
Faculteit Wetenschappen
Vakgroep Toegepaste wiskunde en informatica

Academiejaar 2010-2011

# Multiple imputation for missing and delayed event type data in a competing risks survival setting: Time to Ph.D.-attainment

Machteld Varewyck

Prof. dr. E. Goetghebeur

Proefschrift ingediend tot het behalen van de academische graad van
master in de wiskunde, afstudeerrichting toegepaste wiskunde.

# Preface

Since my Bachelor thesis *The Analysis of Residuals in Logistic Regression* about two years ago, my interest in statistics has even increased. With the support of Prof. dr. Els Goetghebeur I learned to apply and adjust existing methods to the features of the data, followed by writing statistical programs in order to answer the research questions. I first realized the complexity of properly describing and analyzing databases.
The next academic year, I encountered the basic principles of survival analysis during the course *Survival Analysis*, taught by Prof. dr. Els Goetghebeur. Last summer I participated in the project analyzing 'Time to Ph.D.-attainment or withdrawal' under the supervision of Prof. dr. Els Goetghebeur and Katrien Baert. It soon became clear that the available database contained a lot of information, but also a lot of deficiencies impeding the analysis.
In consultation with Prof. dr. Els Goetghebeur we decided to refine the analysis on 'Time to Ph.D.-attainment', correcting for the missing and delayed event type data. In fact, we were working the other way round compared to other theses: we started from a real-life database, trying to develop the best methods to analyze the data, while other theses develop a new method and test its value to data at the end. The consequence is that no prescribed solution is available, so we began by writing a protocol handling these problems. Then we studied the multiple imputation method in order to treat the missing and delayed event type. But to apply the multiple imputation procedure in the competing risks setting we needed to adjust some theoretical methods to the features of the data.
With this thesis, I think we showed the power of the multiple imputation procedure on a real-life database.

We used SAS 9.2 and R 2.12.0 for the statistical results. Detailed source code can be obtained by contacting me at Machteld.Varewyck@UGent.be.

Finally, I also want to seize the opportunity to thank everyone who has contributed to this thesis.

- First and foremost, thank you to my promoter Els Goetghebeur. Not only for introducing me to a fascinating subject, but also for her enthusiasm, support and confidence throughout this thesis. I appreciate that she showed a lot of patience in correcting my mistakes and explaining statistical strategies.
- Thanks to Katrien Baert, Jozefien Buyze and Bart Van Rompaye for the practical guidance in survival analysis and programming. Thanks to Alain Visscher to inform me of the latest developments in this research.
- Thanks to the ECOOM-team especially Katrien De Boyser, Hans Groenvynck and Ronan van Rossem for providing the database and answers in exploring the data.
- Thanks to my close family, supporting me in everything I do, unconditionally: Thanks mom, dad, brothers, for being the most precious people in my life. Thank you for giving me all the opportunities I ever wanted in exploring my interests. Thanks to everyone besides them, being there for me when I needed them most. Thank you Frauke and Ellen for reading and adding your comments to this work.

Ghent, May 2011

Machteld Varewyck

# Toelating tot Bruikleen

De auteur geeft de toelating deze masterproef voor consultatie beschikbaar te stellen en delen van de masterproef te kopiëren voor persoonlijk gebruik.
Elk ander gebruik valt onder de beperkingen van het auteursrecht, in het bijzonder met betrekking tot de verplichting de bron uitdrukkelijk te vermelden bij het aanhalen van resultaten uit deze masterproef.

Gent, mei 2011

Machteld Varewyck

# Nederlandstalige Samenvatting

We beschrijven hier kort de statistische analyse die in deze thesis uitgevoerd werd. Voor de expliciete resultaten en figuren verwijzen we naar de uitgebreide Engelstalige versie.

**Hoofdstuk 1: Algemene Inleiding**

Jaarlijks worden door de overheid aanzienlijke bedragen geïnvesteerd in onderzoek en ontwikkeling, verdeeld over een veelvoud aan onderzoeksonderwerpen en onderzoeksgroepen. Het bestuderen van de efficiëntie van een doctoraatsopleiding is een belangrijke manier om de opbrengst van die inversteringen te bepalen.

Het ECOOM-team (Expertisecentrum Onderzoek en Ontwikkeling Monitoring van de Vlaamse Gemeenschap) heeft ons daartoe twee grote databases ter beschikking gesteld die opvolgingsinformatie bevatten over personen die een doctoraatsopleiding aanvatten in één van de vijf Vlaamse universiteiten tussen 1 oktober 1990 tot 30 september 2009.

We willen analyseren of en hoe specifieke variabelen van een student zoals geslacht, nationaliteit, wetenschapsveld enz. de tijd tot het behalen van een doctoraat beïnvloeden, gedurende de observatieperiode van 1 oktober 1990 tot 30 september 2009. Om deze onderzoeksvraag te beantwoorden moeten we rekening houden met een aantal aspecten:

- De analyse van gesponsorde tijd i.p.v. kalendertijd, aangezien sommige doctoraatsstudenten tijdens hun opleiding voor een bepaalde periode niet gesponsord werden.
- Het behalen van een doctoraatsdiploma (Ph.D.-attainment) en de stopzetting van een doctoraatsopleiding (withdrawal) zijn competing risks, dat wil zeggen dat de ene gebeurtenis de andere verhindert om plaats te vinden.
- Een stopzetting van de doctoraatsopleiding wordt niet expliciet gerapporteerd, enkel het behalen van een doctoraatsdiploma.
- De data reiken geen onderscheid aan tussen een lopende (ongoing - beide opties zijn nog open: uiteindelijk behalen van een doctoraatsdiploma of stopzetting van de doctoraatsopleiding) en stopgezette doctoraatsopleiding. Vermits sommige doctoraatsstudenten hun diploma behalen na een gap (interval van niet-gesponsorde tijd), betekent het einde

van een gesponsorde periode niet noodzakelijk een stopzetting van de doctoraatsopleiding.

- De data zijn rechts-gecensureerd, aangezien niet iedereen zijn/haar doctoraatsopleiding heeft afgerond tegen het einde van onze studie.

Voor onze analyse hebben we twee gecorreleerde datasets afgeleid uit de aangereikte databases van het ECOOM-team. We geven hieronder een korte beschrijving van deze datasets, met behulp van de definities die in sectie 1.3 (p. 3) ingevoerd worden.

1. Een dataset met alle geobserveerde gaps, deze bevat duur van de gap, gaptype (afwezigheid, tussentijd, succes, stopzetting) en covariaten ($\mathbf{Z}$): geslacht, nationaliteit, dominant statuut over de hele doctoraatsopleiding, wetenschapsveld, leeftijd (bij de aanvang van de doctoraatsopleiding), start tijdstip (jaar waarin de doctoraatsopleiding begon) en universiteit (anoniem behandeld).

2. Een dataset met één lijn per doctoraatsstudent, deze bevat duur van de sponsoring tot het eerst voorkomende event, type van event (lopend, doctoraat behaald, stopzetting) en dezelfde covariaten $\mathbf{Z}$.

De beschouwde categorieën voor deze covariaten, alsook een beschrijving van de gaptypes, zijn weergegeven in 1.3 (p. 3). We merken nog op dat enkel het event 'behaald doctoraatsdiploma' direct geobserveerd wordt. Daarom veronderstellen we de volgende assumpties voor de afleiding van de twee datasets:

- Iemand die niet gesponsord geweest is voor een periode langer dan 4 jaar wordt geclassificeerd als iemand die de doctoraatsopleiding stopgezet heeft (ook als die persoon later toch een doctoraatsdiploma behaalt). Een doctoraatsdiploma dat behaald zou worden na gap van meer dan 4 jaar, wordt niet langer gezien als de verwachte opbrengst van de inversteringen en dus niet beschouwd als een resultaat van de gesponsorde tijd.

- Iemand die een gap start na 30 september 2005 en waarbij die gap gecensureerd wordt op het einde van de studie, bevindt zich ofwel in een lopende of stopgezette doctoraatsopleiding en, hoofdzakelijk gebaseerd op informatie verkregen uit de eerste dataset, is het mogelijk om het event imperfect te bepalen (zie hoofdstuk 2).

Deze laatste assumptie veronderstelt dat we een imperfect onderscheid kunnen maken tussen lopende en stopgezette doctoraatsopleidingen voor alle studenten die zich in een gecensureerde gap bevinden van minder dan 4 jaar. Aangezien er geen absolute zekerheid bestaat over dit onderscheid, werd deze classificatie herhaaldelijk uitgevoerd (meervoudige imputatie). Op die manier hebben we verschillende 'vervolledigde' kopies van de tweede dataset geconstrueerd, waarbij alle missing events op een gefundeerde manier ingevuld werden (hoofdstuk 2). Zodra deze tweede dataset geen missing events meer bevat, kunnen we standaard statistische analysetechnieken toepassen uit het competing risks kader (hoofdstuk 3).

**Hoofdstuk 2: Imputatiemodel en Meervoudig Geïmputeerde Datasets**

In het tweede hoofdstuk worden meervoudig geïmputeerde datasets voor doctoraatsstudenten (2.) geconstrueerd, door de ontbrekende events via een imputatiemodel in te vullen. Voor alle studenten die zich op het einde van de studie in een gap van $t < 4$ jaar bevinden, wensen we de volgende kans te schatten

$$P(\text{withdrawer}|\mathbf{Z}, \text{gap} > t \text{ years}).$$

Dit is de kans om een doctoraatsopleiding stop te zetten conditioneel op baseline covariaten van de corresponderende student ($\mathbf{Z}$) en wetende dat de huidige gap minstens $t$ jaar duurt. We zullen deze kans schatten op basis van alle gaps startend voor 1 oktober 2005, omdat hierbij alle stopzetters duidelijk te indentificeren zijn als studenten met een gap langer dan 4 jaar.

We beginnen het hoofdstuk met een aantal assumpties. Zo veronderstellen we een stationair proces, opdat we de informatie verkregen vóór 1 oktober 2005 kunnen gebruiken om de verwachte kans op stopzetting te schatten voor studenten die zich in een gap bevinden op 1 oktober 2009. Vervolgens schetsen we het algemene kader voor meervoudige imputatie. Dan voeren we een korte beschrijvende analyse uit voor alle gaps, voor de gaps startend vóór 1 oktober 2005 en voor de gecensureerde gaps. Daarna introduceren we definities en notaties voor de overlevingsanalyse van de duur van een gap startend vóór 1 oktober 2005.

We kunnen de conditionele kans op het stopzetten van de doctoraatsopleiding als volgt herschrijven

$$P(\text{withdrawer}|\mathbf{Z}, \text{gap} > t \text{ years}) = \left(\frac{S_0(4)}{S_0(t)}\right)^{\exp(\boldsymbol{\beta}^T\mathbf{Z})}. \tag{0.1}$$

Deze gelijkheid geldt op voorwaarde dat het volgende Cox PH model geldt voor de gaptime $t$

$$h(t|\mathbf{Z}) = h_0(t)\exp(\boldsymbol{\beta}^T\mathbf{Z}).$$

Voor meer details verwijzen we naar sectie 2.5 (p. 22). Bovendien geldt de gelijkheid $-\ln(S_0(t)) = H_0(t)$. Om de kans (0.1) te schatten, volstaat het dus om de modelparameters $\boldsymbol{\beta}$ en de baseline cumulatieve hazard functie $H_0(t)$ te schatten. De modelparameters $\boldsymbol{\beta}$ worden geschat door een Cox PH model te bouwen voor $P(\text{gap} > t \text{ years}|\mathbf{Z})$ in sectie 2.6 (p. 23). De baseline cumulatieve hazard functie wordt geschat door de zogenaamde Breslow schatter in sectie 2.7 (p. 25).

Zodra deze kans (0.1) geschat is, kunnen we een imputatiemodel bouwen. We beschrijven twee mogelijkheden voor het imputatiemodel aangeduid door imputatiemethode A (equation (2.19), p. 29) en imputatiemethode B (equation (2.20), p. 29). Voor de tweede methode is het noodzakelijk om de verdeling van de cumulatieve hazard functie te bepalen. Bovendien kunnen

we op basis van die verdeling ook een onder- en bovengrens bepalen voor de onzekerheid van het imputatiemodel bekomen door methode B.

Met behulp van elk imputatiemodel kunnen we dan meervoudige imputatie (multiple imputation) toepassen om de ontbrekende events in te vullen. Dat doen we door de kans $P(\text{withdrawer}|\mathbf{Z}, \text{gap} > t \text{ years})$ te schatten voor alle studenten met een gecensureerde gap. Om rekening te houden met de onzekerheid van het opgebouwde imputatiemodel, construeren we telkens vijf verschillende 'vervolledigde' kopies van de dataset voor doctoraatsstudenten (2.). We vergelijken daarbij kort het aantal stopzettingen bekomen door de verschillende imputatiemethodes.

**Hoofdstuk 3: Competing Risks Analyse**

In het derde hoofdstuk worden de 'vervolledigde' kopies van de dataset voor doctoraatsstudenten (2.) geanalyseerd m.b.v. standaard competing risks procedures. We werken in een competing risks kader aangezien het behalen van een doctoraatsdiploma en het stopzetten van de doctoraatsopleiding competing risks zijn, de ene gebeurtenis verhindert de andere om plaats te vinden. Er bestaan verschillende aanpakken voor dit probleem, maar wij zullen focussen op het cause-specific proportionele hazards model, dat één event tegelijk bestudeert, terwijl het de andere types van events behandelt als gecensureerd.

We voeren eerst een korte beschrijvende analyse uit voor alle doctoraatsstudenten die hun opleiding starten tussen 1 oktober 1990 en 30 september 2009. Vervolgens introduceren we definities en notaties voor de overlevingsanalyse van de gesponsorde duur van een doctoraatsopleiding startend voor 1 oktober 2009 in het competing risks kader.

We bouwen voor elke geïmputeerde dataset van beide imputatiemethodes een regressiemodel. We maken daarvoor het onderscheid tussen een lopende doctoraatsopleiding, een behaald doctoraat en de stopzetting van een opleiding. Merk op dat het soort event (lopende of stopzetting) kan afhangen van imputatie tot imputatie, maar de gesponsorde tijd niet. We bouwen dan een cause-specific proportionele hazards model voor beide uitkomsten: behaald doctoraat of stopzetting.

Vervolgens worden cumulatieve incidentie functies en geassocieerde varianties geschat voor beide uitkomsten als een relatief eenvoudige illustratie van de bekomen resultaten. Daarna worden de combinatieregels voor resultaten van meervoudig geïmputeerde datasets toegepast om elke vijf schattingen per imputatiemethode te combineren. Op die manier bekomen we een gemiddelde schatting en totale variantie voor de cumulatieve incidentie functie. We bespreken enkele cumulatieve incidentie curves voor het behalen van een doctoraatsdiploma in vergelijking met de referentiegroep, één covariaat tegelijk bekijkend. Deze referentiegroep bestaat uit Belgische, mannelijke doctoraatsstudenten gesponsord door andere projecten, met

dominant wetenschapsveld 'wetenschappen', die minder dan 25 jaar oud waren toen ze hun doctoraatsopleiding begonnen tussen 1 oktober 1990 en 30 september 1997, aan een specifieke Vlaamse universiteit.

Tenslotte vergelijken we de resultaten bekomen met deze methode van meervoudige imputatie en een meer naïve methode zonder imputatie uit een eerdere studie. Deze methodes verschillen in de manier waarop ze de ontbrekende uitkomsten behandelen, maar maken allebei gebruik van de cause-specific hazards functie voor de competing risks analyse.

**Hoofdstuk 4: Cox Proportionele Hazards Cure Model**

In dit hoofdstuk wordt een alternatieve methode voorgesteld, het cure model, waarbij niet langer gebruik gemaakt wordt van de cause-specific PH functies in het kader van de competing risks. Deze methode wordt theoretisch uitgewerkt, maar niet concreet toegepast op onze dataset.

Het cure model werd speciaal ontwikkeld voor data waarbij een bepaald deel van de populatie het beschouwde event niet zal meemaken tegen het einde van de observatie. In dat geval zeggen we dat sommige van deze overlevers niet vatbaar (cured) zijn in de zin dat, hoe lang we deze personen ook zouden opvolgen, geen verdere events waargenomen zullen worden. Voor onze analyse zijn deze niet-vatbare studenten de stopzetters omdat we veronderstellen dat, ook al zouden zij extra gesponsorde tijd krijgen, ze nooit een doctoraatsdiploma zullen behalen, omdat ze bijvoorbeeld aangesteld werden als goedkope werkkrachten aan de universiteit zonder doctoraatsdoeleinde.

Enerzijds kunnen we de *prevalentie* schatten, die bepaalt of het doctoraatsdiploma kan behaald worden en dus vatbare van niet-vatbare studenten onderscheidt bij het begin van de observatie. Dat kan door de bouw van een logistisch regressiemodel. Anderzijds kunnen we de *latentie* schatten, die bepaalt wanneer het doctoraatsdiploma zal behaald worden op voorwaarde dat de student de intentie heeft om een doctoraatsdiploma te behalen. Dat kan door de bouw van een Cox PH model.

Het probleem van de missing data stelt zich hier opnieuw, maar nu wordt er gebruik gemaakt van het verwachting-maximalisering (EM) algoritme. Dit is een algemene techniek voor maximum-likelihood schatting in parametrische modellen voor incomplete data.

Tot slot lijsten we enkele theoretische verschillen op ter vergelijking van de meervoudige imputatie - competing risks model methode toegepast in deze thesis en de verwachting maximalisering - cure model methode. Daarbij merken we vooral het verschil in conceptualisatie en parametrisatie van de meervoudige event types op.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Preliminaries

## 1.1 Introduction

Information about sponsored time before Ph.D.-attainment at one of the five Flemish universities is important for the government budget. Yearly, significant amounts are invested in research and development, distributed over a variety of research topics and in favor of a variety of research groups. Therefore, studying the efficiency of the Ph.D.-trajectory is one way to determine the gain of those investments.

The Centre for Research and Development Monitoring (ECOOM) is an inter university consortium with participation of all Flemish universities (K.U.Leuven, UGent, VUB, UA and UHasselt). Its mission is to develop a consistent system of research, development and innovation indicators which have to assist the Flemish government in mapping and monitoring efforts in the Flemish region. This thesis aims to describe the influence of some important indicators on the time to the attainment of a Ph.D.-degree. Differences between the five Flemish universities may not be reported explicitly.

We received two large databases from the ECOOM-team containing information about all individuals who started a Ph.D.-training at one of the five Flemish universities between October 1, 1990 and September 30, 2009, by which time not everyone had finished his/her Ph.D.-training. Students who attained a Ph.D.-degree after October 1, 1990, but started their Ph.D.-training prior to October 1, 1990 are *not* included in the databases. The first database contains one line for each Ph.D.-student and provides general information about his/her career. The second database contains one or more lines for each Ph.D.-student, each line representing a period without changes in classification (e.g. university, absence) and thus providing crucial information about intervals of non-sponsored time.

## 1.2 Description of the Research Question

We wish to analyze whether and how students' characteristics such as gender, nationality, scientific field etc. influence time from the date of first appointment as Ph.D.-student to the attainment of a Ph.D.-degree, during the observation period from October 1, 1990 until September 30, 2009.

Basically, for analyzing time to Ph.D.-attainment we have two options: We can focus on

- the total calender time from the date of first appointment as Ph.D.-student to the attainment of a Ph.D. degree, or
- the sponsored time between the date of first appointment as Ph.D.-student to the attainment of a Ph.D.-degree. Hence, from the total time is deducted the time in gaps (intervals of non-sponsored time).

In consultation with the ECOOM-team, we focus on the second option as sponsored time is clearly of greater interest to the government. The calculation of this sponsored time from the dataset is given below.

We know that not all individuals starting a Ph.D.-training eventually attain a Ph.D.-degree. A substantial part of the Ph.D.-students withdraws early from the training. Withdrawing from the Ph.D.-training prevents the student from attaining a Ph.D.-degree and vice versa the attainment of a Ph.D.-degree prevents the student from withdrawing. So, our study of the sponsored time to Ph.D.-attainment accounting for withdrawers from the training, is framed in a competing risks survival setting since multiple types of events (Ph.D.-attainment or withdrawal) are being observed where one event prevents the other event type from occurring.

Analysis of sponsored time before Ph.D.-attainment based on the reported cases is complicated by the fact that only Ph.D.-attainments are reported as endpoints at a given time and not the withdrawals. A possible indicator for withdrawal from the Ph.D.-training is that the Ph.D.-student is no longer sponsored. Although, for some students the Ph.D.-degree is attained after a gap (interval of non-sponsored time), so the end of a sponsored period needs not necessarily indicate the end of the Ph.D.-training or withdrawal. For our purposes, a Ph.D.-degree which might be reached after more than 4 years of non-sponsored time is no longer seen as the expected yield of the investment and thus considered as no result for the sponsored time. A cut-point $X$ is therefore set at 4 years and all Ph.D.-students having a gap lasting more than $X = 4$ years will be classified as withdrawers. So, withdrawals may be reported long after the last sponsoring time and therefore this type of event has to deal with fixed reporting delay.

Since not everyone had finished his/her Ph.D.-training by September 30, 2009, our data are

right censored. More specifically, two different censoring times are recorded, according to the type of event:

- Withdrawals are observed as a gaptime exceeding 4 years if they occur prior to October 1, 2005.
- Ph.D.-attainments which are observed if they occur prior to October 1, 2009.

A portion of all observed Ph.D.-students is not sponsored by the end of the study and started this gap after September 30, 2005. Ideally, we wish to distinguish withdrawers from ongoing Ph.D.-students who still have both options (eventually attaining a Ph.D.-degree or withdrawing from the Ph.D.-training). Deleting all Ph.D.-students with a gap lasting less than 4 years by the end of the study is no option, as it will not only result in severe information loss but also in bias. To remove this bias, extra information or assumptions are needed. In this project we focus on statistical methods for survival analysis assuming some stationarity of the process over time.

## 1.3 Definitions and Data Derivation

For our analysis purposes, we have derived two correlated data sets from the given databases. We describe the derivation of the following two data sets, using the definitions we will introduce in this section:

1. A data set with all observed gaps containing the gaptime, gaptype (absence, interim, success or withdrawal) and covariates: gender, nationality, dominant statute over the entire Ph.D.-training, scientific field, age (at the start of the Ph.D.-training), start time (year in which the Ph.D.-training began) and university.
2. A data set with one line per Ph.D.-student containing sponsored time to the first event occurring, type of event (ongoing, Ph.D.-attainment or withdrawal) and same covariates.

As mentioned before, the observed data contains direct information about Ph.D.-attainment only and not about withdrawers. Moreover, since for some Ph.D.-students it is not directly observed whether they are ongoing or have withdrawn by the end of the study, we have made some assumptions:

- Someone who has not been sponsored for a period lasting more than the cut-point $X = 4$ years is classified as withdrawer (even if that person attains a Ph.D.-degree later on).
- Someone who started a gap after September 30, 2005 that was censored by the end of the study, is either an ongoer or a withdrawer and, mainly based on information provided by the first data set, this student's event type may be imperfectly determined (see Chapter 2).

The latter assumes that we can imperfectly distinguish withdrawers from ongoers for all Ph.D.-students having a censored gap lasting less than 4 years. As there is no absolute certainty about this distinction, the classification is performed repeatedly (multiple imputation). In that way we construct several 'completed' copies of the second data set with all missing event types filled in, in a well-founded way. Remark that, as we analyze sponsored time instead of calender time, the analyzed time does not depend on whether the missing event type is filled in as ongoing or withdrawal.

In consultation with the ECOOM-team, a vector with covariates of interest ($\mathbf{Z}$) is selected out of all available covariates. We list these baseline covariates with specified categorization [reference category]:

- gender: [male], female,
- dominant statute classification: Assistant lectureship, Competitive scholarship (Flanders), Competitive scholarship (own university), Project funding (FWO, BOF or IUAP), [Project funding (other)],
- dominant scientific field: medicine, humanities, social sciences, applied sciences, [sciences],
- age (at the start of the Ph.D.-training): [$\leq$ 25 years], 26-30, 31-35, 36-40, $>$ 40 years,
- start year (year in which the Ph.D.-training began): [1990-1997], 1997-2004, 2005-2009 (in each case from October 1 until September 30),
- nationality: [Belgian], European Union (excl. Belgium), other,
- university: [anonymous] KUL, UA, UG, UH, VUB.

In sum, the reference group consists of Belgian men funded by other projects, with the dominant scientific field 'sciences', who were less than 25 years old when they started their Ph.D.-training between October 1, 1990 and September 30, 1997 at a specific (known, but unnamed for confidence) Flemish university.

None of these covariates is considered as time dependent for the study of time from starting the Ph.D.-training to attainment or withdrawal. This is also true for the covariate 'start year' since it is known as soon as the individual enters the study.
The categories of the last variable, university, are ordered arbitrarily, because we do not wish to compare the different universities. We do not, however, expect the estimated hazards to be the same for all five Flemish universities.
If a person has spent sponsored time at more than one university, we assign this person to the university where most time was spent. If a person has spent sponsored time in more

than one statute classification or scientific field during the Ph.D.-training, the dominance was determined by the ECOOM-team, based on a decision tree.

In total, we observe 28,871 individuals starting a Ph.D.-training between October 1, 1990 and September 30, 2009, including 475 people with missing values for one or more covariates.

Some Ph.D.-students have a gap (interval of non-sponsored time) during their Ph.D.-training. We encounter 4 different types of gaps, which are defined as follows:

- interim gap: time between two observed appointments as a Ph.D.-student,
- absence gap: time (within an appointment) during which the Ph.D.-student is not active for at least three months due to pregnancy, sick leave or other reasons,
- success gap: time between the last observed appointment and attainment of a Ph.D.-degree before September 30, 2009,
- withdrawal gap: time from the last appointment onwards for those withdrawn from the Ph.D.-training (this information is not directly observed).

Any gap not ended by September 30, 2009 is censored and the corresponding Ph.D.-student is either a withdrawer or an ongoer, i.e. someone who still has both options (eventually attaining a Ph.D.-degree or withdrawing from the Ph.D.-training). According to these definitions, the following equality holds:

$$\begin{aligned} P(\text{ongoer}|\mathbf{Z}, \text{gap} > t \text{ years}) &= P(\text{interim gap}|\mathbf{Z}, \text{gap} > t \text{ years}) \\ &\quad + P(\text{success gap}|\mathbf{Z}, \text{gap} > t \text{ years}) \\ &\quad + P(\text{absence gap}|\mathbf{Z}, \text{gap} > t \text{ years}) \\ &= 1 - P(\text{withdrawer}|\mathbf{Z}, \text{gap} > t \text{ years}), \end{aligned}$$

involving the 4 categories of gaptypes.

We assemble back-to-back gaps, i.e. two or three gaps without a sponsored period in between, including at least one absence. As any person with a non-sponsored gap of more than 4 years is classified as a withdrawal by definition, any subsequent gaps were removed from the data set.

Times (sponsored time, gaptime, etc.) are calculated as follows: normally begin and end date are given, so calculate end date minus begin date plus one day. Then we obtained sponsored time in days, which is no good timescale for these analyses. When we divide by 365, the general number of days per year, we become the sponsored time in years. But the main problem with this strategy are the leap years. So we decided to divide by 365.25 (average number of days per year) in order to correct for a leap year every four years. We preferred this strategy because of the definition of withdrawers (in a gap lasting more than 4 years). In this way we are sure the timespan is given correctly every 4 years.

## 1.4 Our Methodological Approach

In this project, methods for survival data that satisfy the structure described above is being developed. The methods are illustrated by focusing on the Ph.D.-attainment study, but the proposed methodology can equally be applied in other formal settings.

We start by describing the standard analysis method if complete data were available, so if all information in the data set of Ph.D.-students (2.) were directly observed. Since Ph.D.-attainment and withdrawal are competing outcomes, one event prevents the other from occurring, we are working in the competing risks framework. There are several approaches to this problem, but we will focus on the cause-specific proportional hazards model that studies one cause at a time, treating other types of events as censored observations. Therefore, a cause-specific proportional hazards model is built for both competing risks: time to Ph.D.-attainment and time to withdrawal. To present statistical results from a competing risks analysis, it is useful to estimate cumulative incidence functions based on the models for the cause-specific hazards. Since we have to deal with incomplete data, this standard method can not directly be applied.

Chapter 2

In this chapter several 'completed' copies of the data set of Ph.D.-students (2.) are constructed, by filling in the missing outcomes in a well-founded way. For Ph.D.-students ending our observation period with a gap of $t$ years we wish to estimate the probability

$$P(\text{withdrawer}|\mathbf{Z}, \text{gap} > t \text{ years}),$$

of being a withdrawer conditional on knowing the current gap lasts $t$ years or more and baseline covariate values $\mathbf{Z}$. Assuming a stationary process, we base this study on all gaps starting prior to October 1, 2005, so all withdrawers are clearly identified.
Remark that a withdrawer by definition ends his Ph.D.-study by a withdrawal gap. And vice versa every Ph.D.-student who ends his study by a withdrawal gap of more than 4 years is a withdrawer. So the withdrawal gap could be seen as reporting-lag time for withdrawal. In that way this probability can be written as

$$\begin{aligned} P(\text{withdrawer}|\mathbf{Z}, \text{gap} > t \text{ years}) &= P(\text{withdrawal gap}|\mathbf{Z}, \text{gap} > t \text{ years}) \\ &= P(\text{gap} > 4 \text{ years}|\mathbf{Z}, \text{gap} > t \text{ years}). \end{aligned}$$

In Chapter 2 this probability will be estimated to build an imputation model.

Once the imputation model is obtained, we can apply multiple imputation (MI) to fill in the missing event types. So, we estimate $P(\text{withdrawer}|\mathbf{Z}, \text{gap} > t \text{ years})$ for all missing outcomes, taking into account the uncertainty of this probability caused by the model building.

Remember that these missing outcomes are all concerning Ph.D.-students having a gap censored by the end of the study and lasting less than 4 years. Moreover, we construct lower and upper limit data sets in an attempt to represent the boundaries for the uncertainty of the imputation model.

Chapter 3

In that way $D$ - different - completed data sets of Ph.D.-students (2.) are obtained and standard competing risks methods can be applied to the completed data sets. We build $D$ competing risks models based on the $D$ completed data sets, which contain all Ph.D.-students starting a Ph.D.-training between October 1, 1990 and September 30, 2009.
We use the following classification:

0. ongoer: in principle someone who has not yet attained the Ph.D.-degree, and may continue to become either an attainer or a withdrawer. In practice, this is anyone sponsored at September 30, 2009 as well as the Ph.D.-students who are having a gap by September 30, 2009 lasting less than 4 years and classified as ongoer by the MI procedure.

1. Ph.D.-attainer: finished the Ph.D.-training successfully by September 30, 2009.

2. withdrawer: having a gap lasting more than 4 years by September 30, 2009 as well as the Ph.D.-students who are having a gap by September 30, 2009 lasting less than 4 years and classified as withdrawal gap by the MI procedure.

In the main analysis, the outcomes 'attainment of a Ph.D.-degree' and 'withdrawal' are considered competing risks: one event prevents the other from occurring. We aim to estimate the event-free time for both competing risks at specific covariate values, therefore a cause-specific proportional hazards model is built, based on all completed data, for both outcomes to handle the multiple event types.

Next, cumulative incidence functions and associated variances are estimated for both competing risks as a relatively simple illustration of the results. Finally, the rules for combining multiple imputation results are used to combine these $D$ estimates and estimated variances of the cumulative incidence function and calculate the mean estimate and total variance of the cumulative incidence function.

Chapter 4

In this chapter an alternative method, the PH cure model, is introduced for handling the multiple event types. We list some theoretical differences, comparing the multiple imputation - competing risks model approach used in this thesis and the expectation maximization - cure model approach proposed in this chapter.

# Chapter 2

# Imputation Model and Multiple Imputed Data Sets

Our goal in this chapter is to estimate the probability of being a withdrawer for every Ph.D.-student observed with a gap lasting less than 4 years by the end of the study (censored gap). Using the imputation model we estimate the probability of withdrawal, i.e. a gap of ultimately more than 4 years, conditional on the current gap length $t$ ($0 \leq t \leq 4$) and the previously specified list of covariate values $\mathbf{Z}$ in section 1.3 (p. 3). We will assume non-informative censoring here. Hence for $0 \leq t \leq 4$:

$$\begin{aligned} P(\text{withdrawer}|\mathbf{Z}, \text{gap} > t \text{ years}; C = t) &= P(\text{withdrawer}|\mathbf{Z}, \text{gap} > t \text{ years}) \\ &= P(\text{gap} > 4 \text{ years}|\mathbf{Z}, \text{gap} > t \text{ years}) \\ &= \frac{P(\text{gap} > 4 \text{ years}|\mathbf{Z})}{P(\text{gap} > t \text{ years}|\mathbf{Z})} \\ &= \frac{S(4|\mathbf{Z})}{S(t|\mathbf{Z})}, \end{aligned} \tag{2.1}$$

where we used Bayes' rule to rewrite the conditional probability and $S(t|\mathbf{Z})$ represents the survival function at time $t$, conditional on covariate values $\mathbf{Z}$. The above probability will be computed in this chapter.

This analysis is based on the data of all gaps starting prior to October 1, 2005, because all gaps starting within this period have been observed for an additional 4 years. Given our definition of Ph.D. 'success' vs. 'withdrawer', they will thus be recognized for what they are. Specifically any withdrawal gap occurring in this set shows up as a gap not ended by 4 years and hence is clearly identified as a withdrawal.

To perform the analysis, we have made some assumptions:

- We are assuming a homogeneous population of gaps within and between people in this subpopulation, so there is no within person correlation for gaps and we can analyze all

observed gaps as if exchangeable in one large population of gaps.

- We are assuming a (fairly) stationary process, so we can use this information obtained prior to October 1, 2005 to estimate the expected withdrawal status for students with censored gaps by September 30, 2009.
- We are assuming withdrawals were recognized 4 years after the event.

This last assumption is justified because Ph.D.-degrees attained after more than 4 years of non-sponsored time are not seen anymore as a direct result of the sponsorship before the withdrawal gap. So, based on the first assumption and the described subset of all gaps starting between October 1, 1990 and September 30, 2005, we will model the probability of having withdrawn (2.1). Then, based on the second assumption of a fairly stationary process of gaps we will use this imputation model to estimate the probability of having withdrawn for all gaps censored by the end of the study, conditional on knowing their current gap length and the corresponding covariate values of the individual. In that way, we construct multiple imputed - complete - data sets, filling in all missing outcomes.

We start this chapter with a description of the multiple imputation procedure.

## 2.1 Multiple Imputation Procedure

Missing data or more generally incomplete data (e.g. censored data which are not completely missing because we know which interval they fall into) occur in many settings: medical sciences, social sciences etc. In our setting we are confronted with a possibly censored event time as well as missing event type and delayed reporting of withdrawal. This has several implications:

- loss of information and efficiency due to loss of data,
- difficulties in analyzing the data when using standard statistical methods,
- bias due to systematic differences between the observed and unobserved data,
- etc.

The multiple imputation procedure is fully described in [11]: Multiple imputation refers to the procedure of replacing each missing value by a vector of $D \geq 2$ imputed values. The $D$ plausible values are ordered in the sense that $D$ completed data sets can be created from the vectors of imputations: replacing each missing value by the first component in its vector of imputations creates the first completed data set, replacing each missing value by the second component in its vector creates the second completed data set, and so on. Standard complete-data methods are used to analyze each data set to yield 'completed-data' statistics, which are

typically complete-data estimates, $\hat{Q}$ and associated variance-covariance matrices, $U$. When $D$ sets of imputations are repeated random draws from the particular model for nonresponse of the missing values, the $D$ complete-data statistics can be combined to form one inference that properly reflects uncertainty due to nonresponse under that model.

The imputation procedure is a useful method since the imputation and the analysis of the data could be performed separately. It is typically assumed that the data collector and the data analyst are different persons, moreover the data collector could have access to important information helping to impute the missing values. Originally multiple imputation was viewed as being most appropriate in complex surveys that are used to create public-use data sets to be shared by many ultimate users or in complex surveys with standard complete-data analyses that are difficult to modify analytically in the presence of nonresponse, although, over the years, it has proven valuable in other settings as well [13].

We will use so-called regression imputation that replaces missing values by predicted values from a regression of the missing item on items observed for the unit. The $D$ imputations of the missing event type ($Y_{\text{mis}}$) are $D$ repetitions from the predictive distribution of ($Y_{\text{mis}}$), each repetition corresponding to an independent draw of the parameters and missing values. Specifically, we will draw a 0/1 value from the predictive distribution estimated by (2.1) first and then either impute the event 'ongoer' or 'withdrawer'.

For single imputation, only one imputed data set is generated. So standard complete-data methods could be directly performed and no further combination methods are needed. The major disadvantage of this method is that the imputed values are treated as known values and thus not reflecting the uncertainty about the model for nonresponse. As noted in [11], single imputation inference tends to overstate precision because it omits the between-imputation component of the variability.

Multiple imputation shares both advantages of single imputation and rectifies the disadvantages [11]. Specifically, when the $D$ imputations are repetitions under one model for nonresponse, the resulting $D$ complete-data analyses can easily be combined to create an inference that validly reflects sampling variability because of the missing values. The only disadvantage of multiple imputation over single imputation is that it takes more work to create the imputations and analyze the results. The extra work in analyzing the data, however, is really quite modest in today's computing environments, since it basically involves performing the same task $D$ times instead of once.

Some may view multiple imputation as making up data, but [15] provides a good counter argument for this statement. This objection is quite valid for single imputation methods, which treats imputed values not differently from observed ones. Multiple imputation, however, is nothing more than a device for representing missing data uncertainty. Information is not

being invented with multiple imputation any more than with the expectation-maximization algorithm or other well accepted observed data likelihood-based methods, which average over a predictive distribution for ($Y_{\text{mis}}$) by numerical techniques rather than by simulation.

Removing incomplete cases seems much easier than multiple imputation, but it only works well if the discarded cases form a representative and relatively small portion of the entire data set. However, case deletion leads to valid inferences in general only when missing data are missing completely at random (MCAR) in the sense that the probabilities of response do not depend on any data values observed or missing [15]. Multiple imputation assumes missing data to be missing at random (MAR). Although, missing at random (MAR) is a non-testable assumption, it has been pointed out in the literature that we can get very close to MAR if we include enough variables in the imputation model [14].

## 2.2 Descriptive Analysis of the Gaps Data

We perform a descriptive analysis of the gaps data, described in section 1.3 (p. 3).

### 2.2.1 All Gaps

First, we briefly describe our cohort of gaps starting between October 1, 1990 and September 30, 2009.

We consider 4 academic year-epochs of which the last epoch has one year less timespan, namely

- October 1, 1990 → September 30, 1995,
- October 1, 1995 → September 30, 2000,
- October 1, 2000 → September 30, 2005,
- October 1, 2005 → September 30, 2009.

Note that these year-epochs do not correspond to the categories defined for start cohort, suggested by the ECOOM-team, since the latter has one category less.

The number of Ph.D.-students starting per academic year-epoch is represented in table 2.1, considering all students starting between October 1, 1990 and September 30, 2009. The number of students starting per year-epoch increases over time, consequently we expect the number of gaps starting per year also to increase.

| | Epoch of start academic year | | | |
|---|---|---|---|---|
| year-epoch | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| 1990-1995 | 4,695 | 16.26 | 4,695 | 16.26 |
| 1995-2000 | 7,796 | 27.00 | 12,491 | 43.26 |
| 2000-2005 | 8,493 | 29.42 | 20,984 | 72.68 |
| 2005-2009 | 7,887 | 27.32 | 28,871 | 100.00 |

**Table 2.1:** Frequency and percentage of students starting a Ph.D.-training per academic year-epoch, between October 1, 1990 and September 30, 2009.

Figure 2.1 shows the number of each gaptype starting per academic year. E.g. academic year 1990-1991 contains all gaps starting between October 1, 1990 and September 30, 1991.

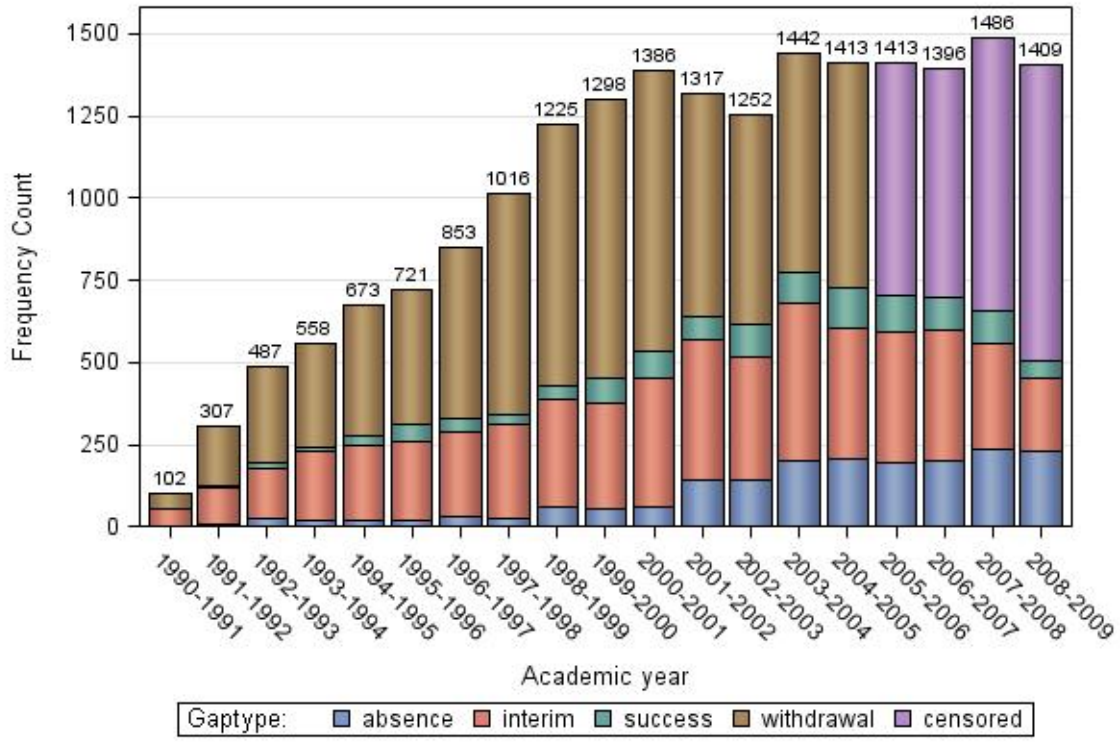


**Figure 2.1:** Frequency of gaps starting per academic year between October 1, 1990 and September 30, 2009.

We note that the total number of gaps starting each year increases until the year 2000, but then remains approximately constant. This could be explained as follows. In the first place, a

general Ph.D.-student who starts a gap, tends to start it after a few years of sponsoring time, especially success gaps. Hence, people starting early in the study haven't been sponsored long enough to start a gap. In the second place this graph expresses frequency counts and as the number of people starting a Ph.D.-training increases, the number of gaps starting each year is also expected to increase.
The total number of withdrawal gaps starting each year appears to stabilize as of 2000 and we work under the assumption of a stationary process of gaps. Obviously censored gaps only occur from October 1, 2005 and no withdrawal gaps occur from then on.

Figure 2.2 shows the frequency and percentage of each gaptype starting within an academic year-epoch. The underlying ratio of gaptypes starting within each academic year-epoch remains approximately constant, supporting the assumption of a stationary process of gaps.

