))[ind,-1]
```

```
    return(data.frame(cbind(getS1,getF1)))
}
```

D.1.3.2 Function `groupCIF`

```
groupCIF<-function(x,Srv,g=6,which.cause=1,u){
   if (missing(u))
        stop("u (time point) must be given")
    s <- !(is.na(x) | is.na(Srv[, 1]) | is.na(Srv[, 2]))
    x <- x[s]
    Srv <- Srv[s, ]
    x[abs(x) < 1e-10] <- 0

    e <- Srv[, 2]
    if (nrow(Srv) != length(x) )
        stop("lengths of x and Srv must match")

    q0 <- cut2(x, g = g,onlycuts=TRUE)
    q<-findInterval(x,q0,all.inside=TRUE)
    cif <- single(g)
    pred <- single(g)
    std.err <- cif
    events <- integer(g)
    numobs <- events
    for (i in 1:g) {
      s <- q == i
        nobs <- sum(s)
        ne <- sum(e[s]==which.cause)
        if (nobs < 2) {
            numobs[i] <- 0
            events[i] <- 0
            pred[i] <- if (nobs == 1)
                mean(x[s], na.rm = TRUE)
            else NA
            cif[i] <- NA
            std.err[i] <- NA
        }
        else {
            pred[i] <- mean(x[s], na.rm = TRUE)
        ob1<-cuminc(Srv[s,1],Srv[s,2])
        tim<-ob1[[which.cause]]$time
        ind<-findInterval(u,tim)
        cif[i]<-ob1[[which.cause]]$est[ind]
            std.err[i] <- sqrt(ob1[[which.cause]]$var[ind]/nobs)
            numobs[i] <- nobs
            events[i] <- ne
        }
    }
    z <- cbind(pred, n = numobs, events = events, CIF = cif, std.err = std.err)
    return(list(est=z,cuts=q0))
}
```

D.2 Multi-state models with R

D.2.1 Functions Pi1cr.0 and Pi1cr.1

```
 Pi1cr.1=function(t1,s1,time1,H2,HRP){

    if(time1[1]==0 & HRP[1]!=0) lambdaRP=c(HRP[1],diff(HRP)) else lambdaRP=c(0,diff(HRP))
    iniT=findInterval(t1,time1)
    finT=findInterval(s1,time1)
    PrT=0
    iT<-iniT+1
    while (iT<=finT){
        PrT=PrT+exp(-(H2[iT]-H2[iniT]))*lambdaRP[iT]
        iT<-iT+1
    }
    return(PrT)
}

Pi1cr.0=function(t1,s1,time1,H1,HP,HR,H2,HRP){
    if(time1[1]==0 & HP[1]!=0) lambdaP=c(HP[1],diff(HP)) else lambdaP=c(0,diff(HP))
    if(time1[1]==0 & HR[1]!=0) lambdaR=c(HR[1],diff(HR)) else lambdaR=c(0,diff(HR))
    iniT=findInterval(t1,time1)
    finT=findInterval(s1,time1)
    PrT=0
    iT<-iniT+1
    while (iT<=finT){
        PrT=PrT+exp(-(H1[iT]-H1[iniT]))*(lambdaP[iT]+lambdaR[iT]*
                                        Pi1cr.2(time1[iT],s1,time1,H2,HRP))
        iT<-iT+1
    }
    return(PrT)
}
```

D.3 Semi-competing risks analysis with **R**

D.3.1 Function **corSCR**