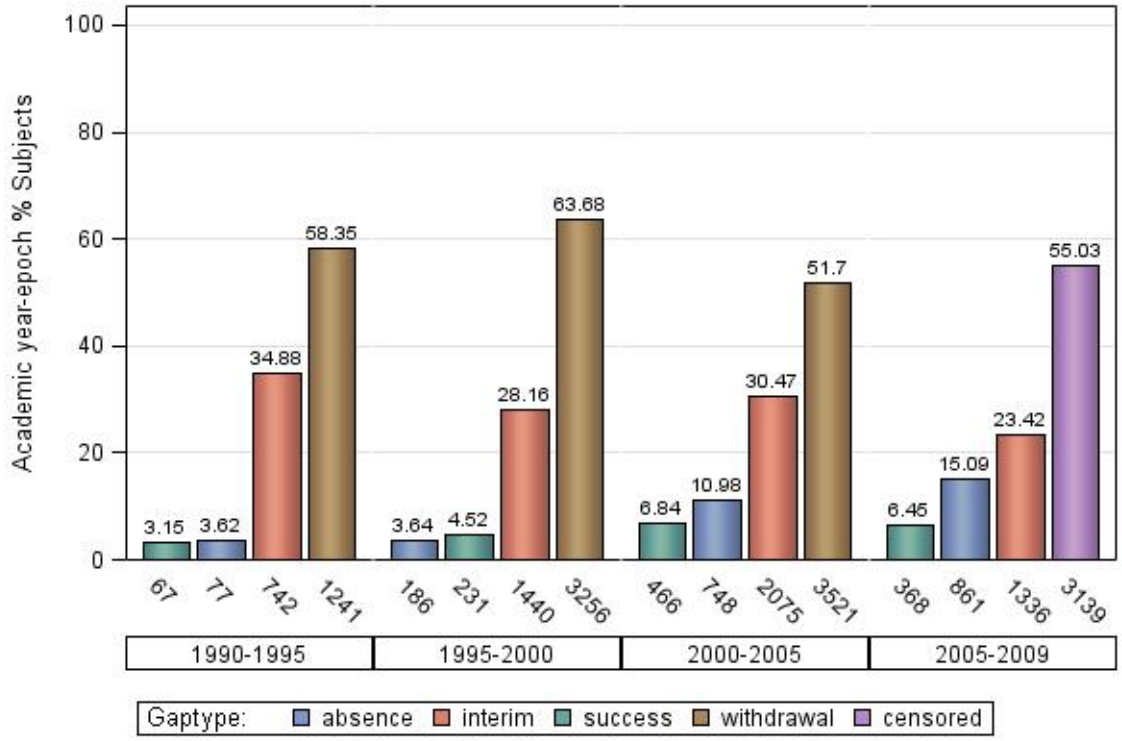


**Figure 2.2:** Frequency and percentage of each gaptype starting per academic year-epoch for all gaps starting between October 1, 1990 and September 30, 2009.

So, the total number of gaps starting each year increases until the year 2000 and the ratio of gaptypes starting within a year-epoch remains approximately constant.

### 2.2.2 Gaps Starting Prior to October 1, 2005

As we will only use the gaps starting prior to October 1, 2005 to build the imputation model, we describe this subset separately in this section.

| Number of gaps | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 1 | 9,240 | 81.86 | 9,240 | 81.86 |
| 2 | 1,544 | 13.68 | 10,784 | 95.54 |
| 3 | 356 | 3.15 | 11,140 | 98.69 |
| 4 | 104 | 0.92 | 11,244 | 99.61 |
| 5 | 31 | 0.27 | 11,275 | 99.88 |
| 6 | 8 | 0.07 | 11,283 | 99.96 |
| 7 | 5 | 0.04 | 11,288 | 100.00 |

**Table 2.2:** Frequency and percentage of gaps per person, over all gaps starting between October 1, 1990 and September 30, 2005.

The total number of Ph.D.-students having a gap that started prior to October 1, 2005, equals 11,288 distributed over 22,775 students starting the Ph.D.-training prior to October 1, 2005. So about half of the Ph.D.-students had at least one gap (absence, interim, success or withdrawal) and 13 students had more than 5 gaps starting between October 1, 1990 and September 30, 2005. About 82% of the students who had a gap during the observation period, had such gap only once.

All assembled gaps starting prior to October 1, 2005 can be categorized by type as follows:

| Gaptype | Frequency | Percent |
|---|---|---|
| absence | 1,011 | 7.20 |
| interim | 4,257 | 30.30 |
| success | 764 | 5.44 |
| withdrawal | 8,018 | 57.07 |

**Table 2.3:** Frequency and percentage of gaps per gaptype, over all gaps starting starting between October 1, 1990 and September 30, 2005.

Based on this subset of all gaps starting prior to October 1, 2005, we estimate the probability of a withdrawal gap for a random gap by

$$\hat{P}(\text{withdrawal gap}) = 57.07\%. \tag{2.2}$$

Figure 2.3 shows the median gap length, first and third quartile of each gaptype starting per academic year-epoch in a boxplot. Median gap length decreases by half a year for success gaps at the beginning of the study, but then remains constant. This decrease can be caused by too few observations at the beginning of the study: the number of success gaps is almost quadrupled from the first to the second year-epoch (see figure 2.2). The same explanation could be used for the large third quartile value in the first year-epoch for success and absence gaps.

Anyway, the gap length of most absence, interim and success gaps is far below the definition boundary of 4 years. So, there is an obvious difference in gap length between absence, interim and success gaps on the one hand and withdrawal gaps, lasting more than 4 years, almost by definition, on the other hand. In figure 2.3 few outlying values are observed near the cut-point of 4 years.

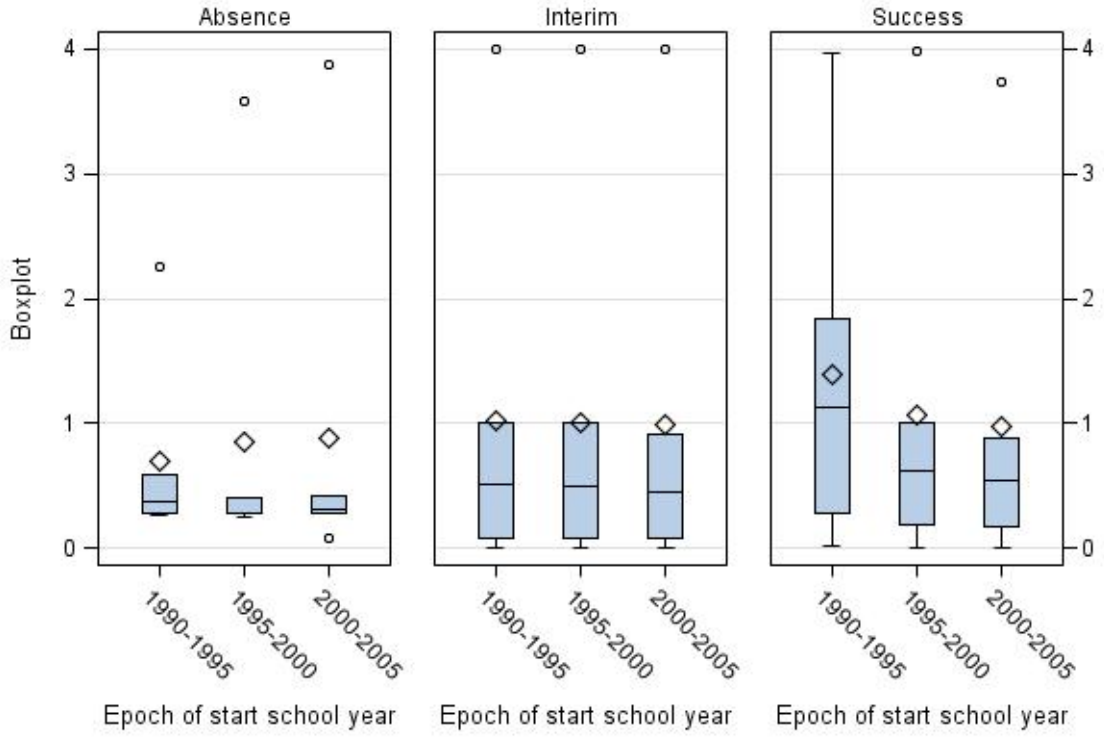


**Figure 2.3:** Boxplot of gap length per gaptype and academic year-epoch for all gaps starting between October 1, 1990 and September 30, 2005.

| Variable | Absence, interim, success gap $\leq 4$ years ($n = 6,032$) | Withdrawal gap $> 4$ years ($n = 8,018$) | Missing values |
|---|---|---|---|
| **Gender** | | | 15 |
| [Male] | 2,775 (39%) | 4,291 (61%) | |
| Female | 3,254 (47%) | 3,715 (53%) | |
| **Dominant scientific field** | | | 87 |
| [sciences] | 1,282 (49%) | 1,335 (51%) | |
| medicine | 1,276 (43%) | 1,691 (57%) | |
| humanities | 1,112 (45%) | 1,350 (55%) | |
| social | 1,193 (40%) | 1,822 (60%) | |
| applied | 1,148 (40%) | 1,754 (60%) | |
| **Nationality** | | | 232 |
| [Belgian] | 4,833 (43%) | 6,460 (57%) | |
| European Union (excl. Belgium) | 521 (39%) | 819 (61%) | |
| Other | 629 (53%) | 556 (47%) | |
| **Dominant statute classification** | | | 0 |
| Assistant lectureship | 1,612 (31%) | 3,554 (69%) | |
| Compet. scholarship (Flanders) | 1,635 (44%) | 2,072 (56%) | |
| Compet. scholarship (own university) | 835 (62%) | 515 (38%) | |
| Project funding (FWO, BOF, IUAP) | 357 (74%) | 125 (26%) | |
| [Project funding (other)] | 1,593 (48%) | 1,752 (52%) | |
| **Age (at start)** | | | 35 |
| [$\leq 25$ years] | 4,036 (44%) | 5,129 (56%) | |
| $26 - 30$ years | 1,226 (42%) | 1,671 (58%) | |
| $31 - 35$ years | 432 (40%) | 649 (60%) | |
| $36 - 40$ years | 204 (44%) | 258 (56%) | |
| $> 40$ years | 131 (32%) | 279 (68%) | |
| **Start cohort** | | | 0 |
| [01/10/1990 - 30/09/1997] | 2,766 (42%) | 3,882 (58%) | |
| 01/10/1997 - 30/09/2004 | 3,158 (44%) | 4,044 (56%) | |
| 01/10/2004 - 30/09/2009 | 108 (54%) | 92 (46%) | |
| **University** | | | 0 |
| [1] | 2,162 (47%) | 2,426 (53%) | |
| 2 | 613 (41%) | 867 (59%) | |
| 3 | 2,378 (40%) | 3,568 (60%) | |
| 4 | 707 (43%) | 948 (57%) | |
| 5 | 172 (45%) | 209 (55%) | |

**Table 2.4:** Classification of all gaps starting prior to October 1, 2005, corresponding to the personal covariate values selected with a view to the imputation model. For each covariate, the [reference category] is indicated.

Table 2.4 presents the number of gaps corresponding to the specified covariate values row by row and row percentages. The two columns make the distinction between gaps lasting less than or equal to 4 years (absence, interim and success gap) and more than 4 years (withdrawal gap). The subset consists of all 14,050 gaps starting between October 1, 1990 and October 1, 2005. We observe more male than female gaps ending in withdrawal, while more female than male absence, interim or success gaps started prior to October 1, 2005. For both genders, the number of withdrawal gaps exceeds the number of non-withdrawal gaps, while this difference is largest for male Ph.D.-students. Regarding the scientific field, for each category more withdrawal than non-withdrawal gaps are found. More withdrawal (61%) than non-withdrawal gaps (39%) are observed for European (excl. Belgian) gaps, while the opposite is observed for non-European gaps (53% vs. 47%). Considering dominant statute classification a gap is more likely to end in withdrawal except for competitive scholarship (own university) and project funding (FWO, BOF, IUAP). For all age groups, the gap is more likely to end in withdrawal rather than non-withdrawal, but this contrast is the most obvious for the oldest age group. The last starting cohort has only few observations because we consider all gaps starting prior to October 1, 2005. Therefore, this last starting cohort provides no reliable information on the difference between withdrawal and non-withdrawal gaps. Based on the university at which most time was spent for the person's gap, we observe that there are substantial differences in the percentage of non-withdrawal vs. withdrawal gaps for each university: e.g. university 1 has 47% vs. 53% such gaps respectively, while for university 3 this is 40% vs. 60% gaps.

These results are purely explorative and describe univariate association only, as they do not take into account possible interaction effects between covariates. Nevertheless we conclude that the significance of all these covariates should be carefully tested while building the imputation model.

### 2.2.3 Censored Gaps

There are 3,139 censored gaps, i.e. gaps starting between October 1, 2005 and September 30, 2009 and not ended by the end of the study. Since each Ph.D.-student can have at most one censored gap, there are 3,139 (11%) Ph.D.-students being censored while having a gap out of 28,871 Ph.D.-students starting between September 30, 1990 and October 1, 2009. Note that a censored Ph.D.-student not necessarily ends the study with a censored gap, he/she can be active (so not in a gap) by the end of the study. The number of censored Ph.D.-students is not uniquely determined, since it depends on the imputation procedure.

Table 2.5 shows the number of gaps censored at given duration of gaptime categorized by years. The length of the observed part of the censored gaps rather seems to be equally distributed.

| | Censored gaps | |
|---|---|---|
| gap length | Frequency | Percent |
| ]0,1] | 904 | 29% |
| ]1,2] | 678 | 22% |
| ]2,3] | 723 | 23% |
| ]3,4] | 834 | 27% |

**Table 2.5:** Frequency and percentage of censored gaps per gap length in years.

In contrast with equation (2.1), we could model the probability of being a withdrawal gap for all gaps starting prior to October 1, 2005, conditional on baseline covariates $\mathbf{Z}$

$$P(\text{withdrawal gap}|\mathbf{Z}),$$

building a logistic regression model for the outcome 'gaptype' that equals 1 if the gap is a withdrawal gap and 0 otherwise (absence, interim or success gap). Assuming a stationary process, we can estimate the forward probability, man or woman who has the greatest chance of withdrawing when having a gap censored by the end of the study? Similarly for other covariate values.

Remark that this descriptive analysis is not taking into account the length of the observed part of the censored gap. Instead of building this logistic regression model, we will estimate the probability of withdrawal for all censored gaps, conditional on the observed covariate values *and* the current gap length, later in this chapter.

## 2.3 Definition and Notation

Generally in survival analysis, we distinguish between the actual time to event, here gap-end ($X_j$), censoring time ($C_j$) and denote the observation time $T_j = \min(X_j, C_j)$ as the time on study for the $j$th observation. Our data set, based on a sample of size $n$, consists of the triple $(T_j, \delta_j, \mathbf{Z}_j)$, $j = 1 \ldots n$, which are assumed to be independent and identically distributed random vectors, where

- $T_j$ is the time on study for the $j$th gap i.e. observed gaptime. By definition there are no censored gaptimes in the subset of all gaps starting between October 1, 1990 and September 30, 2005. As non-withdrawal gap-ends only occur before 4 years and withdrawal gap-ends at 4 years of gaptime, we have

  $$T_j = X_j < 4 \text{ years} \Leftrightarrow \text{absence, interim or success gap,}$$

$T_j = X_j = 4$ years $\Leftrightarrow$ withdrawal gap,

- $\delta_j$ is the event indicator for the $j$th gap

  $\delta_j = 1$ if the gap-end has occurred at $T_j$, specifically, if at that time sponsoring is restarted (interim or absence gap), the Ph.D.-degree is attained (success gap) or the gap length $T_j$ reached 4 years (withdrawal gap), and

  $\delta_j = 0$ if the gaptime is right-censored and hence ongoing at time $T_j$,

- $\mathbf{Z}_j$ is the vector of baseline covariates $(Z_{j1}, \ldots, Z_{jp})^T$ for the $j$th gap which may affect the survival distribution of $X_j$.

To allow for possible ties in the data, suppose that the events occur at $D$ distinct times $t_1 < t_2 < \ldots < t_D$ and that at time $t_i$ there are $d_i$ events. Let $R(t_i)$ denote the set of all individuals who are at risk just prior to time $t_i$, so having a gap ended at time $t \geq t_i$.

## 2.4 The Cox Proportional Hazards Model

We model the distribution of gaptimes in the population of all gaps starting between October 1, 1990 and September 30, 2005 through a Cox proportional hazards (PH) regression model conditional on baseline covariates $\mathbf{Z}$, the hazard of having a gap lasting $t$ years,

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

is modeled in function of the set of covariates $\mathbf{Z} = (Z_1, \ldots, Z_p)$ using a Cox PH regression model:

$$h(t|\mathbf{Z}) = h_0(t) \exp\left(\boldsymbol{\beta}^T \mathbf{Z}\right), \tag{2.3}$$

where $t$ represents the time to event, here gap-end. The hazard function is specified to be the product of an unknown baseline hazard $h_0(t)$ (the hazard for an individual with $\mathbf{z} = \mathbf{0}$) and a log linear factor where the covariates $\mathbf{Z}$ enter via a vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ of unknown regression coefficients. This is called a semiparametric model because a parametric form is assumed only for the covariate effect. The baseline hazard $h_0(t)$ has no specified parametric form, it is left as an arbitrary nonnegative function. The set of possible baseline covariates $\mathbf{Z}$ was listed in section 1.3 (p. 3). As all of them are categorical, they are recoded to dummy variables before they are included in the model. [1]

A key feature of this proportional hazards model is that, when all the covariates are fixed at time 0, the hazard rates of two individuals with distinct values of $\mathbf{Z}$ are proportional over

[1] A dummy variable is one that takes the values 0 or 1 to indicate the absence or presence of some categorical level that may be expected to shift the outcome.

time. To see this, consider two individuals with covariate values $\mathbf{z}_1$ and $\mathbf{z}_2$. The ratio of their hazards is

$$\frac{h(t|\mathbf{z}_1)}{h(t|\mathbf{z}_2)} = \frac{h_0(t)\exp\boldsymbol{\beta}^T\mathbf{z}_1}{h_0(t)\exp\boldsymbol{\beta}^T\mathbf{z}_2} = \exp\left[\boldsymbol{\beta}^T(\mathbf{z}_1 - \mathbf{z}_2)\right], \tag{2.4}$$

which is a constant independent of time. The quantity (2.4) is called the hazard ratio of an individual with risk factor $\mathbf{z}_1$ as compared to an individual with risk factor $\mathbf{z}_2$.

Since time has been measured in days, ties between gaptimes are found in the data. Alternate partial likelihoods have been provided by a variety of authors when there are ties between event times [10]. Cox proposed an extension of the proportional hazards model to discrete time by working with the conditional odds of dying at each time $t_i$ given survival up to that point. Specifically to construct the likelihood, let $D_i$ be the set of all individuals who die at time $t_i$. Then, the discrete partial likelihood is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\left(\sum_{k\in D_i}\boldsymbol{\beta}^T\mathbf{Z_k}\right)}{\sum_{k\in R(t_i)}\exp\left(\boldsymbol{\beta}^T\mathbf{Z}_k\right)}. \tag{2.5}$$

Based on the observation of $n$ independent survival gaptimes $X_1, \ldots, X_n$, associated covariate vectors $\mathbf{z}_1, \ldots, \mathbf{z}_n$, the regression coefficients $\boldsymbol{\beta}$ are estimated by the value $\hat{\boldsymbol{\beta}}$, maximizing this discrete Cox partial likelihood function.

As measuring time in days is quite a strict time unit for analyzing gaptimes, we could also model gaptime based on the continuous partial likelihood instead of the discrete likelihood in (2.5). This continuous adjustment for ties is proposed by Kalbfleisch and Prentice [7] and assumes that tied events are due to the imprecise nature of our measurement, and that there must be some true ordering. In practice both methods give similar results.

We will also need the variance and covariance of $\hat{\boldsymbol{\beta}}$, represented by the estimated covariance matrix $\hat{V}(\hat{\boldsymbol{\beta}})$. This statistic can be estimated by inverting the $(p \times p)$-matrix of the observed information matrix. Representing the log-partial likelihood $\ln(L(\boldsymbol{\beta}))$ by $\ell(\boldsymbol{\beta})$, the observed information matrix $\mathcal{J}$ is calculated as the second derivative of the log-partial likelihood

$$\mathcal{J}(\boldsymbol{\beta}) = -\frac{\partial^2}{\partial\boldsymbol{\beta}^2}\,\ell(\boldsymbol{\beta}).$$

## 2.5 Strategy for Building the Imputation Model

For every gap censored at time $t$ since the gap start, the probability of being a withdrawal gap is estimated by

$$\begin{aligned} P(\text{withdrawal gap}|\mathbf{Z}, \text{gap} > t \text{ years}) &= P(\text{gap} > 4 \text{ years}|\mathbf{Z}, \text{gap} > t \text{ years}) \\ &= \frac{P(\text{gap} > 4 \text{ years}|\mathbf{Z})}{P(\text{gap} > t \text{ years}|\mathbf{Z})} \\ &= \frac{S(4|\mathbf{Z})}{S(t|\mathbf{Z})} \\ &= \frac{S_0(4)^{\exp(\boldsymbol{\beta}^T\mathbf{Z})}}{S_0(t)^{\exp(\boldsymbol{\beta}^T\mathbf{Z})}} \qquad (2.6) \\ &= \left(\frac{S_0(4)}{S_0(t)}\right)^{\exp(\boldsymbol{\beta}^T\mathbf{Z})} \\ &=: p(t;\mathbf{Z}), \qquad (2.7) \end{aligned}$$

knowing $0 \leq t \leq 4$, where $\boldsymbol{\beta}^T\mathbf{Z} = \sum_{k=1}^{p} \beta_k Z_k$. The probability of having a gap lasting more than $t$ years ($0 \leq t \leq 4$ years) knowing the covariate values $\mathbf{Z}$ is given by the survival function $S(t|\mathbf{Z})$. Equation (2.6) holds, provided the following Cox model holds for the gaptimes $t$:

$$h(t|\mathbf{Z}) = h_0(t)\exp(\boldsymbol{\beta}^T\mathbf{Z}).$$

$S_0(t)$ represents the baseline survival of having a gap lasting more than $t$ years ($0 \leq t \leq 4$) and $S_0(4)$ is the same baseline survival function but now evaluated at $t = 4$. More details are explained in the next sections.

Also, using the relationship $-\ln(S_0(t)) = H_0(t)$, where $H_0(t)$ is the baseline cumulative hazard function, we can rewrite equation (2.7) by

$$\begin{aligned} \ln\left[P(\text{withdrawal gap}|\mathbf{Z}, \text{gap} > t \text{ years})\right] &= \ln\left[p(t;\mathbf{Z})\right] \\ &= \ln\left[\left(\frac{S_0(4)}{S_0(t)}\right)^{\exp(\boldsymbol{\beta}^T\mathbf{Z})}\right] \\ &= \exp\left(\boldsymbol{\beta}^T\mathbf{Z}\right)\left[\ln\left(S_0(4)\right) - \ln\left(S_0(t)\right)\right] \\ &= \exp\left(\boldsymbol{\beta}^T\mathbf{Z}\right)\left[H_0(t) - H_0(4)\right], \qquad (2.8) \end{aligned}$$

knowing $0 \leq t \leq 4$. This expression (2.8) is used frequently in the next sections instead of expression (2.7) since it facilitates the variance calculation. The variance of the estimated probability of being a withdrawer expressed in (2.7) is the variance of a ratio, which is quite complicated to calculate. Using the delta method, we can start by finding the asymptotic variance of the natural logarithm of this probability in (2.8), which reduces to the variance of a difference, which is a substantial simplification.

Once we have obtained

- estimates of the ln hazard coefficients $\boldsymbol{\beta}$ (section 2.6) and
- estimates of the baseline cumulative hazard function $H_0(t)$ (section 2.7)

based on all gaps starting prior to October 1, 2005, we can construct the imputation model and estimate the probability of withdrawal for all gaps starting between October 1, 2005 and September 30, 2009 with a given censored gaptime $t$ and set of covariate values $\mathbf{z}^*$.

## 2.6 Building a Cox PH Model for $P(\mathbf{gap} > t \text{ } \mathbf{years}|\mathbf{Z})$

We build a Cox PH regression model for the hazard of having a gap lasting $t$ years, conditional on baseline covariates $\mathbf{Z}$

$$h(t|\mathbf{Z}) = h_0(t) \exp\left(\boldsymbol{\beta}^T \mathbf{Z}\right).$$

An imputation model should be chosen to be (at least approximately) compatible with the analyses to be performed on the imputed data sets. Therefore, the imputation model should be rich enough to preserve the associations or relationships among variables that will be the focus of later investigation. In general, any association that may be important in subsequent analyses should be present in the imputation model. The converse however, is not necessary. Thus, the danger with an imputation model is generally in leaving out predictors rather than including too many [13].

We perform a forward selection of all main effects, at the 5% significance level. In the multivariate analysis the covariates gender, nationality, dominant statute classification, dominant scientific field, age, start time and university were all significant at the 5% significance level.

To obtain a model as rich as possible, we also test for the interactions gender $\times$ dominant statute, gender $\times$ dominant scientific field, nationality $\times$ dominant statute, nationality $\times$ dominant scientific field and gender $\times$ nationality. Because the category 'International non-EU Ph.D.-students' contains too few observations we combine this category with 'EU non-Belgian Ph.D.-students' when considering the interactions, so two categories are formed: Belgian and non-Belgian Ph.D.-students. We include all previously specified interactions, even if they are not explicitly significant at the 5% significance level. For interaction effects the significance level of 5% is quite a severe criterion and we want to build a sufficiently rich model to guarantee proper imputations. Detailed model results are given in the appendix (p. 81).

Due to missing values for some of the explanatory variables, 357 observations were excluded. This exclusion is justified if we assume these missing covariate values are missing completely at random (MCAR), i.e. the probability that a covariate value is missing does not depend on the unobserved value or on the value of any other observed data. Excluding them, 13,693

observations are used for the analysis of which 5,956 gaps had an event prior to 4 years of gaptime (absence, interim or success gap-end) and 7,737 gaps after 4 years of gaptime (withdrawal gap-end). Remember that this distinction between withdrawal and non-withdrawal gaps can be made because we only include gaps starting between October 1, 1990 and September 30, 2005 for the analysis.
An alternative and weaker assumption is that of data missing at random (MAR). In this case we assume that given the observed data, the missingness does not depend on the unobserved data. Simply excluding the observations with missing covariate values could lead to biased inference. Instead, the multiple imputation procedure could be applied to impute these missing covariate values based on the observed data.

Testing the global null hypothesis $\boldsymbol{\beta} = \mathbf{0}$, the Likelihood Ratio test as well as the Score and Wald test result in a $p$-value $< 0.0001$.

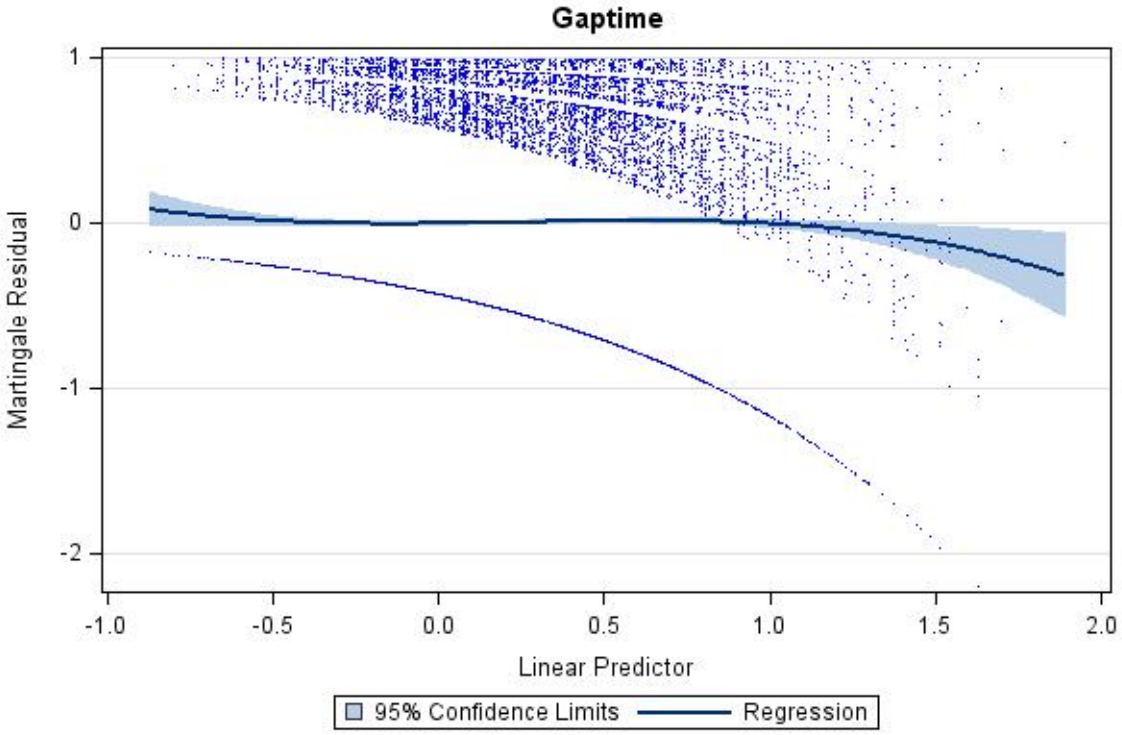


**Figure 2.4:** Martingale residuals of the built Cox PH model for $P(\text{gap} > t \text{ years}|\mathbf{Z})$ versus the prognostic score.

To check survival predicted by the prognostic score $\hat{\boldsymbol{\beta}}^T\mathbf{Z}$ we plot the martingale residuals in figure 2.4. Martingale residuals are defined for the $j$th individual as

$$R_j = \delta_j - \hat{H}(X_j),$$

with range between $-\infty$ and 1 [7]. The residual $R_j$ can be viewed as the difference between the observed number of deaths (0 or 1) for each subject $j$ between time 0 and $X_j$, and the expected number based on the fitted model, $\hat{H}(X_j)$.
The regression line is specified to have a cubic fit and 95% confidence intervals cover 0 except at the end of the range of the linear predictor, where the upper bound falls just below. There is no reason to reject appropriateness of the final model. Similar plots (not shown) against individual predictors equally did not indicate necessary transformation of covariates.

## 2.7 Estimating the Cumulative Hazard Function

We have fit the imputation model to the data to obtain the partial maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and the estimated covariance matrix $\hat{V}(\hat{\boldsymbol{\beta}})$, which are needed for estimating (2.8). In this section we estimate the baseline cumulative hazard function $H_0(t)$ [10].

Let

$$W(t_i, \boldsymbol{\beta}) = \sum_{j \in R(t_i)} \exp\left(\boldsymbol{\beta}^T \mathbf{z}_j\right),$$

where $R(t_i)$ is the risk set containing all gaps at risk or ending at the event time $t_i$, as mentioned in section 2.3. The exponential is taken of all corresponding Ph.D.-student covariate values ($\mathbf{z}_j$), multiplied by the risk coefficients ($\boldsymbol{\beta}$). The baseline cumulative hazard is most often estimated by the step function

$$\hat{H}_0(t) = \sum_{t_i \leq t} \frac{d_i}{W(t_i, \hat{\boldsymbol{\beta}})}, \tag{2.9}$$

the so-called Breslow estimator, which is a step function with jumps at the observed death times. Here $d_i$ represents the number of gaps ending at time $t_i$, allowing for possible ties in the data. An example of this step function is given in figure 2.5.

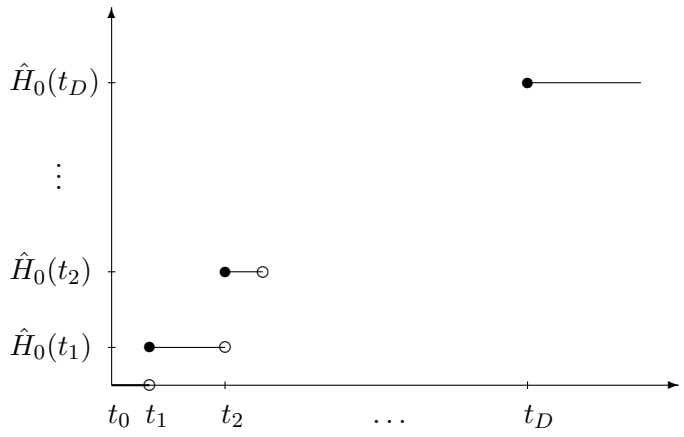


**Figure 2.5:** Example of the baseline cumulative hazard function $\hat{H}_0(t)$.

As mentioned before, the data set of all gaps starting between October 1, 1990 and September 30, 2005 used to calculate $W(t_i, \hat{\boldsymbol{\beta}})$ and $\hat{H}_0(t)$, distinguishes between withdrawal and non-withdrawal gaps by total gaptime. As long as the total gaptime is less than 4 years only non-withdrawal gaps occur (absence, interim and success gap-end) and from then on only withdrawal gaps occur. Hence no censorings are observed. We only need to estimate the baseline cumulative hazard for $0 \leq t \leq 4$ (see (2.8)).

The $n$ observed gaps are ordered by event time and we can suggest the following algorithm for the Breslow estimator in (2.9). Algorithm 1 refers to the data set containing all gaps starting between October 1, 1990 and September 30, 2005. Here cumdeath2 represents the number of gaps who are no longer in the risk set $R(t_i)$ at each event time $t_i$. So if $d_i$ events occurred at time $t_i$, cumdeath2 is increased by $d_i$ at time $t_{i+1}$. This is the cumulative sum of all events, but given 1 event time later. Typically in survival analysis, we frequently summarize over risk sets at different time points. We can calculate these sums by reversing ordered vectors or matrices, taking the cumulative sum, and reversing again, this is done for Wlong.

**Algorithm 1** : Estimating $\hat{H}_0(t)$

exp(bZ) = exp($\hat{\boldsymbol{\beta}}^T Z$)
cumdeath = cumsum(death)
cumdeath2[1] = 0
**for** $j = 2 : D$ **do**
  cumdeath2[$j$] = cumdeath[$j-1$]
**end for**
$W$long = (cumsum(exp(bZ)[$N:1$]))[$N:1$]
$W = W$long[cumdeath2+1]
$H_0$ = cumsum(death/$W$)

The estimates of $\hat{H}_0(t)$ obtained by the algorithm correspond to the ones obtained by SAS PHREG procedure with options TIES=DISCRETE and BASELINE METHOD=CH referring to the Breslow cumulative hazard, handling ties as if they actually occurred at the same time due to the discrete time scale.

When there are no covariates present, the estimator (2.9) reduces to the Nelson-Aalen estimator of the cumulative hazard

$$\tilde{H}(t) = \begin{cases} 0, & \text{if} \quad t < t_1, \\ \sum_{t_i \leq t} \frac{d_i}{r_i}, & \text{if} \quad t_1 \leq t, \end{cases}$$

where $r_i$ is the number of individuals who are at risk at time $t_i$.

For discrete lifetimes, the relationship between the cumulative hazard function and the hazard

function shall be defined by

$$H(t) = \sum_{t_j \leq t} h(t_j). \tag{2.10}$$

Notice that the relationship $S(t) = \exp(-H(t))$ for this definition no longer holds true. Some authors prefer to define the cumulative hazard for discrete lifetimes as

$$H(t) = \sum_{t_j \leq t} -\ln[1 - h(t_j)], \tag{2.11}$$

because the relationship for continuous lifetimes $S(t) = \exp(-H(t))$ will be preserved for discrete lifetimes. If the $h(t_j)$ are small, and this condition is fulfilled in our case, (2.10) will be an approximation of (2.11). We prefer the use of (2.10) because it is directly estimable from a sample of censored or truncated lifetimes.

By straightforward calculations we also get following relationship between the baseline cumulative hazard function and the baseline hazard function

$$\hat{H}_0(t) = \sum_{t_j \leq t} \hat{h}_0(t_j). \tag{2.12}$$

Thanks to Algorithm 1 we can estimate (2.8) by

$$\ln\left[\hat{p}(t; \mathbf{z})\right] = -\exp(\hat{\boldsymbol{\beta}}^T \mathbf{z}) \left[\hat{H}_0(4) - \hat{H}_0(t)\right],$$

getting non-positive values. So, for every gap censored at time $t$ since the gap start, the probability (2.7) of being a withdrawal gap can be estimated by

$$\hat{p}(t; \mathbf{z}) = \exp\left\{-\exp(\hat{\boldsymbol{\beta}}^T \mathbf{z}) \left[\hat{H}_0(4) - \hat{H}_0(t)\right]\right\}, \tag{2.13}$$

getting values in the interval $]0, 1]$.
Detailed calculation of the corresponding gaptime for all censored gaps is given in the appendix (A.1, p. 76).

## 2.8 Multiple Imputed Data Sets

The validity of the multiple imputation method depends on how the imputations were generated. Clearly it is not possible to obtain valid inferences in general if imputations are created arbitrarily. The imputations should, on average, give reasonable predictions for the missing data, and the variability among them must reflect an appropriate degree of uncertainty. Rubin [11] provides technical conditions under which repeated imputation method leads to frequency-valid answers. An imputation method which satisfies these conditions is said to be 'proper'. In this section, we will present two imputation methods and compare the imputation results.

Throughout this section we will work on the subset of all gaps starting between September 30, 2005 and October 1, 2009 which are censored by the end of the study. These gaps are censored at time $t$ since the gap start and have associated covariate values $\mathbf{z}^*$.

We assume the missing outcomes (censored gaps) are missing at random (MAR), i.e. depending only on $(T_j, \mathbf{Z}_j, I(\delta_j > 0))$ and not on $\delta_j$ itself for each gap $j = 1, \dots n$.

### 2.8.1 Imputation Methods

#### Imputation Method A

First, we can impute the missing outcomes by taking draws from the binomial distribution $Bin(1, \hat{p}(t, \mathbf{z}^*))$ with estimated expectation given by (2.13). If for observation $i$ this generated value equals 1, this gap is categorized as withdrawal gap and consequently the corresponding Ph.D.-student as withdrawer. Otherwise his/her gap is categorized as an ongoing gap and the corresponding Ph.D.-student as ongoer, thereby censored by the end of the study. We will refer to these imputations by imputation method A.

#### Imputation Method B

Taking into account the uncertainty of estimating $p(t; \mathbf{z})$ given by (2.13), we may account for the distribution of $\hat{p}(t; \mathbf{z})$ to get proper imputations. In particular we estimate the variance of $\hat{p}(t; \mathbf{z})$ to this end. Additionally, from the estimate and the associated variance estimate a $100(1 - \alpha)\%$ confidence interval for the probability $\hat{p}(t; \mathbf{z})$ for every censored gap may be constructed. We will refer to these imputations by imputation method B.

In both cases we get a completed data set with no more missing outcomes (ongoer, Ph.D.-attainment or withdrawer) for all Ph.D.-students without missing covariate values.

### 2.8.2 Distribution of the Cumulative Hazard Function

The regularity conditions listed in the appendix (A.2.1, p. 77) will be assumed to hold. Based on the asymptotic distribution of the cumulative hazard function, as stated in theorem A.1 (p. 78), we estimate the cumulative hazard function variance.

The estimated variance structure is given by the vector

$$\widehat{\mathrm{Var}}(\hat{H}(t|\mathbf{z}^*)) = \exp(2\hat{\boldsymbol{\beta}}^T \mathbf{z}^*) \left[ Q_1(t) + Q_2(t, t; \mathbf{z}^*) \right], \tag{2.14}$$

and the estimated covariance structure is given by the matrix

$$\widehat{\mathrm{Cov}}[\hat{H}(s|\mathbf{z}^*), \hat{H}(t|\mathbf{z}^*)] = \exp(2\hat{\boldsymbol{\beta}}^T \mathbf{z}^*) \left[ Q_1(s) + Q_2(s, t; \mathbf{z}^*) \right], \tag{2.15}$$

where $0 \leq s \leq t \leq 4$. To estimate the covariance, we need to evaluate $Q_2$ at $s$ equal to the censored gaptime and $t$ equal to 4 years.

Returning to the problem of estimating the variance of $\ln[\hat{p}(t;\mathbf{z}^*)]$, we want to estimate the variance of the difference of two cumulative hazard functions

$$\begin{aligned}\ln[\hat{p}(t;\mathbf{z}^*)] &= \exp\left(\hat{\boldsymbol{\beta}}^T\mathbf{z}^*\right)\left[\hat{H}_0(t)-\hat{H}_0(4)\right]\\ &= \hat{H}(t|\mathbf{z}^*)-\hat{H}(4|\mathbf{z}^*).\end{aligned}$$

Since $G_{\mathbf{z}^*}(t)$ (theorem A.1, p. 78) is a Gaussian process at each time $t$, the difference of the cumulative hazard function at two different time points is bivariate normally distributed. So the distribution of $\ln[\hat{p}(t;\mathbf{z}^*)]$ is given by

$$\hat{H}(t|\mathbf{z}^*)-\hat{H}(4|\mathbf{z}^*) \xrightarrow{\mathcal{L}} N\left(H(t|\mathbf{z}^*)-H(4|\mathbf{z}^*), \widehat{\text{Var}}\left[\hat{H}(t|\mathbf{z}^*)-\hat{H}(4|\mathbf{z}^*)\right]\right), \tag{2.16}$$

where

$$\begin{aligned}\widehat{\text{Var}}\left[\hat{H}(t|\mathbf{z}^*)-\hat{H}(4|\mathbf{z}^*)\right] &= \widehat{\text{Var}}\left(\ln[\hat{p}(t;\mathbf{z}^*)]\right) \qquad (2.17)\\ &= \widehat{\text{Var}}\left(\hat{H}(t|\mathbf{z}^*)\right)+\widehat{\text{Var}}\left(\hat{H}(4|\mathbf{z}^*)\right)-2\,\widehat{\text{Cov}}\left(\hat{H}(t|\mathbf{z}^*),\hat{H}(4|\mathbf{z}^*)\right).\end{aligned}$$

Knowing the distribution of $\ln[\hat{p}(t;\mathbf{z}^*)]$, we generate imputations of method B from this distribution as follows. Take random draws from (2.16), then exponentiate these draws to estimate $\hat{p}(t;\mathbf{z}^*)$. So, actually the distribution of $\hat{p}(t;\mathbf{z}^*)$ is given by

$$\begin{aligned}\exp\left(\hat{H}(t|\mathbf{z}^*)-\hat{H}(4|\mathbf{z}^*)\right) \xrightarrow{\mathcal{L}} \exp\Big\{N\Big(&H(t|\mathbf{z}^*)-H(4|\mathbf{z}^*)\ ,\\ &\widehat{\text{Var}}\left[\hat{H}(t|\mathbf{z}^*)-\hat{H}(4|\mathbf{z}^*)\right]\Big)\Big\}. \qquad (2.18)\end{aligned}$$

**Summary**

This can be summarized as follows. Both imputation methods start by estimating $\ln[p(t;\mathbf{z}^*)]$ in the same way, namely

$$\ln\left[\hat{p}(t;\mathbf{z}^*)\right] = \exp\left(\hat{\boldsymbol{\beta}}^T\mathbf{z}^*\right)\left[\hat{H}_0(t)-\hat{H}_0(4)\right].$$

Then, two imputation models are obtained, depending on how $p(t;\mathbf{z}^*)$ is being estimated.

A. Imputation method A estimates $p(t;\mathbf{z}^*)$ by

$$\hat{p}_A(t;\mathbf{z}^*) = \exp(\ln[\hat{p}(t;\mathbf{z}^*)]). \tag{2.19}$$

B. Imputation method B estimates $p(t;\mathbf{z}^*)$ by exponentiating draws from a normal distribution, so that

$$\hat{p}_B(t;\mathbf{z}^*) \xrightarrow{\mathcal{L}} \exp\left\{N\left(\ln[p(t;\mathbf{z}^*)], \widehat{\text{Var}}\left[\ln[\hat{p}(t;\mathbf{z}^*)]\right]\right)\right\}. \tag{2.20}$$

Next, both imputation methods are taking draws from a binomial distribution, respectively $Bin(1,\hat{p}_A(t;\mathbf{z}^*))$ and $Bin(1,\hat{p}_B(t;\mathbf{z}^*))$ for the imputed outcomes. Finally, we get 5 imputed data sets for both imputation methods.

### 2.8.3 Imputation Results

We can plot the distribution of the imputed outcome, generated by imputation method B, for each censored gap. From (2.18) we can plot the exponential of a normal distribution for $\hat{p}(t; \mathbf{z}^*) = \exp(H(t|\mathbf{z}^*) - H(4|\mathbf{z}^*))$ for each Ph.D.-student with censored gaptime $t$ and covariate values $\mathbf{z}^*$. Figure 2.6 shows the distribution of $\hat{p}(t; \mathbf{z}^*)$ for observation 5, 50, 100, 500, 1500 and 3000, out of 3139 observations, ordered by increasing censored gaptime. This distribution is rather peaked, especially for the observations with large censored gaptimes. Therefore, we expect the imputed values will not vary much. We checked the area under the probability density function equals 1.

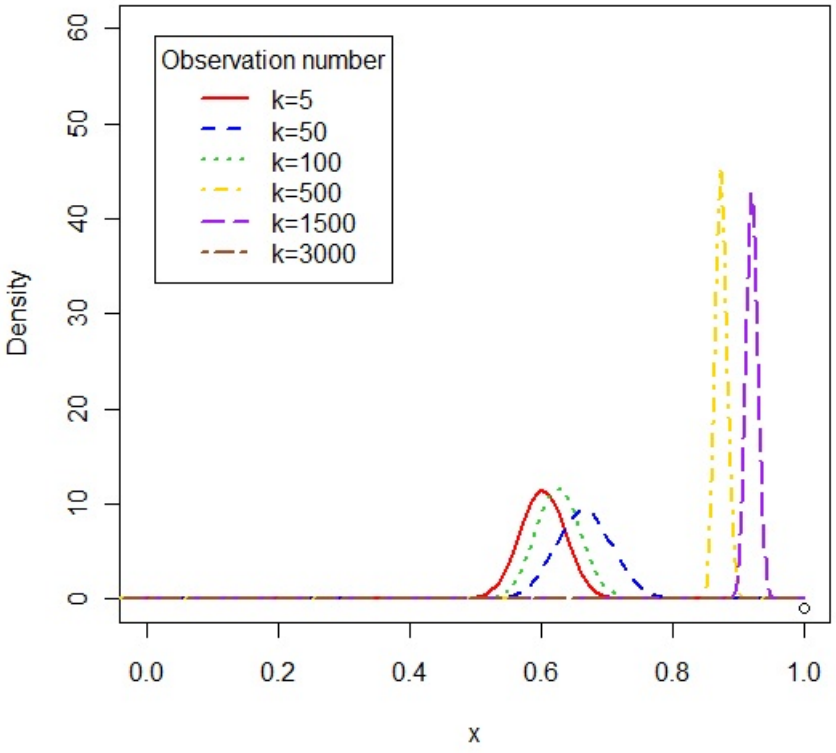


**Figure 2.6:** The density function of the distribution of $\hat{p}(t; \mathbf{z}^*)$ given in (2.18) for 6 observed censored gaps.

We create 5 imputed data sets per imputation method. Once the imputation model has been constructed, the number of imputations can easily be decreased or increased. Unless rates of missing information are unusually high, there tends to be little or no practical benefit to using more than five to ten imputations [15].

Table 2.6 presents the percentage of withdrawal gaps for all imputed censored gaps by imputation method A and imputation method B. There is no specific trend in the difference for the percentage of withdrawal gaps between method A and method B, based on 5 imputations.

| | Percentage of withdrawal gaps | | | | |
|---|---|---|---|---|---|
| Imputation | 1 | 2 | 3 | 4 | 5 |
| Method A | 87.64 | 88.94 | 88.48 | 88.09 | 87.06 |
| Method B | 87.29 | 87.51 | 88.48 | 88.06 | 87.32 |

**Table 2.6:** The percentage of withdrawal gaps for all imputed censored gaps by imputation method A and imputation method B.

## 2.9 Extending Imputation Method B

### 2.9.1 Confidence Intervals for $p(t; \mathbf{z}^*)$

From the distribution of $\ln[\hat{p}(t; \mathbf{z}^*)]$ in (2.16) a $100(1-\alpha)\%$ confidence interval for $p(t; \mathbf{z}^*)$ may be set up in several ways, as pointed in [4]. We work out two methods in detail.

First, we construct a $100(1-\alpha)\%$ confidence interval for $\ln[p(t; \mathbf{z}^*)] = H(t|\mathbf{z}^*) - H(4|\mathbf{z}^*)$. As stated in (2.16) the estimate for $H(t|\mathbf{z}^*) - H(4|\mathbf{z}^*)$ is asymptotically normally distributed with the correct mean and variance given by (2.17):

$$\hat{H}(t|\mathbf{z}^*) - \hat{H}(4|\mathbf{z}^*) \xrightarrow{\mathcal{L}} N\left(H(t|\mathbf{z}^*) - H(4|\mathbf{z}^*), \widehat{\text{Var}}\left[\hat{H}(t|\mathbf{z}^*) - \hat{H}(4|\mathbf{z}^*)\right]\right)$$

The simplest choice for a confidence interval for $\ln[p(t; \mathbf{z}^*)]$ is the linear one obtained simply as

$$\left(\hat{H}(t|\mathbf{z}^*) - \hat{H}(4|\mathbf{z}^*)\right) \pm c_{\alpha/2}\sqrt{\widehat{\text{Var}}\left[\hat{H}(t|\mathbf{z}^*) - \hat{H}(4|\mathbf{z}^*)\right]},$$

where $c_{\alpha/2}$ is the upper $\alpha/2$ fractile of the standard normal distribution. We obtain a $100(1-\alpha)\%$ confidence interval for $p(t; \mathbf{z}^*) = \exp\left(H(t|\mathbf{z}^*) - H(4|\mathbf{z}^*)\right)$ by exponentiating the lower and upper bound of this last confidence interval.

However, it may be advantageous to use a transformed, non-symmetrical confidence interval - one possibility being based on a log transformation of the cumulative hazard function [4]. The $100(1-\alpha)\%$ confidence interval for $\ln(H(t|\mathbf{z}^*) - H(4|\mathbf{z}^*))$ is, using the delta-method

$$\ln\left(\hat{H}(t|\mathbf{z}^*) - \hat{H}(4|\mathbf{z}^*)\right) \pm c_{\alpha/2}\frac{\sqrt{\widehat{\text{Var}}\left[\hat{H}(t|\mathbf{z}^*) - \hat{H}(4|\mathbf{z}^*)\right]}}{\left(\hat{H}(t|\mathbf{z}^*) - \hat{H}(4|\mathbf{z}^*)\right)}.$$

So the $100(1-\alpha)\%$ log-transformed confidence interval for $H(t|\mathbf{z}^*) - H(4|\mathbf{z}^*)$ is

$$\left(\hat{H}(t|\mathbf{z}^*) - \hat{H}(4|\mathbf{z}^*)\right) \exp\left\{ \pm c_{\alpha/2} \frac{\sqrt{\widehat{\text{Var}}\left[\hat{H}(t|\mathbf{z}^*) - \hat{H}(4|\mathbf{z}^*)\right]}}{\left(\hat{H}(t|\mathbf{z}^*) - \hat{H}(4|\mathbf{z}^*)\right)} \right\}, \tag{2.21}$$

where $c_{\alpha/2}$ is the upper $\alpha/2$ fractile of the standard normal distribution. Next, we construct an approximate $100(1-\alpha)\%$ confidence interval for $p(t;\mathbf{z}^*) = \exp\left(H(t|\mathbf{z}^*) - H(4|\mathbf{z}^*)\right)$ by exponentiating the lower and upper bound of this last confidence interval (2.21)

$$\exp\left\{ \left(\hat{H}(t|\mathbf{z}^*) - \hat{H}(4|\mathbf{z}^*)\right) \exp\left\{ \pm c_{\alpha/2} \frac{\sqrt{\widehat{\text{Var}}\left[\hat{H}(t|\mathbf{z}^*) - \hat{H}(4|\mathbf{z}^*)\right]}}{\left(\hat{H}(t|\mathbf{z}^*) - \hat{H}(4|\mathbf{z}^*)\right)} \right\} \right\},$$

where $c_{\alpha/2}$ is the upper $\alpha/2$ fractile of the standard normal distribution. If $t = 4$ the denominator equals 0, but in that case we know the outcome for each censored gap, namely withdrawal, so no lower and upper bound are needed.

### 2.9.2 Lower Bound and Upper Bound for Imputation Method B

Remember that for imputation method B, we computed the distribution of $\ln[\hat{p}(t;\mathbf{z})]$. Now, we aim to generate an imputation lower and upper bound for imputation method B. First, we compute the distribution of the lower and upper bound of the constructed confidence interval for $\ln[p(t;\mathbf{z})]$ in (2.21) by using the delta-method.

We define the function

$$g(\theta) = \theta \exp\left\{ -c_{\alpha/2} \frac{\sqrt{\widehat{\text{Var}}(\theta)}}{\theta} \right\}.$$

The lower bound for $\ln[p(t;\mathbf{z})] = H(t|\mathbf{z}^*) - H(4|\mathbf{z}^*)$ is

$$\begin{aligned} \ln[\hat{p}_{\text{lower}}(t;\mathbf{z}^*)] &= \left(\hat{H}(t|\mathbf{z}^*) - \hat{H}(4|\mathbf{z}^*)\right) \exp\left\{ -c_{\alpha/2} \frac{\sqrt{\widehat{\text{Var}}\left[\hat{H}(t|\mathbf{z}^*) - \hat{H}(4|\mathbf{z}^*)\right]}}{\left(\hat{H}(t|\mathbf{z}^*) - \hat{H}(4|\mathbf{z}^*)\right)} \right\} \\ &= g(\theta), \end{aligned}$$

where $\theta = \hat{H}(t|\mathbf{z}^*) - \hat{H}(4|\mathbf{z}^*)$. Then

$$g'(\theta) = \exp\left\{ -c_{\alpha/2} \frac{\sqrt{\widehat{\text{Var}}(\theta)}}{\theta} \right\} + \theta \exp\left\{ -c_{\alpha/2} \frac{\sqrt{\widehat{\text{Var}}(\theta)}}{\theta} \right\} \left( \frac{c_{\alpha/2} \sqrt{\widehat{\text{Var}}(\theta)}}{\theta^2} \right).$$

So, by applying the delta-method, the distribution of the lower bound is given by

$$\ln[\hat{p}_{\text{lower}}(t;\mathbf{z}^*)] \xrightarrow{\mathcal{L}} N\left(\ln[p_{\text{lower}}(t;\mathbf{z}^*)], \widehat{\text{Var}}(\theta)(g'(\theta))^2\right),$$

and that is all we need to know to perform imputation method B on the lower bound of the confidence interval. We estimate $p_{B,\text{lower}}(t;\mathbf{z}^*)]$ by exponentiating draws from this normal distribution. Imputed outcomes are obtained by taking draws from the binomial distribution $Bin(1, \hat{p}_{B,\text{lower}}(t;\mathbf{z}^*))$.

Analogously, the upper bound for $\ln[p(t;\mathbf{z})] = H(t|\mathbf{z}^*) - H(4|\mathbf{z}^*)$ is

$$\ln[\hat{p}_{\text{upper}}(t;\mathbf{z}^*)] \xrightarrow{\mathcal{L}} N\left(\ln[p_{\text{upper}}(t;\mathbf{z}^*)], \widehat{\text{Var}}(\theta)(h'(\theta))^2\right),$$

where

$$h(\theta) = \theta \exp\left\{c_{\alpha/2}\frac{\sqrt{\widehat{\text{Var}}(\theta)}}{\theta}\right\}.$$

Table 2.7 presents the percentage of withdrawal gaps for all imputed censored gaps by imputation method B on the estimated mean, lower and upper bound (95% CI for $\ln[p(t;\mathbf{z})]$). As expected the percentages of withdrawal gaps for the imputation lower and upper bound seem to be a lower and upper bound for the imputation mean by imputation method B, based on 5 imputations. As well as for the estimated mean $\hat{p}_B(t;\mathbf{z}^*)$, the results for the lower bound $\hat{p}_{B,\text{lower}}(t;\mathbf{z}^*)$ and upper bound $\hat{p}_{B,\text{upper}}(t;\mathbf{z}^*)$ seem not to differ much over the 5 imputations we performed.

| | Percentage of withdrawal gaps | | | | |
|---|---|---|---|---|---|
| Imputation | 1 | 2 | 3 | 4 | 5 |
| Method B - lower bound | 86.74 | 87.16 | 87.06 | 86.25 | 86.67 |
| Method B | 87.29 | 87.51 | 88.48 | 88.06 | 87.32 |
| Method B - upper bound | 90.55 | 90.97 | 91.49 | 90.88 | 90.39 |

**Table 2.7:** The percentage of withdrawal gaps for all imputed censored gaps by imputation method B on the estimated mean, lower and upper bound (95% CI for $\ln[p(t;\mathbf{z})]$).

## 2.10 Discussion

In this chapter we applied the multiple imputation method in order to handle the missing and delayed event type data.

In our case the imputation model is built, based on fully observed data. Therefore, we made the assumption of a (fairly) stationary process, so we can use the information obtained prior to October 1, 2005 to estimate the expected withdrawal status for students with censored

gaps by September 30, 2009. Thanks to this assumption we obtained a closed form solution for the imputation model, estimating

$$P(\text{withdrawer}|\mathbf{Z}, \text{gap} > t \text{ years}; C = t) = p(t; \mathbf{Z}).$$

On the other hand, if we would build the imputation model based on all observed gaps starting between October 1, 1990 and September 30, 2009, another method is required for estimating model parameters based on the observed data only. For example the expectation-maximization (EM) algorithm can be applied, which is a technique for maximum-likelihood estimates in parametric models for incomplete data. This iterative procedure is described in Chapter 4.

Once we have estimated this probability $p(t; \mathbf{Z})$, imputed data sets can be constructed in several ways. We developed two methods: Imputation method A takes draws from the binomial distribution $Bin(1; \hat{p}_A(t; \mathbf{z}^*))$ with $p_A(t; \mathbf{z}^*)$ estimated in equation (2.19) on p. 29, while imputation method B takes draws from the binomial distribution $Bin(1; \hat{p}_B(t; \mathbf{z}^*))$ with $p_B(t; \mathbf{z}^*)$ estimated in equation (2.20) on p. 29. Comparing the number of censored gaps imputed as withdrawal by imputation method A and imputation method B in table 2.6 (p. 31), no big differences nor a specific trend are noticed.

Since we estimated the distribution of $\ln[p(t; \mathbf{z}^*)]$ in (2.16), a lower and upper bound can be constructed for imputation method B. Obviously, the lower and upper bound for method B are non-symmetrical as we used a transformation function to calculate the 95% confidence interval boundaries. We have, however, only considered one possible transformation, and there may exist better ones [4]. Also, we have only considered one possible estimator for the variance of the cumulative hazard function in equation (2.17) on p. 29, but there are alternatives available: one alternative is presented in [8] and compared to the estimator we used, and another estimator based on the counting process theory is presented in [1].

In the descriptive analysis for the subset of all gaps starting prior to October 1, 2005, we estimated the probability of a withdrawal gap for a random gap is 57% by equation (2.2) on p. 16. The counterpart of this probability can be calculated for each imputation. Remark that the estimated percentages of withdrawal gaps in table 2.6 (p. 31) and table 2.7 (p. 33) do not include the portion of gaps started after September 30, 2005 and ended prior to October 1, 2009. These are non-censored gaps, so without an imputed outcome and are always categorized as non-withdrawal. Taking these gaps into account, we calculate the percentage of withdrawal gaps for the subset of all gaps started between October 1, 2005 and September 30, 2009 and obtain table 2.8.

| | Percentage of withdrawal gaps | | | | |
|---|---|---|---|---|---|
| Imputation | 1 | 2 | 3 | 4 | 5 |
| Method A | 54.08 | 54.91 | 54.63 | 54.15 | 53.87 |
| Method B - lower bound | 53.28 | 53.38 | 53.49 | 52.75 | 53.38 |
| Method B | 53.54 | 53.91 | 54.51 | 54.23 | 53.51 |
| Method B - upper bound | 55.70 | 55.84 | 56.50 | 55.80 | 55.59 |

**Table 2.8:** The percentage of withdrawal gaps for all gaps started between October 1, 2005 and September 30, 2009 by imputation method A and imputation method B (mean, lower and upper bound).

The percentages in table 2.8 are all lower than 57%, although this difference is not large, it indicates that the assumption of a stationary process of the gaps is probably violated.

# Chapter 3

# Competing Risks Analysis

All individuals enter the study at a particular time point between October 1, 1990 and September 30, 2009, according to their date of first appointment to the Ph.D.-study. Every Ph.D.-student is followed from this entry time until the event of interest occurs (Ph.D.-attainment or withdrawal, whichever comes first) or until he/she is censored. The outcomes Ph.D.-attainment and withdrawal are competing risks.
Students who have not yet obtained their Ph.D.-degree by the end of the study are classified either as a withdrawer or as an ongoer. The latter classification contains observed (all those who are not in a gap by the end of the study are ongoers) and imputed values.

Throughout this chapter we will perform analysis on the Ph.D.-students data set, described in section 1.3 (p. 3). In Chapter 2 we obtained multiple imputed copies of this data set for both imputation methods (imputation A and imputation B). Now we aim to draw statistical inference from these data sets. Multiple imputation inference from 5 imputed data sets involves three distinct phases:

- The missing data are filled in 5 times to generate 5 'completed' data sets.
- The 5 'completed' data sets are analyzed by using standard procedures.
- The results from the 5 'completed' data sets are combined for the inference.

The first step has already been developed in the previous chapter. In this chapter we will fulfill the next two steps. We start by building a cause-specific PH model for both types of event (Ph.D.-attainment and withdrawal) and estimate the cumulative incidence functions for all imputed data sets by using standard procedures. Next, we combine these results and estimate pointwise confidence intervals for the cumulative incidence function.

## 3.1 Descriptive Analysis of the Ph.D.-students Data

First, we briefly describe all 28,871 students starting a Ph.D.-training between October 1, 1990 and September 30, 2009.

From table 3.1, we observe more male (15,827) than female students (13,028) starting a Ph.D.-training during the observation period. Regarding the scientific field, the highest proportion of Ph.D.-students is found in the medicine (24%), followed by applied sciences and sciences (both 22%). Compared to these scientific fields, social sciences (18%) and humanities (15%) have less Ph.D.-students starting during the observation period. Most Ph.D.-students have Belgian nationality (82%), followed by European EU and others (both 9%). Considering the dominant statute classification most students follow the Ph.D.-training in the assistent lectureship (29%) or other project funding (27%) statute. Accounting for 5% of all Ph.D.-students, the lowest proportion of Ph.D.-students is obviously found in the project funding (FWO, BOF or IUAP). Most students starting the Ph.D.-training are less than 25 years of age (71%), and we are detecting a trend of lower proportion of Ph.D-students starting the training with increasing age. There are 26% of all Ph.D.-students starting during the first start cohort (1990-1997), 41% during the second (1997-2004) and 33% during the last start cohort (2004-2009). We are not comparing the amount of Ph.D.-students in the five Flemish universities explicitly. The covariate nationality has most missing values (245).

Remember that the reference group consists of Belgian men funded by other projects, with the dominant scientific field 'sciences', who were less than 25 years old when they started their Ph.D.-training between October 1, 1990 and September 30, 1997 at a specific Flemish university (see section 1.3 p. 3). There are 29 Ph.D.-students in this reference group out of 28,871 in the study.

## 3.2 Definition and Notation

Once more, we are working in a survival setting, but now on the database of all observed Ph.D.-students instead of observed gaps (Chapter 2). Similar notation is used in this setting.

We are interested in sponsored time until one of the two events occur, Ph.D.-attainment or withdrawal, whichever comes first. Only one of these competing risks is actually observed and is called the type of event. Denote the sponsored time to the event 'Ph.D.-attainment' by $X_p$ and sponsored time to the event 'withdrawal' by $X_w$. In other words, $X_k$ ($k = p$ or $w$), represents the random sponsored time when the Ph.D.-student is exposed to the $k$th risk only. However, in real life both risks act simultaneously and we only observe the shortest of both event times. Denote the censored sponsoring time by the random variable $C$. The number of students starting a Ph.D.-training between October 1, 1990 and September 30, 2009, is

| Variable | Frequency | Percent | Missing values |
|---|---|---|---|
| **Gender** | | | 16 |
| [Male] | 15,827 | 55% | |
| Female | 13,028 | 45% | |
| **Dominant scientific field** | | | 148 |
| [sciences] | 6,234 | 22% | |
| medicine | 6,787 | 24% | |
| humanities | 4,188 | 15% | |
| social | 5,179 | 18% | |
| applied | 6,335 | 22% | |
| **Nationality** | | | 245 |
| [Belgian] | 23,438 | 82% | |
| European Union (excl. Belgium) | 2,639 | 9% | |
| Other | 2,549 | 9% | |
| **Dominant statute classification** | | | 0 |
| Assistant lectureship | 8,368 | 29% | |
| Compet. scholarship (Flanders) | 5,806 | 20% | |
| Compet. scholarship (own university) | 5,455 | 19% | |
| Project funding (FWO, BOF, IUAP) | 1,311 | 5% | |
| Project funding (other) | 7,931 | 27% | |
| **Age (at start)** | | | 80 |
| [$\leq$ 25 years] | 20,526 | 71% | |
| $26 - 30$ years | 5,308 | 18% | |
| $31 - 35$ years | 1,674 | 6% | |
| $36 - 40$ years | 667 | 2% | |
| $>$ 40 years | 616 | 2% | |
| **Start cohort** | | | 0 |
| [01/10/1990 - 30/09/1997] | 7,379 | 26% | |
| 01/10/1997 - 30/09/2004 | 11,857 | 41% | |
| 01/10/2004 - 30/09/2009 | 9,635 | 33% | |

**Table 3.1:** Classification of all Ph.D.-students starting between October 1, 1990 and September 30, 2009, corresponding to the personal covariate values used in the competing risks setting (excl. university). For each covariate, the [reference category] is indicated.

denoted by $n$. We define sponsored time as the total calendar time minus the time in gaps. More specifically, the time, event indicators and baseline covariates for the $j$th Ph.D.-student, $j = 1 \dots n$, are defined as follows:

- sponsored time $T_j = \min(X_{p,j}, X_{w,j}, C_j)$: Total time from study entry to final observation minus, in the first place, time in absence, interim, success and withdrawal gaps and, in the second place, time spent in statute classification 7b, 8 or R. Statute 7b, 8 and R represent respectively other junior statutes without a Ph.D.-attainment purpose, volunteers and a residual category. No distinction is made between full-time and part-time commitments.

- event indicators for the $j$th Ph.D.-student

$$\delta_{p,j} = \begin{cases} 0 & \text{if ongoing or withdrawal} \quad \Leftrightarrow \quad T_j = C_j \text{ or } X_{w,j} \\ 1 & \text{if Ph.D.-attainment} \quad \Leftrightarrow \quad T_j = X_{p,j} \end{cases}$$

$$\delta_{w,j} = \begin{cases} 0 & \text{if ongoing or Ph.D.-attainment} \quad \Leftrightarrow \quad T_j = C_j \text{ or } X_{p,j} \\ 1 & \text{if withdrawal} \quad \Leftrightarrow \quad T_j = X_{w,j} \end{cases}$$

- the set of baseline covariates in both models is denoted by the union $\mathbf{Z} = (\mathbf{Z}_p, \mathbf{Z}_w)$ of two subsets, where

  $\mathbf{Z}_{p,j}$ is a set of prognostic factors for the $j$th Ph.D.-student in the cause-specific hazards model with outcome of interest 'Ph.D.-attainment'

  $\mathbf{Z}_{w,j}$ is a set of prognostic factors for the $j$th Ph.D.-student in the cause-specific hazards model with outcome of interest 'withdrawal'.

Two event indicators $\delta_{p,j}$ and $\delta_{w,j}$ are defined, the former to perform a competing risks analysis with outcome of interest 'Ph.D.-attainment' and the latter to perform a competing risks analysis with outcome of interest 'withdrawal'. We assume that the random vectors $(T_j, \delta_{p,j}, \delta_{w,j}, \mathbf{Z}_j)$ are independent and identically distributed for $j = 1 \dots n$. We suppose that the competing events Ph.D.-attainment and withdrawal occur at $D$ distinct times $t_1 < t_2 < \dots < t_D$. We allow for possible ties in the data and at time $t_i$ $(i = 1, \dots, D)$ there are $d_{p,i}$ Ph.D.-attainments and $d_{w,i}$ withdrawals, where only $d_{w,i}$ depends on the specific imputation set obtained under imputation method A or B.

Remark that the defined time to event $T_j$ is not influenced by the imputation value, since we are considering sponsored time. If in contrast total calendar time were observed, the time to event $T_j$ would depend on the imputed outcome withdrawer or ongoer. The withdrawer event time would be the observed time until the last sponsoring before withdrawal (excluding the last gaptime, since this is reporting delay) and the ongoer event time would be the observed time until censoring (including the last gaptime). Hence for the $j$th Ph.D.-student with missing outcome, if he is imputed as ongoer, his event time would always be larger than if he were imputed as withdrawer.

## 3.3 Model Building in the Competing Risks Setting

### 3.3.1 The Cause-specific PH model

To build regression models in this competing risks setting we distinguish between Ph.D.-attainers, withdrawers and ongoers. We encounter different scenarios by the end of the study, which was set at September 30, 2009.

1. The junior researcher has attained a Ph.D.-degree by September 30, 2009 at an observed time point and had no gaps lasting more than 4 years. For our purposes, a Ph.D.-degree which might be reached after more than 4 years of non-sponsored time is no longer seen as the expected yield of the investment and thus considered as no result for the sponsored time.

2. The junior researcher has not been sponsored for a period lasting more than 4 years during the observation period from October 1, 1990 until September 30, 2009.
   This person will be classified as a withdrawer even if he/she attains a Ph.D.-degree before September 30, 2009.

3. The junior researcher has not (yet) attained a Ph.D.-degree by September 30, 2009, had no gaps lasting more than 4 years and

   (a) is sponsored at September 30, 2009.

   (b) is not sponsored, so in a gap by September 30, 2009 and classified as ongoer by the imputation procedure.

   (c) is not sponsored, so in a gap by September 30, 2009 and classified as withdrawer by the imputation procedure.

People in scenario 1 obviously have the event 'Ph.D.-attainment'. Since the database only contains direct information about Ph.D.-attainment and not about withdrawers, we have made some assumptions.

- We assume people in scenario 2 and 3c have the event 'withdrawal' and total sponsored time is sponsored time until the start of the withdrawal gap.

- People in scenario 3a and 3b we consider as not having had any event by the end of the study, they are censored for both competing risks.

Moreover we are assuming non-informative censoring: We assume that the potential censoring time is unrelated to the potential event time for both competing risks (Ph.D.-attainment and withdrawal from the Ph.D.-training), conditionally on the considered set of covariates $\mathbf{Z}_p$ respectively $\mathbf{Z}_w$ in the cause-specific Cox PH models. This assumption would be violated, for

example, if Ph.D.-students with poor prognosis were routinely censored (e.g. entered later into the study).

We aim to estimate the probability of attaining a Ph.D.-degree, respectively withdrawal from the Ph.D.-training, within a certain amount of sponsored time after starting the Ph.D.-training. The cause-specific hazard of attaining a Ph.D.-degree ($p$), respectively withdrawal from the Ph.D.-training ($w$), at a certain time point $t$,

$$h_k(t|\mathbf{Z}_k) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, \delta_k = 1 | T \geq t, \mathbf{Z}_k)}{\Delta t},$$

$k = p$ or $w$, will be modeled in function of a subset of prognostic factors $\mathbf{Z}_k$ using a Cox proportional hazards model

$$h_k(t|\mathbf{Z}_k) = h_{0,k}(t) \exp(\boldsymbol{\beta}_k^T \mathbf{Z}_k),$$

where $t$ represents the sponsored time to the event of interest ($k = p$ or $w$), while the other events are handled as competing risks i.e. they stay in the risk set until their observed event and then are censored.

For the outcome of interest Ph.D.-attainment we distinguish between Ph.D.-attainers and students who have not (yet) attained a Ph.D.-degree, i.e. ongoers and withdrawers. So, ongoers and withdrawers are both censored for the event Ph.D.-attainment ($\delta_p = 0$). Since no distinction is made between ongoers and withdrawers, and all Ph.D.-attainments are clearly observed if they occur prior to the end of the study, no imputation is needed for analyzing the cause-specific hazard for Ph.D.-attainment. A cause-specific PH model can be fit immediately. Withdrawals from the Ph.D.-training, however, are not directly observed as endpoints at a given time, so missing outcomes are multiple imputed to build a cause-specific PH model for withdrawal.
The following sets of cause-specific PH models are built:

- Cause-specific PH model with outcome of interest 'Ph.D.-attainment', attained straightforward.
- Cause-specific PH model with outcome of interest 'withdrawal', for each of the 5 imputed data sets attained by imputation method A.
- Cause-specific PH model with outcome of interest 'withdrawal', for each of the 5 imputed data sets attained by imputation method B.

There are 475 out of 28,871 observations having missing covariate values. Excluding them, there are 28,396 observations used for the analysis. This exclusion is justified if we assume these missing covariate values are missing completely at random (MCAR), i.e. the probability

that a covariate value is missing does not depend on the unobserved value or on the value of any other observed data.
Based on the weaker missing at random assumption (MAR), we could prevent excluding 475 observations with missing covariate values for the analysis, by using multiple imputation to impute these missing covariate values and preventing loss of information or potential bias.

In section 2.6 (p. 23) we built a Cox PH model for gaptime, including all main effects and interaction effects that could be of interest for the competing risks analysis. This because we should not include effects in the competing risks models that were not included in the imputation model. Including effects in the competing risks models that were not taken into account for the imputation values would give improper results.
Now, we perform a backward selection at the 5% significance level, starting from all main effects and interaction effects of interest: the baseline covariates listed in section 1.3 (p. 3) and interactions gender × dominant statute, gender × dominant scientific field, nationality × dominant statute, nationality × dominant scientific field and gender × nationality. When considering the interactions, two categories are formed for nationality: Belgian and non-Belgian Ph.D.-students, because of too few observations for some nationality categories. For all other covariates we use the same categories as listed previously (section 1.3, p. 3).

It is very probable that - above the baseline information we included - the subsequence of certain appointments (e.g. change in statute classification, gaps) in a Ph.D.-training contains information on the time to attainment of a Ph.D.-degree. If we wish to take this information into account in the cause-specific hazards model, there will be some complications. For instance, it will be much harder to estimate the cumulative incidence function for a specific scenario of e.g. subsequent statute classifications, especially when we have to combine that with time-dependent information on e.g. previous gaps [2].

### 3.3.2 Building a Cause-specific PH Model for Time to Ph.D.-attainment

The cause-specific hazard of attaining a Ph.D.-degree at a certain time point $t$,

$$h_p(t|\mathbf{Z}_p) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, \delta_p = 1 | T \geq t, \mathbf{Z}_p)}{\Delta t},$$

is modeled in function of a subset of prognostic factors $\mathbf{Z}_p$ using a Cox proportional hazards model

$$h_p(t|\mathbf{Z}_p) = h_{0,p}(t) \exp(\boldsymbol{\beta}_p^T \mathbf{Z}_p),$$

where $t$ represents the sponsored time to the event of interest, here attainment of a Ph.D.-degree, while withdrawals are handled as competing risks: they stay in the risk set until their observed withdrawal and then are censored. The cause-specific hazard function $h_p$ calculates at each time $t$ the conditional probability that a Ph.D.-student with covariate values $\mathbf{z}$ instantaneously attains a Ph.D.-degree, given that the student was at risk just before time

$t$. Note that we are not working on an imputed data set to model the cause-specific hazard for Ph.D.-attainment.

We built a cause-specific hazards model for Ph.D.-attainment, based on the triples of information $(T_j, \delta_{p,j}, \mathbf{Z}_j), j = 1 \dots n$. For the outcome of interest 'Ph.D.-attainment' the interaction effects nationality × dominant scientific field and gender × nationality were not significant at the 5% significance level and therefore excluded from the model. This resulted in the subset of prognostic factors $\mathbf{Z}_p$. Detailed output is given in the appendix (p. 85).

### 3.3.3 Building a Cause-specific PH Model for Time to Withdrawal

Similarly, the cause-specific hazard for the competing event withdrawal from the Ph.D.-training at a certain time point $t$,

$$h_w(t|\mathbf{Z}_w) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, \delta_w = 1 | T \geq t, \mathbf{Z}_w)}{\Delta t},$$

is also modeled with a Cox proportional hazards model

$$h_w(t|\mathbf{Z}_w) = h_{0,w}(t) \exp(\boldsymbol{\beta}_w^T \mathbf{Z}_w),$$

where $t$ represents the sponsored time to the event of interest, here withdrawal from the Ph.D.-training, while Ph.D.-attainments are handled as competing risks: they stay in the risk set until their observed Ph.D.-attainment and then are censored. Recall that the event 'withdrawal' is not observed directly and its 'observation' depends strongly on our assumptions related to the cut-point of $X = 4$ years. The cause-specific hazard function $h_w$ calculates at each time $t$ the conditional probability that a Ph.D.-student with covariate values $\mathbf{z}$ instantaneously withdraws from the Ph.D.-training, given that the student was at risk just before time $t$.

We built a cause-specific hazards model for withdrawal, based on the triples of information $(T_j, \delta_{w,j}, \mathbf{Z}_j), j = 1 \dots n$. For the outcome of interest 'withdrawal' the backward selection resulted in the same set of included covariates for both imputation methods A and B. The main effect gender and the interaction effect gender × dominant statute were not significant at the 5% significance level. As no main effect should be excluded unless all related interaction effects are excluded, we first exclude the interaction effects with gender one by one. Excluding these interaction effects, the main effect gender turns out to be significant at the 5% significance level, so gender is not excluded from the final model. Summarizing, all main effects are included in the model as well as the interaction effects nationality × dominant statute and nationality × dominant scientific field. This resulted in the subset of prognostic factors $\mathbf{Z}_w$. Detailed output is given in the appendix (p. 89).

## 3.4 The Cumulative Incidence Function

The naively derived 'survival function' $S_p(t|\mathbf{Z}_p) = \exp\left(-\int_0^t h_p(u|\mathbf{Z}_p)\mathrm{d}u\right)$ cannot be interpreted as a standard survival function, as it would only have this interpretation in the counterfactual situation that the competing event of withdrawal cannot occur. In contrast, the cumulative incidence function (CIF) expresses what percentage of starters can be expected to have obtained their Ph.D.-degree by year $t$, respectively to have withdrawn from the Ph.D.-training by year $t$, conditional on baseline covariates $\mathbf{z}^*$:

$$\begin{aligned} I_p(t|\mathbf{z}^*) = P(T \leq t, \delta_p = 1|\mathbf{z}^*) &= \int_0^t h_p(u|\mathbf{z}_p^*)S(u|\mathbf{z}^*)\mathrm{d}u \\ I_w(t|\mathbf{z}^*) = P(T \leq t, \delta_w = 1|\mathbf{z}^*) &= \int_0^t h_w(u|\mathbf{z}_w^*)S(u|\mathbf{z}^*)\mathrm{d}u, \end{aligned} \tag{3.1}$$

where $\mathbf{z}^* = (\mathbf{z}_p^*, \mathbf{z}_w^*)$ represents a specific value of a vector of covariates for which we want to estimate the effects, e.g. $\mathbf{z}_1^* = (1, 0, 0, \ldots, 0)^T$, $\mathbf{z}_2^* = (0, 1, 0, \ldots, 0)^T$ etc. Not all combinations of values are necessarily observed in the data set of all Ph.D.-students starting between October 1, 1990 and September 30, 2009. In that way we compare the estimated covariate effect to the reference group with covariate values $\mathbf{z}^* = (0, 0, 0, \ldots, 0)^T$. As denoted in section 1.3 (p. 3), this reference group consists of Belgian men funded by other projects, with the dominant scientific field 'sciences', who were less than 25 years old when they started their Ph.D.-training between October 1, 1990 and September 30, 1997 at a specific Flemish university.

In the competing risks setting, the overall survival function $S(t|\mathbf{z}^*)$ expresses the survival distribution for the minimum of all event times, knowing $\mathbf{z}^*$ [7]. So $T$ is not the time to event of interest (which is possibly unobserved), but the time to the first event occurring (Ph.D.-attainment or withdrawal). This overall survival function is given by

$$\begin{aligned} S(t|\mathbf{z}^*) &= P(T > t|\mathbf{z}^*) \\ &= \exp\left\{-H(t|\mathbf{z}^*)\right\} \\ &= \exp\left\{-\left[H_p(t|\mathbf{z}_p^*) + H_w(t|\mathbf{z}_w^*)\right]\right\} \\ &= \exp\left\{-\left[H_{0,p}(t)\exp(\boldsymbol{\beta}_p^T\mathbf{z}_p^*) + H_{0,w}(t)\exp(\boldsymbol{\beta}_w^T\mathbf{z}_w^*)\right]\right\}, \end{aligned}$$

where $H_{0,p}(t)$ represents the baseline cumulative hazard and $\boldsymbol{\beta}_p^T\mathbf{z}_p^*$ the prognostic score for the model with outcome of interest Ph.D.-attainment, similarly for the outcome of interest withdrawal.

The cumulative incidence function $I_p(t|\mathbf{z}^*)$ gives at any time point $t$ the probability that a Ph.D.-student with covariate values $\mathbf{z}^*$ will attain the Ph.D.-degree within $t$ years of sponsored time after starting the Ph.D.-training, accounting for the fact that some students will

withdraw from the Ph.D.-training. Similarly, the cumulative incidence function $I_w(t|\mathbf{z}^*)$ gives at any time point $t$ the probability that a Ph.D.-student with covariate values $\mathbf{z}^*$ will have withdrawn from the Ph.D.-training within $t$ years of sponsored time after starting the Ph.D.-training, accounting for the fact that some students will attain a Ph.D.-degree. The interpretation of this cumulative incidence function is relatively simple and the results reflect the observable situation. Effects from covariates on the other hand are correctly interpretable from the cause-specific hazards.

### 3.4.1 Estimating the Cumulative Incidence Function

We wish to estimate the cumulative incidence functions in (3.1). The discrete cumulative incidence function for the outcome of interest 'Ph.D.-attainment' and 'withdrawal' can respectively be estimated by [5]

$$\hat{I}_p(t|\mathbf{z}^*) = \sum_{t_i \leq t} \hat{S}(t_i|\mathbf{z}^*) \left[\hat{H}_{0,p}(t_i) - \hat{H}_{0,p}(t_{i-1})\right] \exp\left(\hat{\boldsymbol{\beta}}_p^T \mathbf{z}_p^*\right) \tag{3.2}$$

$$\hat{I}_w(t|\mathbf{z}^*) = \sum_{t_i \leq t} \hat{S}(t_i|\mathbf{z}^*) \left[\hat{H}_{0,w}(t_i) - \hat{H}_{0,w}(t_{i-1})\right] \exp\left(\hat{\boldsymbol{\beta}}_w^T \mathbf{z}_w^*\right), \tag{3.3}$$

where the overall survival function at any time point $t_i, i = 1 \dots D$, can be consistently estimated by

$$\hat{S}(t_i|\mathbf{z}^*) = \exp\left\{-\left[\hat{H}_{0,p}(t_i)\exp(\hat{\boldsymbol{\beta}}_p^T \mathbf{z}_p^*) + \hat{H}_{0,w}(t_i)\exp(\hat{\boldsymbol{\beta}}_w^T \mathbf{z}_w^*)\right]\right\}. \tag{3.4}$$

We should expect the cumulative incidence function for time to Ph.D.-attainment (3.2) is not depending on the imputation method, but the opposite is true. From (3.4) we see the overall survival function depends on the estimated $\hat{\boldsymbol{\beta}}_w$ and in that way on the imputation method.

We perform these calculations explicitly for the outcome of interest 'Ph.D.-attainment' ($p$), but they can be repeated for 'withdrawal' ($w$) by exchanging $p$ and $w$. These quantities are estimated for all imputations ($1 \dots 5$) and both imputation methods (imputation A and imputation B). Afterwards the results from each 5 imputations are combined by the rules for combining results from imputed data sets.

We first estimate the discrete baseline cumulative hazard functions $H_{0,p}(t)$ and $H_{0,w}(t)$ in the same way as in section 2.7 (p. 25). Let

$$W(t_i, \boldsymbol{\beta}_p) = \sum_{j \in R(t_i)} \exp\left(\boldsymbol{\beta}_p^T \mathbf{z}_{p,j}\right),$$

where $R(t_i)$ represents the set of all individuals at risk at time $t_i$ ($i = 1 \dots D$). Note that this set is independent of the considered outcome (Ph.D.-attainment vs. withdrawal) or imputation method (imputation A vs. imputation B). Here the covariate values $\mathbf{z}_{p,j}$ are

effectively corresponding to the $j$th observed Ph.D.-student in the study in contrast to the covariate values $\mathbf{z}_p^*$.

Let $\hat{\boldsymbol{\beta}}_p$ be the maximum partial likelihood estimate for $\boldsymbol{\beta}_p$, then the discrete baseline cumulative hazard is estimated by the Breslow estimator

$$\hat{H}_{0,p}(t) = \sum_{t_i \leq t} \frac{d_{p,i}}{W(t_i, \hat{\boldsymbol{\beta}}_p)}$$

where $d_{p,i}$ represents the number of Ph.D.-attainments at event time $t_i$. Remember that for the competing event 'withdrawal', the number of withdrawals $d_{w,i}$ at event time $t_i$ is depending on the imputation.

### 3.4.2 Distribution of the Cumulative Incidence Function

Based on the theory in [5] we discuss the distribution of the cumulative incidence function. Again, the event 'Ph.D.-attainment' ($p$) can be replaced by the competing event 'withdrawal' ($w$).

It can be shown that the distribution of the process

$$U_p(t|\mathbf{z}^*) = \sqrt{n} \left\{ \hat{I}_p(t|\mathbf{z}^*) - I_p(t|\mathbf{z}^*) \right\},$$

can be approximated by that of a Gaussian process whose realizations can be easily generated through simulation. The process $U_p(t|\mathbf{z}^*)$ is asymptotically equivalent to a process $\tilde{U}_p(t|\mathbf{z}^*)$ defined in (A.4) in [5], which converges weakly to a zero-mean Gaussian process. The covariance function of $\tilde{U}_p$ is denoted by $\xi_p(s,t|\mathbf{z}^*)$ [5] and a consistent estimator $\hat{\xi}_p(t,t|\mathbf{z}^*)$ for the variance function $\xi_p(t,t|\mathbf{z}^*)$ at time $t$ can be obtained by replacing all the theoretical quantities given there with their empirical counterparts. The explicit formula for the estimated variance function $\hat{\xi}_p(t,t|\mathbf{z}^*)$ is given below.