```
corSCR<-function(A,a,b,v=T,mQ=F...){
   n<-length(A[[1]])
   R<-matrix(0,nrow=n,ncol=n)
   S<-matrix(0,nrow=n,ncol=n)
   D<-matrix(0,nrow=n,ncol=n)
   d<-matrix(0,nrow=n,ncol=n)
   d0<-matrix(0,nrow=n,ncol=n)
   W<-matrix(0,nrow=n,ncol=n)
   for(j in 2:n){
       for(i in 1:(j-1)){
           S[i,j]<-min(min(A$X[i],A$X[j]),min(A$Y[i],A$Y[j]))
           R[i,j]<-min(A$Y[i],A$Y[j])
           d[i,j]<-conc(i,j,A)
           D[i,j]<-comp(i,j,A)
           W[i,j]<-Wab(S[i,j],R[i,j],a,b,A)
       }
   }
   con<-sum(W*D*d)/sum(W*D*(1-d))
   #variance
   if(v==TRUE){
       I0<-0
       for(j in 2:n){
           for(i in 1:(j-1)){
               I0<-I0+(W[i,j]*D[i,j]*(1+con)^(-2))
           }
       }
       I<-I0*n^(-2)
       Q<-matrix(0,nrow=n,ncol=n)
       for(j in 2:n){
           for(i in 1:(j-1)){
               Q[i,j]<-W[i,j]*D[i,j]*(d[i,j]-(con/(1+con)))
           }
       }
       J0<-0
       for(m in 3:n){
           for(k in 1:(m-2)){
               for(l in (k+1):(m-1)){
                   J0<-J0+(Q[k,l]*Q[k,m]+Q[k,l]*Q[l,m]+Q[l,m]*Q[k,m])
               }
           }
       }
       J<-(2/(n^3))*J0
       sig<-(1/n)*(1/I^2)*J
       } else (sig<-NA)
   ifelse(mQ==F,return(list(con.index=con, se=sqrt(sig), a=a,b=b, I=I)),
       return(list(con.index=con, se=sqrt(sig), a=a,b=b, I=I,Q=Q)))
}
```

D.3.2 Function **margSCR**

```
marg.SCR<-function(A,u,m=1,alp=0.05,ic=F){
   con.ind<-u$con.index
   n<-length(A$X)
   Kz<-survfit(Surv(A$X,A$dz)~1,type='kaplan-meier')
   K2<-survfit(Surv(A$Y,A$d2)~1,type='kaplan-meier')
   time<-unique(sort(c(0,Kz$time,K2$time)))
   nt<-length(time)
   Sz<-numeric(length(time))
   S2<-numeric(length(time))
   indz<-findInterval(time,Kz$time)
   ind2<-findInterval(time,K2$time)
   for(i in 1:length(time)){
       if (indz[i]==0) Sz[i]<-1
       else            Sz[i]<-Kz$surv[indz[i]]
       if (ind2[i]==0) S2[i]<-1
       else            S2[i]<-K2$surv[ind2[i]]
       }
   S1<-g(Sz,S2,con.ind)  #lines(time,S1tz,type='s',col=5)
   #S1tz<-((Sz^(1-con.ind)) - (S2^(1-con.ind)) + 1)^(1/(1-con.ind))
   aux<-ifelse(sum(is.na(S1))>0,min(which(is.na(S1)==T)),9999999)
   s1<-c(0,diff(S1))
   t1<-min(max(which(((Sz^(1-con.ind) - S2^(1-con.ind)) >-1 ) & (0<=S1) & (S1<=1))),
       aux-1)

   S1.e<-numeric(length=t1)
   S1.e[1]<-S1[1]

   for(i in 2:t1){
       S1.e[i]<-min(S1[1:i])
   }

   #Variance
   if(ic==T){
       o<-sapply(time,sigmt,time,A,Sz,S2,u)
       Sig<-as.numeric(o[1,])
       Sig.est<-as.numeric(o[2,])
       if(m==1){
         LF<-numeric(t1)
         UF<-numeric(t1)
         for(i in 1:t1){
           if(S1[i]==1) {
             LF[i]<-1
             UF[i]<-1
           }else{
             LF[i]<- im1(m1(S1.e[i])-(n^(-0.5))*dm1(S1.e[i])*
                                        (Sig.est[i]^(0.5))*qnorm(1-2*alp))
             UF[i]<- im1(m1(S1.e[i])+(n^(-0.5))*dm1(S1.e[i])*
                                        (Sig.est[i]^(0.5))*qnorm(1-2*alp))
           }
         }
       }
```