Define the random scalar $S^{(0)}$, $p$-vector $S^{(1)}$ and $(p \times p)$-matrix $S^{(2)}$ by

$$\begin{aligned}
S^{(0)}(t,\boldsymbol{\beta}) &= \frac{1}{n} \sum_{j \in R(t)} \exp(\boldsymbol{\beta}^T \mathbf{Z}_j) = \frac{1}{n} W(t,\boldsymbol{\beta}) \\
S^{(1)}(t,\boldsymbol{\beta}) &= \frac{1}{n} \sum_{j \in R(t)} \exp(\boldsymbol{\beta}^T \mathbf{Z}_j) \mathbf{Z}_j \\
S^{(2)}(t,\boldsymbol{\beta}) &= \frac{1}{n} \sum_{j \in R(t)} \exp(\boldsymbol{\beta}^T \mathbf{Z}_j) \mathbf{Z}_j \mathbf{Z}_j^T,
\end{aligned}$$

and the $p$-vector $\bar{Z}$ is defined by

$$\bar{Z}(t,\boldsymbol{\beta}) = \frac{S^{(1)}(t,\boldsymbol{\beta})}{S^{(0)}(t,\boldsymbol{\beta})}.$$

Then a consistent estimator for $\xi_p$ is

$$\begin{aligned}\hat{\xi}_p(t,t|\mathbf{z}^*) &= \frac{1}{n}\sum_{t_i\le t} d_{i,p}\left[\hat{S}(t_i|\mathbf{z}^*) - \left\{\hat{I}_p(t|\mathbf{z}^*) - \hat{I}_p(t_i|\mathbf{z}^*)\right\}\right]^2 \left(\frac{\exp(\hat{\boldsymbol{\beta}}_p^T\mathbf{z}^*)}{S^{(0)}(t_i,\hat{\boldsymbol{\beta}}_p)}\right)^2 \\ &\quad + \frac{1}{n}\sum_{t_i\le t} d_{i,w}\left\{\hat{I}_p(t|\mathbf{z}^*) - \hat{I}_p(t_i|\mathbf{z}^*)\right\}^2 \left(\frac{\exp(\hat{\boldsymbol{\beta}}_w^T\mathbf{z}^*)}{S^{(0)}(t_i,\hat{\boldsymbol{\beta}}_w)}\right)^2 \\ &\quad + \left\{\hat{\phi}_p(t|\mathbf{z}^*) - \hat{\psi}_{pp}(t|\mathbf{z}^*)\right\}^T \hat{\Omega}_p^{-1}\left\{\hat{\phi}_p(t|\mathbf{z}^*) - \hat{\psi}_{pp}(t|\mathbf{z}^*)\right\} \\ &\quad + \hat{\psi}_{pw}^T(t|\mathbf{z}^*)\hat{\Omega}_w^{-1}\hat{\psi}_{pw}(t|\mathbf{z}^*), \end{aligned} \tag{3.5}$$

where the $p$-vectors $\hat{\phi}$ and $\hat{\psi}$, and $(p\times p)$-matrix $\hat{\Omega}$ are defined by

$$\begin{aligned}\hat{\phi}_k(t|\mathbf{z}^*) &= \sum_{t_i\le t}\hat{S}(t_i|\mathbf{z}^*)\left\{\mathbf{z}^* - \bar{Z}(t_i,\hat{\boldsymbol{\beta}}_k)\right\}\left[\hat{H}_k(t_i|\mathbf{z}^*) - \hat{H}_k(t_{i-1}|\mathbf{z}^*)\right] \\ \hat{\psi}_{k\ell}(t|\mathbf{z}^*) &= \sum_{t_i\le t}\left\{\hat{I}_k(t|\mathbf{z}^*) - \hat{I}_k(t_i|\mathbf{z}^*)\right\}\left\{\mathbf{z}^* - \bar{Z}(t_i,\hat{\boldsymbol{\beta}}_\ell)\right\}\left[\hat{H}_\ell(t_i|\mathbf{z}^*) - \hat{H}_\ell(t_{i-1}|\mathbf{z}^*)\right] \\ \hat{\Omega}_k &= \frac{1}{n}\sum_{t_i\le t_D}\left\{\frac{S^{(2)}(t_i,\hat{\boldsymbol{\beta}}_k)}{S^{(0)}(t_i,\hat{\boldsymbol{\beta}}_k)} - \bar{Z}(t_i,\hat{\boldsymbol{\beta}}_k)\bar{Z}(t_i,\hat{\boldsymbol{\beta}}_k)^T\right\} d_{k,i},\end{aligned}$$

with parameters $k$ and $\ell$ indicating the competing events 'withdrawal' $(w)$ or 'Ph.D.-attainment' $(p)$.

## 3.5 Combining Results from Multiple Imputed Data Sets

As mentioned at the beginning of this chapter, the last phase of the multiple imputation procedure is a well chosen combination of the results from the imputed data sets. Rubin [11] presented the following method for combining multiple imputation results.

Let $I_k$ ($k = p$ or $w$) denote the generic scalar quantity to be estimated, more specifically the cumulative incidence function for outcome of interest 'Ph.D.-attainment' or 'withdrawal'. Let $Y$ denote the intended data, part of which is observed ($Y_{\text{obs}}$) and part of which is missing ($Y_{\text{mis}}$). Let $\hat{I}_k = \hat{I}(Y_{\text{obs}}, Y_{\text{mis}})$ denote the statistic that would be used to estimate $I_k$ if complete data were available, and let $\xi = \xi(Y_{\text{obs}}, Y_{\text{mis}})$ be its squared standard error.

Per imputation method we have 5 independent simulated versions or imputations $Y_{\text{mis}}^{(1)}, \ldots, Y_{\text{mis}}^{(5)}$. From these 5 imputed data sets, 5 different sets of the point and variance estimates for $I_k$ can be computed. Suppose that $\hat{I}_k^{(\ell)} = \hat{I}_k(Y_{\text{obs}}, Y_{\text{mis}}^{(\ell)})$ and $\hat{\xi}^{(\ell)} = \hat{\xi}(Y_{\text{obs}}, Y_{\text{mis}}^{(\ell)})$ are the point and variance estimates, respectively, from the $\ell$th imputed data set, $\ell = 1, \ldots, 5$. Then the combined point estimate for $I_k$ from multiple imputation is simply the average

$$\bar{I}_k = \frac{1}{5}\sum_{\ell=1}^{5}\hat{I}_k^{(\ell)},$$

$k = p$ or $w$. To obtain a standard error for $\bar{I}_k$, one must calculate the between-imputation variance

$$B = \frac{1}{4}\sum_{\ell=1}^{5}(\hat{I}_k^{(\ell)} - \bar{I}_k)^2,$$

and the within-imputation variance, which is the average of the naive variances estimated from each of the 5 complete-data estimates

$$W = \frac{1}{5}\sum_{\ell=1}^{5}\hat{\xi}^{(\ell)}.$$

Then the estimated total variance associated with $\bar{I}_k$ is

$$V = \left(1 + \frac{1}{5}\right)B + W.$$

The statistic $(\bar{I}_k - I_k)/\sqrt{V}$ is approximately distributed as a Student's $t$-distribution with $\nu$ degrees of freedom [11], where

$$\nu = 4\left[1 + \frac{W}{(1+1/5)B}\right]^2.$$

The degrees of freedom $\nu$ depend on the number of imputations, which is set to 5, and the ratio

$$r = \frac{(1+1/5)B}{W}.$$

The ratio $r$ measures the relative increase in variance due to missing data [14]. Notice that if $Y_{\text{mis}}$ carried no information about $I_k$, then the imputed data estimates $\hat{I}_k^{(\ell)}$ would be identical, total variance $V$ would reduce to $W$, and the values of $r$ and $B$ are both zero.
The fraction of missing information in the system is

$$\frac{r}{1+r} = \frac{(1+1/5)B}{W + (1+1/5)B}.$$

It turns out a better estimate of this quantity is

$$\hat{\lambda} = \frac{r + 2/(\nu+3)}{r+1}.$$

Both statistics $r$ and $\lambda$ are helpful diagnostics for assessing how the missing data contribute to the uncertainty about $I_k$.

With a large number of imputations or a small value of $r$, the degrees of freedom $\nu$ will be large and the distribution of $(\bar{I}_k - I_k)/\sqrt{V}$ will be approximately normal.

In the same way the parameter estimates $\hat{\boldsymbol{\beta}}$ can be combined, as well as their associated variance. The results are given in the appendix for both imputation methods A (p. 121) and B (p. 125). No big differences between both imputation methods are detected in the combined parameter estimates and corresponding 95% confidence interval.

## 3.6 Cumulative Incidence Plots

All figures in this section were plotted using the data sets obtained by imputation method B. Analogously these cumulative incidence functions can be plotted using the data sets obtained by imputation method A. There are no big differences noticed between both imputation methods in the plotted cumulative incidence functions, neither in the calculated $p$-values for the parameter estimates and hypothesis tests below.

### 3.6.1 Pointwise Confidence Intervals for the Combined Cumulative Incidence Functions

Pointwise confidence intervals for $I_k(t|\mathbf{z}^*)$ can be constructed. Since $U_k(t|\mathbf{z}^*)$ can be approximated by a Gaussian process [5], the combined estimate $\bar{I}_k(t|\mathbf{z}^*)$ is asymptotically normally distributed with mean $I_k(t|\mathbf{z}^*)$ and variance $\hat{V}(t|\mathbf{z}^*)/n$

$$\bar{I}_k(t|\mathbf{z}^*) \xrightarrow{\mathcal{L}} N\left(I_k(t|\mathbf{z}^*), \frac{\hat{V}(t|\mathbf{z}^*)}{n}\right).$$

However since $I_k(t|\mathbf{z}^*)$ is bounded by 0 and 1, one may obtain interval estimates for $I_k$ based on a transformation of $\bar{I}_k$. To this end, consider the process

$$G_k(t|\mathbf{z}^*) = \sqrt{n}\left[g\{\bar{I}_k(t|\mathbf{z}^*)\} - g\{I_k(t|\mathbf{z}^*)\}\right].$$

Here, $g$ is a known function whose derivative $g'$ is continuous and nonzero. For example, we let $g(y) = \ln(-\ln(y))$. It follows from the delta-method that the process $G_k(t|\mathbf{z}^*)$ is asymptotically equivalent to $g'\{\bar{I}_k(t|\mathbf{z}^*)\}U_k(t|\mathbf{z}^*)$.

An approximate pointwise $100(1-\alpha)\%$ confidence interval for $I_k(t|\mathbf{z}^*)$ is

$$g^{-1}\left[g\left\{\bar{I}_k(t|\mathbf{z}^*)\right\} \pm \frac{1}{\sqrt{n}} g'\left\{\bar{I}_k(t|\mathbf{z}^*)\right\}\hat{V}^{1/2}(t|\mathbf{z}^*)c_{\alpha/2}\right]$$

where $c_\alpha$ is the $100\alpha$ upper percentage point of the standard normal distribution and $g'(y) = 1/(y\ln(y))$.

This theory can be extended to calculate confidence bands instead of confidence intervals for the cumulative incidence functions [5].

### 3.6.2 Outcome of Interest: Ph.D.-attainment

First, we estimate the combined cumulative incidence functions and associated 95% pointwise confidence intervals for the probability of attaining a Ph.D.-degree by year $t$, conditional on baseline covariates $\mathbf{z}^*$ and accounting for the fact that some students withdraw from the Ph.D.-training.

In each figure we let vary one category of covariate values, keeping the others constant to the ones of the reference group. Remember that this reference group consists of Belgian male Ph.D.-students funded by other projects, with the dominant scientific field 'sciences', who were less than 25 years old when they started their Ph.D.-training between October 1, 1990 and September 30, 1997 at a specific Flemish university. We do not estimate the cumulative incidence function and associated 95% pointwise confidence interval for a combination of covariate values (e.g. $\mathbf{z}^* = (1, 1, 0, \ldots, 0)^T$). In that case, the estimated covariance of the parameter estimates $\hat{\boldsymbol{\beta}}_p$ and the estimated baseline cumulative hazard function $\hat{H}_{0,p}(t)$ is required, in order to construct 95% pointwise confidence intervals.

### Dominant statute classification

Regarding 'dominant statute classification', figure 3.1 shows the estimated probability of attaining a Ph.D.-degree after $t$ years of sponsoring, accounting for the fact that some students withdraw from the training and conditional on baseline covariates. Only the covariate values of 'dominant scientific field' differ from those of the reference group.

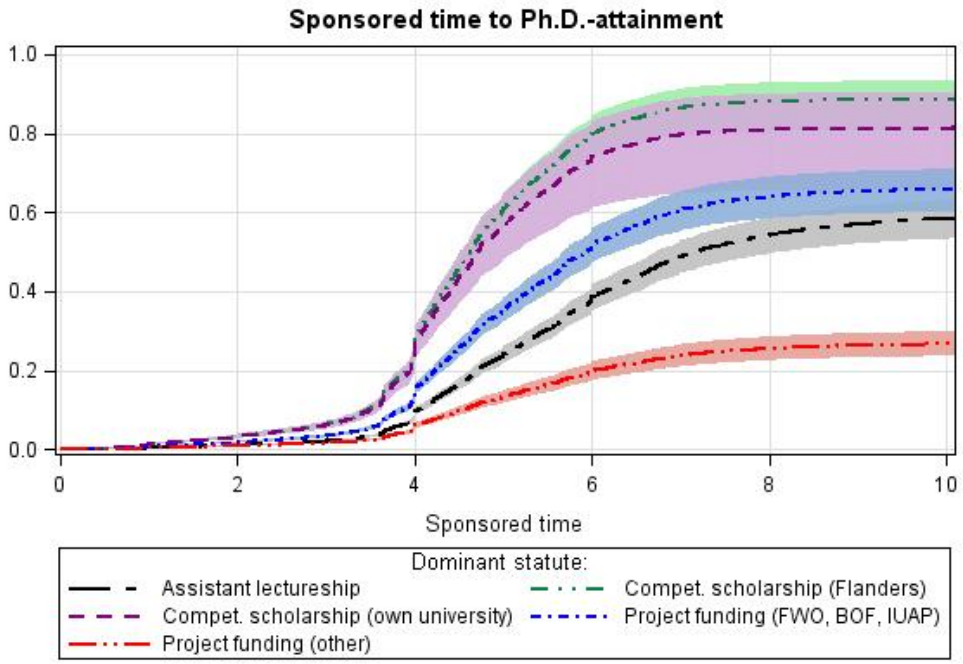


**Figure 3.1:** The combined cumulative incidence function estimates and associated 95% pointwise confidence intervals for covariate 'dominant statute classification' and outcome Ph.D.-attainment - imputation method B.

Although the cumulative incidence curves do not cross, the estimated confidence intervals do.

We perform a hypothesis test with $\boldsymbol{\beta}_p = (\beta_{\text{domstat2},p}, \beta_{\text{domstat3},p}, \beta_{\text{domstat4},p}, \beta_{\text{domstat5},p})$

$$\begin{cases} H_0 : \boldsymbol{\beta}_p = \mathbf{0} \\ H_A : \boldsymbol{\beta}_p \neq \mathbf{0}. \end{cases}$$

This gives a $p$-value $< 0.0001$, indicating the estimated effect of 'dominant statute classification' is significant at the 5% significance level.

The lowest estimated probability of Ph.D.-attainment corresponds to 'Project funding (other)', while the highest probabilities are estimated for 'Competitive scholarship (Flanders)' and 'Competitive scholarship (own university)'. Belgian male students, with the dominant scientific field 'sciences', who were less than 25 years old when they started their Ph.D.-training between October 1, 1990 and September 30, 1997 at a certain Flemish university, will have attained the Ph.D.-degree after 8 years of sponsoring with a probability of 88% (95% CI $[0.81, 0.93]$) respectively 26% (95% CI $[0.23, 0.29]$) funded by 'Competitive scholarship (Flanders)' respectively 'Project funding (other)', accounting for the fact that some students will have withdrawn from the Ph.D.-training. Detailed results for the combined cumulative incidence function estimates and associated 95% pointwise confidence intervals are given in the appendix (table C.1, p. 134).

Note the little jump after 4 years of sponsored time, introducing an increase in the estimated probability from year 4 until 8. Most contracts expire after 4 or 6 years, but a lot of students need 2 years additional sponsored time to attain the Ph.D.-degree. For some students, even this additional sponsored time is not sufficient. After 8 years of sponsoring the estimated probability of Ph.D.-attainment stays approximately constant, except for the assistant lectureship. This could be explained by the fact that assistants lose more time due to other activities beside the Ph.D.-training, so it takes more years to complete the Ph.D.-training. Competitive scholarship (own university) has wide confidence interval estimates, particularly for large numbers of sponsored years ($t = 8$ : 95% CI $[0.65, 0.90]$). This indicates a sharp decline in the number of Ph.D.-students at risk between 4 and 6 years within this dominant statute classification.

**Gender**

Regarding 'gender', figure 3.2 shows the estimated probability of attaining a Ph.D.-degree after $t$ years of sponsoring, accounting for the fact that some students withdraw from the training and conditional on baseline covariates. Only the covariate values of 'gender' differ from those of the reference group.

Obviously, the estimated cumulative incidence curves do not cross, but for large values the estimated confidence intervals seem to overlap. We test the null hypothesis $\beta_{\text{gender},p} = 0$, obtaining a $p$-value equal to 0.0089. This indicates that the estimated effect of 'gender' is

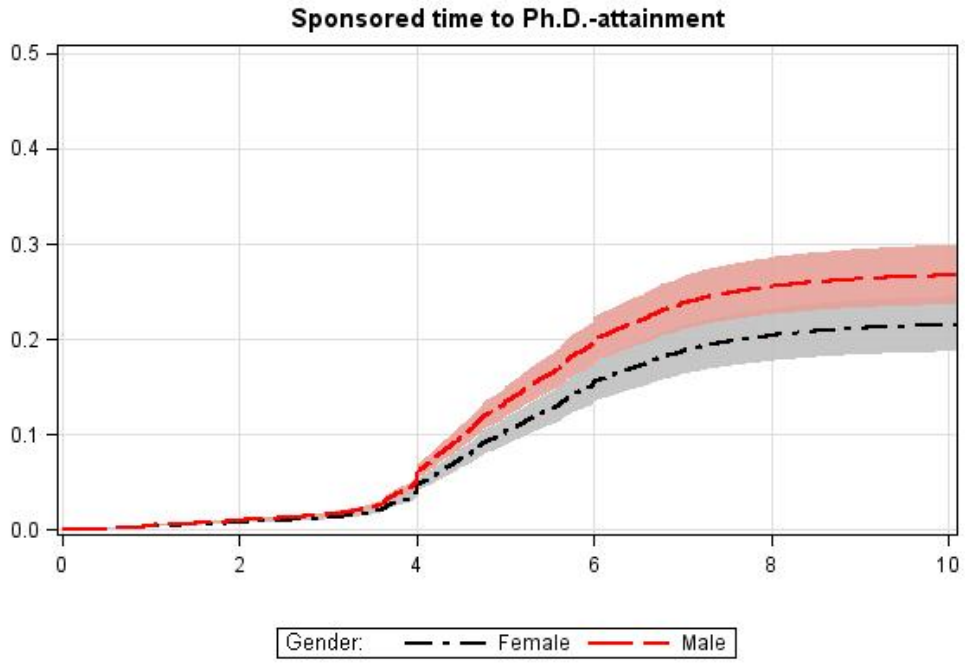


**Figure 3.2:** The combined cumulative incidence function estimates and associated 95% pointwise confidence intervals for covariate 'gender' and outcome Ph.D.-attainment - imputation method B.

significant at the 5% significance level. Considering Belgian Ph.D.-students, funded by other projects with the dominant scientific field 'sciences', who were less than 25 years old when they started their Ph.D.-training between October 1, 1990 and September 30, 1997 at a certain Flemish university, we estimated 26% (95% CI $[0.23, 0.29]$) male compared to 20% (95% CI $[0.18, 0.23]$) female students will have attained the Ph.D.-degree after 8 years of sponsoring, accounting for the fact that some students will have withdrawn from the Ph.D.-training.

### Dominant scientific field

Regarding 'dominant scientific field', figure 3.3 shows the estimated probability of attaining a Ph.D.-degree after $t$ years of sponsoring, accounting for the fact that some students withdraw from the training and conditional on baseline covariates. Only the covariate values of 'dominant scientific field' differ from those of the reference group.

Although the estimated cumulative incidence curves do not cross, the estimated confidence intervals do overlap. We test the null hypothesis

$$\boldsymbol{\beta}_p = (\beta_{\text{domclusmed},p}, \beta_{\text{domclushum},p}, \beta_{\text{domclussoc},p}, \beta_{\text{domclustoe},p}) = \mathbf{0}.$$

This results in a $p$-value $< 0.0001$, indicating the estimated effect of 'dominant scientific field' is significant at the 5% significance level. Considering Belgian male Ph.D.-students, funded

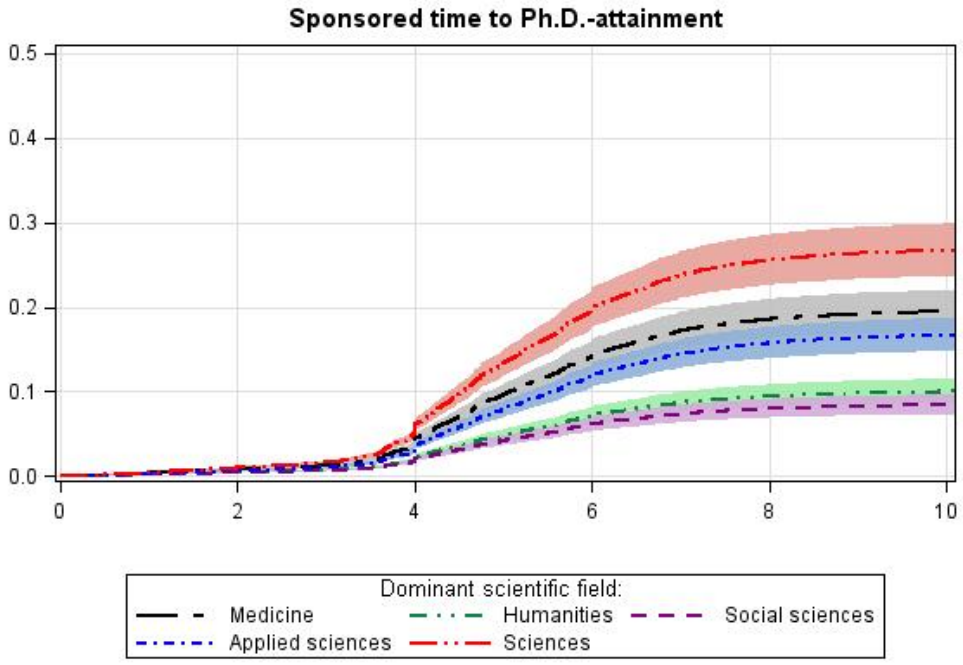


**Figure 3.3:** The combined cumulative incidence function estimates and associated 95% pointwise confidence intervals for covariate 'dominant scientific field' and outcome Ph.D.-attainment - imputation method B.

by other projects, who were less than 25 years old when they started their Ph.D.-training between October 1, 1990 and September 30, 1997 at a certain Flemish university, we estimated 26% (95% CI $[0.23, 0.29]$) students with the dominant scientific field 'sciences' compared to 8% (95% CI $[0.07, 0.09]$) with the dominant scientific field 'social sciences' will have attained the Ph.D.-degree after 8 years of sponsoring, accounting for the fact that some students will have withdrawn from the Ph.D.-training.

### Age (at start)

Regarding 'age at the start of the Ph.D.-training', figure 3.4 shows the estimated probability of attaining a Ph.D.-degree after $t$ years of sponsoring, accounting for the fact that some students withdraw from the training and conditional on baseline covariates. Only the covariate values of 'age at start' differ from those of the reference group.

Only the lowest and highest estimated cumulative incidence curve are clearly separated. So, we test the null hypothesis

$$\boldsymbol{\beta}_p = (\beta_{\text{leeft2},p}, \beta_{\text{leeft3},p}, \beta_{\text{leeft4},p}, \beta_{\text{leeft5},p}) = \mathbf{0}.$$

This results in a $p$-value $< 0.0001$, indicating the estimated effect of 'age at start' is significant at the 5% significance level. Regarding the estimated cumulative incidence curves, we suggest

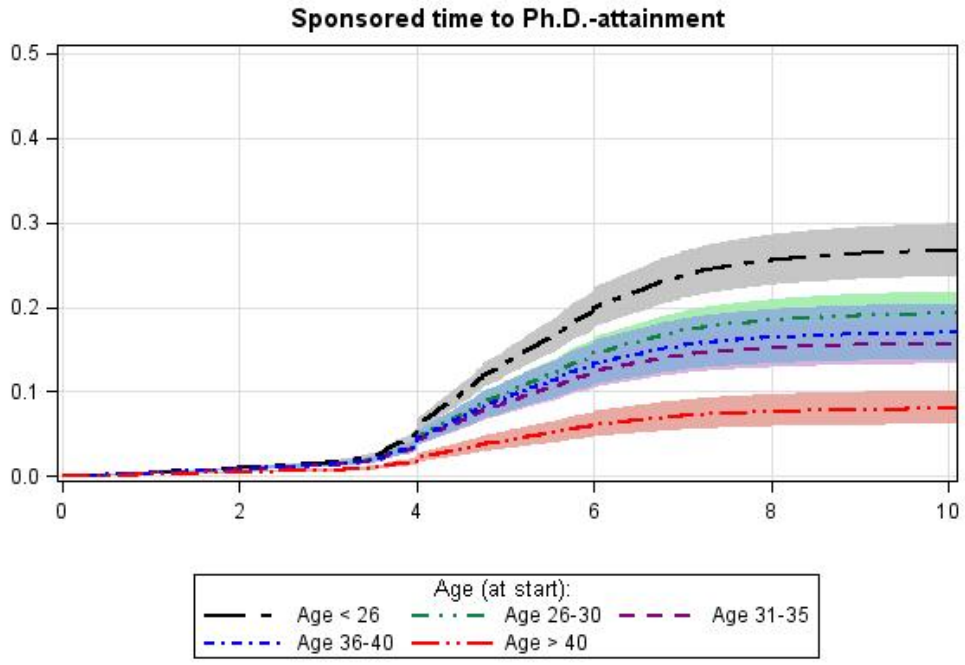


**Figure 3.4:** The combined cumulative incidence function estimates and associated 95% pointwise confidence intervals for covariate 'age (at start)' and outcome Ph.D.-attainment - imputation method B.

a trend that lower age at the start of the Ph.D.-training implies higher probability of attaining a Ph.D.-degree. To obtain statistical evidence for this hypothesis a specific trend test should be performed. Considering Belgian male Ph.D.-students, funded by other projects with the dominant scientific field 'sciences' and who started their Ph.D.-training between October 1, 1990 and September 30, 1997 at a certain Flemish university, we estimated 26% (95% CI $[0.23, 0.29]$) students who were less than 25 years old at the start of their Ph.D.-training compared to 8% (95% CI $[0.06, 0.10]$) who were more than 40 years old at the start, will have attained the Ph.D.-degree after 8 years of sponsoring, accounting for the fact that some students will have withdrawn from the Ph.D.-training.

### Start cohort

Regarding 'start cohort', figure 3.5 shows the estimated probability of attaining a Ph.D.-degree after $t$ years of sponsoring, accounting for the fact that some students withdraw from the training and conditional on baseline covariates. Only the covariate values of 'start cohort' differ from those of the reference group.

We can not distinguish between the estimated cumulative incidence curves until 4 years of sponsored time. From then on, the estimated cumulative incidence curves do not cross, but

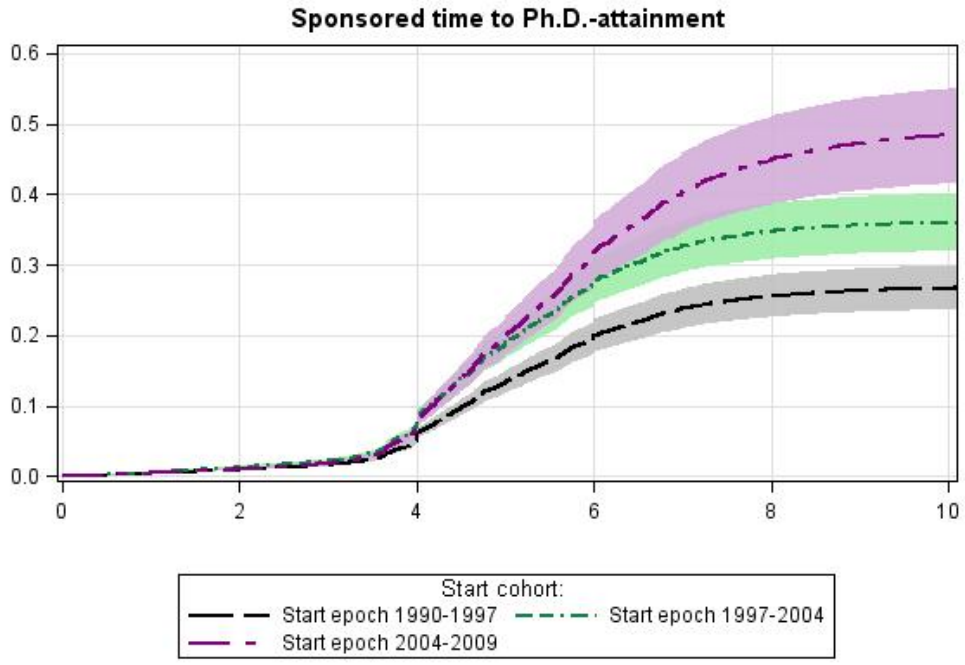


**Figure 3.5:** The combined cumulative incidence function estimates and associated 95% pointwise confidence intervals for covariate 'start cohort' and outcome Ph.D.-attainment - imputation method B.

the associated 95% confidence intervals do overlap. We test the null hypothesis

$$\boldsymbol{\beta}_p = (\beta_{\text{start2},p}, \beta_{\text{start3},p}) = \mathbf{0}.$$

This gives a $p$-value $< 0.0001$, indicating the estimated effect of 'start cohort' is significant at the 5% significance level. Regarding the estimated cumulative incidence curves, we suggest a trend that students starting the Ph.D.-training later have higher probability of attaining a Ph.D.-degree. To obtain statistical evidence for this hypothesis a specific trend test should be performed. Considering Belgian male Ph.D.-students, funded by other projects with the dominant scientific field 'sciences', who were less than 25 years old when they started their Ph.D.-training at a certain Flemish university, we estimated 26% (95% CI $[0.23, 0.29]$) students who started their Ph.D.-training between October 1, 1990 and September 30, 1997 compared to 45% (95% CI $[0.39, 0.51]$) who started between October 1, 2004 and September 30, 2009 will have attained the Ph.D.-degree after 8 years of sponsoring, accounting for the fact that some students will have withdrawn from the Ph.D.-training. The 95% confidence intervals for the last start cohort have the widest range. Actually, the cumulative incidence curve for Ph.D.-students who started between October 1, 2004 and September 30, 2009 is not reliable because of insufficient observation time.

### Nationality

Regarding 'nationality', figure 3.6 shows the estimated probability of attaining a Ph.D.-degree after $t$ years of sponsoring, accounting for the fact that some students withdraw from the training and conditional on baseline covariates. Only the covariate values of 'nationality' differ from those of the reference group.

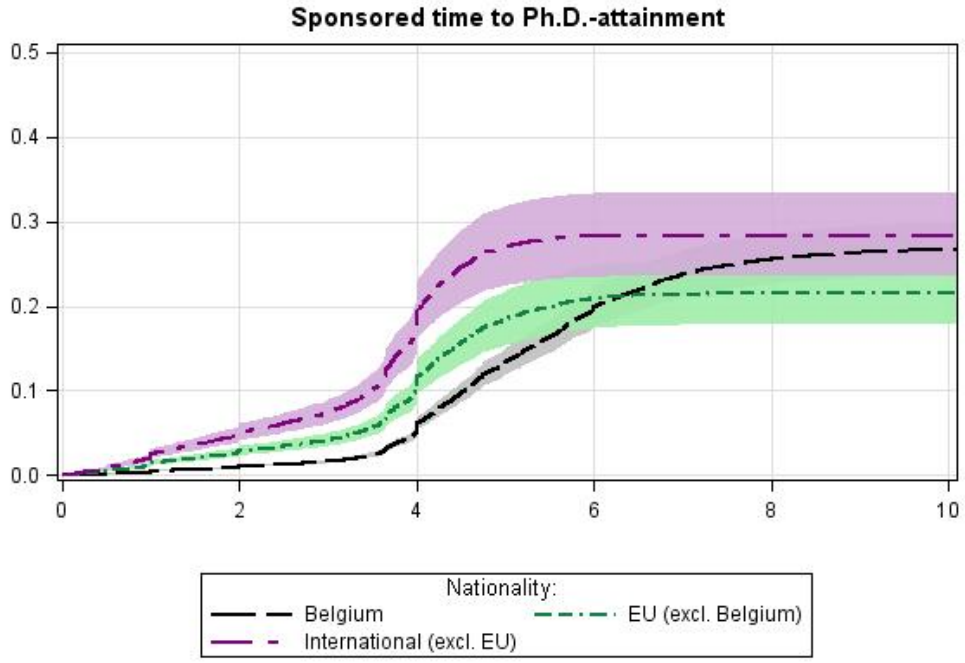


**Figure 3.6:** The combined cumulative incidence function estimates and associated 95% pointwise confidence intervals for covariate 'nationality' and outcome Ph.D.-attainment - imputation method B.

The estimated cumulative incidence curves clearly cross. After 6 years of sponsoring, the estimated cumulative incidence curves for international (excl. EU) and European EU (excl. Belgium) Ph.D.-students stay approximately constant, in contrast to the estimated cumulative incidence curve for Belgian Ph.D.-students. Foreign students have a higher probability to end their Ph.D.-training before the contract expires after 4 or 6 years. The widest 95% pointwise confidence intervals are estimated for the international (excl. EU) and European EU (excl. Belgium) Ph.D.-students, due to less observations. We test the null hypothesis

$$\boldsymbol{\beta}_p = (\beta_{\text{natEurEU},p}, \beta_{\text{natAnd},p}) = \mathbf{0}.$$

This gives a $p$-value $< 0.0001$, indicating the estimated effect of 'nationality' is significant at the 5% significance level. Considering male Ph.D.-students, funded by other projects with the dominant scientific field 'sciences', who were less than 25 years old when they started

their Ph.D.-training between October 1, 1990 and September 30, 1997 at a certain Flemish university, we estimated 26% (95% CI $[0.23, 0.29]$) Belgian students compared to 21% (95% CI $[0.18, 0.25]$) European EU (excl. Belgium) will have attained the Ph.D.-degree after 8 years of sponsoring, accounting for the fact that some students will have withdrawn from the Ph.D.-training. Belgian Ph.D.-students start out the slowest, but end up with a higher percentage of Ph.D.-attainments than other EU-students. Remark that the mixture of PH functions (cause-specific PH model for Ph.D.-attainment and withdrawal) makes it possible for the cumulative incidence curves to cross, as no stochastic ordering is imposed to the cumulative incidence curves.

### 3.6.3 Outcome of Interest: Withdrawal

Similarly the cumulative incidence functions and associated 95% pointwise confidence intervals can be estimated for the probability of withdrawal by year $t$, conditional on baseline covariates $\mathbf{z}^*$ and accounting for the fact that some students attain the Ph.D.-degree.

To calculate $p$-values for the parameter estimates and hypothesis tests, combined point estimates and variances are used. We can calculate these combined point estimates, in the same way as we combined the estimated cumulative incidence functions in section 3.5 (p. 47). The results are given in the appendix for both imputation methods A (p. 121) and B (p. 125).

#### Dominant statute classification

Regarding 'dominant statute classification', figure 3.7 shows the estimated probability of withdrawing from the Ph.D.-training after $t$ years of sponsoring, accounting for the fact that some students attain the Ph.D.-degree and conditional on baseline covariates. Only the covariate values of 'dominant statute classification' differ from those of the reference group.

The estimated cumulative incidence curves and associated 95% pointwise confidence intervals are clearly separated, except for 'Assistant lectureship' and 'Project funding (FWO, BOF or IUAP)'. Again, we test the null hypothesis

$$\boldsymbol{\beta}_w = (\beta_{\text{domstat2},w}, \beta_{\text{domstat3},w}, \beta_{\text{domstat4},w}, \beta_{\text{domstat5},w}) = \mathbf{0}.$$

This gives a $p$-value $< 0.0001$, indicating the estimated effect of 'dominant statute classification' is significant at the 5% significance level. Belgian male students, with the dominant scientific field 'sciences', who were less than 25 years old when they started their Ph.D.-training between October 1, 1990 and September 30, 1997 at a certain Flemish university, will have withdrawn from the Ph.D.-training after 2 years of sponsoring with a probability of 41% (95% CI $[0.37, 0.44]$) if funded by 'Project funding (other)' compared to 6% (95% CI $[0.05, 0.07]$) if funded by 'Competitive scholarship (Flanders)', accounting for the fact that

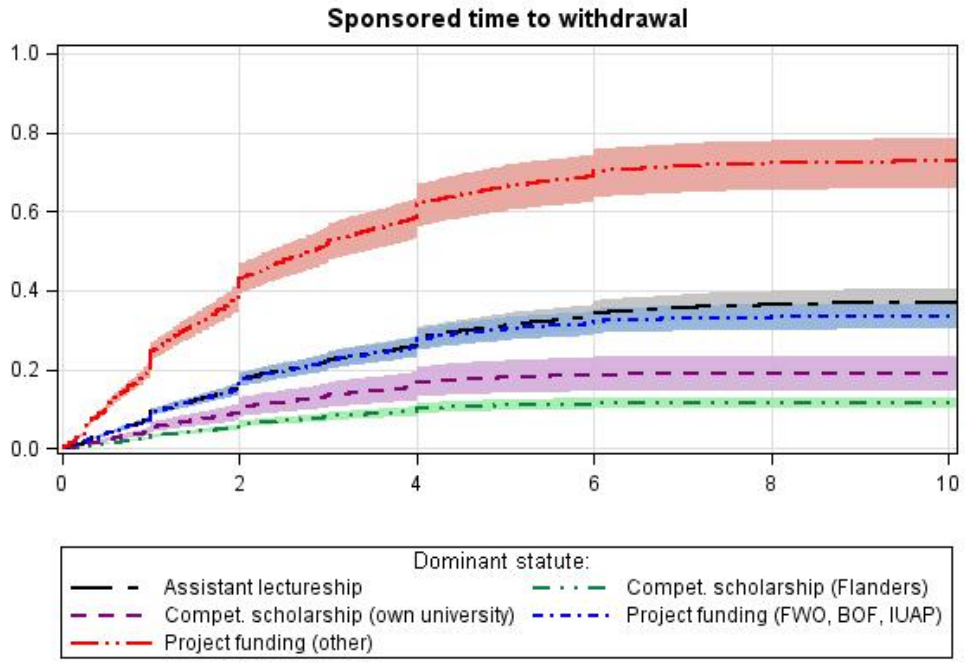


**Figure 3.7:** The combined cumulative incidence function estimates and associated 95% pointwise confidence intervals for covariate 'dominant statute classification' and outcome withdrawal - imputation method B.

some students will have attained the Ph.D.-degree. Detailed results for the combined cumulative incidence function estimates and associated 95% pointwise confidence intervals are given in the appendix, table C.2 (p. 138).

From the standpoint of the government, the lowest curves are regarded as best, preferably rising sharply. This because sponsored time that ends in withdrawal is not a good investment and should be ended as soon as possible. After 8 years of sponsored time the estimated probability stays approximately constant. Note the jumps after 1, 2, 4 and 6 years of sponsored time after starting the Ph.D.-training. This is explained by the fact that many sponsored contracts expire after such time intervals.

The figures for the combined cumulative incidence function estimates and associated 95% pointwise confidence interval for outcome withdrawal and covariates 'gender', 'dominant scientific field', 'age (at start)', 'start cluster' and 'nationality' are given in the appendix (p. 133).

## 3.7 An Experimental Comparison of Imputed and Non-imputed Data Analysis

The type of event is missing for all Ph.D.-students starting a gap between October 1, 2005 and September 30, 2009, that is censored by the end of the study. These Ph.D.-students, having a censored gap, are either ongoers or withdrawers. There are several approaches to handle these missing data.

### 3.7.1 Compare Imputed and Non-imputed Data Procedures

Remember that in this thesis we built an imputation model based on all gaps starting between October 1, 1990 and September 30, 2005. In that way, we estimated the probability of a censored gap to be a withdrawal gap conditional on baseline covariate values and the current gap length (section 2.5, p. 22). This resulted in imputed data sets.

In earlier research [19], this database has already been analyzed by using an alternative approach for the censored gaps. This more naive method focuses on the determination of an optimal cut-point $X$, that mimics the actual distinction between ongoers and withdrawers, for all Ph.D.-students having a censored gap [2]. All Ph.D.-students having a censored gap lasting less than $X$ years are categorized as ongoers, all Ph.D.-students having a censored gap lasting more than $X$ years are classified as withdrawers. Clearly in reality, the distinction between these two groups is not only explained by their current gap length, but this information is not taken into account for this analysis.

We determine the cut-point value $X$ based on the analysis of all gaps starting prior to October 1, 2005 as any withdrawal gap occurring in this set shows up as a gap exceeding 4 years and hence is clearly identified. We explicitly estimate the type I and type II error that result from a decision rule whereby someone with a gaptime censored at $c > x$ years is classified as withdrawal, for varying cut-point values $x$. Finally, this cut-point value was determined to be $X = 2$ years [19]. Thus, someone who has not been sponsored for a period lasting more than 4 years is classified as withdrawal, as well as any Ph.D.-student who started a non-sponsored gap after September 30, 2005 lasting more than 2 years and censored by the end of the study. The obtained data set is called 'non-imputed'.

### 3.7.2 Compare Imputed and Non-imputed Data Results

First, we compare the percentage of withdrawal gaps in the subset of all gaps started between October 1, 2005 and September 30, 2009 (5,704 observations) by the imputed data sets (table 2.8, p. 35) and non-imputed data set. The imputed subsets have an average proportion of 54% withdrawal gaps, while the non-imputed subset has a proportion of 27% withdrawal gaps. The proportion of withdrawal gaps in the non-imputed subset is also much lower than

the probability of a withdrawal gap for a random gap estimated to be 57% by equation (2.2) on p. 16.

For both the imputed and non-imputed data sets a cause-specific hazards model for the outcome of interest 'Ph.D.-attainment' and 'withdrawal' is built. Of course, for the outcome of interest 'Ph.D.-attainment', there is no difference between both methods, because no distinction is made between withdrawers and ongoers for the analysis. We created two types of imputed data sets, one by imputation method A and one by imputation method B, but as no big differences are detected between both imputation methods, we will focus on imputation method B for this comparison. So we compare the results from the cause-specific hazards model attained from the imputed (method B) and non-imputed data sets for the outcome 'withdrawal'.

On both data sets a backward selection is performed at the 5% significance level, starting from all main effects and interaction effects of interest: the baseline covariates listed in section 1.3 (p. 3) and interactions gender $\times$ dominant statute, gender $\times$ dominant scientific field, nationality $\times$ dominant statute, nationality $\times$ dominant scientific field and gender $\times$ nationality. As we mentioned before, when considering the interactions, two categories are formed for nationality: Belgian and non-Belgian Ph.D.-students, because of too few observations for some nationality categories. For all other covariates we use the same categories as listed previously (section 1.3, p. 3). It turns out that the same subset of prognostic factors is included in both cause-specific hazards models with outcome of interest 'withdrawal'. Combined parameter estimates for the cause-specific hazards model with outcome of interest 'withdrawal' on the multiple imputed data sets (method B), are given in the appendix (p. 125). Parameter estimates for the cause-specific hazards model with outcome of interest 'withdrawal' on the non-imputed data set are also given in the appendix (p. 129).

No big differences between both methods are noticed for the parameter estimates in the cause-specific hazards model, except for the dummy variable 'start3', indicating Ph.D.-students starting their training between October 1, 2004 and September 30, 2009. The parameter estimate for start3 is $-0.75$ (95% CI $[-0.81, -0.69]$) respectively $-1.54$ (95% CI $[-1.61, -1.46]$) for the analysis on the imputed vs. non-imputed data set. Note that the 95% confidence intervals do not overlap, indicating that at least one of both point estimates is seriously biased. These parameter estimates can be interpreted as follows. Based on the imputed data set, the cause-specific hazards ratio of withdrawing for a Ph.D.-student starting his training during the last cohort 2004-2009 compared to a Ph.D.-student starting his training during the first cohort 1990-1997, is $\exp(-0.75) = 0.47$ (95% CI $[0.44, 0.50]$), given constant values of all other covariates and accounting for the competing outcome Ph.D.-attainment. The latest starters have lower probability of withdrawing from the Ph.D.-training. Based on the non-imputed data set, the cause-specific hazards ratio of withdrawing for a Ph.D.-student starting his train-

ing during the last cohort 2004-2009 compared to a Ph.D.-student starting his training during the first cohort 1990-1997, is $\exp(-1.54) = 0.21$ (95% CI $[0.20, 0.23]$), given constant values of all other covariates and accounting for the competing outcome Ph.D.-attainment. So, also in this case the latest starters have lower probability of withdrawing from the Ph.D.-training, although, the effect of 'start3' on withdrawing is estimated to be smaller on the imputed than on the non-imputed data set. This because less Ph.D.-students are categorized as withdrawer in the latest start cohort of the non-imputed data set compared to the imputed data set.

The censored gaps are handled differently in the imputed and non-imputed data sets. We compared the results from the cause-specific hazard function for 'withdrawal' attained from the imputed (method B) and non-imputed data sets and detected a notable difference in the parameter estimate for 'start3'. The estimated effect, of starting the Ph.D.-training during the last start cohort (2004-2009) compared to the first start cohort (1990-1997), on withdrawing is overestimated on the non-imputed data set, because of handling the missing data in a naive way.

## 3.8 Discussion

We wish to analyze whether and how student characteristics such as gender, nationality, scientific field etc. influence sponsored time to Ph.D.-attainment, during the observation period from October 1, 1990 until September 30, 2009.

In Chapter 2 we generated multiple imputed data sets so we can apply standard statistical complete-data methods and combine the obtained results. Since Ph.D.-attainment and withdrawal are competing outcomes, we are working in a competing risks setting.

There has been a lively debate in the literature about the best way to attack the problem of competing risks [7]. We used the approach based on cause-specific PH functions, but this function does not have a direct interpretation in terms of survival probabilities for the particular type of event. Additionally, many authors have noted that the effect of a covariate on the cause-specific hazard function of a particular type of event may be very different from the effect of the covariate on the corresponding cumulative incidence function [6]. An alternative approach for competing risks is to use a semiparametric proportional hazards model for the subdistribution of a competing risk, presented in [6].

A disadvantage of our approach is that, although we are only interested in estimating the cumulative incidence function for the outcome Ph.D.-attainment, the cause-specific PH function for both type of events has to be estimated. It would be interesting to investigate whether such a strong model assumption can be relaxed [5].

Once we have combined the estimated cumulative incidence results, corresponding pointwise

confidence intervals are obtained through the use of a transformation. We have, however, only considered one possible transformation, and there may exist better ones. The width of these confidence intervals adds valuable information to the plots of the cumulative incidence curves. Although, our impression is that even if the 95% confidence intervals for the estimated cumulative incidence curves are overlapping, the estimated effect of the category of covariates can be significant at the 5% significance level.

The estimated cumulative incidence curves for Ph.D.-attainment and different covariate values do not differ much until about 4 years of sponsoring, because very few Ph.D.-degrees are attained before 4 years of sponsoring. The largest differences between the estimated cumulative incidence curves for Ph.D.-attainment are observed for 'dominant statute classification'. Nevertheless, all main effects of interest (gender, nationality, dominant statute classification, dominant scientific field, age and start year) are tested to be significant at the 5% significance level. Two peculiarities are detected from the estimated cumulative incidence curves, when sponsored time is large: Large standard errors for the cumulative incidence function for 'Competitive scholarship (own university)' are estimated and, the lower bound of the 95% confidence interval for the oldest age group decreases for increasing sponsored time. Both peculiarities can be explained by too few Ph.D.-students who are at risk for 'Competitive scholarship (own university)' respectively 'Age $>$ 40 years (at start)' when sponsored time is large.

Comparing the imputed and non-imputed data results, we noticed a substantial difference in the estimated cause-specific hazards ratio of withdrawing for people starting in the last start cohort (2004-2009) compared to people starting in the first start cohort (1990-1997). Moreover, the proportion of withdrawal gaps in the subset of all gaps started between October 1, 2005 and September 30, 2009, deviates substantially from the estimated proportion in the fully observed subset of all gaps started prior to October 1, 2005. Probably the non-imputed data set leads to biased statistical estimates, because of handling the missing data in a naive way.

# Chapter 4

# Cox Proportional Hazards Cure Model

## 4.1 Introduction

In survival analysis, it is usually assumed that if complete follow-up were possible for all individuals, each would eventually experience the event of interest, in our case Ph.D.-attainment. Sometimes, however, the data come from a population where a substantial proportion of the individuals do not experience the event by the end of any observation period. In that case, some of these survivors are actually 'cured' in the sense that, even after an extended follow-up, no further events are observed. For our analysis, we called those cured people the withdrawers.

We aim to answer the question 'How much sponsored time is needed for Ph.D.-attainment to occur?' We assume that even if we gave the withdrawers extra sponsored time, they would never attain their Ph.D.-degree, they are cured for the Ph.D.-attainment event. A heuristic justification for this assumption is that in the past, some Ph.D.-students may have chosen the Ph.D.-training, merely for the purpose of having an interesting job with no intention of striving for a Ph.D.-degree. The use of standard survival analysis for this data may be inappropriate since not all the individuals are susceptible.

In a cure model, the population is considered from the outset as a mixture of susceptible and non-susceptible (cured) individuals. The objective is usually to study the proportion of cured individuals and for the non-cured subpopulation the survival distribution, and the effect of any covariates on both. We are interested in

- whether the event Ph.D.-attainment can occur, which we call *prevalence* and
- when the Ph.D.-attainment will occur, given that it can occur, which we call *latency*

How the covariates influence the proportion of cured individuals would be viewed as most important, but there is also interest in how they relate to the time to occurrence.

It is clear that the cure model should not be used indiscriminately. There must be good empirical evidence of a non-susceptible population. If it is believed that a proportion of individuals will not experience the event of interest, then it may be appropriate to fit models that explicitly allow for the cure fraction to be estimated and directly modeled.

We start by constructing the proportional hazards cure model. Next, we describe the expectation-maximization (EM) method to estimate the model parameters. Finally, we compare the competing risks model approach we used in this thesis to the cure model. We discuss this theory based on article [16]. Remark that this cure model analysis is not practically applied to our data.

## 4.2 The Proportional Hazards Cure Model

We start by introducing some definitions and notation to construct a PH cure model.

Let $Y$ be the indicator function

$$Y = \begin{cases} 0 & \text{not aiming to attain the Ph.D.-degree} \\ 1 & \text{eventually experience the event Ph.D.-attainment,} \end{cases}$$

with $p = P(Y = 1)$. So, Ph.D.-students have probability $1 - p$ to have no intention to attain the Ph.D.-degree, i.e. withdrawing, and then are labeled as $Y = 0$. The value of $Y$ is unobserved as long as the Ph.D.-student has not experienced any event (Ph.D.-attainment or withdrawal). Let $T$ denote the sponsored time to Ph.D.-attainment, defined only when $Y = 1$, with density $f(t|Y = 1)$ and survival function $S(t|Y = 1)$. For a censored individual, $Y$ is not observed.

The marginal survival function of $T$ is

$$\begin{aligned} &S(t) = (1 - p) + p \cdot S(t|Y = 1) && \text{for } t < \infty \\ &S(t) \to (1 - p) && \text{for } t \to \infty, \end{aligned}$$

and expresses the probability of not having attained a Ph.D.-degree after $t$ years of sponsoring. So, $1 - S(t)$ is the percentage of Ph.D.-students who have attained a Ph.D.-degree by year $t$. As the sponsored time $t$ goes to infinity, $S(t)$ expresses the proportion of Ph.D.-students who are cured for the event Ph.D.-attainment.

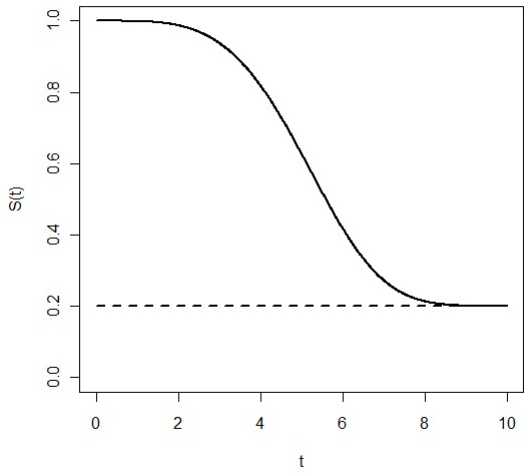


**Figure 4.1:** Example of the marginal survival function $S(t)$ with $p = 0.2$.

We assume an independent, non-informative, random censoring model and that censoring is statistically independent of $Y$. If the assumption of independent censoring is not fulfilled, adjustments can be applied to this method [12].

We aim to estimate the prevalence and latency, based on the observed and unobserved data. A logistic regression model can be used for the prevalence

$$p(\mathbf{x}) = P(Y = 1; \mathbf{x}) = \frac{\exp(\mathbf{b}^T\mathbf{x})}{1 + \exp(\mathbf{b}^T\mathbf{x})},$$

where the covariate vector $\mathbf{x}$ includes the intercept. The logit of $p(\mathbf{x})$

$$\text{logit}(p(\mathbf{x})) = \ln\left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})}\right) = \mathbf{b}^T\mathbf{x},$$

is linear in its model parameters $\mathbf{b}$ and has a range $]-\infty, \infty[$, depending on the values of $\mathbf{x}$. A logit transformation is used to yield an estimated probability $\hat{p}(\mathbf{x})$ that takes values between 0 and 1. The latency $S(t|Y = 1)$ can be modeled by using a Cox PH model with hazard function

$$h(t|Y = 1; \mathbf{z}) = h_0(t|Y = 1)\exp(\boldsymbol{\beta}^T\mathbf{z}),$$

where $\mathbf{z}$ is a vector of covariates other than the intercept and $h_0(t|Y = 1)$ is the conditional baseline hazard function. The hazard function $h(t|Y = 1; \mathbf{z})$ expresses the conditional probability that a student with covariate values $\mathbf{z}$ attains a Ph.D.-degree after $t$ years of sponsoring, knowing that he has not attained a Ph.D.-degree until just before $t$ years of sponsoring and that he is aiming to attain the Ph.D.-degree.

Through $\mathbf{b}$ and $\boldsymbol{\beta}$, the model is able to separate the covariates' effects on the prevalence and the latency and, in that sense, provide a flexible class of models when there is *a priori* belief in a non-susceptible group. The conditional cumulative hazard function is

$$H(t|Y=1;\mathbf{z}) = H_0(t|Y=1)\exp(\boldsymbol{\beta}^T\mathbf{z}),$$

where

$$H_0(t|Y=1) = \int_0^t h_0(u|Y=1)\ \mathrm{d}u.$$

The conditional survival function or latency is

$$S(t|Y=1;\mathbf{z}) = S_0(t|Y=1)^{\exp(\boldsymbol{\beta}^T\mathbf{z})},$$

where $S_0(t|Y=1)$ is the conditional baseline survival function. The conditional survival function $S(t|Y=1;\mathbf{z})$ expresses the probability that a student with covariate values $\mathbf{z}$ has not attained a Ph.D.-degree after $t$ years of sponsoring, knowing that he intends to attain the Ph.D.-degree.

Remark that a mixture of PH functions is no longer proportional, and in fact, for a binary covariate, a PH cure model can have marginal survival curves that cross. However, the standard PH model is a special case of a PH cure model in which $p(\mathbf{x}) = 1$ for all $\mathbf{x}$.

## 4.3 Estimation of the Model Parameters

### 4.3.1 Maximum Likelihood Estimation

We are working in a survival setting, on the database of all students starting a Ph.D.-training between October 1, 1990 and September 30, 2009. To construct the observed and complete data full likelihood function, we first introduce some standard notation.

We distinguish between the actual time to Ph.D.-attainment ($X_i$), censoring time ($C_i$) and denote $T_i = \min(X_i, C_i)$ as the time on the study for the $i$th Ph.D.-student. The observed data, based on a sample of size $n$, consists of the triple $(T_i, \delta_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, which are assumed to be independent and identically distributed random vectors, where

- $T_i$ is the observed sponsored time for the $i$th Ph.D.-student, i.e. total event or censoring time,
- $\delta_i$ is the event indicator for the $i$th Ph.D.-student

  $\delta_i = 0$ if the sponsored time is right-censored and hence the $i$th Ph.D.-student is ongoing at time $T_i$ or has withdrawn by time $T_i$

  $\delta_i = 1$ if $T_i$ is uncensored, more specifically the $i$th Ph.D.-student has attained a Ph.D.-degree

- $\mathbf{Z}_i$ is a vector of baseline covariates for the $i$th Ph.D.-student which may affect the conditional survival distribution of $T_i$.

Let $t_i$ be the value taken by the random variable $T_i$ and $\mathbf{z}_i$ be the value taken by the random variable $\mathbf{Z}_i$. For convenience, we let the covariates of the logistic model $\mathbf{x}_i = (1, \mathbf{z}_i^T)^T$, although the covariates in $\mathbf{x}_i$ and $\mathbf{z}_i$ do not have to be identical. Denote the $k$ distinct event times by $t_{(1)} < \ldots < t_{(k)}$. It follows that, if $\delta_i = 1$, $y_i = 1$ and, if $\delta_i = 0$, $y_i$ is unobserved, where $y_i$ is the value taken by the random variable $Y_i$.

The likelihood contribution of individual $i$ is

$$\begin{array}{ll} p_i f(t_i|Y=1;\mathbf{z}_i) & \text{for } \delta_i = 1 \\ (1-p_i) + p_i S(t_i|Y=1;\mathbf{z}_i) & \text{for } \delta_i = 0, \end{array}$$

where $p_i = P(Y_i = 1; \mathbf{x}_i)$. We know that $f(t_i|Y=1;\mathbf{z}_i) = h(t_i|Y=1;\mathbf{z}_i)S(t_i|Y=1;\mathbf{z}_i)$ and $S(t_i|Y=1;\mathbf{z}_i) = \exp(-H(t_i|Y=1;\mathbf{z}_i))$. For the PH cure model, the observed full likelihood is then

$$\begin{aligned} L(\mathbf{b}, \boldsymbol{\beta}, H_0) &= \prod_{i=1}^{n} \left[p_i h_0(t_i|Y=1)\exp(\boldsymbol{\beta}^T\mathbf{z}_i)\exp(-H_0(t_i|Y=1)\exp(\boldsymbol{\beta}^T\mathbf{z}_i))\right]^{\delta_i} \\ &\quad \times \left[(1-p_i) + p_i\exp(-H_0(t_i|Y=1)\exp(\boldsymbol{\beta}^T\mathbf{z}_i))\right]^{1-\delta_i}. \end{aligned}$$

We want to obtain the estimates $\hat{\mathbf{b}}$ and $\hat{\boldsymbol{\beta}}$ that maximize $L(\mathbf{b}, \boldsymbol{\beta}, H_0)$. In the ordinary Cox PH model, the standard analysis is to maximize the partial likelihood function. Since for each Ph.D.-student the value $y_i$ is not known as long as he/she did not attain the Ph.D.-degree or withdraw from the Ph.D.-training, the partial likelihood contains unobserved data. Therefore, the EM algorithm is proposed for the Cox PH cure model.

### 4.3.2 The EM Algorithm

Denote the complete data by $(T_i, \delta_i, \mathbf{Z}_i, Y_i)$, $i = 1, \ldots, n$, which includes the observed data and the unobserved $Y_i$'s. The likelihood contribution of individual $i$ is

$$\begin{array}{ll} (1-p_i) & \text{for } Y_i = 0, \text{ so } \delta_i = 0 \\ p_i S(t_i|Y=1;\mathbf{z}_i) & \text{for } Y_i = 1 \text{ and } \delta_i = 0 \\ p_i f(t_i|Y=1;\mathbf{z}_i) & \text{for } Y_i = 1 \text{ and } \delta_i = 1 \end{array}$$

So the complete data full likelihood is

$$\begin{aligned} L_C(\mathbf{b}, \boldsymbol{\beta}, H_0; y) &= \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \prod_{i=1}^{n} \left[h_0(t_i|Y=1)\exp(\boldsymbol{\beta}^T\mathbf{z}_i)\right]^{\delta_i y_i} \\ &\quad \times \exp\left(-y_i H_0(t_i|Y=1)\exp(\boldsymbol{\beta}^T\mathbf{z}_i)\right) \\ &= L_1(\mathbf{b}; y) L_2(\boldsymbol{\beta}, H_0; y), \end{aligned} \tag{4.1}$$

where $y$ is the vector of $y_i$ values. The likelihood factors into a logistic ($L_1$) and a PH component ($L_2$).

The expectation-maximization (EM) algorithm is a technique for maximum likelihood estimation (MLE) in parametric models for incomplete data. This algorithm in essence converts a difficult incomplete-data problem into a sequence of pseudo-complete data problems [18]. The E step of the EM algorithm imputes values for the unobserved portion of the data, using the observed data and estimates of the model parameters from the previous M step. The imputed values are used to obtain a pseudo-complete data log-likelihood function. The model parameters are reestimated in each M step by maximizing the pseudo-complete data log-likelihood, treating the imputed data as though they were observed. The details are given below.

**The E step**

The E step takes the expectation of the log-likelihood $\ln(L_C(\mathbf{b}, \boldsymbol{\beta}, H_0; y)) = l_C(\mathbf{b}, \boldsymbol{\beta}, H_0; y)$ with respect to the distribution of the unobserved $y_i$'s, given the current parameter estimates and the observed data $O$, where $O = \{\text{observed } Y_i\text{'s}, (T_i, \delta_i, \mathbf{Z}_i); i = 1, \ldots, n\}$. Note that, for censored cases ($\delta_i = 0$), the $y_i$'s are linear terms in the complete data log-likelihood so that we only need to compute

$$\begin{aligned}
\pi_i^{(m)} &= \mathrm{E}[Y_i|\theta^{(m)}, O] \\
&= P\left(Y_i = 1|X_i > t_i, \delta_i = 0, \mathbf{z}_i; \theta^{(m)}\right) \\
&= \frac{P\left(Y_i = 1, \delta_i = 0|X_i > t_i, \mathbf{z}_i; \theta^{(m)}\right)}{P\left(\delta_i = 1|X_i > t_i, \mathbf{z}_i; \theta^{(m)}\right)} \\
&= \left.\frac{P(Y_i = 1; \mathbf{b}) S_0(t_i|Y = 1)^{\exp(\boldsymbol{\beta}^T \mathbf{z}_i)}}{\left[1 - P(Y_i = 1; \mathbf{b}) + P(Y_i = 1; \mathbf{b}) S_0(t_i|Y = 1)^{\exp(\boldsymbol{\beta}^T \mathbf{z}_i)}\right]}\right|_{\theta = \theta^{(m)}}
\end{aligned}$$

for censored cases, where $\theta = (\mathbf{b}, \boldsymbol{\beta}, H_0)$, $\theta^{(m)}$ denotes the current parameter estimates at the $m$th iteration, and $S_0(t_i|Y = 1) = \exp(-H_0(t_i|Y = 1))$. For uncensored $i$, $\delta_i = y_i = 1$ and $\mathrm{E}[Y_i|\theta^{(m)}, O] = y_i = 1$. Thus, the E step replaces the $y_i$'s in (4.1) with

$$w_i^{(m)} = \begin{cases} 1 & \text{if individual } i \text{ is uncensored } (\delta_i = 1) \\ \pi_i^{(m)} & \text{if individual } i \text{ is censored } (\delta_i = 0). \end{cases}$$

Denote the expected log-likelihood by

$$\tilde{l}_C(\mathbf{b}, \boldsymbol{\beta}, H_0; w^{(m)}) = \tilde{l}_1(\mathbf{b}; w^{(m)}) + \tilde{l}_2(\boldsymbol{\beta}, H_0; w^{(m)}), \tag{4.2}$$

where $w^{(m)} = \left\{w_i^{(m)}; i = 1, \ldots, n\right\}$. Note that, for $T_i$ censored, so $\delta_i = 0$, the weight $w_i^{(m)}$ represents 'the probability' for individual $i$ to belong to the susceptible group ($y_i = 1$), conditional on the observed data and current parameter estimates.

### The M step

The M step of the algorithm involves the maximization of the expected log-likelihood $\tilde{l}_C$ (4.2) with respect to $\mathbf{b}$, $\boldsymbol{\beta}$ and the function $H_0$, given $w^{(m)}$. This expected log-likelihood is the sum of two functions, of which the first one can be easily maximized. To deal with the nuisance function $H_0(t|Y = 1)$ in the second term, an additional maximization step in the M step is performed, using profile likelihood techniques. The profile likelihood technique estimates parameters by maximizing the likelihood function for given (fixed) estimated values of the other parameters, so we obtain conditional maximum likelihood estimates.

Two methods from the Cox PH model can be extended: the Breslow estimator for $H_0(t|Y = 1)$, as we did in (2.9) on p. 25, and the Kalbfleisch and Prentice estimator for $S_0(t|Y = 1)$, also known as the product-limit estimator. The latter is based on a nonparametric full likelihood construction that produces the generalized MLE for $S_0(t|Y = 1)$. This method is preferred here because the Kalbfleisch and Prentice estimator is able to send $S_0(t_{(k)}|Y = 1)$ to zero, where $t_{(k)}$ is the last event time. This zero-tail constraint is important in order to obtain a good estimate for $\mathbf{b}$ and $\boldsymbol{\beta}$ [16].

We define the following sets of Ph.D.-students:

- $D_i$ is the set of individuals experiencing an event at time $t_{(i)}$,
- $C_i$ is the set of individuals censored in $[t_{(i)}, t_{(i+1)}[$, $i = 0, 1, \ldots, k$ and
- $R_i$ is the set of individuals at risk at time $t_{(i-1)}$.

The complete-data likelihood is [16]

$$L_2(\boldsymbol{\beta}, \boldsymbol{\alpha}; y) = \prod_{i=0}^{k} \left[ \prod_{\ell \in D_i} \left\{ \lambda(t_{(i)}; \mathbf{z}_\ell) S_0(t_{(i)}^-|Y = 1)^{\exp(\boldsymbol{\beta}^T \mathbf{z}_\ell)} \right\} \times \prod_{\ell \in C_i} S_0(t_{(i)}|Y = 1)^{y_\ell \exp(\boldsymbol{\beta}^T \mathbf{z}_\ell)} \right],$$

where a discrete PH model is assumed and $S_0(t|Y = 1)$ has the product-limit form

$$S_0(t|Y = 1) = \prod_{j: t_{(j)} \leq t} \alpha_j,$$

with $S_0\left(t_{(i)}^-|Y = 1\right) = S_0\left(t_{(i-1)}|Y = 1\right)$.
The $\alpha$'s are nonnegative parameters at each of the $k$ distinct event times with $\alpha_0 = 1$ and

$$\lambda(t_{(i)}; \mathbf{z}) = 1 - \frac{S(t_{(i)})}{S(t_{(i-1)})} = 1 - \alpha_i^{\exp(\boldsymbol{\beta}^T \mathbf{z})}$$

is the hazard function given $\mathbf{z}$.

Rearranging terms, applying the E step and knowing $R_{i+1} = R_i - C_i - D_i$, we obtain

$$\begin{aligned}
\tilde{L}_2\left(\boldsymbol{\beta}, \boldsymbol{\alpha}; w^{(m)}\right) &= \prod_{i=1}^{k}\left[\prod_{\ell\in D_i}\left\{\left(1-\alpha_i^{\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}\right)\prod_{j:t_{(j)}\le t_{(i-1)}}\alpha_j^{w_\ell^{(m)}\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}\right\}\right. \\
&\quad \left.\times\prod_{\ell\in C_i}\left(\prod_{j:t_{(j)}\le t_{(i)}}\alpha_j^{w_\ell^{(m)}\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}\right)\right] \\
&= \prod_{i=1}^{k}\left[\prod_{\ell\in D_i}\left\{1-\alpha_i^{\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}\right\}\times\prod_{\ell\in(D_i+C_i)}\prod_{j:t_{(j)}\le t_{(i-1)}}\alpha_j^{w_\ell^{(m)}\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}\right. \\
&\quad \left.\times\prod_{\ell\in C_i}\alpha_i^{w_\ell^{(m)}\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}\right] \\
&= \prod_{i=1}^{k}\left[\prod_{\ell\in D_i}\left\{1-\alpha_i^{\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}\right\}\prod_{\ell\in(R_i-D_i)}\alpha_i^{w_\ell^{(m)}\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}\right].
\end{aligned}$$

The expected log-likelihood is then given by

$$\tilde{l}_2\left(\boldsymbol{\beta}, \boldsymbol{\alpha}; w^{(m)}\right) = \sum_{i=1}^{k}\left[\sum_{\ell\in D_i}\ln\left(1-\alpha_i^{\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}\right) + \sum_{\ell\in(R_i-D_i)} w_\ell^{(m)}\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)\ln(\alpha_i)\right]. \quad (4.3)$$

Applying the partial derivative to the expected log-likelihood given in (4.3), we obtain the score statistic, which has the form

$$\frac{\partial\tilde{l}_2}{\partial\alpha_i} = \sum_{i=1}^{k}\left[\sum_{\ell\in D_i}\frac{-\alpha_i^{(\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)-1)}\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}{1-\alpha_i^{\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}} + \sum_{\ell\in(R_i-D_i)} w_\ell^{(m)}\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)\frac{1}{\alpha_i}\right].$$

Given $\boldsymbol{\beta}$, we obtain independent estimating equations for each $\alpha_i$,

$$\begin{aligned}
\frac{\partial\tilde{l}_2}{\partial\alpha_i} &= 0 \\
\Leftrightarrow \quad \sum_{\ell\in D_i}\frac{-\alpha_i^{(\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)-1)}\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}{1-\alpha_i^{\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}} + \sum_{\ell\in R_i}\frac{w_\ell^{(m)}\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}{\alpha_i} & \\
- \sum_{\ell\in D_i}\frac{w_\ell^{(m)}\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}{\alpha_i} &= 0 \\
\Leftrightarrow \quad \sum_{\ell\in D_i}\left[\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell) + \frac{\alpha_i^{\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}{1-\alpha_i^{\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}}\right] &= \sum_{\ell\in R_i} w_\ell^{(m)}\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell) \\
\Leftrightarrow \quad \sum_{\ell\in D_i}\left(\frac{\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}{1-\alpha_i^{\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell)}}\right) &= \sum_{\ell\in R_i} w_\ell^{(m)}\exp(\boldsymbol{\beta}^T\mathbf{z}_\ell),
\end{aligned}$$

where $i = 1, \ldots, k$. The solution for $\alpha_i$ is not of closed form except when there are no ties at $t_{(i)}$, in which case the MLE of $\alpha_i$ given $\boldsymbol{\beta}$ is

$$\tilde{\alpha}_i = \left(1 - \frac{\exp(\boldsymbol{\beta}^{\boldsymbol{T}} \mathbf{z}_{\boldsymbol{\ell}})}{\sum_{\ell \in R_i} w_l^{(m)} \exp(\boldsymbol{\beta}^T \mathbf{z}_\ell)}\right)^{\exp\left(-\boldsymbol{\beta}^T \mathbf{z}_{(i)}\right)}$$

We then substitute $\tilde{\alpha}_i$ into $\tilde{L}_2\left(\boldsymbol{\beta}, \boldsymbol{\alpha}; w^{(m)}\right)$, obtain a nonparametric profile likelihood of $\boldsymbol{\beta}$ and obtain its MLE. But when there are ties, the MLEs for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ must be jointly obtained from $\tilde{L}_2\left(\boldsymbol{\beta}, \boldsymbol{\alpha}; w^{(m)}\right)$. This requires the maximization of a potentially very high-dimensional function. Note that since $w_\ell^{(m)}$ depends on $S_0(t_\ell | Y = 1)$, the baseline function is involved in the estimation of $\mathbf{b}$ and $\boldsymbol{\beta}$.

A Newton Raphson procedure can be used to maximize $\tilde{l}_1(\mathbf{b}; w^{(m)})$ to find $\hat{\mathbf{b}}$. A simultaneous Newton Raphson on $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ using $\tilde{l}_2(\boldsymbol{\beta}, \boldsymbol{\alpha}; w^{(m)})$ is, however, sensitive to starting values and will easily fail to converge. The method that is found to be most efficient is the two-step Newton Raphson in the grouped PH model wherein the updates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are obtained alternately.

## 4.4 A Theoretical Comparison of the Competing Risks Model and Cure Model

We list some theoretical differences, comparing the multiple imputation - competing risks model approach used in this thesis and the expectation maximization - cure model approach proposed in this chapter.

A first distinction between both approaches is based on the initial concept. The cure model is based on *a priori* belief in a non-susceptible group, so the probability of having the intention to attain a Ph.D.-degree or not can be estimated based on baseline covariate values only. The multiple imputation method we used, however, allows that this distinction becomes clear during the observation period: a Ph.D.-student is categorized as withdrawer if he has a gap lasting more than 4 years or when the current gap is censored by the end of the study, then an imperfect categorization is based on both current gap length and baseline covariate values.

As both methods differ in conceptualization, it is clear that the parametric modeling is different too and different parametric assumptions are supposed. The competing risks approach estimates a cause-specific PH model for both outcomes: Ph.D.-attainment and withdrawal. The cure model, on the other hand, consists of a logistic regression model estimating the prevalence and a Cox PH model estimating the latency.

As soon as the partial likelihood function is computed and provided the non-informative censoring assumption is fulfilled, both approaches can roughly be handled in the same way. In

our case, however, the competing risks approach uses the multiple imputation (MI) procedure to handle the missing data, while the cure model uses the expectation-maximization (EM) algorithm. In theory, the MI procedure and the EM algorithm asymptotically result in the same parameter estimates. Both missing-data methods summarize a likelihood function which has been averaged over a predictive distribution for the missing values [15].

Suppose the imputation model in the competing risks approach was built, based on the data of all gaps starting prior to October 1, 2009 instead of October 1, 2005, so no closed solution form would be available for equation (2.1) on p. 9

$$P(\text{withdrawer}|\mathbf{Z}, \text{gap} > t \text{ years}; C = t).$$

Then the EM algorithm could be applied to estimate the model parameters. In that way, we show the competing risks approach and the cure model can both use the EM algorithm. Although, the model that leads to the E step would be fundamentally different for both approaches.

- Imputation model: For the $i$th Ph.D.-student we would estimate

  $$P(\text{withdrawer}|\mathbf{z}_i, \text{gap} > t_i \text{ years}, C_i = t_i; \theta^{(m)}),$$

  where $t_i$ is the observed gaptime.

- Prevalence: For the $i$th Ph.D.-student we estimated

  $$\begin{aligned} \pi_i^{(m)} &= P\left(Y_i = 1 | X_i > t_i, \delta_i = 0, \mathbf{z}_i; \theta^{(m)}\right) \\ &= 1 - P\left(\text{withdrawer}|\mathbf{z}_i, X_i > t_i, C_i = t_i; \theta^{(m)}\right), \end{aligned}$$

  where $t_i$ is the observed sponsored time.

The current parameter estimates are denoted by $\theta^{(m)}$. As we mentioned before, the prevalence is based on baseline covariate values only, while the imputation model is based on both baseline covariate values and the current gap length.
In this way, the imputation model is built using the EM algorithm, but that does not exclude the use of the MI procedure for imputing the missing outcomes. Vice versa the cure model approach can impute the missing outcomes using the logistic regression model for the prevalence, although this model is built by iteration of the EM algorithm. So, it is obvious that switching these missing-data methods only increases the computational complexity.

A major advantage of the multiple imputation procedure is that it allows us to separately deal with the missing and delayed event type data, and with the survival analysis, thus, avoiding the difficulties of joint estimation [3]. The distribution of the missing data is estimated explicitly for the imputation model. Next, we multiple impute the missing data based on

the imputation model and in that way provide a number of completed data sets. On these completed data sets standard competing risks methods can be applied to estimate scalar quantities. Finally, these results are combined in a well-founded way for the inference.

A major advantage of the cure model is that only one PH model is built to estimate the latency, so the cure model corrects automatically for the two competing events 'Ph.D.-attainment' and 'withdrawal'.

We can conclude that theoretical differences between both approaches are mainly based on the conceptualization and parameterization of the multiple event types, rather than on the missing-data method. In theory, both the MI procedure and the EM algorithm can be applied in the competing risks model and cure model approach and asymptotically the same parameter estimates are obtained, however, there is a difference in computational complexity.

# Chapter 5

# Conclusions

The analysis of sponsored time to Ph.D.-attainment is considered as an important criterion in controlling the efficiency of the government investments. However, this analysis is impeded by missing and delayed event type data in a competing risks survival setting.

We applied the multiple imputation procedure to handle missing data. This procedure has gained popularity since the first publications between 1977 and 1987 [13]. It has proven to be very useful in the context for which it was envisioned, namely where the data collector and the data analyst are different persons. Beside this advantage, the popularity can be explained by the fact that it becomes relatively easy to create multiple imputed files using modern computing software and analyze the completed data sets with standard statistical procedures. Although, the development of user-friendly software for analyzing multiple imputed data is still needed.

In the competing risks setting, we built a cause-specific PH model for both outcomes 'Ph.D.-attainment' and 'withdrawal'. An advantage of this double model building is that the set of prognostic factors can differ for both outcomes and that a mixture of these proportional hazards functions is no longer proportional. As a relatively simple illustration of the results, cumulative incidence curves and associated 95% confidence intervals are plotted for both outcomes.

We estimated the variance of the cumulative incidence function, but a similar estimator could be constructed for the covariance as suggested in [5]. An estimator for the covariance of the cumulative incidence function would allow us to estimate pointwise confidence intervals for the cumulative incidence function for a combination of covariate values (e.g. $\mathbf{z}^* = (1, 1, 0, \ldots, 0)^T$). Now, we are strongly restricted to the reference group. The formulas for the variance and covariance of the cumulative incidence function could also be useful in many other settings. Further research is needed to check their accuracy by performing simulations.

As an alternative to the method we used in this thesis, the cure model is described. This model is based on *a priori* belief in a group of Ph.D.-students with no the intention of attaining a Ph.D.-degree. This is rather a strong assumption to be fulfilled, although it would provide important information to the government.

Concerning the research question as stated in section 1.2 (p. 2) we may conclude that the Ph.D.-student's gender, dominant statute classification, dominant scientific field, age (at the start of the Ph.D.-training), start year (year in which the Ph.D.-training began), nationality and university all influence the sponsored time to attainment of a Ph.D.-degree. The largest differences in the probability of attaining a Ph.D.-degree while accounting for the withdrawers, are estimated between covariate categories for 'dominant statute classification', 'dominant scientific field' and 'age (at start)'. Although, most contracts expire after 4 or 6 years, a lot of students need 2 years additional sponsored time to attain the Ph.D.-degree. For some students, even this additional sponsored time is not sufficient, but generally after 8 years of sponsoring hardly any more students attain a Ph.D.-degree.

The multiple imputation - competing risks model approach we applied, is only one possible method for handling missing and delayed event type data in this setting with multiple types of events. Although, compared to the naively non-imputed approach our analysis seems to prevent the loss of information and biased inference due to loss of data.

# Appendix A

# The Imputation Model

## A.1 Calculation of the Corresponding Gaptime

Suppose now that there are $n^*$ censored gaps which all need an imputed outcome. The $j$th censored gap has censoring time $t_j^*$ and covariate values $\mathbf{z}_j^*$. To estimate the baseline cumulative hazard for censored gaps at $t_j^*$, we first have to calculate the corresponding event time $t_{\text{maxindex},j}$ from all gaps starting between October 1, 1990 and September 30, 2005 so that

$$t_{\text{maxindex},j} \leq t_j^* < t_{\text{maxindex}+1,j},$$

and then estimate $\hat{H}_0(t_j^*)$ by $\hat{H}_0(t_{\text{maxindex},j})$. This is represented by figure A.1.

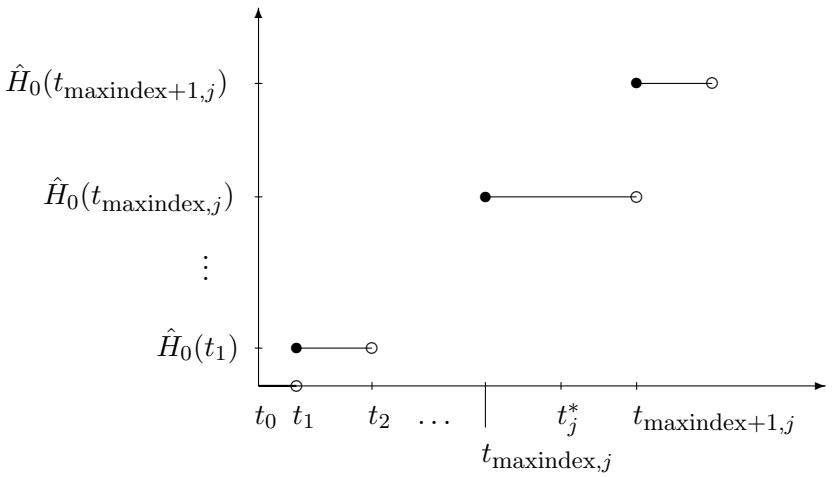


**Figure A.1:** Example of the baseline cumulative hazard function $\hat{H}_0(t_j^*)$ at a censored gap time $t_j^*$ .

**Algorithm 2** : Calculating corresponding event times

```
i = 0
for j = 1 : n* do
  while t*_j ≥ t_{i+1} and i < D do
    i = i + 1
  end while
  t_{maxindex,j} = t*_i
end for
```

## A.2 Variance Estimation

### A.2.1 Regularity Conditions

We aim to derive the asymptotic properties of the estimates $\boldsymbol{\beta}$ and $\hat{H}$ from the Cox PH model. To this end some regularity conditions are assumed to hold, which are listed in [1] and repeated below.

Since we are interested in asymptotic properties, we shall in fact consider a sequence of models, indexed by $(n) = (1), (2), \ldots$. For simplicity we are dropping a superfix $(n)$ almost everywhere; only $\boldsymbol{\beta}$ and $h_0$ are fixed (i.e. independent of $n$). Unless otherwise stated, all limits are taken as $n \to \infty$. We observe $n$ individuals in each of the $(n)$ models. We give regularity conditions on the time interval $[0, 1]$, but the results can be extended to processes on $[0, \infty[$. For a vector $a$, $|a| = \left(\sum a_i^2\right)^{1/2} = (a^T a)^{1/2}$. For a matrix $A$ or a vector $a$, $||A|| = \sup_{i,j} |a_{ij}|$ and $||a|| = \sup_i |a_i|$. Some further important definitions are:

$$
\begin{aligned}
S^{(0)}(\boldsymbol{\beta}, t) &= \frac{1}{n} \sum_{j \in R(t)} \exp(\beta^T \mathbf{Z}_j), \\
S^{(1)}(\boldsymbol{\beta}, t) &= \frac{1}{n} \sum_{j \in R(t)} \mathbf{Z}_j \exp(\beta^T \mathbf{Z}_j), \\
S^{(2)}(\boldsymbol{\beta}, t) &= \frac{1}{n} \sum_{j \in R(t)} \mathbf{Z}_j \mathbf{Z}_j^T \exp(\beta^T \mathbf{Z}_j), \\
E(\boldsymbol{\beta}, t) &= \frac{S^{(1)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)}, \\
V(\boldsymbol{\beta}, t) &= \frac{S^{(2)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)} - E(\boldsymbol{\beta}, t) E(\boldsymbol{\beta}, t)^T.
\end{aligned}
$$

Note that $S^{(0)}$ is a scalar, $S^{(1)}$ and $E$ are $p$-vectors and $S^{(2)}$ and $V$ are $(p \times p)$-matrices. These quantities can be interpreted as follows. Suppose at time $t$, we select an individual $i$ out of those individuals under observation (at risk) with probabilities proportional to $\exp(\boldsymbol{\beta}^T \mathbf{Z}_i)$. Then $E(\boldsymbol{\beta}, t)$ and $V(\boldsymbol{\beta}, t)$ are the expectation and variance respectively of the covariate vector $\mathbf{Z}_i$ of the individual selected. $S^{(0)}$, $S^{(1)}$ and $S^{(2)}$ are roughly to be interpreted as a norming

factor, a sum and a sum of squares respectively.
The following list of (mild) regularity conditions (see Andersen & Gill, [1]) will be assumed to hold throughout this section. There are a number of redundancies in them, but in this way we hope to avoid too many technical distractions.

1. (Finite interval). $\int_0^1 h_0(t)dt < \infty$.

2. (Asymptotic stability). There exists a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}$ and scalar, vector and matrix functions $s^{(0)}$, $s^{(1)}$ and $s^{(2)}$ defined on $\mathcal{B} \times [0,1]$ such that for $j = 0, 1, 2$

$$\sup_{t\in[0,1],\boldsymbol{\beta}\in\mathcal{B}} \left\| S^{(j)}(\boldsymbol{\beta}, t) - s^{(j)}(\boldsymbol{\beta}, t) \right\| \xrightarrow{P} 0.$$

3. (Lindeberg condition). There exists $\delta > 0$ such that

$$n^{-1/2} \sup_{i\in R(t),t} |\mathbf{Z}_i| I\{\boldsymbol{\beta}^T \mathbf{Z}_i > -\delta|\mathbf{Z}_i|\} \xrightarrow{P} 0.$$

4. (Asymptotic regularity conditions). Let $\mathcal{B}$, $s^{(0)}$, $s^{(1)}$ and $s^{(2)}$ be as in condition 2 and define $e = s^{(1)}/s^{(0)}$ and $v = s^{(2)}/s^{(0)} - ee^T$. For all $\boldsymbol{\beta} \in \mathcal{B}$, $t \in [0,1]$:

$$s^{(1)}(\boldsymbol{\beta}, t) = \frac{\partial}{\partial \boldsymbol{\beta}} s^{(0)}(\boldsymbol{\beta}, t), \quad s^{(2)}(\boldsymbol{\beta}, t) = \frac{\partial^2}{\partial \boldsymbol{\beta}^2} s^{(0)}(\boldsymbol{\beta}, t),$$

   $s^{(0)}(\cdot, t)$, $s^{(1)}(\cdot, t)$ and $s^{(2)}(\cdot, t)$ are continuous functions of $\boldsymbol{\beta} \in \mathcal{B}$, uniformly in $t \in [0,1]$, $s^{(0)}$, $s^{(1)}$ and $s^{(2)}$ are bounded on $\mathcal{B} \times [0,1]$; $s^{(0)}$ is bounded away from zero on $\mathcal{B} \times [0,1]$, and the matrix

$$\Sigma = \int_0^1 v(\boldsymbol{\beta}, t) s^{(0)}(\boldsymbol{\beta}, t) h_0(t) dt$$

   is positive definite.

Note that the partial derivative conditions on $s^{(0)}$, $s^{(1)}$ and $s^{(2)}$ are satisfied by $S^{(0)}$, $S^{(1)}$ and $S^{(2)}$; and that $\Sigma$ is automatically positive semidefinite. Furthermore the interval $[0,1]$ in the conditions may everywhere be replaced by the set $\{t : h_0(t) > 0\}$.

### A.2.2 Asymptotic Properties

We start by giving the definition of a Gaussian process [9].