```
    }

    tim<-time[1:t1]
    S1t<-S1[1:t1]
    if(ic==T) return(list(time=tim,S1.e=S1.e,Sig=Sig,Sig.est=Sig.est, LS1=LF,US1=UF))
    else
        return(list(time=tim, S1.e=S1.e))
}
```

D.3.3 Internal functions

The following functions are employed by the previous procedures and will be soon available at `http://www-eio06.upc.es/research/grass/`. The notation used follows the notation employed in the paper from (Fine *et al.*, 2001).

- **Function `Wab`**: It computes the bivariate weighted random function Wab(x,y) specified in the estimating equation.
- **Function `comp`:** It computes the indicator of comparability O_{ij}^R for a pair (i, j) of observed right-censored semi-competing risks data subjects.
- **Function `conc`:** It computes the concordance indicator Δ_{ij} for a pair (i, j) of observed right-censored semi-competing risks data subjects.
- **Function `pi.f`:** Function appearing in the martingale representation of the estimates of $S_T(t) = P(T < t)$ and $S_2(t) = P(T_2 > t)$.
- **Function `M.f`:** Martingales involved in the martingale representation of the estimates of of $S_T(t) = P(T < t)$ and $S_2(t) = P(T_2 > t)$.
- **Functions `g, g1, g2, g3`:** Functional of the plug-in estimator, and partial derivatives with respect to a,b and c.
- **Function `V`:** Units of the summary forming the U-statistic $J_x(t)$ (used to determine the asymptotic variance of $\widehat{S}_1(t)$.
- **Function `sigmt`:** Consistent estimator of the covariance function of $\sqrt{}(n)(\widehat{S}_1(t) - S_1(t))$, evaluated at s=t.
- **Functions `m1,dm1,im1`:** Functions used to compute confidence intervals for hatS1.

D.4 Interval-censored semi-competing risks analysis with **R**

D.4.1 Function **algICSCR**

```
algICSCR<-function(dSCRi,e.alp.m, sh1.0,sc1.0,p0=T,p1=T,survC,datLR,strat,breaks,
time=time)
{
n<-length(dSCRi[[1]])
I<-t(combn(c(1:n),2))
C.0<-mapply(compIC,I[,1],I[,2],MoreArgs=list(A=dSCRi))
C.01<-mapply(compIC.1,I[,1],I[,2],MoreArgs=list(A=dSCRi))
fila<-C.01%%6
fila[fila==0]<-6
colum<-(C.01-1)%/%6+1
mat<-table(fila,colum)
nc1<-sum(C.01==3)
nc3<-sum(C.01==15)

max.iter<-20
tol<-0.001

#STRATEGY 1 ###################################################################
if(p0){
    iter<-1
    sh1<-sh1.0
    sc1<-sc1.0
    valp.0<-numeric(length=max.iter+1)
    vS1.0<-vector('list',length=max.iter+1)
    alp0<-999
    alp<-e.alp.m
    while((iter<=max.iter)&(abs(alp-alp0)>tol)){
        p.nd<-mapply(nD,I[,1],I[,2],MoreArgs=list(A=dSCRi,alp=alp,
            sh1=sh1,sc1=sc1))
        e.nd1<-sum(p.nd[C.01==4])
        e.nd3<-sum(p.nd[C.01==16])
        pes0.ij<-(nc1+nc3)/(nc1+nc3+e.nd1+e.nd3)
        Dij.0<-C.0/ifelse(C.01==3 | C.01==15, pes0.ij,1)
        alp0<-alp
        Zij<-f.Zij(alp0,dSCRi,sh1=sh1,sc1=sc1,I,Dij.0)
        alp<-sum(W*Dij.0*Zij)/sum(W*Dij.0*(1-Zij))
        eS1<-iS1(time,alp,sh1=sh1,sc1=sc1)
        valp.0[iter]<-alp
        vS1.0[[iter]]<-eS1
        y0<-log(-log(eS1))
        y<-y0[!is.infinite(y0)]
        t<-time[!is.infinite(y0)]
        fit1<-lsfit(log(t),y)
        sh1<-fit1$coefficients[2]
        sc1<-exp(fit1$coefficients[1])
        iter<-iter+1
    }
    e.alp.0=valp.0[[max(which(valp.0!=0))]]
```

```
        S1.p0=vS1.0[[max(which(valp.0!=0))]]
50
        p.nd<-mapply(nD,I[,1],I[,2],MoreArgs=list(A=dSCRi,alp=alp,sh1=sh1,sc1=sc1))
        e.nd1<-sum(p.nd[C.01==4])
        e.nd3<-sum(p.nd[C.01==16])
        pes0.ij<-(nc1+nc3)/(nc1+nc3+e.nd1+e.nd3)
55      Dij.0<-C.0/ifelse(C.01==3 | C.01==15, pes0.ij,1)
        Zij<-f.Zij(alp0,dSCRi,sh1=sh1,sc1=sc1,I,Dij.0)

        jalp<-numeric(length=n)
        jS1<-matrix(nrow=length(time),ncol=n)
60      rownames(jS1)<-round(time,1)
        for(i in 1:n){
            indJ<-which(I[,1]!=i & I[,2]!=i)
            jalp[i]<-sum(Dij.0[indJ]*Zij[indJ])/sum(Dij.0[indJ]*(1-Zij[indJ]))
            jS1[,i]<-iS1(time,jalp[i],sh1=sh1,sc1=sc1)
65      }
        e.jalp<-mean(jalp,na.rm=T)
        var.p0<-(n-1)*mean((jalp-e.jalp)^2)
        e.jS1<-apply(jS1,1,mean)
        var.S1.p0<-(n-1)*apply((jS1-e.jS1)^2,1,mean)
70
    }else{
        e.alp.0=NA
        var.p0=NA
        S1.p0=rep(NA,length(time))
75      var.S1.p0=rep(NA,length(time))
    }

    #STRATEGY 2 ###################################################################
    if(p1){
80      iter<-1
        sh1<-sh1.0
        sc1<-sc1.0
        valp.01<-numeric(length=max.iter+1)
        vS1.01<-vector('list',length=max.iter+1)
85      alp0<-999
        alp<-e.alp.m
        pes01.ij<-rep(1,length(C.0))
        while((iter<=max.iter)&(abs(alp-alp0)>tol)&(alp>1)){
            pes01.ij[C.0==1]<-mapply(pesC,i=I[C.0==1,1],j=I[C.0==1,2],
90              MoreArgs=list(A=dSCRi,alp=alp,sh1=sh1,sc1=sc1,survC=survC,
                datLR=datLR,strat=strat,breaks=breaks))
            Dij.01<-C.0/pes01.ij
            alp0<-alp
            Zij<-f.Zij(alp0,dSCRi,sh1=sh1,sc1=sc1,I,Dij.01)
95          alp<-sum(W*Dij.01*Zij)/sum(W*Dij.01*(1-Zij))

            eS1<-iS1(time,alp,sh1=sh1,sc1=sc1)
            valp.01[iter]<-alp
            vS1.01[[iter]]<-eS1
100         y0<-log(-log(eS1))
```