**Definition A.1.** Given the probability space $(\Omega, \mathcal{F}, P)$, an $\mathbb{R}^d$-valued stochastic process $X = \{X_t; 0 \le t < \infty\}$ is called *Gaussian* if, for any integer $k \ge 1$ and real numbers $0 \le t_1 < t_2 < \cdots < \infty$, the random vector $(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ has a joint normal distribution.

The finite-dimensional distributions of a Gaussian process $X$ are determined by its expectation vector $m(t) := E[X_t]; t \geq 0$, and its covariance matrix

$$\rho(s,t) := E\left[(X_s - m(s))(X_t - m(t))^T\right]$$

where $s, t \geq 0$. If $m(t) \equiv 0; t \geq 0$, we say that $X$ is a zero-mean Gaussian process.

We extend the results from [17] to state the distribution of the cumulative hazard function in next theorem.

**Theorem A.1.** *Under the set of (mild) regularity conditions (A.2.1, p. 77) the random function* $\sqrt{n}\left[\hat{H}_0(t)\exp(\hat{\boldsymbol{\beta}}^T\mathbf{z}^*) - H_0(t)\exp(\boldsymbol{\beta}^T\mathbf{z}^*)\right]$ *converges weakly to a Gaussian process* $G_{\mathbf{z}^*}(t)$ *which has mean 0 and covariance structure that can be estimated by*

$$\begin{aligned}\widehat{Cov}(G_{\mathbf{z}^*}(s), G_{\mathbf{z}^*}(t)) &= \exp(2\hat{\boldsymbol{\beta}}^T\mathbf{z}^*)\left[nQ_1(s) + \mathbf{Q}_3(s;\mathbf{z}^*)^T\hat{\Sigma}^2_{\hat{\boldsymbol{\beta}}}\mathbf{Q}_3(t;\mathbf{z}^*)\right] \\ &= \exp(2\hat{\boldsymbol{\beta}}^T\mathbf{z}^*)\left[nQ_1(s) + nQ_2(s,t;\mathbf{z}^*\right]\end{aligned}$$

*where*

$$\begin{aligned}Q_1(s) &= \sum_{t_i \leq s}\frac{d_i}{W(t_i,\hat{\boldsymbol{\beta}})^2}, \\ Q_2(s,t;\mathbf{z}^*) &= \mathbf{Q}_3(s;\mathbf{z}^*)^T\hat{V}(\hat{\boldsymbol{\beta}})\mathbf{Q}_3(t;\mathbf{z}^*) \\ &= \frac{1}{n}\mathbf{Q}_3(s;\mathbf{z}^*)^T\Sigma_{\hat{\boldsymbol{\beta}}}\mathbf{Q}_3(t;\mathbf{z}^*), \\ \mathbf{Q}_3(s;\mathbf{z}^*) &= (Q_3(s;\mathbf{z}^*)_1, \ldots, Q_3(s;\mathbf{z}^*)_p), \\ Q_3(s;\mathbf{z}^*)_k &= \sum_{t_i \leq s}\left[\frac{W^{(k)}(t_i,\hat{\boldsymbol{\beta}})}{W(t_i,\hat{\boldsymbol{\beta}})} - z_k^*\right]\cdot\left[\frac{d_i}{W(t_i,\hat{\boldsymbol{\beta}})}\right] \qquad k = 1,\ldots,p, \\ W^{(k)}(t_i,\hat{\boldsymbol{\beta}}) &= \sum_{j\in R(t_i)} z_{jk}\exp(\hat{\boldsymbol{\beta}}^T\mathbf{z}_j),\end{aligned}$$

*knowing* $s \leq t$ *and* $\hat{\Sigma}^2_{\hat{\boldsymbol{\beta}}}$ *is the estimated asymptotic variance-covariance matrix of* $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$*, while* $\hat{V}(\hat{\boldsymbol{\beta}})$ *is the estimated variance-covariance matrix of* $\hat{\boldsymbol{\beta}}$ *derived by Cox.*

These quantities may be interpreted as follows [10]. $Q_1(t)$ may be interpreted as the estimated variance of $\hat{H}_0(t)$ if $\boldsymbol{\beta}$ were known. $Q_2$ reflects the uncertainty in the estimation process added by estimating $\boldsymbol{\beta}$. Here, $\mathbf{Q}_3(t;\mathbf{z}^*)$ is large when $\mathbf{z}^*$ is far from the average covariate in the risk set. Using this variance estimate, pointwise confidence intervals for the cumulative hazard function can be constructed for $H(t|\mathbf{Z} = \mathbf{z}^*)$.

Since $G_{\mathbf{z}^*}(t)$ is a Gaussian process at each time $t$, the estimated cumulative hazard function $\hat{H}(t|\mathbf{z}^*) = \hat{H}_0(t)\exp(\hat{\boldsymbol{\beta}}^T\mathbf{z}^*)$ is multivariate normally distributed.

$$\hat{H}_0(t)\exp(\hat{\boldsymbol{\beta}}^T\mathbf{z}^*) \xrightarrow{\mathcal{L}} N\left(H_0(t)\exp(\boldsymbol{\beta}^T\mathbf{z}^*), \frac{\widehat{\mathrm{Var}}(G_{\mathbf{z}^*}(t))}{n}\right)$$

and

$$\begin{aligned}\frac{\widehat{\mathrm{Var}}(G_{\mathbf{z}^*}(t))}{n} &= \widehat{\mathrm{Var}}(\hat{H}(t|\mathbf{z}^*)) \\ &= \exp(2\hat{\boldsymbol{\beta}}^T\mathbf{z}^*)\left[Q_1(t) + Q_2(t,t;\mathbf{z}^*)\right].\end{aligned}$$

The algorithm for calculating $Q_1$ is quite similar to that for calculating $\hat{H}_0$ and the algorithm for calculating $W^{(k)}(t_i, \hat{\boldsymbol{\beta}})$ is quite similar to that for calculating $W(t_i, \hat{\boldsymbol{\beta}})$ in Algorithm 1. However, the algorithm for the $p$-vector $\mathbf{Q}_3(t;\mathbf{z}^*)$ is given by Algorithm 3. Note that for each censored gap $j$ with covariate values $\mathbf{z}_j^*$, we only need to calculate this value $\mathbf{Q}_3$ at the censoring time $t = t_j^*$ and the end time $t = 4$.

**Algorithm 3** : Estimating $\mathbf{Q}_3(t_j^*;\mathbf{z}_j^*)$ and $\mathbf{Q}_3(4;\mathbf{z}_j^*)$

**for** $j = 1 : n^*$ **do**
  ZCensMat = matrix(rep($\mathbf{z}^*$[$j$,],$D$),nrow = $D$, byrow = T)
  help = ($Wk/W$ - ZCensMat) $\times$ (death/$W$)
  Q3long = apply(as.matrix(help),2,cumsum)
  Q3[$j$,] = Q3long[maxindex[$j$],]
  Q3End[$j$,] = Q3long[$D$,]
**end for**

That way, we can estimate $Q_2$ at the censoring times to estimate the variance. Again for each censored gap, we only need to calculate these values at the censoring time $t = t_j^*$ and the end time $t = 4$, reducing the number of calculations.

# Appendix B

# Detailed Model Output (SAS)

***Gaptime*** ***1***

***The PHREG Procedure***

| *Model Information* | |
|---|---|
| *Data Set* | LIB.GAPSRL |
| *Dependent Variable* | time |
| *Censoring Variable* | cens |
| *Censoring Value(s)* | 1 |
| *Ties Handling* | DISCRETE |

| | |
|---|---|
| *Number of Observations Read* | 14050 |
| *Number of Observations Used* | 13693 |

*Summary of the Number of Event and Censored Values*

| *Total* | *Event* | *Censored* | *Percent Censored* |
|---|---|---|---|
| 13693 | 5956 | 7737 | 56.50 |

| *Convergence Status* |
|---|
| Convergence criterion (GCONV=1E-8) satisfied. |

*Model Fit Statistics*

| *Criterion* | *Without Covariates* | *With Covariates* |
|---|---|---|
| *-2 LOG L* | 80556.939 | 79513.633 |
| *AIC* | 80556.939 | 79589.633 |
| *SBC* | 80556.939 | 79843.935 |

*Testing Global Null Hypothesis: BETA=0*

| *Test* | *Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *Likelihood Ratio* | 1043.3059 | 38 | <.0001 |
| *Score* | 1116.5030 | 38 | <.0001 |
| *Wald* | 1063.1683 | 38 | <.0001 |

***Gaptime*** ***2***

***The PHREG Procedure***

*Analysis of Maximum Likelihood Estimates*

| *Parameter* | *DF* | *Parameter Estimate* | *Standard Error* | *Chi-Square* | *Pr > ChiSq* | *Hazard Ratio* | *95% Hazard Ratio Confidence Limits* | | *Label* |
|---|---|---|---|---|---|---|---|---|---|
| *GESLACHT* | 1 | 0.33587 | 0.07647 | 19.2890 | <.0001 | 1.399 | 1.204 | 1.625 | Female |
| *domstat2* | 1 | 0.46417 | 0.06031 | 59.2351 | <.0001 | 1.591 | 1.413 | 1.790 | Assistant |
| *domstat3* | 1 | 1.05193 | 0.06622 | 252.3188 | <.0001 | 2.863 | 2.515 | 3.260 | Scholarship (Flanders) |
| *domstat4* | 1 | 1.05095 | 0.12007 | 76.6160 | <.0001 | 2.860 | 2.261 | 3.619 | Scholarship (univers.) |
| *domstat5* | 1 | 0.53790 | 0.06040 | 79.2976 | <.0001 | 1.712 | 1.521 | 1.928 | Project (FWO/BOF/IUAP) |
| *domclusmed* | 1 | -0.20256 | 0.06779 | 8.9297 | 0.0028 | 0.817 | 0.715 | 0.933 | Medical |
| *domclushum* | 1 | -0.13195 | 0.06736 | 3.8367 | 0.0501 | 0.876 | 0.768 | 1.000 | Humanities |
| *domclussoc* | 1 | -0.23156 | 0.06840 | 11.4618 | 0.0007 | 0.793 | 0.694 | 0.907 | Social |
| *domclustoe* | 1 | -0.34990 | 0.06285 | 30.9924 | <.0001 | 0.705 | 0.623 | 0.797 | Applied |
| *leeft2* | 1 | -0.04105 | 0.03635 | 1.2751 | 0.2588 | 0.960 | 0.894 | 1.031 | Age 26-30 (at start) |
| *leeft3* | 1 | -0.11601 | 0.05627 | 4.2507 | 0.0392 | 0.890 | 0.797 | 0.994 | Age 31-35 (at start) |
| *leeft4* | 1 | -0.03773 | 0.07699 | 0.2402 | 0.6241 | 0.963 | 0.828 | 1.120 | Age 36-40 (at start) |
| *leeft5* | 1 | -0.33429 | 0.09309 | 12.8972 | 0.0003 | 0.716 | 0.596 | 0.859 | Age 41+ (at start) |
| *start2* | 1 | 0.06877 | 0.02761 | 6.2047 | 0.0127 | 1.071 | 1.015 | 1.131 | Start cohort 1997-2004 |
| *start3* | 1 | 0.40367 | 0.09950 | 16.4584 | <.0001 | 1.497 | 1.232 | 1.820 | Start cohort 2004-2009 |
| *natEurEU* | 1 | -0.13493 | 0.09862 | 1.8720 | 0.1712 | 0.874 | 0.720 | 1.060 | EU (excl. Belgium) |
| *natAnd* | 1 | 0.18368 | 0.09870 | 3.4634 | 0.0627 | 1.202 | 0.990 | 1.458 | International (non-EU) |
| *univ2* | 1 | -0.33165 | 0.04683 | 50.1642 | <.0001 | 0.718 | 0.655 | 0.787 | |
| *univ3* | 1 | -0.25820 | 0.03110 | 68.9253 | <.0001 | 0.772 | 0.727 | 0.821 | |
| *univ4* | 1 | -0.14613 | 0.04570 | 10.2250 | 0.0014 | 0.864 | 0.790 | 0.945 | |
| *univ5* | 1 | 0.06094 | 0.08081 | 0.5687 | 0.4508 | 1.063 | 0.907 | 1.245 | |
| *fml_domstat2* | 1 | 0.01760 | 0.07496 | 0.0551 | 0.8144 | 1.018 | 0.879 | 1.179 | Interaction w. gender |
| *fml_domstat3* | 1 | -0.20894 | 0.09033 | 5.3510 | 0.0207 | 0.811 | 0.680 | 0.969 | Interaction w. gender |
| *fml_domstat4* | 1 | -0.16837 | 0.12546 | 1.8011 | 0.1796 | 0.845 | 0.661 | 1.081 | Interaction w. gender |
| *fml_domstat5* | 1 | 0.05849 | 0.07337 | 0.6354 | 0.4254 | 1.060 | 0.918 | 1.224 | Interaction w. gender |
| *fml_domclusmed* | 1 | -0.12017 | 0.08350 | 2.0710 | 0.1501 | 0.887 | 0.753 | 1.044 | Interaction w. gender |
| *fml_domclushum* | 1 | -0.01963 | 0.08660 | 0.0514 | 0.8207 | 0.981 | 0.827 | 1.162 | Interaction w. gender |
| *fml_domclussoc* | 1 | -0.04250 | 0.08697 | 0.2388 | 0.6251 | 0.958 | 0.808 | 1.137 | Interaction w. gender |
| *fml_domclustoe* | 1 | 0.06002 | 0.08461 | 0.5032 | 0.4781 | 1.062 | 0.900 | 1.253 | Interaction w. gender |
| *natNBel_domstat2* | 1 | 0.07402 | 0.13930 | 0.2823 | 0.5952 | 1.077 | 0.820 | 1.415 | Interaction w. nationality |
| *natNBel_domstat3* | 1 | 0.13070 | 0.25775 | 0.2571 | 0.6121 | 1.140 | 0.688 | 1.889 | Interaction w. nationality |
| *natNBel_domstat4* | 1 | 0.27879 | 0.13098 | 4.5308 | 0.0333 | 1.322 | 1.022 | 1.708 | Interaction w. nationality |
| *natNBel_domstat5* | 1 | -0.03509 | 0.08344 | 0.1768 | 0.6741 | 0.966 | 0.820 | 1.137 | Interaction w. nationality |
| *natNBel_domclusmed* | 1 | 0.28790 | 0.10019 | 8.2581 | 0.0041 | 1.334 | 1.096 | 1.623 | Interaction w. nationality |
| *natNBel_domclushum* | 1 | -0.07005 | 0.12031 | 0.3390 | 0.5604 | 0.932 | 0.736 | 1.180 | Interaction w. nationality |

*Analysis of Maximum Likelihood Estimates*

| Parameter | DF | Parameter Estimate | Standard Error | Chi-Square | Pr > ChiSq | Hazard Ratio | 95% Hazard Ratio Confidence Limits | | Label |
|---|---|---|---|---|---|---|---|---|---|
| natNBel_domclussoc | 1 | 0.08147 | 0.12712 | 0.4107 | 0.5216 | 1.085 | 0.846 | 1.392 | Interaction w. nationality |
| natNBel_domclustoe | 1 | 0.43786 | 0.09436 | 21.5308 | <.0001 | 1.549 | 1.288 | 1.864 | Interaction w. nationality |
| fml_natNBel | 1 | -0.08306 | 0.07623 | 1.1875 | 0.2758 | 0.920 | 0.793 | 1.069 | |

*Linear Hypotheses Testing Results*

| Label | Wald Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| gender | 19.2890 | 1 | <.0001 |
| domstat | 285.3971 | 4 | <.0001 |
| domclus | 33.2731 | 4 | <.0001 |
| leeft | 15.3025 | 4 | 0.0041 |
| start | 19.9151 | 2 | <.0001 |
| nat | 25.6315 | 2 | <.0001 |
| univ | 96.8168 | 4 | <.0001 |
| fml_domstat | 11.5732 | 4 | 0.0208 |
| fml_domclus | 4.7093 | 4 | 0.3184 |
| nat_domstat | 6.3842 | 4 | 0.1722 |
| nat_domclus | 32.2118 | 4 | <.0001 |
| fml_natNBel | 1.1875 | 1 | 0.2758 |

**Sponsored time competing risks analysis** **1**
**Outcome of interest: Ph.D.-attainment**

**The PHREG Procedure**

| Model Information | |
|---|---|
| Data Set | WORK.IMPUTATIEA_1 |
| Dependent Variable | sponsTime |
| Censoring Variable | cens1 |
| Censoring Value(s) | 0 |
| Ties Handling | DISCRETE |

| | |
|---|---|
| Number of Observations Read | 28871 |
| Number of Observations Used | 28396 |

Summary of the Number of Event and Censored Values

| Total | Event | Censored | Percent Censored |
|---|---|---|---|
| 28396 | 9277 | 19119 | 67.33 |

| Convergence Status |
|---|
| Convergence criterion (GCONV=1E-8) satisfied. |

Model Fit Statistics

| Criterion | Without Covariates | With Covariates |
|---|---|---|
| -2 LOG L | 141511.89 | 137591.06 |
| AIC | 141511.89 | 137657.06 |
| SBC | 141511.89 | 137892.52 |

Testing Global Null Hypothesis: BETA=0

| Test | Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| Likelihood Ratio | 3920.8387 | 33 | <.0001 |
| Score | 4310.8738 | 33 | <.0001 |
| Wald | 3850.0000 | 33 | <.0001 |

***Sponsored time competing risks analysis*** **2**
***Outcome of interest: Ph.D.-attainment***

***The PHREG Procedure***

*Analysis of Maximum Likelihood Estimates*

| *Parameter* | *DF* | *Parameter Estimate* | *Standard Error* | *Chi-Square* | *Pr > ChiSq* | *Hazard Ratio* | *95% Hazard Ratio Confidence Limits* | | *Label* |
|---|---|---|---|---|---|---|---|---|---|
| *GESLACHT* | 1 | -0.19845 | 0.07582 | 6.8503 | 0.0089 | 0.820 | 0.707 | 0.951 | Female |
| *domstat2* | 1 | -0.03214 | 0.05918 | 0.2949 | 0.5871 | 0.968 | 0.862 | 1.087 | Assistant |
| *domstat3* | 1 | 0.99138 | 0.05537 | 320.6324 | <.0001 | 2.695 | 2.418 | 3.004 | Scholarship (Flanders) |
| *domstat4* | 1 | 1.00797 | 0.08487 | 141.0582 | <.0001 | 2.740 | 2.320 | 3.236 | Scholarship (univers.) |
| *domstat5* | 1 | 0.47492 | 0.05694 | 69.5640 | <.0001 | 1.608 | 1.438 | 1.798 | Project (FWO/BOF/IUAP) |
| *domclusmed* | 1 | -0.16553 | 0.03955 | 17.5153 | <.0001 | 0.847 | 0.784 | 0.916 | Medical |
| *domclushum* | 1 | -0.58491 | 0.04793 | 148.9406 | <.0001 | 0.557 | 0.507 | 0.612 | Humanities |
| *domclussoc* | 1 | -0.61569 | 0.04971 | 153.4012 | <.0001 | 0.540 | 0.490 | 0.596 | Social |
| *domclustoe* | 1 | -0.33695 | 0.03569 | 89.1305 | <.0001 | 0.714 | 0.666 | 0.766 | Applied |
| *leeft2* | 1 | -0.10632 | 0.03418 | 9.6742 | 0.0019 | 0.899 | 0.841 | 0.961 | Age 26-30 (at start) |
| *leeft3* | 1 | -0.02264 | 0.05730 | 0.1561 | 0.6928 | 0.978 | 0.874 | 1.094 | Age 31-35 (at start) |
| *leeft4* | 1 | -0.01419 | 0.08198 | 0.0300 | 0.8625 | 0.986 | 0.840 | 1.158 | Age 36-40 (at start) |
| *leeft5* | 1 | -0.49498 | 0.11392 | 18.8784 | <.0001 | 0.610 | 0.488 | 0.762 | Age 41+ (at start) |
| *start2* | 1 | 0.21010 | 0.02348 | 80.0446 | <.0001 | 1.234 | 1.178 | 1.292 | Start cohort 1997-2004 |
| *start3* | 1 | -0.11122 | 0.04790 | 5.3902 | 0.0202 | 0.895 | 0.815 | 0.983 | Start cohort 2004-2009 |
| *natEurEU* | 1 | 1.22242 | 0.07677 | 253.5192 | <.0001 | 3.395 | 2.921 | 3.947 | EU (excl. Belgium) |
| *natAnd* | 1 | 1.77746 | 0.07770 | 523.3677 | <.0001 | 5.915 | 5.079 | 6.888 | International (non-EU) |
| *univ2* | 1 | -0.35403 | 0.03888 | 82.8963 | <.0001 | 0.702 | 0.650 | 0.757 | |
| *univ3* | 1 | 0.21591 | 0.02521 | 73.3371 | <.0001 | 1.241 | 1.181 | 1.304 | |
| *univ4* | 1 | -0.09658 | 0.03851 | 6.2884 | 0.0122 | 0.908 | 0.842 | 0.979 | |
| *univ5* | 1 | 0.04923 | 0.06577 | 0.5604 | 0.4541 | 1.050 | 0.923 | 1.195 | |
| *fml_domstat2* | 1 | 0.08036 | 0.08331 | 0.9304 | 0.3348 | 1.084 | 0.920 | 1.276 | Interaction w. gender |
| *fml_domstat3* | 1 | 0.10614 | 0.07645 | 1.9277 | 0.1650 | 1.112 | 0.957 | 1.292 | Interaction w. gender |
| *fml_domstat4* | 1 | 0.02555 | 0.10606 | 0.0580 | 0.8096 | 1.026 | 0.833 | 1.263 | Interaction w. gender |
| *fml_domstat5* | 1 | 0.24351 | 0.07856 | 9.6088 | 0.0019 | 1.276 | 1.094 | 1.488 | Interaction w. gender |
| *fml_domclusmed* | 1 | -0.04407 | 0.05819 | 0.5737 | 0.4488 | 0.957 | 0.854 | 1.072 | Interaction w. gender |
| *fml_domclushum* | 1 | -0.26116 | 0.07520 | 12.0593 | 0.0005 | 0.770 | 0.665 | 0.892 | Interaction w. gender |
| *fml_domclussoc* | 1 | -0.10142 | 0.07645 | 1.7600 | 0.1846 | 0.904 | 0.778 | 1.050 | Interaction w. gender |
| *fml_domclustoe* | 1 | 0.13260 | 0.06311 | 4.4148 | 0.0356 | 1.142 | 1.009 | 1.292 | Interaction w. gender |
| *natNBel_domstat2* | 1 | -1.24767 | 0.14097 | 78.3325 | <.0001 | 0.287 | 0.218 | 0.379 | Interaction w. nationality |
| *natNBel_domstat3* | 1 | -1.31863 | 0.17012 | 60.0787 | <.0001 | 0.268 | 0.192 | 0.373 | Interaction w. nationality |
| *natNBel_domstat4* | 1 | -0.62069 | 0.10754 | 33.3108 | <.0001 | 0.538 | 0.435 | 0.664 | Interaction w. nationality |
| *natNBel_domstat5* | 1 | -0.76037 | 0.08396 | 82.0228 | <.0001 | 0.467 | 0.397 | 0.551 | Interaction w. nationality |

**The PHREG Procedure**

*Linear Hypotheses Testing Results*

| Label | Wald Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| gender | 6.8503 | 1 | 0.0089 |
| domstat | 833.1768 | 4 | <.0001 |
| domclus | 254.9287 | 4 | <.0001 |
| leeft | 26.5832 | 4 | <.0001 |
| start | 115.5760 | 2 | <.0001 |
| nat | 523.3691 | 2 | <.0001 |
| univ | 268.7948 | 4 | <.0001 |
| fml_domstat | 14.6365 | 4 | 0.0055 |
| fml_domclus | 26.0206 | 4 | <.0001 |
| nat_domstat | 135.3900 | 4 | <.0001 |

***Sponsored time competing risks analysis*** ***1***
***Outcome of interest: Withdrawal***
***Imputation method A***

***The PHREG Procedure***

***_Imputation_=1***

| *Model Information* | |
|---|---|
| *Data Set* | LIB.IMPUTATIEA |
| *Dependent Variable* | sponsTime |
| *Censoring Variable* | cens2 |
| *Censoring Value(s)* | 0 |
| *Ties Handling* | DISCRETE |

| | |
|---|---|
| *Number of Observations Read* | 28871 |
| *Number of Observations Used* | 28396 |

*Summary of the Number of Event and Censored Values*

| *Total* | *Event* | *Censored* | *Percent Censored* |
|---|---|---|---|
| 28396 | 10446 | 17950 | 63.21 |

| *Convergence Status* |
|---|
| Convergence criterion (GCONV=1E-8) satisfied. |

*Model Fit Statistics*

| *Criterion* | *Without Covariates* | *With Covariates* |
|---|---|---|
| *-2 LOG L* | 157210.27 | 151057.73 |
| *AIC* | 157210.27 | 151115.73 |
| *SBC* | 157210.27 | 151326.09 |

*Testing Global Null Hypothesis: BETA=0*

| *Test* | *Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *Likelihood Ratio* | 6152.5363 | 29 | <.0001 |
| *Score* | 6576.4083 | 29 | <.0001 |
| *Wald* | 5543.3817 | 29 | <.0001 |

***Sponsored time competing risks analysis*** ***2***
***Outcome of interest: Withdrawal***
***Imputation method A***

***The PHREG Procedure***

***_Imputation_=1***

*Analysis of Maximum Likelihood Estimates*

| *Parameter* | *DF* | *Parameter Estimate* | *Standard Error* | *Chi-Square* | *Pr > ChiSq* | *Hazard Ratio* | *95% Hazard Ratio Confidence Limits* | | *Label* |
|---|---|---|---|---|---|---|---|---|---|
| *GESLACHT* | 1 | 0.10410 | 0.02035 | 26.1625 | <.0001 | 1.110 | 1.066 | 1.155 | Female |
| *domstat2* | 1 | -1.06215 | 0.02818 | 1420.4166 | <.0001 | 0.346 | 0.327 | 0.365 | Assistant |
| *domstat3* | 1 | -2.12046 | 0.04432 | 2288.7256 | <.0001 | 0.120 | 0.110 | 0.131 | Scholarship (Flanders) |
| *domstat4* | 1 | -1.57768 | 0.08857 | 317.3314 | <.0001 | 0.206 | 0.174 | 0.246 | Scholarship (univers.) |
| *domstat5* | 1 | -1.05764 | 0.02965 | 1272.5414 | <.0001 | 0.347 | 0.328 | 0.368 | Project (FWO/BOF/IUAP) |
| *domclusmed* | 1 | 0.17861 | 0.03837 | 21.6701 | <.0001 | 1.196 | 1.109 | 1.289 | Medical |
| *domclushum* | 1 | 0.44141 | 0.03954 | 124.6220 | <.0001 | 1.555 | 1.439 | 1.680 | Humanities |
| *domclussoc* | 1 | 0.52260 | 0.03664 | 203.4125 | <.0001 | 1.686 | 1.570 | 1.812 | Social |
| *domclustoe* | 1 | 0.21962 | 0.03847 | 32.5941 | <.0001 | 1.246 | 1.155 | 1.343 | Applied |
| *leeft2* | 1 | 0.21448 | 0.02649 | 65.5571 | <.0001 | 1.239 | 1.177 | 1.305 | Age 26-30 (at start) |
| *leeft3* | 1 | 0.36045 | 0.03988 | 81.6981 | <.0001 | 1.434 | 1.326 | 1.551 | Age 31-35 (at start) |
| *leeft4* | 1 | 0.33487 | 0.05949 | 31.6819 | <.0001 | 1.398 | 1.244 | 1.571 | Age 36-40 (at start) |
| *leeft5* | 1 | 0.59884 | 0.05679 | 111.1836 | <.0001 | 1.820 | 1.628 | 2.034 | Age 41+ (at start) |
| *start2* | 1 | -0.22100 | 0.02242 | 97.1501 | <.0001 | 0.802 | 0.767 | 0.838 | Start cohort 1997-2004 |
| *start3* | 1 | -0.72322 | 0.03118 | 537.9361 | <.0001 | 0.485 | 0.456 | 0.516 | Start cohort 2004-2009 |
| *natEurEU* | 1 | 0.48687 | 0.06305 | 59.6279 | <.0001 | 1.627 | 1.438 | 1.841 | EU (excl. Belgium) |
| *natAnd* | 1 | 0.40963 | 0.06850 | 35.7630 | <.0001 | 1.506 | 1.317 | 1.723 | International (non-EU) |
| *univ2* | 1 | 0.08045 | 0.03408 | 5.5730 | 0.0182 | 1.084 | 1.014 | 1.159 | |
| *univ3* | 1 | 0.02639 | 0.02393 | 1.2156 | 0.2702 | 1.027 | 0.980 | 1.076 | |
| *univ4* | 1 | -0.04255 | 0.03511 | 1.4685 | 0.2256 | 0.958 | 0.895 | 1.027 | |
| *univ5* | 1 | -0.37337 | 0.06329 | 34.8062 | <.0001 | 0.688 | 0.608 | 0.779 | |
| *natNBel_domstat2* | 1 | -0.07097 | 0.10644 | 0.4446 | 0.5049 | 0.931 | 0.756 | 1.148 | Interaction w. nationality |
| *natNBel_domstat3* | 1 | -0.22441 | 0.30726 | 0.5334 | 0.4652 | 0.799 | 0.438 | 1.459 | Interaction w. nationality |
| *natNBel_domstat4* | 1 | -0.19558 | 0.14015 | 1.9474 | 0.1629 | 0.822 | 0.625 | 1.082 | Interaction w. nationality |
| *natNBel_domstat5* | 1 | 0.28787 | 0.05816 | 24.4981 | <.0001 | 1.334 | 1.190 | 1.495 | Interaction w. nationality |
| *natNBel_domclusmed* | 1 | -0.30158 | 0.07610 | 15.7061 | <.0001 | 0.740 | 0.637 | 0.859 | Interaction w. nationality |
| *natNBel_domclushum* | 1 | -0.45506 | 0.09161 | 24.6731 | <.0001 | 0.634 | 0.530 | 0.759 | Interaction w. nationality |
| *natNBel_domclussoc* | 1 | -0.52809 | 0.08925 | 35.0119 | <.0001 | 0.590 | 0.495 | 0.702 | Interaction w. nationality |
| *natNBel_domclustoe* | 1 | -0.45640 | 0.07274 | 39.3704 | <.0001 | 0.634 | 0.549 | 0.731 | Interaction w. nationality |

***The PHREG Procedure***

***_Imputation_=1***

*Linear Hypotheses Testing Results*

| *Label* | *Wald Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *gender* | 26.1625 | 1 | <.0001 |
| *domstat* | 3366.7209 | 4 | <.0001 |
| *domclus* | 268.1290 | 4 | <.0001 |
| *leeft* | 202.2093 | 4 | <.0001 |
| *start* | 538.5079 | 2 | <.0001 |
| *nat* | 59.7926 | 2 | <.0001 |
| *univ* | 50.1135 | 4 | <.0001 |
| *nat_domstat* | 33.4138 | 4 | <.0001 |
| *nat_domclus* | 55.3680 | 4 | <.0001 |

***Sponsored time competing risks analysis***
***Outcome of interest: Withdrawal***
***Imputation method A***

***The PHREG Procedure***

***_Imputation_=2***

| *Model Information* | |
|---|---|
| *Data Set* | LIB.IMPUTATIEA |
| *Dependent Variable* | sponsTime |
| *Censoring Variable* | cens2 |
| *Censoring Value(s)* | 0 |
| *Ties Handling* | DISCRETE |

| | |
|---|---|
| *Number of Observations Read* | 28871 |
| *Number of Observations Used* | 28396 |

*Summary of the Number of Event and Censored Values*

| *Total* | *Event* | *Censored* | *Percent Censored* |
|---|---|---|---|
| 28396 | 10486 | 17910 | 63.07 |

| *Convergence Status* |
|---|
| Convergence criterion (GCONV=1E-8) satisfied. |

*Model Fit Statistics*

| *Criterion* | *Without Covariates* | *With Covariates* |
|---|---|---|
| *-2 LOG L* | 157873.25 | 151723.42 |
| *AIC* | 157873.25 | 151781.42 |
| *SBC* | 157873.25 | 151991.90 |

*Testing Global Null Hypothesis: BETA=0*

| *Test* | *Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *Likelihood Ratio* | 6149.8298 | 29 | <.0001 |
| *Score* | 6559.8810 | 29 | <.0001 |
| *Wald* | 5522.9435 | 29 | <.0001 |

***Sponsored time competing risks analysis*** ***5***
***Outcome of interest: Withdrawal***
***Imputation method A***

***The PHREG Procedure***

***_Imputation_=2***

*Analysis of Maximum Likelihood Estimates*

| *Parameter* | *DF* | *Parameter Estimate* | *Standard Error* | *Chi-Square* | *Pr > ChiSq* | *Hazard Ratio* | *95% Hazard Ratio Confidence Limits* | | *Label* |
|---|---|---|---|---|---|---|---|---|---|
| *GESLACHT* | 1 | 0.11334 | 0.02031 | 31.1482 | <.0001 | 1.120 | 1.076 | 1.165 | Female |
| *domstat2* | 1 | -1.05657 | 0.02811 | 1412.3182 | <.0001 | 0.348 | 0.329 | 0.367 | Assistant |
| *domstat3* | 1 | -2.13073 | 0.04447 | 2295.7086 | <.0001 | 0.119 | 0.109 | 0.130 | Scholarship (Flanders) |
| *domstat4* | 1 | -1.56043 | 0.08795 | 314.8059 | <.0001 | 0.210 | 0.177 | 0.250 | Scholarship (univers.) |
| *domstat5* | 1 | -1.05670 | 0.02958 | 1276.4826 | <.0001 | 0.348 | 0.328 | 0.368 | Project (FWO/BOF/IUAP) |
| *domclusmed* | 1 | 0.18156 | 0.03831 | 22.4571 | <.0001 | 1.199 | 1.112 | 1.293 | Medical |
| *domclushum* | 1 | 0.43529 | 0.03953 | 121.2456 | <.0001 | 1.545 | 1.430 | 1.670 | Humanities |
| *domclussoc* | 1 | 0.52035 | 0.03660 | 202.0863 | <.0001 | 1.683 | 1.566 | 1.808 | Social |
| *domclustoe* | 1 | 0.22633 | 0.03841 | 34.7256 | <.0001 | 1.254 | 1.163 | 1.352 | Applied |
| *leeft2* | 1 | 0.21697 | 0.02643 | 67.3936 | <.0001 | 1.242 | 1.180 | 1.308 | Age 26-30 (at start) |
| *leeft3* | 1 | 0.36796 | 0.03971 | 85.8812 | <.0001 | 1.445 | 1.337 | 1.562 | Age 31-35 (at start) |
| *leeft4* | 1 | 0.33122 | 0.05947 | 31.0148 | <.0001 | 1.393 | 1.239 | 1.565 | Age 36-40 (at start) |
| *leeft5* | 1 | 0.59033 | 0.05685 | 107.8399 | <.0001 | 1.805 | 1.614 | 2.017 | Age 41+ (at start) |
| *start2* | 1 | -0.21846 | 0.02241 | 94.9948 | <.0001 | 0.804 | 0.769 | 0.840 | Start cohort 1997-2004 |
| *start3* | 1 | -0.70048 | 0.03097 | 511.5668 | <.0001 | 0.496 | 0.467 | 0.527 | Start cohort 2004-2009 |
| *natEurEU* | 1 | 0.46270 | 0.06316 | 53.6615 | <.0001 | 1.588 | 1.403 | 1.798 | EU (excl. Belgium) |
| *natAnd* | 1 | 0.40235 | 0.06845 | 34.5507 | <.0001 | 1.495 | 1.308 | 1.710 | International (non-EU) |
| *univ2* | 1 | 0.08800 | 0.03406 | 6.6741 | 0.0098 | 1.092 | 1.021 | 1.167 | |
| *univ3* | 1 | 0.04209 | 0.02391 | 3.0983 | 0.0784 | 1.043 | 0.995 | 1.093 | |
| *univ4* | 1 | -0.03376 | 0.03506 | 0.9270 | 0.3357 | 0.967 | 0.903 | 1.036 | |
| *univ5* | 1 | -0.36826 | 0.06329 | 33.8583 | <.0001 | 0.692 | 0.611 | 0.783 | |
| *natNBel_domstat2* | 1 | -0.04083 | 0.10466 | 0.1522 | 0.6964 | 0.960 | 0.782 | 1.179 | Interaction w. nationality |
| *natNBel_domstat3* | 1 | -0.20238 | 0.30729 | 0.4338 | 0.5101 | 0.817 | 0.447 | 1.492 | Interaction w. nationality |
| *natNBel_domstat4* | 1 | -0.22292 | 0.13978 | 2.5433 | 0.1108 | 0.800 | 0.608 | 1.052 | Interaction w. nationality |
| *natNBel_domstat5* | 1 | 0.29899 | 0.05807 | 26.5130 | <.0001 | 1.349 | 1.203 | 1.511 | Interaction w. nationality |
| *natNBel_domclusmed* | 1 | -0.31004 | 0.07622 | 16.5467 | <.0001 | 0.733 | 0.632 | 0.852 | Interaction w. nationality |
| *natNBel_domclushum* | 1 | -0.42799 | 0.09109 | 22.0767 | <.0001 | 0.652 | 0.545 | 0.779 | Interaction w. nationality |
| *natNBel_domclussoc* | 1 | -0.46276 | 0.08775 | 27.8122 | <.0001 | 0.630 | 0.530 | 0.748 | Interaction w. nationality |
| *natNBel_domclustoe* | 1 | -0.47863 | 0.07298 | 43.0118 | <.0001 | 0.620 | 0.537 | 0.715 | Interaction w. nationality |

***Sponsored time competing risks analysis*** ***6***
***Outcome of interest: Withdrawal***
***Imputation method A***

***The PHREG Procedure***

***_Imputation_=2***

*Linear Hypotheses Testing Results*

| *Label* | *Wald Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *gender* | 31.1482 | 1 | <.0001 |
| *domstat* | 3373.9073 | 4 | <.0001 |
| *domclus* | 261.9270 | 4 | <.0001 |
| *leeft* | 203.0278 | 4 | <.0001 |
| *start* | 512.0055 | 2 | <.0001 |
| *nat* | 54.1204 | 2 | <.0001 |
| *univ* | 52.7193 | 4 | <.0001 |
| *nat_domstat* | 35.7557 | 4 | <.0001 |
| *nat_domclus* | 52.5063 | 4 | <.0001 |

***Sponsored time competing risks analysis***
***Outcome of interest: Withdrawal***
***Imputation method A***

***The PHREG Procedure***

***_Imputation_=3***

| *Model Information* | |
|---|---|
| *Data Set* | LIB.IMPUTATIEA |
| *Dependent Variable* | sponsTime |
| *Censoring Variable* | cens2 |
| *Censoring Value(s)* | 0 |
| *Ties Handling* | DISCRETE |

| | |
|---|---|
| *Number of Observations Read* | 28871 |
| *Number of Observations Used* | 28396 |

*Summary of the Number of Event and Censored Values*

| *Total* | *Event* | *Censored* | *Percent Censored* |
|---|---|---|---|
| 28396 | 10472 | 17924 | 63.12 |

| *Convergence Status* |
|---|
| Convergence criterion (GCONV=1E-8) satisfied. |

*Model Fit Statistics*

| *Criterion* | *Without Covariates* | *With Covariates* |
|---|---|---|
| *-2 LOG L* | 157595.81 | 151444.46 |
| *AIC* | 157595.81 | 151502.46 |
| *SBC* | 157595.81 | 151712.90 |

*Testing Global Null Hypothesis: BETA=0*

| *Test* | *Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *Likelihood Ratio* | 6151.3499 | 29 | <.0001 |
| *Score* | 6585.2998 | 29 | <.0001 |
| *Wald* | 5548.7597 | 29 | <.0001 |

***Sponsored time competing risks analysis*** ***8***
***Outcome of interest: Withdrawal***
***Imputation method A***

***The PHREG Procedure***

***_Imputation_=3***

*Analysis of Maximum Likelihood Estimates*

| *Parameter* | *DF* | *Parameter Estimate* | *Standard Error* | *Chi-Square* | *Pr > ChiSq* | *Hazard Ratio* | *95% Hazard Ratio Confidence Limits* | | *Label* |
|---|---|---|---|---|---|---|---|---|---|
| *GESLACHT* | 1 | 0.11643 | 0.02032 | 32.8310 | <.0001 | 1.123 | 1.080 | 1.169 | Female |
| *domstat2* | 1 | -1.06070 | 0.02815 | 1419.6960 | <.0001 | 0.346 | 0.328 | 0.366 | Assistant |
| *domstat3* | 1 | -2.11841 | 0.04425 | 2291.8063 | <.0001 | 0.120 | 0.110 | 0.131 | Scholarship (Flanders) |
| *domstat4* | 1 | -1.58455 | 0.08886 | 317.9622 | <.0001 | 0.205 | 0.172 | 0.244 | Scholarship (univers.) |
| *domstat5* | 1 | -1.05788 | 0.02960 | 1276.9906 | <.0001 | 0.347 | 0.328 | 0.368 | Project (FWO/BOF/IUAP) |
| *domclusmed* | 1 | 0.17484 | 0.03822 | 20.9296 | <.0001 | 1.191 | 1.105 | 1.284 | Medical |
| *domclushum* | 1 | 0.43042 | 0.03944 | 119.0845 | <.0001 | 1.538 | 1.423 | 1.662 | Humanities |
| *domclussoc* | 1 | 0.51103 | 0.03654 | 195.5482 | <.0001 | 1.667 | 1.552 | 1.791 | Social |
| *domclustoe* | 1 | 0.20660 | 0.03840 | 28.9452 | <.0001 | 1.229 | 1.140 | 1.326 | Applied |
| *leeft2* | 1 | 0.21481 | 0.02647 | 65.8362 | <.0001 | 1.240 | 1.177 | 1.306 | Age 26-30 (at start) |
| *leeft3* | 1 | 0.37459 | 0.03966 | 89.2253 | <.0001 | 1.454 | 1.346 | 1.572 | Age 31-35 (at start) |
| *leeft4* | 1 | 0.34166 | 0.05932 | 33.1767 | <.0001 | 1.407 | 1.253 | 1.581 | Age 36-40 (at start) |
| *leeft5* | 1 | 0.59358 | 0.05685 | 108.9986 | <.0001 | 1.810 | 1.620 | 2.024 | Age 41+ (at start) |
| *start2* | 1 | -0.22134 | 0.02242 | 97.4470 | <.0001 | 0.801 | 0.767 | 0.837 | Start cohort 1997-2004 |
| *start3* | 1 | -0.70401 | 0.03099 | 516.0505 | <.0001 | 0.495 | 0.465 | 0.526 | Start cohort 2004-2009 |
| *natEurEU* | 1 | 0.47163 | 0.06300 | 56.0494 | <.0001 | 1.603 | 1.416 | 1.813 | EU (excl. Belgium) |
| *natAnd* | 1 | 0.40582 | 0.06833 | 35.2692 | <.0001 | 1.501 | 1.312 | 1.716 | International (non-EU) |
| *univ2* | 1 | 0.08192 | 0.03408 | 5.7773 | 0.0162 | 1.085 | 1.015 | 1.160 | |
| *univ3* | 1 | 0.03636 | 0.02391 | 2.3132 | 0.1283 | 1.037 | 0.990 | 1.087 | |
| *univ4* | 1 | -0.04874 | 0.03515 | 1.9227 | 0.1656 | 0.952 | 0.889 | 1.020 | |
| *univ5* | 1 | -0.36439 | 0.06299 | 33.4632 | <.0001 | 0.695 | 0.614 | 0.786 | |
| *natNBel_domstat2* | 1 | -0.04031 | 0.10465 | 0.1484 | 0.7001 | 0.960 | 0.782 | 1.179 | Interaction w. nationality |
| *natNBel_domstat3* | 1 | -0.13733 | 0.29466 | 0.2172 | 0.6412 | 0.872 | 0.489 | 1.553 | Interaction w. nationality |
| *natNBel_domstat4* | 1 | -0.24639 | 0.14187 | 3.0163 | 0.0824 | 0.782 | 0.592 | 1.032 | Interaction w. nationality |
| *natNBel_domstat5* | 1 | 0.27544 | 0.05819 | 22.4078 | <.0001 | 1.317 | 1.175 | 1.476 | Interaction w. nationality |
| *natNBel_domclusmed* | 1 | -0.30684 | 0.07612 | 16.2505 | <.0001 | 0.736 | 0.634 | 0.854 | Interaction w. nationality |
| *natNBel_domclushum* | 1 | -0.43590 | 0.09138 | 22.7542 | <.0001 | 0.647 | 0.541 | 0.774 | Interaction w. nationality |
| *natNBel_domclussoc* | 1 | -0.47733 | 0.08821 | 29.2840 | <.0001 | 0.620 | 0.522 | 0.738 | Interaction w. nationality |
| *natNBel_domclustoe* | 1 | -0.45631 | 0.07287 | 39.2140 | <.0001 | 0.634 | 0.549 | 0.731 | Interaction w. nationality |