```
            y<-y0[!is.infinite(y0)]
            t<-time[!is.infinite(y0)]
            fit1<-lsfit(log(t),y)
            sh1<-fit1$coefficients[2]
            sc1<-exp(fit1$coefficients[1])
            iter<-iter+1
        }
        e.alp.01=valp.01[max(which(valp.01!=0))]
        S1.p01=vS1.01[[max(which(valp.01!=0))]]

        pes01.ij[C.0==1]<-mapply(pesC,I[C.0==1,1],I[C.0==1,2],MoreArgs=
            list(A=dSCRi,alp=e.alp.01,sh1=sh1,sc1=sc1,survC,datLR,strat,breaks=breaks))
        Dij.01<-C.0/pes01.ij
        Zij<-f.Zij(e.alp.01,dSCRi,sh1=sh1,sc1=sc1,I,Dij.01)

        jalp<-numeric(length=n)
        jS1<-matrix(nrow=length(time),ncol=n)
        rownames(jS1)<-round(time,1)

        for(i in 1:n){
            indJ<-which(I[,1]!=i & I[,2]!=i)
            jalp[i]<-sum(Dij.01[indJ]*Zij[indJ])/sum(Dij.01[indJ]*(1-Zij[indJ]))
            jS1[,i]<-iS1(time,jalp[i],sh1=sh1,sc1=sc1)
        }
        e.jalp<-mean(jalp,na.rm=T)
        var.p01<-(n-1)*mean((jalp-e.jalp)^2)
        e.jS1<-apply(jS1,1,mean)
        var.S1.p01<-(n-1)*apply((jS1-e.jS1)^2,1,mean)

    }else{
        e.alp.01=NA
        var.p01=NA
        S1.p01=rep(NA,length(time))
        var.S1.p01<-rep(NA,length(time))
     }
    return(list(mat=mat,e.alp.0=e.alp.0,var.p0=var.p0,e.alp.01=e.alp.01,var.p01=
        var.p01,S1.p0=S1.p0,var.S1.p0=var.S1.p0,S1.p01=S1.p01,var.S1.p01=var.S1.p01))
    }
```

D.4.2 Functions `compIC` and `compIC.1`

```
compIC<-function(i,j,A)
{
min.R=min(A$R[i],A$R[j])
min.Y=min(A$Y[i],A$Y[j])
tmp<-0
A1<-(A$d1[i]+A$d1[j]>0)& (A$d2[i]+A$d2[j]>0)
A2<-if(A$d1[i]+A$d1[j]==1) min.R<=min.Y else T
A3<-if(A$d2[i]+A$d2[j]==1)
    A$d2[i]*A$Y[i]+A$d2[j]*A$Y[j]<(1-A$d2[i])*A$Y[i]+(1-A$d2[j])*A$Y[j] else T
return(A1*A2*A3)
}
```

```
compIC.1<-function(i,j,A)
{
rw<-0
min.R=min(A$R[i],A$R[j])
min.Y=min(A$Y[i],A$Y[j])
if(A$d1[i]+A$d1[j]==2) rw=1
if(A$d1[i]+A$d1[j]==1){
    if(A$d1[i]*A$Y[i]+A$d1[j]*A$Y[j]<(1-A$d1[i])*A$Y[i]+(1-A$d1[j])*A$Y[j]) rw=2
    else{
        if(min.R<=min.Y) rw=3
        else{
            if((A$d1[i]*A$L[i]+A$d1[j]*A$L[j]<=(1-A$d1[i])*A$L[i]+(1-A$d1[j])*A$L[j])) rw=4
            if((A$d1[i]*A$L[i]+A$d1[j]*A$L[j]>(1-A$d1[i])*A$L[i]+(1-A$d1[j])*A$L[j] ) ) rw=5
        }
    }
}
if(A$d1[i]+A$d1[j]==0) rw=6

if(A$d2[i]+A$d2[j]==2) cl=1
if(A$d2[i]+A$d2[j]==1) {
    if(A$d1[i]+A$d1[j]==2 | A$d1[i]+A$d1[j]==0){
        if(A$d2[i]*A$Y[i]+A$d2[j]*A$Y[j]<=(1-A$d2[i])*A$Y[i]+(1-A$d2[j])*A$Y[j]) cl=2
        if(A$d2[i]*A$Y[i]+A$d2[j]*A$Y[j]>(1-A$d2[i])*A$Y[i]+(1-A$d2[j])*A$Y[j]) cl=3
    }
    if(A$d1[i]+A$d1[j]==1){
        if(A$d1[i]*A$d2[i]+A$d1[j]*A$d2[j]==1) cl=2
        if(A$d1[i]*(1-A$d2[i])+A$d1[j]*(1-A$d2[j])==1) cl=3
    }
}
if(A$d2[i]+A$d2[j]==0) cl=4
return(cod=(cl-1)*6+rw)
}
```

D.4.3 Function nD

```
nD<-function(i,j,A,alp,sh1,sc1)
{
IND<-c(i,j)
if(A$d1[i]+A$d1[j]==1)
{
obs<-IND[which(c(A$d1[i],A$d1[j])==1)]
cen<-IND[which(c(A$d1[i],A$d1[j])==0)]
obs.R<-A$R[obs]
obs.L<-A$L[obs]
cen.L<-A$L[cen]
obs.Y<-A$Y[obs]
cen.Y<-A$Y[cen]
if(A$d1[i]+A$d1[j]==1 & obs.Y>cen.Y & obs.R>cen.L & obs.L<cen.L)
{
  if(A$d2[i]+A$d2[j]==2)
    Int<-integrate(f,obs.L,cen.Y,t=obs.Y,alp=alp,sh1=sh1,sc1=sc1)$value/
```

```
            integrate(f,obs.L,obs.R,t=obs.Y,alp=alp,sh1=sh1,sc1=sc1)$value
  else if(A$d2[i]+A$d2[j]==1 & A$d2[obs]==0)
    Int<-integrate(Q0,obs.L,cen.Y,t=obs.Y,alp=alp,sh1=sh1,sc1=sc1)$value/
            integrate(Q0,obs.L,obs.R,t=obs.Y,alp=alp,sh1=sh1,sc1=sc1)$value
  else {Int<-1}
}
else {Int<-1}
} else{Int<-1}
return(Int)
}
```

D.4.4 Function `pesC`

```
pesC<-function(i,j,A,alp,sh1,sc1,survC,datLR,strat,breaks)
{
 Rmin=min(A$R[i],A$R[j])
 Lmin=min(A$L[i],A$L[j])
 Ymin<-min(A$Y[i],A$Y[j])
 Y.obs<-A$d1[i]*A$Y[i]+A$d1[j]*A$Y[j]
 Y.cens<-(1-A$d1[i])*A$Y[i]+(1-A$d1[j])*A$Y[j]

 if(A$d2[i]+A$d2[j]>0 &Rmin<=Ymin){
  if(A$d1[i]+A$d1[j]==2) Int<-(sC(Ymin,survC)^2)
   if(A$d1[i]+A$d1[j]==1){
     if(Y.obs<Y.cens) Int<-(sC(Ymin,survC)^2) else {
      ind<-findInterval(Y.obs,breaks,rightmost.closed=T)+1
      datLR.st<-datLR[strat==ind,]
      datLR.st1<-datLR[strat>=ind,]
      num<-integrate(fun4.a,Lmin,Rmin,y=Ymin,alp=alp,sh1=sh1,sc1=sc1,
                datLR.st=datLR.st,datLR.st1=datLR.st1)$value
      den<-integrate(fun4.b,Lmin,Rmin,y=Ymin,alp=alp,sh1=sh1,sc1=sc1,
                datLR.st=datLR.st,datLR.st1=datLR.st1)$value
      Int<-(num*(sC(Ymin,survC)^2))/den
    }
  }
  if(A$d1[i]+A$d1[j]==0) Int<-1
  } else Int<-1
return(Int)
}
```

D.4.5 Function `f.Zij`