***Sponsored time competing risks analysis*** ***9***
***Outcome of interest: Withdrawal***
***Imputation method A***

***The PHREG Procedure***

***_Imputation_=3***

*Linear Hypotheses Testing Results*

| *Label* | *Wald Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *gender* | 32.8310 | 1 | <.0001 |
| *domstat* | 3373.8118 | 4 | <.0001 |
| *domclus* | 258.6799 | 4 | <.0001 |
| *leeft* | 206.6288 | 4 | <.0001 |
| *start* | 516.3605 | 2 | <.0001 |
| *nat* | 56.4166 | 2 | <.0001 |
| *univ* | 51.5982 | 4 | <.0001 |
| *nat_domstat* | 31.2369 | 4 | <.0001 |
| *nat_domclus* | 51.0003 | 4 | <.0001 |

***Sponsored time competing risks analysis*** ***10***
***Outcome of interest: Withdrawal***
***Imputation method A***

***The PHREG Procedure***

***_Imputation_=4***

| *Model Information* | |
|---|---|
| *Data Set* | LIB.IMPUTATIEA |
| *Dependent Variable* | sponsTime |
| *Censoring Variable* | cens2 |
| *Censoring Value(s)* | 0 |
| *Ties Handling* | DISCRETE |

| | |
|---|---|
| *Number of Observations Read* | 28871 |
| *Number of Observations Used* | 28396 |

*Summary of the Number of Event and Censored Values*

| *Total* | *Event* | *Censored* | *Percent Censored* |
|---|---|---|---|
| 28396 | 10460 | 17936 | 63.16 |

| *Convergence Status* |
|---|
| Convergence criterion (GCONV=1E-8) satisfied. |

*Model Fit Statistics*

| *Criterion* | *Without Covariates* | *With Covariates* |
|---|---|---|
| *-2 LOG L* | 157447.57 | 151300.90 |
| *AIC* | 157447.57 | 151358.90 |
| *SBC* | 157447.57 | 151569.30 |

*Testing Global Null Hypothesis: BETA=0*

| *Test* | *Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *Likelihood Ratio* | 6146.6737 | 29 | <.0001 |
| *Score* | 6568.0016 | 29 | <.0001 |
| *Wald* | 5535.3086 | 29 | <.0001 |

***Sponsored time competing risks analysis*** ***11***
***Outcome of interest: Withdrawal***
***Imputation method A***

***The PHREG Procedure***

***_Imputation_=4***

*Analysis of Maximum Likelihood Estimates*

| *Parameter* | *DF* | *Parameter Estimate* | *Standard Error* | *Chi-Square* | *Pr > ChiSq* | *Hazard Ratio* | *95% Hazard Ratio Confidence Limits* | | *Label* |
|---|---|---|---|---|---|---|---|---|---|
| *GESLACHT* | 1 | 0.10708 | 0.02034 | 27.7248 | <.0001 | 1.113 | 1.070 | 1.158 | Female |
| *domstat2* | 1 | -1.06009 | 0.02818 | 1415.2436 | <.0001 | 0.346 | 0.328 | 0.366 | Assistant |
| *domstat3* | 1 | -2.12217 | 0.04435 | 2290.0370 | <.0001 | 0.120 | 0.110 | 0.131 | Scholarship (Flanders) |
| *domstat4* | 1 | -1.56927 | 0.08826 | 316.1671 | <.0001 | 0.208 | 0.175 | 0.248 | Scholarship (univers.) |
| *domstat5* | 1 | -1.05118 | 0.02959 | 1262.2434 | <.0001 | 0.350 | 0.330 | 0.370 | Project (FWO/BOF/IUAP) |
| *domclusmed* | 1 | 0.17559 | 0.03828 | 21.0444 | <.0001 | 1.192 | 1.106 | 1.285 | Medical |
| *domclushum* | 1 | 0.43291 | 0.03948 | 120.2637 | <.0001 | 1.542 | 1.427 | 1.666 | Humanities |
| *domclussoc* | 1 | 0.50869 | 0.03662 | 192.9900 | <.0001 | 1.663 | 1.548 | 1.787 | Social |
| *domclustoe* | 1 | 0.21854 | 0.03836 | 32.4528 | <.0001 | 1.244 | 1.154 | 1.341 | Applied |
| *leeft2* | 1 | 0.21265 | 0.02650 | 64.4082 | <.0001 | 1.237 | 1.174 | 1.303 | Age 26-30 (at start) |
| *leeft3* | 1 | 0.36887 | 0.03974 | 86.1685 | <.0001 | 1.446 | 1.338 | 1.563 | Age 31-35 (at start) |
| *leeft4* | 1 | 0.34380 | 0.05923 | 33.6878 | <.0001 | 1.410 | 1.256 | 1.584 | Age 36-40 (at start) |
| *leeft5* | 1 | 0.59211 | 0.05686 | 108.4381 | <.0001 | 1.808 | 1.617 | 2.021 | Age 41+ (at start) |
| *start2* | 1 | -0.22092 | 0.02241 | 97.1824 | <.0001 | 0.802 | 0.767 | 0.838 | Start cohort 1997-2004 |
| *start3* | 1 | -0.71787 | 0.03112 | 532.0325 | <.0001 | 0.488 | 0.459 | 0.518 | Start cohort 2004-2009 |
| *natEurEU* | 1 | 0.47724 | 0.06308 | 57.2318 | <.0001 | 1.612 | 1.424 | 1.824 | EU (excl. Belgium) |
| *natAnd* | 1 | 0.40453 | 0.06849 | 34.8811 | <.0001 | 1.499 | 1.310 | 1.714 | International (non-EU) |
| *univ2* | 1 | 0.08797 | 0.03399 | 6.6975 | 0.0097 | 1.092 | 1.022 | 1.167 | |
| *univ3* | 1 | 0.03028 | 0.02392 | 1.6030 | 0.2055 | 1.031 | 0.984 | 1.080 | |
| *univ4* | 1 | -0.04600 | 0.03514 | 1.7135 | 0.1905 | 0.955 | 0.891 | 1.023 | |
| *univ5* | 1 | -0.38890 | 0.06368 | 37.2945 | <.0001 | 0.678 | 0.598 | 0.768 | |
| *natNBel_domstat2* | 1 | -0.05138 | 0.10465 | 0.2410 | 0.6235 | 0.950 | 0.774 | 1.166 | Interaction w. nationality |
| *natNBel_domstat3* | 1 | -0.13671 | 0.29467 | 0.2152 | 0.6427 | 0.872 | 0.490 | 1.554 | Interaction w. nationality |
| *natNBel_domstat4* | 1 | -0.28073 | 0.14231 | 3.8912 | 0.0485 | 0.755 | 0.571 | 0.998 | Interaction w. nationality |
| *natNBel_domstat5* | 1 | 0.27350 | 0.05814 | 22.1309 | <.0001 | 1.315 | 1.173 | 1.473 | Interaction w. nationality |
| *natNBel_domclusmed* | 1 | -0.30526 | 0.07627 | 16.0183 | <.0001 | 0.737 | 0.635 | 0.856 | Interaction w. nationality |
| *natNBel_domclushum* | 1 | -0.43024 | 0.09142 | 22.1502 | <.0001 | 0.650 | 0.544 | 0.778 | Interaction w. nationality |
| *natNBel_domclussoc* | 1 | -0.45411 | 0.08798 | 26.6403 | <.0001 | 0.635 | 0.534 | 0.755 | Interaction w. nationality |
| *natNBel_domclustoe* | 1 | -0.46197 | 0.07290 | 40.1599 | <.0001 | 0.630 | 0.546 | 0.727 | Interaction w. nationality |

**The PHREG Procedure**

**_Imputation_=4**

*Linear Hypotheses Testing Results*

| Label | Wald Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| gender | 27.7248 | 1 | <.0001 |
| domstat | 3358.4534 | 4 | <.0001 |
| domclus | 253.8228 | 4 | <.0001 |
| leeft | 203.5142 | 4 | <.0001 |
| start | 532.5174 | 2 | <.0001 |
| nat | 57.4485 | 2 | <.0001 |
| univ | 55.3436 | 4 | <.0001 |
| nat_domstat | 32.4429 | 4 | <.0001 |
| nat_domclus | 49.9891 | 4 | <.0001 |

***Sponsored time competing risks analysis***
***Outcome of interest: Withdrawal***
***Imputation method A***

***The PHREG Procedure***

***_Imputation_=5***

| *Model Information* | |
|---|---|
| *Data Set* | LIB.IMPUTATIEA |
| *Dependent Variable* | sponsTime |
| *Censoring Variable* | cens2 |
| *Censoring Value(s)* | 0 |
| *Ties Handling* | DISCRETE |

| | |
|---|---|
| *Number of Observations Read* | 28871 |
| *Number of Observations Used* | 28396 |

*Summary of the Number of Event and Censored Values*

| *Total* | *Event* | *Censored* | *Percent Censored* |
|---|---|---|---|
| 28396 | 10428 | 17968 | 63.28 |

| *Convergence Status* |
|---|
| Convergence criterion (GCONV=1E-8) satisfied. |

*Model Fit Statistics*

| *Criterion* | *Without Covariates* | *With Covariates* |
|---|---|---|
| *-2 LOG L* | 156956.82 | 150842.40 |
| *AIC* | 156956.82 | 150900.40 |
| *SBC* | 156956.82 | 151110.72 |

*Testing Global Null Hypothesis: BETA=0*

| *Test* | *Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *Likelihood Ratio* | 6114.4164 | 29 | <.0001 |
| *Score* | 6534.8025 | 29 | <.0001 |
| *Wald* | 5513.6529 | 29 | <.0001 |

***Sponsored time competing risks analysis*** ***14***
***Outcome of interest: Withdrawal***
***Imputation method A***

***The PHREG Procedure***

***_Imputation_=5***

*Analysis of Maximum Likelihood Estimates*

| *Parameter* | *DF* | *Parameter Estimate* | *Standard Error* | *Chi-Square* | *Pr > ChiSq* | *Hazard Ratio* | *95% Hazard Ratio Confidence Limits* | | *Label* |
|---|---|---|---|---|---|---|---|---|---|
| *GESLACHT* | 1 | 0.10918 | 0.02037 | 28.7372 | <.0001 | 1.115 | 1.072 | 1.161 | Female |
| *domstat2* | 1 | -1.05779 | 0.02819 | 1408.1257 | <.0001 | 0.347 | 0.329 | 0.367 | Assistant |
| *domstat3* | 1 | -2.11491 | 0.04427 | 2282.4636 | <.0001 | 0.121 | 0.111 | 0.132 | Scholarship (Flanders) |
| *domstat4* | 1 | -1.54880 | 0.08767 | 312.1128 | <.0001 | 0.213 | 0.179 | 0.252 | Scholarship (univers.) |
| *domstat5* | 1 | -1.05748 | 0.02968 | 1269.5086 | <.0001 | 0.347 | 0.328 | 0.368 | Project (FWO/BOF/IUAP) |
| *domclusmed* | 1 | 0.19084 | 0.03831 | 24.8133 | <.0001 | 1.210 | 1.123 | 1.305 | Medical |
| *domclushum* | 1 | 0.43666 | 0.03960 | 121.6136 | <.0001 | 1.548 | 1.432 | 1.672 | Humanities |
| *domclussoc* | 1 | 0.51338 | 0.03672 | 195.4448 | <.0001 | 1.671 | 1.555 | 1.796 | Social |
| *domclustoe* | 1 | 0.22342 | 0.03848 | 33.7145 | <.0001 | 1.250 | 1.160 | 1.348 | Applied |
| *leeft2* | 1 | 0.21714 | 0.02650 | 67.1606 | <.0001 | 1.243 | 1.180 | 1.309 | Age 26-30 (at start) |
| *leeft3* | 1 | 0.35444 | 0.04003 | 78.3923 | <.0001 | 1.425 | 1.318 | 1.542 | Age 31-35 (at start) |
| *leeft4* | 1 | 0.35039 | 0.05915 | 35.0862 | <.0001 | 1.420 | 1.264 | 1.594 | Age 36-40 (at start) |
| *leeft5* | 1 | 0.58528 | 0.05709 | 105.1005 | <.0001 | 1.795 | 1.605 | 2.008 | Age 41+ (at start) |
| *start2* | 1 | -0.22297 | 0.02243 | 98.8002 | <.0001 | 0.800 | 0.766 | 0.836 | Start cohort 1997-2004 |
| *start3* | 1 | -0.72271 | 0.03119 | 536.7618 | <.0001 | 0.485 | 0.457 | 0.516 | Start cohort 2004-2009 |
| *natEurEU* | 1 | 0.48530 | 0.06325 | 58.8687 | <.0001 | 1.625 | 1.435 | 1.839 | EU (excl. Belgium) |
| *natAnd* | 1 | 0.40453 | 0.06876 | 34.6170 | <.0001 | 1.499 | 1.310 | 1.715 | International (non-EU) |
| *univ2* | 1 | 0.08147 | 0.03414 | 5.6932 | 0.0170 | 1.085 | 1.015 | 1.160 | |
| *univ3* | 1 | 0.03473 | 0.02396 | 2.1018 | 0.1471 | 1.035 | 0.988 | 1.085 | |
| *univ4* | 1 | -0.04026 | 0.03516 | 1.3112 | 0.2522 | 0.961 | 0.897 | 1.029 | |
| *univ5* | 1 | -0.37150 | 0.06340 | 34.3406 | <.0001 | 0.690 | 0.609 | 0.781 | |
| *natNBel_domstat2* | 1 | -0.03915 | 0.10437 | 0.1407 | 0.7076 | 0.962 | 0.784 | 1.180 | Interaction w. nationality |
| *natNBel_domstat3* | 1 | -0.14161 | 0.29467 | 0.2310 | 0.6308 | 0.868 | 0.487 | 1.546 | Interaction w. nationality |
| *natNBel_domstat4* | 1 | -0.31224 | 0.14284 | 4.7783 | 0.0288 | 0.732 | 0.553 | 0.968 | Interaction w. nationality |
| *natNBel_domstat5* | 1 | 0.27617 | 0.05838 | 22.3798 | <.0001 | 1.318 | 1.176 | 1.478 | Interaction w. nationality |
| *natNBel_domclusmed* | 1 | -0.32562 | 0.07655 | 18.0931 | <.0001 | 0.722 | 0.621 | 0.839 | Interaction w. nationality |
| *natNBel_domclushum* | 1 | -0.43815 | 0.09177 | 22.7928 | <.0001 | 0.645 | 0.539 | 0.772 | Interaction w. nationality |
| *natNBel_domclussoc* | 1 | -0.46137 | 0.08821 | 27.3551 | <.0001 | 0.630 | 0.530 | 0.749 | Interaction w. nationality |
| *natNBel_domclustoe* | 1 | -0.46736 | 0.07313 | 40.8395 | <.0001 | 0.627 | 0.543 | 0.723 | Interaction w. nationality |

***The PHREG Procedure***

***_Imputation_=5***

*Linear Hypotheses Testing Results*

| *Label* | *Wald Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *gender* | 28.7372 | 1 | <.0001 |
| *domstat* | 3348.0425 | 4 | <.0001 |
| *domclus* | 251.3495 | 4 | <.0001 |
| *leeft* | 198.0135 | 4 | <.0001 |
| *start* | 537.1575 | 2 | <.0001 |
| *nat* | 58.9720 | 2 | <.0001 |
| *univ* | 51.0888 | 4 | <.0001 |
| *nat_domstat* | 33.6535 | 4 | <.0001 |
| *nat_domclus* | 51.0565 | 4 | <.0001 |

***Sponsored time competing risks analysis*** ***1***
***Outcome of interest: Withdrawal***
***Imputation method B***

***The PHREG Procedure***

***_Imputation_=1***

| *Model Information* | |
|---|---|
| *Data Set* | LIB.IMPUTATIEB |
| *Dependent Variable* | sponsTime |
| *Censoring Variable* | cens2 |
| *Censoring Value(s)* | 0 |
| *Ties Handling* | DISCRETE |

| | |
|---|---|
| *Number of Observations Read* | 28871 |
| *Number of Observations Used* | 28396 |

*Summary of the Number of Event and Censored Values*

| *Total* | *Event* | *Censored* | *Percent Censored* |
|---|---|---|---|
| 28396 | 10435 | 17961 | 63.25 |

| *Convergence Status* |
|---|
| Convergence criterion (GCONV=1E-8) satisfied. |

*Model Fit Statistics*

| *Criterion* | *Without Covariates* | *With Covariates* |
|---|---|---|
| *-2 LOG L* | 157245.09 | 150921.93 |
| *AIC* | 157245.09 | 150979.93 |
| *SBC* | 157245.09 | 151190.26 |

*Testing Global Null Hypothesis: BETA=0*

| *Test* | *Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *Likelihood Ratio* | 6323.1625 | 29 | <.0001 |
| *Score* | 6762.5823 | 29 | <.0001 |
| *Wald* | 5684.8835 | 29 | <.0001 |

***Sponsored time competing risks analysis*** ***2***
***Outcome of interest: Withdrawal***
***Imputation method B***

***The PHREG Procedure***

***_Imputation_=1***

*Analysis of Maximum Likelihood Estimates*

| *Parameter* | *DF* | *Parameter Estimate* | *Standard Error* | *Chi-Square* | *Pr > ChiSq* | *Hazard Ratio* | *95% Hazard Ratio Confidence Limits* | | *Label* |
|---|---|---|---|---|---|---|---|---|---|
| *GESLACHT* | 1 | 0.10105 | 0.02036 | 24.6261 | <.0001 | 1.106 | 1.063 | 1.151 | Female |
| *domstat2* | 1 | -1.06822 | 0.02813 | 1442.3240 | <.0001 | 0.344 | 0.325 | 0.363 | Assistant |
| *domstat3* | 1 | -2.14734 | 0.04470 | 2307.2665 | <.0001 | 0.117 | 0.107 | 0.127 | Scholarship (Flanders) |
| *domstat4* | 1 | -1.58419 | 0.08887 | 317.7784 | <.0001 | 0.205 | 0.172 | 0.244 | Scholarship (univers.) |
| *domstat5* | 1 | -1.07543 | 0.02970 | 1310.7265 | <.0001 | 0.341 | 0.322 | 0.362 | Project (FWO/BOF/IUAP) |
| *domclusmed* | 1 | 0.19359 | 0.03851 | 25.2750 | <.0001 | 1.214 | 1.125 | 1.309 | Medical |
| *domclushum* | 1 | 0.44904 | 0.03970 | 127.9399 | <.0001 | 1.567 | 1.450 | 1.694 | Humanities |
| *domclussoc* | 1 | 0.53507 | 0.03676 | 211.9241 | <.0001 | 1.708 | 1.589 | 1.835 | Social |
| *domclustoe* | 1 | 0.23588 | 0.03857 | 37.4097 | <.0001 | 1.266 | 1.174 | 1.365 | Applied |
| *leeft2* | 1 | 0.21012 | 0.02654 | 62.6938 | <.0001 | 1.234 | 1.171 | 1.300 | Age 26-30 (at start) |
| *leeft3* | 1 | 0.37693 | 0.03963 | 90.4551 | <.0001 | 1.458 | 1.349 | 1.576 | Age 31-35 (at start) |
| *leeft4* | 1 | 0.32979 | 0.05957 | 30.6488 | <.0001 | 1.391 | 1.237 | 1.563 | Age 36-40 (at start) |
| *leeft5* | 1 | 0.59596 | 0.05672 | 110.3845 | <.0001 | 1.815 | 1.624 | 2.028 | Age 41+ (at start) |
| *start2* | 1 | -0.21248 | 0.02238 | 90.1681 | <.0001 | 0.809 | 0.774 | 0.845 | Start cohort 1997-2004 |
| *start3* | 1 | -0.75735 | 0.03150 | 578.2392 | <.0001 | 0.469 | 0.441 | 0.499 | Start cohort 2004-2009 |
| *natEurEU* | 1 | 0.49551 | 0.06316 | 61.5457 | <.0001 | 1.641 | 1.450 | 1.858 | EU (excl. Belgium) |
| *natAnd* | 1 | 0.41506 | 0.06869 | 36.5142 | <.0001 | 1.514 | 1.324 | 1.733 | International (non-EU) |
| *univ2* | 1 | 0.08985 | 0.03419 | 6.9046 | 0.0086 | 1.094 | 1.023 | 1.170 | |
| *univ3* | 1 | 0.04635 | 0.02398 | 3.7358 | 0.0533 | 1.047 | 0.999 | 1.098 | |
| *univ4* | 1 | -0.03437 | 0.03521 | 0.9533 | 0.3289 | 0.966 | 0.902 | 1.035 | |
| *univ5* | 1 | -0.34227 | 0.06273 | 29.7677 | <.0001 | 0.710 | 0.628 | 0.803 | |
| *natNBel_domstat2* | 1 | -0.07579 | 0.10569 | 0.5142 | 0.4733 | 0.927 | 0.754 | 1.140 | Interaction w. nationality |
| *natNBel_domstat3* | 1 | -0.11231 | 0.29473 | 0.1452 | 0.7032 | 0.894 | 0.502 | 1.593 | Interaction w. nationality |
| *natNBel_domstat4* | 1 | -0.27950 | 0.14315 | 3.8126 | 0.0509 | 0.756 | 0.571 | 1.001 | Interaction w. nationality |
| *natNBel_domstat5* | 1 | 0.28009 | 0.05838 | 23.0140 | <.0001 | 1.323 | 1.180 | 1.484 | Interaction w. nationality |
| *natNBel_domclusmed* | 1 | -0.32628 | 0.07655 | 18.1648 | <.0001 | 0.722 | 0.621 | 0.838 | Interaction w. nationality |
| *natNBel_domclushum* | 1 | -0.45386 | 0.09187 | 24.4043 | <.0001 | 0.635 | 0.531 | 0.760 | Interaction w. nationality |
| *natNBel_domclussoc* | 1 | -0.47521 | 0.08806 | 29.1252 | <.0001 | 0.622 | 0.523 | 0.739 | Interaction w. nationality |
| *natNBel_domclustoe* | 1 | -0.48623 | 0.07314 | 44.1907 | <.0001 | 0.615 | 0.533 | 0.710 | Interaction w. nationality |

***Sponsored time competing risks analysis*** ***3***
***Outcome of interest: Withdrawal***
***Imputation method B***

***The PHREG Procedure***

***_Imputation_=1***

*Linear Hypotheses Testing Results*

| *Label* | *Wald Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *gender* | 24.6261 | 1 | <.0001 |
| *domstat* | 3427.5501 | 4 | <.0001 |
| *domclus* | 272.2641 | 4 | <.0001 |
| *leeft* | 205.4185 | 4 | <.0001 |
| *start* | 580.4095 | 2 | <.0001 |
| *nat* | 61.6847 | 2 | <.0001 |
| *univ* | 49.4924 | 4 | <.0001 |
| *nat_domstat* | 33.8503 | 4 | <.0001 |
| *nat_domclus* | 54.9004 | 4 | <.0001 |

***Sponsored time competing risks analysis***
***Outcome of interest: Withdrawal***
***Imputation method B***

***The PHREG Procedure***

***_Imputation_=2***

| *Model Information* | |
|---|---|
| *Data Set* | LIB.IMPUTATIEB |
| *Dependent Variable* | sponsTime |
| *Censoring Variable* | cens2 |
| *Censoring Value(s)* | 0 |
| *Ties Handling* | DISCRETE |

| | |
|---|---|
| *Number of Observations Read* | 28871 |
| *Number of Observations Used* | 28396 |

*Summary of the Number of Event and Censored Values*

| *Total* | *Event* | *Censored* | *Percent Censored* |
|---|---|---|---|
| 28396 | 10442 | 17954 | 63.23 |

| *Convergence Status* |
|---|
| Convergence criterion (GCONV=1E-8) satisfied. |

*Model Fit Statistics*

| *Criterion* | *Without Covariates* | *With Covariates* |
|---|---|---|
| *-2 LOG L* | 157307.98 | 150929.80 |
| *AIC* | 157307.98 | 150987.80 |
| *SBC* | 157307.98 | 151198.15 |

*Testing Global Null Hypothesis: BETA=0*

| *Test* | *Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *Likelihood Ratio* | 6378.1862 | 29 | <.0001 |
| *Score* | 6817.7898 | 29 | <.0001 |
| *Wald* | 5715.5785 | 29 | <.0001 |

***Sponsored time competing risks analysis*** ***5***
***Outcome of interest: Withdrawal***
***Imputation method B***

***The PHREG Procedure***

***_Imputation_=2***

*Analysis of Maximum Likelihood Estimates*

| *Parameter* | *DF* | *Parameter Estimate* | *Standard Error* | *Chi-Square* | *Pr > ChiSq* | *Hazard Ratio* | *95% Hazard Ratio Confidence Limits* | | *Label* |
|---|---|---|---|---|---|---|---|---|---|
| *GESLACHT* | 1 | 0.10311 | 0.02035 | 25.6607 | <.0001 | 1.109 | 1.065 | 1.154 | Female |
| *domstat2* | 1 | -1.07896 | 0.02811 | 1473.1261 | <.0001 | 0.340 | 0.322 | 0.359 | Assistant |
| *domstat3* | 1 | -2.16631 | 0.04486 | 2332.4683 | <.0001 | 0.115 | 0.105 | 0.125 | Scholarship (Flanders) |
| *domstat4* | 1 | -1.64765 | 0.09110 | 327.1132 | <.0001 | 0.193 | 0.161 | 0.230 | Scholarship (univers.) |
| *domstat5* | 1 | -1.08133 | 0.02965 | 1330.3995 | <.0001 | 0.339 | 0.320 | 0.359 | Project (FWO/BOF/IUAP) |
| *domclusmed* | 1 | 0.18988 | 0.03841 | 24.4341 | <.0001 | 1.209 | 1.121 | 1.304 | Medical |
| *domclushum* | 1 | 0.44636 | 0.03961 | 127.0202 | <.0001 | 1.563 | 1.446 | 1.689 | Humanities |
| *domclussoc* | 1 | 0.52451 | 0.03670 | 204.2229 | <.0001 | 1.690 | 1.572 | 1.816 | Social |
| *domclustoe* | 1 | 0.22473 | 0.03850 | 34.0673 | <.0001 | 1.252 | 1.161 | 1.350 | Applied |
| *leeft2* | 1 | 0.20569 | 0.02653 | 60.1081 | <.0001 | 1.228 | 1.166 | 1.294 | Age 26-30 (at start) |
| *leeft3* | 1 | 0.36332 | 0.03976 | 83.4887 | <.0001 | 1.438 | 1.330 | 1.555 | Age 31-35 (at start) |
| *leeft4* | 1 | 0.33073 | 0.05947 | 30.9264 | <.0001 | 1.392 | 1.239 | 1.564 | Age 36-40 (at start) |
| *leeft5* | 1 | 0.58672 | 0.05678 | 106.7612 | <.0001 | 1.798 | 1.609 | 2.010 | Age 41+ (at start) |
| *start2* | 1 | -0.21364 | 0.02237 | 91.2023 | <.0001 | 0.808 | 0.773 | 0.844 | Start cohort 1997-2004 |
| *start3* | 1 | -0.75244 | 0.03143 | 573.2973 | <.0001 | 0.471 | 0.443 | 0.501 | Start cohort 2004-2009 |
| *natEurEU* | 1 | 0.49069 | 0.06307 | 60.5305 | <.0001 | 1.633 | 1.444 | 1.848 | EU (excl. Belgium) |
| *natAnd* | 1 | 0.39803 | 0.06871 | 33.5558 | <.0001 | 1.489 | 1.301 | 1.704 | International (non-EU) |
| *univ2* | 1 | 0.08454 | 0.03420 | 6.1101 | 0.0134 | 1.088 | 1.018 | 1.164 | |
| *univ3* | 1 | 0.04100 | 0.02394 | 2.9333 | 0.0868 | 1.042 | 0.994 | 1.092 | |
| *univ4* | 1 | -0.04615 | 0.03522 | 1.7174 | 0.1900 | 0.955 | 0.891 | 1.023 | |
| *univ5* | 1 | -0.36249 | 0.06301 | 33.1003 | <.0001 | 0.696 | 0.615 | 0.787 | |
| *natNBel_domstat2* | 1 | -0.06104 | 0.10533 | 0.3358 | 0.5622 | 0.941 | 0.765 | 1.157 | Interaction w. nationality |
| *natNBel_domstat3* | 1 | -0.18523 | 0.30733 | 0.3633 | 0.5467 | 0.831 | 0.455 | 1.518 | Interaction w. nationality |
| *natNBel_domstat4* | 1 | -0.26726 | 0.14675 | 3.3167 | 0.0686 | 0.765 | 0.574 | 1.021 | Interaction w. nationality |
| *natNBel_domstat5* | 1 | 0.29543 | 0.05828 | 25.6935 | <.0001 | 1.344 | 1.199 | 1.506 | Interaction w. nationality |
| *natNBel_domclusmed* | 1 | -0.34154 | 0.07675 | 19.8036 | <.0001 | 0.711 | 0.611 | 0.826 | Interaction w. nationality |
| *natNBel_domclushum* | 1 | -0.44951 | 0.09179 | 23.9840 | <.0001 | 0.638 | 0.533 | 0.764 | Interaction w. nationality |
| *natNBel_domclussoc* | 1 | -0.44917 | 0.08762 | 26.2788 | <.0001 | 0.638 | 0.537 | 0.758 | Interaction w. nationality |
| *natNBel_domclustoe* | 1 | -0.47807 | 0.07311 | 42.7608 | <.0001 | 0.620 | 0.537 | 0.715 | Interaction w. nationality |

***Sponsored time competing risks analysis*** ***6***
***Outcome of interest: Withdrawal***
***Imputation method B***

***The PHREG Procedure***

***_Imputation_=2***

*Linear Hypotheses Testing Results*

| *Label* | *Wald Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *gender* | 25.6607 | 1 | <.0001 |
| *domstat* | 3492.0856 | 4 | <.0001 |
| *domclus* | 265.8042 | 4 | <.0001 |
| *leeft* | 196.6077 | 4 | <.0001 |
| *start* | 575.1761 | 2 | <.0001 |
| *nat* | 60.5377 | 2 | <.0001 |
| *univ* | 52.3623 | 4 | <.0001 |
| *nat_domstat* | 36.1110 | 4 | <.0001 |
| *nat_domclus* | 52.4825 | 4 | <.0001 |

***Sponsored time competing risks analysis*** 
***Outcome of interest: Withdrawal***
***Imputation method B***

***The PHREG Procedure***

***_Imputation_=3***

| *Model Information* | |
|---|---|
| *Data Set* | LIB.IMPUTATIEB |
| *Dependent Variable* | sponsTime |
| *Censoring Variable* | cens2 |
| *Censoring Value(s)* | 0 |
| *Ties Handling* | DISCRETE |

| | |
|---|---|
| *Number of Observations Read* | 28871 |
| *Number of Observations Used* | 28396 |

*Summary of the Number of Event and Censored Values*

| *Total* | *Event* | *Censored* | *Percent Censored* |
|---|---|---|---|
| 28396 | 10472 | 17924 | 63.12 |

| *Convergence Status* |
|---|
| Convergence criterion (GCONV=1E-8) satisfied. |

*Model Fit Statistics*

| *Criterion* | *Without Covariates* | *With Covariates* |
|---|---|---|
| *-2 LOG L* | 157734.27 | 151449.24 |
| *AIC* | 157734.27 | 151507.24 |
| *SBC* | 157734.27 | 151717.68 |

*Testing Global Null Hypothesis: BETA=0*

| *Test* | *Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *Likelihood Ratio* | 6285.0250 | 29 | <.0001 |
| *Score* | 6704.4848 | 29 | <.0001 |
| *Wald* | 5633.5952 | 29 | <.0001 |

***Sponsored time competing risks analysis*** ***8***
***Outcome of interest: Withdrawal***
***Imputation method B***

***The PHREG Procedure***

***_Imputation_=3***

*Analysis of Maximum Likelihood Estimates*

| *Parameter* | *DF* | *Parameter Estimate* | *Standard Error* | *Chi-Square* | *Pr > ChiSq* | *Hazard Ratio* | *95% Hazard Ratio Confidence Limits* | | *Label* |
|---|---|---|---|---|---|---|---|---|---|
| *GESLACHT* | 1 | 0.09445 | 0.02033 | 21.5855 | <.0001 | 1.099 | 1.056 | 1.144 | Female |
| *domstat2* | 1 | -1.06174 | 0.02808 | 1430.1977 | <.0001 | 0.346 | 0.327 | 0.365 | Assistant |
| *domstat3* | 1 | -2.15087 | 0.04476 | 2309.3762 | <.0001 | 0.116 | 0.107 | 0.127 | Scholarship (Flanders) |
| *domstat4* | 1 | -1.59546 | 0.08917 | 320.1481 | <.0001 | 0.203 | 0.170 | 0.242 | Scholarship (univers.) |
| *domstat5* | 1 | -1.06362 | 0.02958 | 1293.0987 | <.0001 | 0.345 | 0.326 | 0.366 | Project (FWO/BOF/IUAP) |
| *domclusmed* | 1 | 0.18763 | 0.03839 | 23.8908 | <.0001 | 1.206 | 1.119 | 1.301 | Medical |
| *domclushum* | 1 | 0.45009 | 0.03951 | 129.7821 | <.0001 | 1.568 | 1.452 | 1.695 | Humanities |
| *domclussoc* | 1 | 0.52573 | 0.03666 | 205.6950 | <.0001 | 1.692 | 1.574 | 1.818 | Social |
| *domclustoe* | 1 | 0.22663 | 0.03846 | 34.7241 | <.0001 | 1.254 | 1.163 | 1.353 | Applied |
| *leeft2* | 1 | 0.22032 | 0.02642 | 69.5422 | <.0001 | 1.246 | 1.184 | 1.313 | Age 26-30 (at start) |
| *leeft3* | 1 | 0.37211 | 0.03968 | 87.9236 | <.0001 | 1.451 | 1.342 | 1.568 | Age 31-35 (at start) |
| *leeft4* | 1 | 0.33061 | 0.05955 | 30.8200 | <.0001 | 1.392 | 1.238 | 1.564 | Age 36-40 (at start) |
| *leeft5* | 1 | 0.59225 | 0.05678 | 108.7979 | <.0001 | 1.808 | 1.618 | 2.021 | Age 41+ (at start) |
| *start2* | 1 | -0.20928 | 0.02236 | 87.5713 | <.0001 | 0.811 | 0.776 | 0.848 | Start cohort 1997-2004 |
| *start3* | 1 | -0.74077 | 0.03135 | 558.3342 | <.0001 | 0.477 | 0.448 | 0.507 | Start cohort 2004-2009 |
| *natEurEU* | 1 | 0.47675 | 0.06327 | 56.7745 | <.0001 | 1.611 | 1.423 | 1.824 | EU (excl. Belgium) |
| *natAnd* | 1 | 0.39314 | 0.06878 | 32.6722 | <.0001 | 1.482 | 1.295 | 1.695 | International (non-EU) |
| *univ2* | 1 | 0.09013 | 0.03407 | 6.9971 | 0.0082 | 1.094 | 1.024 | 1.170 | |
| *univ3* | 1 | 0.04296 | 0.02392 | 3.2258 | 0.0725 | 1.044 | 0.996 | 1.094 | |
| *univ4* | 1 | -0.04616 | 0.03524 | 1.7152 | 0.1903 | 0.955 | 0.891 | 1.023 | |
| *univ5* | 1 | -0.34833 | 0.06271 | 30.8508 | <.0001 | 0.706 | 0.624 | 0.798 | |
| *natNBel_domstat2* | 1 | -0.09651 | 0.10638 | 0.8231 | 0.3643 | 0.908 | 0.737 | 1.119 | Interaction w. nationality |
| *natNBel_domstat3* | 1 | -0.10955 | 0.29473 | 0.1382 | 0.7101 | 0.896 | 0.503 | 1.597 | Interaction w. nationality |
| *natNBel_domstat4* | 1 | -0.24959 | 0.14250 | 3.0678 | 0.0799 | 0.779 | 0.589 | 1.030 | Interaction w. nationality |
| *natNBel_domstat5* | 1 | 0.26501 | 0.05834 | 20.6324 | <.0001 | 1.303 | 1.163 | 1.461 | Interaction w. nationality |
| *natNBel_domclusmed* | 1 | -0.31895 | 0.07680 | 17.2487 | <.0001 | 0.727 | 0.625 | 0.845 | Interaction w. nationality |
| *natNBel_domclushum* | 1 | -0.43125 | 0.09169 | 22.1211 | <.0001 | 0.650 | 0.543 | 0.778 | Interaction w. nationality |
| *natNBel_domclussoc* | 1 | -0.45994 | 0.08832 | 27.1207 | <.0001 | 0.631 | 0.531 | 0.751 | Interaction w. nationality |
| *natNBel_domclustoe* | 1 | -0.44908 | 0.07301 | 37.8378 | <.0001 | 0.638 | 0.553 | 0.736 | Interaction w. nationality |

***Sponsored time competing risks analysis*** ***9***
***Outcome of interest: Withdrawal***
***Imputation method B***

***The PHREG Procedure***

***_Imputation_=3***

*Linear Hypotheses Testing Results*

| *Label* | *Wald Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *gender* | 21.5855 | 1 | <.0001 |
| *domstat* | 3413.2136 | 4 | <.0001 |
| *domclus* | 269.7559 | 4 | <.0001 |
| *leeft* | 206.3009 | 4 | <.0001 |
| *start* | 560.5149 | 2 | <.0001 |
| *nat* | 56.8215 | 2 | <.0001 |
| *univ* | 51.3815 | 4 | <.0001 |
| *nat_domstat* | 30.6481 | 4 | <.0001 |
| *nat_domclus* | 48.4580 | 4 | <.0001 |

***Sponsored time competing risks analysis*** ***10***
***Outcome of interest: Withdrawal***
***Imputation method B***

***The PHREG Procedure***

***_Imputation_=4***

| *Model Information* | |
|---|---|
| *Data Set* | LIB.IMPUTATIEB |
| *Dependent Variable* | sponsTime |
| *Censoring Variable* | cens2 |
| *Censoring Value(s)* | 0 |
| *Ties Handling* | DISCRETE |

| | |
|---|---|
| *Number of Observations Read* | 28871 |
| *Number of Observations Used* | 28396 |

*Summary of the Number of Event and Censored Values*

| *Total* | *Event* | *Censored* | *Percent Censored* |
|---|---|---|---|
| 28396 | 10459 | 17937 | 63.17 |

| *Convergence Status* |
|---|
| Convergence criterion (GCONV=1E-8) satisfied. |

*Model Fit Statistics*

| *Criterion* | *Without Covariates* | *With Covariates* |
|---|---|---|
| *-2 LOG L* | 157680.09 | 151333.90 |
| *AIC* | 157680.09 | 151391.90 |
| *SBC* | 157680.09 | 151602.30 |

*Testing Global Null Hypothesis: BETA=0*

| *Test* | *Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *Likelihood Ratio* | 6346.1920 | 29 | <.0001 |
| *Score* | 6768.0331 | 29 | <.0001 |
| *Wald* | 5677.4021 | 29 | <.0001 |

***Sponsored time competing risks analysis*** ***11***
***Outcome of interest: Withdrawal***
***Imputation method B***

***The PHREG Procedure***

***_Imputation_=4***

*Analysis of Maximum Likelihood Estimates*

| *Parameter* | *DF* | *Parameter Estimate* | *Standard Error* | *Chi-Square* | *Pr > ChiSq* | *Hazard Ratio* | *95% Hazard Ratio Confidence Limits* | | *Label* |
|---|---|---|---|---|---|---|---|---|---|
| *GESLACHT* | 1 | 0.10469 | 0.02034 | 26.4938 | <.0001 | 1.110 | 1.067 | 1.156 | Female |
| *domstat2* | 1 | -1.06227 | 0.02807 | 1431.7663 | <.0001 | 0.346 | 0.327 | 0.365 | Assistant |
| *domstat3* | 1 | -2.16621 | 0.04499 | 2318.1821 | <.0001 | 0.115 | 0.105 | 0.125 | Scholarship (Flanders) |
| *domstat4* | 1 | -1.62315 | 0.09045 | 322.0587 | <.0001 | 0.197 | 0.165 | 0.236 | Scholarship (univers.) |
| *domstat5* | 1 | -1.06642 | 0.02958 | 1299.3337 | <.0001 | 0.344 | 0.325 | 0.365 | Project (FWO/BOF/IUAP) |
| *domclusmed* | 1 | 0.17400 | 0.03842 | 20.5112 | <.0001 | 1.190 | 1.104 | 1.283 | Medical |
| *domclushum* | 1 | 0.43894 | 0.03952 | 123.3570 | <.0001 | 1.551 | 1.435 | 1.676 | Humanities |
| *domclussoc* | 1 | 0.51692 | 0.03664 | 199.0719 | <.0001 | 1.677 | 1.561 | 1.802 | Social |
| *domclustoe* | 1 | 0.23189 | 0.03838 | 36.5092 | <.0001 | 1.261 | 1.170 | 1.359 | Applied |
| *leeft2* | 1 | 0.21851 | 0.02646 | 68.1934 | <.0001 | 1.244 | 1.181 | 1.310 | Age 26-30 (at start) |
| *leeft3* | 1 | 0.37976 | 0.03960 | 91.9667 | <.0001 | 1.462 | 1.353 | 1.580 | Age 31-35 (at start) |
| *leeft4* | 1 | 0.32580 | 0.05973 | 29.7541 | <.0001 | 1.385 | 1.232 | 1.557 | Age 36-40 (at start) |
| *leeft5* | 1 | 0.58820 | 0.05693 | 106.7482 | <.0001 | 1.801 | 1.611 | 2.013 | Age 41+ (at start) |
| *start2* | 1 | -0.20878 | 0.02236 | 87.1560 | <.0001 | 0.812 | 0.777 | 0.848 | Start cohort 1997-2004 |
| *start3* | 1 | -0.74969 | 0.03142 | 569.1291 | <.0001 | 0.473 | 0.444 | 0.503 | Start cohort 2004-2009 |
| *natEurEU* | 1 | 0.48438 | 0.06303 | 59.0652 | <.0001 | 1.623 | 1.435 | 1.837 | EU (excl. Belgium) |
| *natAnd* | 1 | 0.40916 | 0.06851 | 35.6640 | <.0001 | 1.506 | 1.316 | 1.722 | International (non-EU) |
| *univ2* | 1 | 0.09529 | 0.03412 | 7.7999 | 0.0052 | 1.100 | 1.029 | 1.176 | |
| *univ3* | 1 | 0.04921 | 0.02395 | 4.2222 | 0.0399 | 1.050 | 1.002 | 1.101 | |
| *univ4* | 1 | -0.03371 | 0.03521 | 0.9169 | 0.3383 | 0.967 | 0.902 | 1.036 | |
| *univ5* | 1 | -0.36342 | 0.06321 | 33.0565 | <.0001 | 0.695 | 0.614 | 0.787 | |
| *natNBel_domstat2* | 1 | -0.12397 | 0.10751 | 1.3295 | 0.2489 | 0.883 | 0.716 | 1.091 | Interaction w. nationality |
| *natNBel_domstat3* | 1 | -0.09575 | 0.29477 | 0.1055 | 0.7453 | 0.909 | 0.510 | 1.619 | Interaction w. nationality |
| *natNBel_domstat4* | 1 | -0.24076 | 0.14372 | 2.8065 | 0.0939 | 0.786 | 0.593 | 1.042 | Interaction w. nationality |
| *natNBel_domstat5* | 1 | 0.26237 | 0.05833 | 20.2299 | <.0001 | 1.300 | 1.160 | 1.457 | Interaction w. nationality |
| *natNBel_domclusmed* | 1 | -0.32289 | 0.07667 | 17.7355 | <.0001 | 0.724 | 0.623 | 0.841 | Interaction w. nationality |
| *natNBel_domclushum* | 1 | -0.42189 | 0.09132 | 21.3433 | <.0001 | 0.656 | 0.548 | 0.784 | Interaction w. nationality |
| *natNBel_domclussoc* | 1 | -0.48859 | 0.08875 | 30.3080 | <.0001 | 0.613 | 0.516 | 0.730 | Interaction w. nationality |
| *natNBel_domclustoe* | 1 | -0.47515 | 0.07286 | 42.5252 | <.0001 | 0.622 | 0.539 | 0.717 | Interaction w. nationality |