```
f.Zij<-function(alp,A,sh1,sc1,J,D)
{
  n<-length(A$L)
  Zij.0<-rep(0,nrow(J))
  Pi<-rep(0,nrow(J))
  Pj<-rep(0,nrow(J))
  Zij<-rep(0,nrow(J))
  Zij.0[D!=0]<-mapply(pij,J[D!=0,1],J[D!=0,2],MoreArgs=list(A=A,alp=alp,sh1=sh1,sc1=sc1))
  Pi[D!=0]<-pHi(J[D!=0,1],A,alp=alp,sh1=sh1,sc1=sc1)
```

```
  Pj[D!=0]<-pHi(J[D!=0,2],A,alp=alp,sh1=sh1,sc1=sc1)
  ind2<-which((Pi!=0)&(Pj!=0))
  Zij[ind2]<-Zij.0[ind2]/(Pi[ind2]*Pj[ind2])
  return(Zij)
}
```

D.4.6 Function `iS1`

```
iS1<-function(time,alp=0.01,sh1,sc1)
{
  eS1.0<-((Sz(time)^(1-alp)) - (S2(time)^(1-alp)) + 1)^(1/(1-alp))
  eS1<-eS1.0
  eS1[eS1.0>1]<-1
  aux<-min(which(is.na(eS1)==T & is.na(Sz(time))==F),9999)
  s1<-c(0,diff(eS1))
  t1<-min(max(which(((Sz(time)^(1-alp) - S2(time)^(1-alp)) >-1 ) & (0<=eS1) & (eS1<=1))),aux-1)

  S1.e<-numeric(length=t1)
  S1.e[1]<-eS1[1]
  for(i in 2:t1){
    if(is.nan(eS1[i])) S1.e[i]<-NaN
    else {S1.e[i]<-min(eS1[1:i],na.rm=TRUE)}
  }
  d<-length(time)-t1
  return(c(S1.e,rep(S1.e[t1],d)))
}
```

D.4.7 Internal functions

The following functions are employed by the previous procedures and will be soon available and documented at `http://www-eio06.upc.es/research/grass/`.

- **Functions `S1, f1, S, H, G, f, Q0`**: Functions describing the parametric form of
  - `S1`: Survival function of T_1 (local parametric Weibull fit).
  - `f1`: Density function of T_1 (local parametric Weibull fit).
  - `S`: Joint survival function of (T_1, T_2) (Clayton's copula).
  - `H`: Minus partial derivative of S with respect to the second component (T_2) (Clayton).
  - `G`: Minus partial derivative of S with respect to the first component (T_1) (Clayton).
  - `f`: Joint density function of (T_1, T_2) (Clayton model).
- **Functions Q0, I1, I2**: Functions returning arithmetic operations with the previous functions:
  - `Q0`: It returns `-G`.
  - `I1`: It returns `H*f`.
  - `I2`: It returns `S*f`.

- **Functions `pHi, pHij`:** Functions returning, under Clayton's copula model,
  - `pHi`: the probability of the observed data for individual i, $P(T_1 \in (a,b], T_2 = y_i)$ or $P(T_1 \in (a,b], T_2 > y_i)$,
  - `pHij`: the probability of the observed pair (i,j), `PHij=PHi*PHj`.
- **Function `pij`:** It computes the probability of being concordant and of observing the data $\mathcal{H}_{ij}$ for the pair (i,j), $P(\Delta_{ij} = 1, \mathcal{H}_{ij})$.
- **Functions `pij.1, pij.2`:** Functions returning the probability of being concordant given that $\delta_{2i} = 1$ and $\delta_{2j} = 1$ ($P_1(i,j)$, page 103) and the probability of being concordant given that $\delta_{2i} = 1$ and $\delta_{2j} = 0$ ($P_2(i,j)$, page 103).

D.4.8 Functions `rclay.exp` and `fsimulICSCR3`

These functions are used to generate censored bivariate data following Clayton's copula model for the survival function with exponential margins.

```
rclay.exp<-function(N,alp,lam1,lam2){
 v1<-runif(N)
 v2<-runif(N)
 u1<-v1
 u2<-((v1^(1-alp))*((v2^((1-alp)/alp))-1)+1)^(1/(1-alp))
 x<-(-1/lam1)*log(u1)
 y<-(-1/lam2)*log(u2)
 return(samp=cbind(x,y))
}

simulICSCR.3<-function(n,alp,rate1=1,rate2=1,cmax=5,nvisit=20,pvisit=0.70,
        sed=NA,C=T,pf=0){
 if(alp<=1) return('ERROR: the association parameter must be > 1')
 else {
  library(Epi)

  #generation of bivariate data
  if(!is.na(sed)) set.seed(sed)
  samp <- rclay.exp(N=n,alp=alp,lam1=rate1,lam2=rate2)
  T1<-samp[,1]
  T2<-samp[,2]

  #generation of censoring variable
  if(C==T) cens<-rep(cmax,n)
  else {
    if(!is.na(sed)) set.seed(round(sed/alp)+sed)
    cens<-runif(n,(pf)*cmax,cmax)
  }

  #generation of intervals of observation
  amp<-cmax/(nvisit)
  V<-matrix(nrow=n,ncol=nvisit+1)
  V[,1]<-rep(1,n) #all subjects are observed at baseline
  if(!is.na(sed)) set.seed(round(sed/alp)+sed)
  for(i in 2:(nvisit+1)){
    V[,i]<-rbinom(n,1,pvisit)
  }

  #generation of semi-competing risks data
```

```
  X<-pmin(T1,T2,cens)
  d1.s<-1*(T1<=pmin(T2,cens))
  Y<-pmin(T2,cens)
  d2.s<-1*(T2<=cens)
  Tz<-pmin(pmin(T1,T2),cens)
  dz<-1*(pmin(T1,T2)<cens)
  A<-as.data.frame(cbind(T1,T2,cens,X,d1.s,Y,d2.s,Tz,dz))

  #generation of interval-censored data

  I<-seq(0,cmax,by=amp)
  L<-numeric(length=n)
  R<-numeric(length=n)
  d1<-d1.s
  d2<-d2.s
  for(i in 1:n){
    if(d1.s[i]==1){
      max1<-which(I<=T1[i])
      max2<-max(max1[V[i,max1]==1])
      L[i]<-max(I[max2],0.01)
      min1<-which((I>T1[i])&(V[i,]==1))
      if(length(min1)>0) min2<- min(min1)
      else min2<-0
      R[i]<-min(I[min2],Y[i])
    }
    if(d1.s[i]==0){
      L[i]<-Y[i]
      R[i]<-Inf
    }
  }
  B<-cbind(A,L,R,d1,d2)

  #summary statistics
  dist<-B$R-B$L
  tab1<-stat.table(list('Intermediate'=d1),contents=list(count(),percent(d1)),data=B)
  tab2<-stat.table(list('Final'=d2),contents=list(count(),percent(d2)),data=B)
  tab3<-stat.table(list('Intermediate'=d1,'Final'=d2),contents=list(count(),
            percent(d1,d2)),data=B)
  tab4<-summary(dist[!is.infinite(dist)])

  a=c(n,rate1,rate2,alp,pvisit,tab1[2,2],ifelse(dim(tab2)[2]==2,tab2[2,2],tab2[2,1]),
     ifelse(dim(tab2)[2]==2,tab2[2,1],0),ifelse(dim(tab3)[3]==2,tab3[2,1,1],0),
     ifelse(dim(tab3)[3]==2,tab3[2,1,2],tab3[2,1,1]),
     ifelse(dim(tab3)[3]==2,tab3[2,2,1],0),ifelse(dim(tab3)[3]==2,tab3[2,2,2],
                tab3[2,2,1]), tab4[c(1,3,4,6)])
  names(a)<-c('n','rate1','rate2','alfa','pvisit','%d1','%d2','%Ind. cens','%00',
                    '%Dep.cens','%10','%11','Min','Median','Mean','Max')
  return(list(a=a,b=V, D=B))
  detach("package:Epi")

}
```