***Sponsored time competing risks analysis*** ***12***
***Outcome of interest: Withdrawal***
***Imputation method B***

***The PHREG Procedure***

***_Imputation_=4***

*Linear Hypotheses Testing Results*

| *Label* | *Wald Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *gender* | 26.4938 | 1 | <.0001 |
| *domstat* | 3433.8503 | 4 | <.0001 |
| *domclus* | 261.9042 | 4 | <.0001 |
| *leeft* | 206.1467 | 4 | <.0001 |
| *start* | 571.6501 | 2 | <.0001 |
| *nat* | 59.2699 | 2 | <.0001 |
| *univ* | 54.8443 | 4 | <.0001 |
| *nat_domstat* | 30.6327 | 4 | <.0001 |
| *nat_domclus* | 53.2432 | 4 | <.0001 |

***Sponsored time competing risks analysis***
***Outcome of interest: Withdrawal***
***Imputation method B***

***The PHREG Procedure***

***_Imputation_=5***

| *Model Information* | |
|---|---|
| *Data Set* | LIB.IMPUTATIEB |
| *Dependent Variable* | sponsTime |
| *Censoring Variable* | cens2 |
| *Censoring Value(s)* | 0 |
| *Ties Handling* | DISCRETE |

| | |
|---|---|
| *Number of Observations Read* | 28871 |
| *Number of Observations Used* | 28396 |

*Summary of the Number of Event and Censored Values*

| *Total* | *Event* | *Censored* | *Percent Censored* |
|---|---|---|---|
| 28396 | 10436 | 17960 | 63.25 |

| *Convergence Status* |
|---|
| Convergence criterion (GCONV=1E-8) satisfied. |

*Model Fit Statistics*

| *Criterion* | *Without Covariates* | *With Covariates* |
|---|---|---|
| *-2 LOG L* | 157212.22 | 150896.51 |
| *AIC* | 157212.22 | 150954.51 |
| *SBC* | 157212.22 | 151164.85 |

*Testing Global Null Hypothesis: BETA=0*

| *Test* | *Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *Likelihood Ratio* | 6315.7062 | 29 | <.0001 |
| *Score* | 6755.2368 | 29 | <.0001 |
| *Wald* | 5683.5512 | 29 | <.0001 |

***Sponsored time competing risks analysis*** ***14***
***Outcome of interest: Withdrawal***
***Imputation method B***

***The PHREG Procedure***

***_Imputation_=5***

*Analysis of Maximum Likelihood Estimates*

| *Parameter* | *DF* | *Parameter Estimate* | *Standard Error* | *Chi-Square* | *Pr > ChiSq* | *Hazard Ratio* | *95% Hazard Ratio Confidence Limits* | | *Label* |
|---|---|---|---|---|---|---|---|---|---|
| *GESLACHT* | 1 | 0.09782 | 0.02036 | 23.0786 | <.0001 | 1.103 | 1.060 | 1.148 | Female |
| *domstat2* | 1 | -1.06882 | 0.02812 | 1444.9416 | <.0001 | 0.343 | 0.325 | 0.363 | Assistant |
| *domstat3* | 1 | -2.13982 | 0.04455 | 2307.2126 | <.0001 | 0.118 | 0.108 | 0.128 | Scholarship (Flanders) |
| *domstat4* | 1 | -1.62737 | 0.09045 | 323.7395 | <.0001 | 0.196 | 0.165 | 0.235 | Scholarship (univers.) |
| *domstat5* | 1 | -1.07373 | 0.02966 | 1310.5143 | <.0001 | 0.342 | 0.322 | 0.362 | Project (FWO/BOF/IUAP) |
| *domclusmed* | 1 | 0.18539 | 0.03846 | 23.2380 | <.0001 | 1.204 | 1.116 | 1.298 | Medical |
| *domclushum* | 1 | 0.44648 | 0.03961 | 127.0811 | <.0001 | 1.563 | 1.446 | 1.689 | Humanities |
| *domclussoc* | 1 | 0.52720 | 0.03671 | 206.2409 | <.0001 | 1.694 | 1.577 | 1.821 | Social |
| *domclustoe* | 1 | 0.23245 | 0.03847 | 36.5017 | <.0001 | 1.262 | 1.170 | 1.361 | Applied |
| *leeft2* | 1 | 0.21894 | 0.02649 | 68.3051 | <.0001 | 1.245 | 1.182 | 1.311 | Age 26-30 (at start) |
| *leeft3* | 1 | 0.38404 | 0.03960 | 94.0404 | <.0001 | 1.468 | 1.359 | 1.587 | Age 31-35 (at start) |
| *leeft4* | 1 | 0.32644 | 0.05982 | 29.7824 | <.0001 | 1.386 | 1.233 | 1.558 | Age 36-40 (at start) |
| *leeft5* | 1 | 0.59871 | 0.05672 | 111.4275 | <.0001 | 1.820 | 1.628 | 2.034 | Age 41+ (at start) |
| *start2* | 1 | -0.20917 | 0.02237 | 87.4324 | <.0001 | 0.811 | 0.776 | 0.848 | Start cohort 1997-2004 |
| *start3* | 1 | -0.75382 | 0.03147 | 573.8611 | <.0001 | 0.471 | 0.442 | 0.501 | Start cohort 2004-2009 |
| *natEurEU* | 1 | 0.48657 | 0.06323 | 59.2110 | <.0001 | 1.627 | 1.437 | 1.841 | EU (excl. Belgium) |
| *natAnd* | 1 | 0.37699 | 0.06901 | 29.8452 | <.0001 | 1.458 | 1.273 | 1.669 | International (non-EU) |
| *univ2* | 1 | 0.10036 | 0.03406 | 8.6805 | 0.0032 | 1.106 | 1.034 | 1.182 | |
| *univ3* | 1 | 0.04439 | 0.02397 | 3.4311 | 0.0640 | 1.045 | 0.997 | 1.096 | |
| *univ4* | 1 | -0.04271 | 0.03528 | 1.4657 | 0.2260 | 0.958 | 0.894 | 1.027 | |
| *univ5* | 1 | -0.36091 | 0.06321 | 32.6010 | <.0001 | 0.697 | 0.616 | 0.789 | |
| *natNBel_domstat2* | 1 | -0.05321 | 0.10504 | 0.2566 | 0.6125 | 0.948 | 0.772 | 1.165 | Interaction w. nationality |
| *natNBel_domstat3* | 1 | -0.12413 | 0.29470 | 0.1774 | 0.6736 | 0.883 | 0.496 | 1.574 | Interaction w. nationality |
| *natNBel_domstat4* | 1 | -0.26228 | 0.14548 | 3.2500 | 0.0714 | 0.769 | 0.578 | 1.023 | Interaction w. nationality |
| *natNBel_domstat5* | 1 | 0.27538 | 0.05852 | 22.1403 | <.0001 | 1.317 | 1.174 | 1.477 | Interaction w. nationality |
| *natNBel_domclusmed* | 1 | -0.32357 | 0.07689 | 17.7102 | <.0001 | 0.724 | 0.622 | 0.841 | Interaction w. nationality |
| *natNBel_domclushum* | 1 | -0.40647 | 0.09109 | 19.9097 | <.0001 | 0.666 | 0.557 | 0.796 | Interaction w. nationality |
| *natNBel_domclussoc* | 1 | -0.49942 | 0.08907 | 31.4377 | <.0001 | 0.607 | 0.510 | 0.723 | Interaction w. nationality |
| *natNBel_domclustoe* | 1 | -0.47173 | 0.07323 | 41.4899 | <.0001 | 0.624 | 0.540 | 0.720 | Interaction w. nationality |

***Sponsored time competing risks analysis***
***Outcome of interest: Withdrawal***
***Imputation method B***

***The PHREG Procedure***

***_Imputation_=5***

*Linear Hypotheses Testing Results*

| *Label* | *Wald Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *gender* | 23.0786 | 1 | <.0001 |
| *domstat* | 3433.4023 | 4 | <.0001 |
| *domclus* | 268.5416 | 4 | <.0001 |
| *leeft* | 211.4234 | 4 | <.0001 |
| *start* | 576.3791 | 2 | <.0001 |
| *nat* | 59.3075 | 2 | <.0001 |
| *univ* | 55.1972 | 4 | <.0001 |
| *nat_domstat* | 31.4293 | 4 | <.0001 |
| *nat_domclus* | 52.5553 | 4 | <.0001 |

**Sponsored time competing risks analysis**
**Outcome of interest: Withdrawal**
**Imputation method A**

**The MIANALYZE Procedure**

| Model Information | |
|---|---|
| Data Set | LIB.ESTWITHDRA |
| Number of Imputations | 5 |

Variance Information

| Parameter | Variance Between | Variance Within | Variance Total | DF | Relative Increase in Variance | Fraction Missing Information | Relative Efficiency |
|---|---|---|---|---|---|---|---|
| geslacht | 0.000024127 | 0.000414 | 0.000443 | 934.57 | 0.070002 | 0.067416 | 0.986696 |
| domstat2 | 0.000005089 | 0.000793 | 0.000799 | 68513 | 0.007700 | 0.007670 | 0.998468 |
| domstat3 | 0.000034909 | 0.001965 | 0.002007 | 9183.6 | 0.021315 | 0.021083 | 0.995801 |
| domstat4 | 0.000199 | 0.007790 | 0.008028 | 4530.5 | 0.030624 | 0.030142 | 0.994008 |
| domstat5 | 0.000007996 | 0.000877 | 0.000887 | 34170 | 0.010938 | 0.010877 | 0.997829 |
| domclusmed | 0.000041877 | 0.001467 | 0.001517 | 3644.7 | 0.034263 | 0.033658 | 0.993313 |
| domclushum | 0.000017186 | 0.001562 | 0.001582 | 23544 | 0.013207 | 0.013118 | 0.997383 |
| domclussoc | 0.000036077 | 0.001341 | 0.001385 | 4092.4 | 0.032273 | 0.031737 | 0.993693 |
| domclustoe | 0.000056888 | 0.001476 | 0.001545 | 2047.8 | 0.046240 | 0.045128 | 0.991055 |
| leeft2 | 0.000003512 | 0.000701 | 0.000705 | 112047 | 0.006011 | 0.005993 | 0.998803 |
| leeft3 | 0.000061877 | 0.001584 | 0.001658 | 1995.5 | 0.046870 | 0.045728 | 0.990937 |
| leeft4 | 0.000056933 | 0.003521 | 0.003589 | 11038 | 0.019406 | 0.019214 | 0.996172 |
| leeft5 | 0.000024301 | 0.003236 | 0.003266 | 50160 | 0.009010 | 0.008969 | 0.998209 |
| start2 | 0.000002609 | 0.000503 | 0.000506 | 104401 | 0.006228 | 0.006209 | 0.998760 |
| start3 | 0.000114 | 0.000967 | 0.001104 | 258.64 | 0.142022 | 0.131054 | 0.974459 |
| natEurEU | 0.000099845 | 0.003983 | 0.004103 | 4689.9 | 0.030083 | 0.029618 | 0.994111 |
| natAnd | 0.000007218 | 0.004693 | 0.004702 | 1.18E6 | 0.001846 | 0.001844 | 0.999631 |
| univ2 | 0.000013767 | 0.001161 | 0.001177 | 20318 | 0.014231 | 0.014128 | 0.997182 |
| univ3 | 0.000035846 | 0.000572 | 0.000615 | 818.67 | 0.075153 | 0.072164 | 0.985773 |
| univ4 | 0.000033106 | 0.001234 | 0.001274 | 4110.7 | 0.032199 | 0.031665 | 0.993707 |
| univ5 | 0.000087882 | 0.004011 | 0.004116 | 6093.3 | 0.026295 | 0.025941 | 0.994839 |
| natNBel_domstat2 | 0.000182 | 0.011015 | 0.011233 | 10624 | 0.019788 | 0.019589 | 0.996098 |
| natNBel_domstat3 | 0.001745 | 0.089864 | 0.091958 | 7716.4 | 0.023298 | 0.023021 | 0.995417 |
| natNBel_domstat4 | 0.002128 | 0.019993 | 0.022547 | 311.75 | 0.127743 | 0.118907 | 0.976771 |
| natNBel_domstat5 | 0.000118 | 0.003386 | 0.003527 | 2485.2 | 0.041796 | 0.040891 | 0.991888 |
| natNBel_domclusmed | 0.000086829 | 0.005814 | 0.005918 | 12906 | 0.017921 | 0.017757 | 0.996461 |
| natNBel_domclushum | 0.000114 | 0.008364 | 0.008500 | 15546 | 0.016302 | 0.016167 | 0.996777 |
| natNBel_domclussoc | 0.000895 | 0.007794 | 0.008868 | 272.54 | 0.137848 | 0.127527 | 0.975129 |
| natNBel_domclustoe | 0.000086593 | 0.005318 | 0.005422 | 10890 | 0.019540 | 0.019346 | 0.996146 |

**The MIANALYZE Procedure**

*Parameter Estimates*

| Parameter | Estimate | Std Error | 95% Confidence Limits | | DF | Minimum | Maximum |
|---|---|---|---|---|---|---|---|
| geslacht | 0.110029 | 0.021037 | 0.06874 | 0.15131 | 934.57 | 0.104103 | 0.116434 |
| domstat2 | -1.059460 | 0.028271 | -1.11487 | -1.00405 | 68513 | -1.062152 | -1.056567 |
| domstat3 | -2.121338 | 0.044802 | -2.20916 | -2.03352 | 9183.6 | -2.130735 | -2.114913 |
| domstat4 | -1.568148 | 0.089602 | -1.74381 | -1.39248 | 4530.5 | -1.584552 | -1.548798 |
| domstat5 | -1.056181 | 0.029781 | -1.11455 | -0.99781 | 34170 | -1.057885 | -1.051184 |
| domclusmed | 0.180287 | 0.038948 | 0.10393 | 0.25665 | 3644.7 | 0.174836 | 0.190838 |
| domclushum | 0.435339 | 0.039778 | 0.35737 | 0.51331 | 23544 | 0.430420 | 0.441413 |
| domclussoc | 0.515212 | 0.037212 | 0.44226 | 0.58817 | 4092.4 | 0.508694 | 0.522603 |
| domclustoe | 0.218901 | 0.039302 | 0.14183 | 0.29598 | 2047.8 | 0.206600 | 0.226328 |
| leeft2 | 0.215209 | 0.026557 | 0.16316 | 0.26726 | 112047 | 0.212647 | 0.217136 |
| leeft3 | 0.365262 | 0.040724 | 0.28540 | 0.44513 | 1995.5 | 0.354441 | 0.374589 |
| leeft4 | 0.340387 | 0.059907 | 0.22296 | 0.45782 | 11038 | 0.331222 | 0.350388 |
| leeft5 | 0.592027 | 0.057145 | 0.48002 | 0.70403 | 50160 | 0.585277 | 0.598835 |
| start2 | -0.220939 | 0.022490 | -0.26502 | -0.17686 | 104401 | -0.222972 | -0.218462 |
| start3 | -0.713657 | 0.033227 | -0.77909 | -0.64823 | 258.64 | -0.723216 | -0.700481 |
| natEurEU | 0.476748 | 0.064051 | 0.35118 | 0.60232 | 4689.9 | 0.462703 | 0.486873 |
| natAnd | 0.405373 | 0.068570 | 0.27098 | 0.53977 | 1.18E6 | 0.402353 | 0.409631 |
| univ2 | 0.083961 | 0.034313 | 0.01670 | 0.15122 | 20318 | 0.080451 | 0.088000 |
| univ3 | 0.033969 | 0.024807 | -0.01472 | 0.08266 | 818.67 | 0.026386 | 0.042092 |
| univ4 | -0.042262 | 0.035687 | -0.11223 | 0.02770 | 4110.7 | -0.048742 | -0.033757 |
| univ5 | -0.373283 | 0.064156 | -0.49905 | -0.24751 | 6093.3 | -0.388902 | -0.364387 |
| natNBel_domstat2 | -0.048528 | 0.105988 | -0.25628 | 0.15923 | 10624 | -0.070971 | -0.039147 |
| natNBel_domstat3 | -0.168489 | 0.303246 | -0.76293 | 0.42595 | 7716.4 | -0.224410 | -0.136707 |
| natNBel_domstat4 | -0.251571 | 0.150156 | -0.54702 | 0.04388 | 311.75 | -0.312236 | -0.195579 |
| natNBel_domstat5 | 0.282395 | 0.059390 | 0.16594 | 0.39885 | 2485.2 | 0.273496 | 0.298993 |
| natNBel_domclusmed | -0.309867 | 0.076931 | -0.46066 | -0.15907 | 12906 | -0.325620 | -0.301578 |
| natNBel_domclushum | -0.437469 | 0.092198 | -0.61819 | -0.25675 | 15546 | -0.455064 | -0.427993 |
| natNBel_domclussoc | -0.476733 | 0.094170 | -0.66213 | -0.29134 | 272.54 | -0.528091 | -0.454110 |
| natNBel_domclustoe | -0.464134 | 0.073633 | -0.60847 | -0.31980 | 10890 | -0.478635 | -0.456305 |

***Sponsored time competing risks analysis***
***Outcome of interest: Withdrawal***
***Imputation method A***

***The MIANALYZE Procedure***

*Parameter Estimates*

| *Parameter* | *Theta0* | *t for H0: Parameter=Theta0* | *Pr > \|t\|* |
|---|---|---|---|
| *geslacht* | 0 | 5.23 | <.0001 |
| *domstat2* | 0 | -37.47 | <.0001 |
| *domstat3* | 0 | -47.35 | <.0001 |
| *domstat4* | 0 | -17.50 | <.0001 |
| *domstat5* | 0 | -35.47 | <.0001 |
| *domclusmed* | 0 | 4.63 | <.0001 |
| *domclushum* | 0 | 10.94 | <.0001 |
| *domclussoc* | 0 | 13.85 | <.0001 |
| *domclustoe* | 0 | 5.57 | <.0001 |
| *leeft2* | 0 | 8.10 | <.0001 |
| *leeft3* | 0 | 8.97 | <.0001 |
| *leeft4* | 0 | 5.68 | <.0001 |
| *leeft5* | 0 | 10.36 | <.0001 |
| *start2* | 0 | -9.82 | <.0001 |
| *start3* | 0 | -21.48 | <.0001 |
| *natEurEU* | 0 | 7.44 | <.0001 |
| *natAnd* | 0 | 5.91 | <.0001 |
| *univ2* | 0 | 2.45 | 0.0144 |
| *univ3* | 0 | 1.37 | 0.1713 |
| *univ4* | 0 | -1.18 | 0.2364 |
| *univ5* | 0 | -5.82 | <.0001 |
| *natNBel_domstat2* | 0 | -0.46 | 0.6471 |
| *natNBel_domstat3* | 0 | -0.56 | 0.5785 |
| *natNBel_domstat4* | 0 | -1.68 | 0.0949 |
| *natNBel_domstat5* | 0 | 4.75 | <.0001 |
| *natNBel_domclusmed* | 0 | -4.03 | <.0001 |
| *natNBel_domclushum* | 0 | -4.74 | <.0001 |
| *natNBel_domclussoc* | 0 | -5.06 | <.0001 |
| *natNBel_domclustoe* | 0 | -6.30 | <.0001 |

***Sponsored time competing risks analysis*** ***1***
***Outcome of interest: Withdrawal***
***Imputation method B***

***The MIANALYZE Procedure***

| *Model Information* | |
|---|---|
| *Data Set* | LIB.ESTWITHDRB |
| *Number of Imputations* | 5 |

*Variance Information*

| *Parameter* | *Variance Between* | *Variance Within* | *Variance Total* | *DF* | *Relative Increase in Variance* | *Fraction Missing Information* | *Relative Efficiency* |
|---|---|---|---|---|---|---|---|
| *geslacht* | 0.000017008 | 0.000414 | 0.000435 | 1813.2 | 0.049284 | 0.048018 | 0.990488 |
| *domstat2* | 0.000048212 | 0.000790 | 0.000848 | 858.39 | 0.073264 | 0.070427 | 0.986110 |
| *domstat3* | 0.000139 | 0.002004 | 0.002171 | 678.21 | 0.083186 | 0.079508 | 0.984347 |
| *domstat4* | 0.000654 | 0.008102 | 0.008886 | 513.34 | 0.096820 | 0.091805 | 0.981970 |
| *domstat5* | 0.000050751 | 0.000878 | 0.000939 | 951.17 | 0.069346 | 0.066809 | 0.986814 |
| *domclusmed* | 0.000054890 | 0.001477 | 0.001543 | 2195.9 | 0.044583 | 0.043551 | 0.991365 |
| *domclushum* | 0.000019013 | 0.001567 | 0.001590 | 19427 | 0.014558 | 0.014451 | 0.997118 |
| *domclussoc* | 0.000042092 | 0.001346 | 0.001397 | 3059.1 | 0.037517 | 0.036790 | 0.992696 |
| *domclustoe* | 0.000020697 | 0.001480 | 0.001505 | 14693 | 0.016776 | 0.016633 | 0.996684 |
| *leeft2* | 0.000041538 | 0.000702 | 0.000751 | 909.08 | 0.071045 | 0.068380 | 0.986508 |
| *leeft3* | 0.000063111 | 0.001573 | 0.001648 | 1894.9 | 0.048157 | 0.046950 | 0.990697 |
| *leeft4* | 0.000005615 | 0.003556 | 0.003562 | 1.12E6 | 0.001895 | 0.001893 | 0.999621 |
| *leeft5* | 0.000025576 | 0.003225 | 0.003255 | 45005 | 0.009517 | 0.009472 | 0.998109 |
| *start2* | 0.000004962 | 0.000500 | 0.000506 | 28924 | 0.011900 | 0.011828 | 0.997640 |
| *start3* | 0.000039156 | 0.000988 | 0.001035 | 1940.8 | 0.047558 | 0.046381 | 0.990809 |
| *natEurEU* | 0.000049467 | 0.003988 | 0.004048 | 18598 | 0.014884 | 0.014771 | 0.997054 |
| *natAnd* | 0.000220 | 0.004725 | 0.004989 | 1430.3 | 0.055836 | 0.054205 | 0.989275 |
| *univ2* | 0.000036120 | 0.001165 | 0.001208 | 3108.1 | 0.037209 | 0.036494 | 0.992754 |
| *univ3* | 0.000009962 | 0.000574 | 0.000586 | 9598.8 | 0.020839 | 0.020618 | 0.995893 |
| *univ4* | 0.000038104 | 0.001241 | 0.001287 | 3168.6 | 0.036839 | 0.036138 | 0.992824 |
| *univ5* | 0.000091820 | 0.003966 | 0.004076 | 5473.8 | 0.027784 | 0.027388 | 0.994552 |
| *natNBel_domstat2* | 0.000820 | 0.011235 | 0.012218 | 617.21 | 0.087552 | 0.083469 | 0.983580 |
| *natNBel_domstat3* | 0.001221 | 0.088383 | 0.089848 | 15047 | 0.016574 | 0.016435 | 0.996724 |
| *natNBel_domstat4* | 0.000229 | 0.020830 | 0.021105 | 23555 | 0.013203 | 0.013115 | 0.997384 |
| *natNBel_domstat5* | 0.000175 | 0.003407 | 0.003618 | 1185.4 | 0.061672 | 0.059675 | 0.988206 |
| *natNBel_domclusmed* | 0.000076168 | 0.005888 | 0.005979 | 17118 | 0.015524 | 0.015401 | 0.996929 |
| *natNBel_domclushum* | 0.000384 | 0.008382 | 0.008843 | 1469.7 | 0.055041 | 0.053457 | 0.989422 |
| *natNBel_domclussoc* | 0.000418 | 0.007808 | 0.008310 | 1096.2 | 0.064291 | 0.062117 | 0.987729 |
| *natNBel_domclustoe* | 0.000194 | 0.005339 | 0.005572 | 2298.5 | 0.043533 | 0.042550 | 0.991562 |

***Sponsored time competing risks analysis***
***Outcome of interest: Withdrawal***
***Imputation method B***

***The MIANALYZE Procedure***

*Parameter Estimates*

| *Parameter* | *Estimate* | *Std Error* | *95% Confidence Limits* | | *DF* | *Minimum* | *Maximum* |
|---|---|---|---|---|---|---|---|
| *geslacht* | 0.100226 | 0.020845 | 0.05934 | 0.14111 | 1813.2 | 0.094455 | 0.104691 |
| *domstat2* | -1.068004 | 0.029112 | -1.12514 | -1.01086 | 858.39 | -1.078962 | -1.061744 |
| *domstat3* | -2.154110 | 0.046597 | -2.24560 | -2.06262 | 678.21 | -2.166308 | -2.139821 |
| *domstat4* | -1.615563 | 0.094266 | -1.80076 | -1.43037 | 513.34 | -1.647648 | -1.584189 |
| *domstat5* | -1.072105 | 0.030645 | -1.13224 | -1.01197 | 951.17 | -1.081326 | -1.063623 |
| *domclusmed* | 0.186099 | 0.039285 | 0.10906 | 0.26314 | 2195.9 | 0.174005 | 0.193590 |
| *domclushum* | 0.446182 | 0.039875 | 0.36802 | 0.52434 | 19427 | 0.438938 | 0.450090 |
| *domclussoc* | 0.525886 | 0.037374 | 0.45260 | 0.59917 | 3059.1 | 0.516920 | 0.535069 |
| *domclustoe* | 0.230317 | 0.038798 | 0.15427 | 0.30637 | 14693 | 0.224731 | 0.235881 |
| *leeft2* | 0.214714 | 0.027412 | 0.16092 | 0.26851 | 909.08 | 0.205689 | 0.220315 |
| *leeft3* | 0.375231 | 0.040600 | 0.29561 | 0.45486 | 1894.9 | 0.363324 | 0.384036 |
| *leeft4* | 0.328674 | 0.059685 | 0.21169 | 0.44565 | 1.12E6 | 0.325800 | 0.330729 |
| *leeft5* | 0.592367 | 0.057057 | 0.48054 | 0.70420 | 45005 | 0.586722 | 0.598706 |
| *start2* | -0.210667 | 0.022501 | -0.25477 | -0.16656 | 28924 | -0.213635 | -0.208777 |
| *start3* | -0.750813 | 0.032171 | -0.81391 | -0.68772 | 1940.8 | -0.757352 | -0.740768 |
| *natEurEU* | 0.486781 | 0.063621 | 0.36208 | 0.61148 | 18598 | 0.476748 | 0.495506 |
| *natAnd* | 0.398476 | 0.070633 | 0.25992 | 0.53703 | 1430.3 | 0.376988 | 0.415060 |
| *univ2* | 0.092034 | 0.034759 | 0.02388 | 0.16019 | 3108.1 | 0.084537 | 0.100357 |
| *univ3* | 0.044783 | 0.024199 | -0.00265 | 0.09222 | 9598.8 | 0.041000 | 0.049210 |
| *univ4* | -0.040622 | 0.035874 | -0.11096 | 0.02972 | 3168.6 | -0.046159 | -0.033711 |
| *univ5* | -0.355484 | 0.063843 | -0.48064 | -0.23033 | 5473.8 | -0.363420 | -0.342274 |
| *natNBel_domstat2* | -0.082103 | 0.110537 | -0.29918 | 0.13497 | 617.21 | -0.123967 | -0.053205 |
| *natNBel_domstat3* | -0.125394 | 0.299747 | -0.71293 | 0.46215 | 15047 | -0.185229 | -0.095746 |
| *natNBel_domstat4* | -0.259878 | 0.145277 | -0.54463 | 0.02487 | 23555 | -0.279504 | -0.240763 |
| *natNBel_domstat5* | 0.275655 | 0.060147 | 0.15765 | 0.39366 | 1185.4 | 0.262372 | 0.295428 |
| *natNBel_domclusmed* | -0.326647 | 0.077326 | -0.47821 | -0.17508 | 17118 | -0.341540 | -0.318954 |
| *natNBel_domclushum* | -0.432597 | 0.094040 | -0.61706 | -0.24813 | 1469.7 | -0.453863 | -0.406465 |
| *natNBel_domclussoc* | -0.474468 | 0.091161 | -0.65334 | -0.29560 | 1096.2 | -0.499424 | -0.449174 |
| *natNBel_domclustoe* | -0.472050 | 0.074645 | -0.61843 | -0.32567 | 2298.5 | -0.486228 | -0.449075 |

***Sponsored time competing risks analysis***
***Outcome of interest: Withdrawal***
***Imputation method B***

***The MIANALYZE Procedure***

*Parameter Estimates*

| *Parameter* | *Theta0* | *t for H0: Parameter=Theta0* | *Pr > \|t\|* |
|---|---|---|---|
| *geslacht* | 0 | 4.81 | <.0001 |
| *domstat2* | 0 | -36.69 | <.0001 |
| *domstat3* | 0 | -46.23 | <.0001 |
| *domstat4* | 0 | -17.14 | <.0001 |
| *domstat5* | 0 | -34.98 | <.0001 |
| *domclusmed* | 0 | 4.74 | <.0001 |
| *domclushum* | 0 | 11.19 | <.0001 |
| *domclussoc* | 0 | 14.07 | <.0001 |
| *domclustoe* | 0 | 5.94 | <.0001 |
| *leeft2* | 0 | 7.83 | <.0001 |
| *leeft3* | 0 | 9.24 | <.0001 |
| *leeft4* | 0 | 5.51 | <.0001 |
| *leeft5* | 0 | 10.38 | <.0001 |
| *start2* | 0 | -9.36 | <.0001 |
| *start3* | 0 | -23.34 | <.0001 |
| *natEurEU* | 0 | 7.65 | <.0001 |
| *natAnd* | 0 | 5.64 | <.0001 |
| *univ2* | 0 | 2.65 | 0.0081 |
| *univ3* | 0 | 1.85 | 0.0643 |
| *univ4* | 0 | -1.13 | 0.2576 |
| *univ5* | 0 | -5.57 | <.0001 |
| *natNBel_domstat2* | 0 | -0.74 | 0.4579 |
| *natNBel_domstat3* | 0 | -0.42 | 0.6757 |
| *natNBel_domstat4* | 0 | -1.79 | 0.0737 |
| *natNBel_domstat5* | 0 | 4.58 | <.0001 |
| *natNBel_domclusmed* | 0 | -4.22 | <.0001 |
| *natNBel_domclushum* | 0 | -4.60 | <.0001 |
| *natNBel_domclussoc* | 0 | -5.20 | <.0001 |
| *natNBel_domclustoe* | 0 | -6.32 | <.0001 |

**Sponsored time competing risks analysis** **1**
**Outcome of interest: Withdrawal**
**Non-imputed data**

**The PHREG Procedure**

| Model Information | |
|---|---|
| Data Set | LIB.SPONSNONIMPUTED |
| Dependent Variable | sponsTime |
| Censoring Variable | cens2 |
| Censoring Value(s) | 0 |
| Ties Handling | DISCRETE |

| | |
|---|---|
| Number of Observations Read | 30965 |
| Number of Observations Used | 28396 |

Summary of the Number of Event and Censored Values

| Total | Event | Censored | Percent Censored |
|---|---|---|---|
| 28396 | 9272 | 19124 | 67.35 |

| Convergence Status |
|---|
| Convergence criterion (GCONV=1E-8) satisfied. |

Model Fit Statistics

| Criterion | Without Covariates | With Covariates |
|---|---|---|
| -2 LOG L | 140859.20 | 134101.34 |
| AIC | 140859.20 | 134159.34 |
| SBC | 140859.20 | 134366.25 |

Testing Global Null Hypothesis: BETA=0

| Test | Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| Likelihood Ratio | 6757.8629 | 29 | <.0001 |
| Score | 6845.0627 | 29 | <.0001 |
| Wald | 5835.9576 | 29 | <.0001 |

***Sponsored time competing risks analysis*** ***2***
***Outcome of interest: Withdrawal***
***Non-imputed data***

***The PHREG Procedure***

*Analysis of Maximum Likelihood Estimates*

| *Parameter* | *DF* | *Parameter Estimate* | *Standard Error* | *Chi-Square* | *Pr > ChiSq* | *Hazard Ratio* | *95% Hazard Ratio Confidence Limits* | | *Label* |
|---|---|---|---|---|---|---|---|---|---|
| *GESLACHT* | 1 | 0.10927 | 0.02160 | 25.5839 | <.0001 | 1.115 | 1.069 | 1.164 | Female |
| *domstat2* | 1 | -1.04226 | 0.02932 | 1263.7230 | <.0001 | 0.353 | 0.333 | 0.374 | Assistant |
| *domstat3* | 1 | -2.12827 | 0.04705 | 2046.4362 | <.0001 | 0.119 | 0.109 | 0.131 | Scholarship (Flanders) |
| *domstat4* | 1 | -1.57645 | 0.09581 | 270.7229 | <.0001 | 0.207 | 0.171 | 0.249 | Scholarship (univers.) |
| *domstat5* | 1 | -1.07160 | 0.03173 | 1140.6085 | <.0001 | 0.342 | 0.322 | 0.364 | Project (FWO/BOF/IUAP) |
| *domclusmed* | 1 | 0.18923 | 0.04072 | 21.6015 | <.0001 | 1.208 | 1.116 | 1.309 | Medical |
| *domclushum* | 1 | 0.45743 | 0.04170 | 120.3246 | <.0001 | 1.580 | 1.456 | 1.715 | Humanities |
| *domclussoc* | 1 | 0.53305 | 0.03882 | 188.5817 | <.0001 | 1.704 | 1.579 | 1.839 | Social |
| *domclustoe* | 1 | 0.23558 | 0.04059 | 33.6874 | <.0001 | 1.266 | 1.169 | 1.370 | Applied |
| *leeft2* | 1 | 0.20339 | 0.02825 | 51.8212 | <.0001 | 1.226 | 1.160 | 1.295 | Age 26-30 (at start) |
| *leeft3* | 1 | 0.34904 | 0.04230 | 68.0880 | <.0001 | 1.418 | 1.305 | 1.540 | Age 31-35 (at start) |
| *leeft4* | 1 | 0.30904 | 0.06305 | 24.0262 | <.0001 | 1.362 | 1.204 | 1.541 | Age 36-40 (at start) |
| *leeft5* | 1 | 0.60683 | 0.05973 | 103.2046 | <.0001 | 1.835 | 1.632 | 2.062 | Age 41+ (at start) |
| *start2* | 1 | -0.28052 | 0.02275 | 152.0504 | <.0001 | 0.755 | 0.722 | 0.790 | Start cohort 1997-2004 |
| *start3* | 1 | -1.53555 | 0.04086 | 1412.1126 | <.0001 | 0.215 | 0.199 | 0.233 | Start cohort 2004-2009 |
| *natEurEU* | 1 | 0.54136 | 0.06722 | 64.8672 | <.0001 | 1.718 | 1.506 | 1.960 | EU (excl. Belgium) |
| *natAnd* | 1 | 0.49923 | 0.07363 | 45.9669 | <.0001 | 1.647 | 1.426 | 1.903 | International (non-EU) |
| *univ2* | 1 | 0.06572 | 0.03647 | 3.2483 | 0.0715 | 1.068 | 0.994 | 1.147 | |
| *univ3* | 1 | 0.03696 | 0.02532 | 2.1307 | 0.1444 | 1.038 | 0.987 | 1.090 | |
| *univ4* | 1 | -0.07482 | 0.03781 | 3.9168 | 0.0478 | 0.928 | 0.862 | 0.999 | |
| *univ5* | 1 | -0.35150 | 0.06645 | 27.9803 | <.0001 | 0.704 | 0.618 | 0.802 | |
| *natNBel_domstat2* | 1 | -0.15411 | 0.11372 | 1.8366 | 0.1754 | 0.857 | 0.686 | 1.071 | Interaction w. nationality |
| *natNBel_domstat3* | 1 | 0.05880 | 0.29559 | 0.0396 | 0.8423 | 1.061 | 0.594 | 1.893 | Interaction w. nationality |
| *natNBel_domstat4* | 1 | -0.30712 | 0.15520 | 3.9159 | 0.0478 | 0.736 | 0.543 | 0.997 | Interaction w. nationality |
| *natNBel_domstat5* | 1 | 0.26230 | 0.06304 | 17.3144 | <.0001 | 1.300 | 1.149 | 1.471 | Interaction w. nationality |
| *natNBel_domclusmed* | 1 | -0.36772 | 0.08241 | 19.9099 | <.0001 | 0.692 | 0.589 | 0.814 | Interaction w. nationality |
| *natNBel_domclushum* | 1 | -0.47709 | 0.09771 | 23.8432 | <.0001 | 0.621 | 0.512 | 0.752 | Interaction w. nationality |
| *natNBel_domclussoc* | 1 | -0.57440 | 0.09733 | 34.8262 | <.0001 | 0.563 | 0.465 | 0.681 | Interaction w. nationality |
| *natNBel_domclustoe* | 1 | -0.50133 | 0.07798 | 41.3344 | <.0001 | 0.606 | 0.520 | 0.706 | Interaction w. nationality |

***The PHREG Procedure***

*Linear Hypotheses Testing Results*

| *Label* | *Wald Chi-Square* | *DF* | *Pr > ChiSq* |
|---|---|---|---|
| *gender* | 25.5839 | 1 | <.0001 |
| *domstat* | 2996.0060 | 4 | <.0001 |
| *domclus* | 246.9495 | 4 | <.0001 |
| *leeft* | 173.4519 | 4 | <.0001 |
| *start* | 1412.6761 | 2 | <.0001 |
| *nat* | 66.6252 | 2 | <.0001 |
| *univ* | 44.9273 | 4 | <.0001 |
| *nat_domstat* | 29.2710 | 4 | <.0001 |
| *nat_domclus* | 56.5856 | 4 | <.0001 |

# Appendix C

# Cumulative Incidence Curves

| | $\hat{I}_p(t\|\mathbf{z}^*)$ and associated 95% confidence interval | | | | | |
|---|---|---|---|---|---|---|
| Variable | $t=2$ | $t=4$ | $t=6$ | $t=8$ | $t=10$ | $p$-value |
| **Gender** | | | | | | 0.0089 |
| [Male] | 0.01 [0.01, 0.01] | 0.06 [0.05, 0.07] | 0.20 [0.17, 0.22] | 0.26 [0.23, 0.29] | 0.27 [0.24, 0.30] | |
| Female | 0.01 [0.01, 0.01] | 0.05 [0.04, 0.05] | 0.15 [0.13, 0.17] | 0.20 [0.18, 0.23] | 0.22 [0.19, 0.24] | |
| **Dominant scientific field** | | | | | | $< 0.0001$ |
| [sciences] | 0.01 [0.01, 0.01] | 0.06 [0.05, 0.07] | 0.20 [0.17, 0.22] | 0.26 [0.23, 0.29] | 0.27 [0.24, 0.30] | |
| medicine | 0.01 [0.01, 0.01] | 0.04 [0.04, 0.05] | 0.14 [0.12, 0.16] | 0.19 [0.16, 0.21] | 0.19 [0.17, 0.22] | |
| humanities | 0 [0, 0.01] | 0.02 [0.02, 0.03] | 0.07 [0.06, 0.08] | 0.09 [0.08, 0.11] | 0.10 [0.09, 0.11] | |
| social | 0 [0, 0.01] | 0.02 [0.02, 0.02] | 0.06 [0.05, 0.07] | 0.08 [0.07, 0.09] | 0.08 [0.07, 0.10] | |
| applied | 0.01 [0.01, 0.01] | 0.04 [0.03, 0.04] | 0.12 [0.10, 0.13] | 0.16 [0.14, 0.18] | 0.17 [0.15, 0.19] | |
| **Nationality** | | | | | | $< 0.0001$ |
| [Belgian] | 0.01 [0.01, 0.01] | 0.06 [0.05, 0.07] | 0.20 [0.17, 0.22] | 0.26 [0.23, 0.29] | 0.27 [0.24, 0.30] | |
| European Union (excl. Belgium) | 0.03 [0.02, 0.03] | 0.11 [0.10, 0.14] | 0.21 [0.17, 0.25] | 0.21 [0.18, 0.25] | 0.21 [0.18, 0.25] | |
| Other | 0.05 [0.04, 0.06] | 0.19 [0.16, 0.23] | 0.28 [0.23, 0.33] | 0.28 [0.24, 0.33] | 0.28 [0.24, 0.33] | |
| **Dominant statute classification** | | | | | | $< 0.0001$ |
| Assistant lectureship | 0.01 [0.01, 0.01] | 0.10 [0.09, 0.10] | 0.38 [0.35, 0.41] | 0.54 [0.50, 0.59] | 0.58 [0.53, 0.63] | |
| Compet. scholarship (Flanders) | 0.03 [0.03, 0.04] | 0.28 [0.26, 0.30] | 0.79 [0.73, 0.84] | 0.88 [0.81, 0.93] | 0.89 [0.81, 0.93] | |
| Compet. scholarship (own university) | 0.03 [0.03, 0.04] | 0.27 [0.23, 0.31] | 0.73 [0.61, 0.83] | 0.81 [0.65, 0.90] | 0.81 [0.66, 0.90] | |
| Project funding (FWO, BOF, IUAP) | 0.02 [0.02, 0.02] | 0.15 [0.14, 0.17] | 0.51 [0.47, 0.56] | 0.64 [0.58, 0.69] | 0.66 [0.60, 0.71] | |
| [Project funding (other)] | 0.01 [0.01, 0.01] | 0.06 [0.05, 0.07] | 0.20 [0.17, 0.22] | 0.26 [0.23, 0.29] | 0.27 [0.24, 0.30] | |
| **Age (at start)** | | | | | | $< 0.0001$ |
| [$\leq 25$ years] | 0.01 [0.01, 0.01] | 0.06 [0.05, 0.07] | 0.20 [0.17, 0.22] | 0.26 [0.23, 0.29] | 0.27 [0.24, 0.30] | |
| $26-30$ years | 0.01 [0.01, 0.01] | 0.05 [0.04, 0.05] | 0.14 [0.13, 0.16] | 0.18 [0.16, 0.21] | 0.19 [0.17, 0.22] | |
| $31-35$ years | 0.01 [0.01, 0.01] | 0.04 [0.04, 0.05] | 0.12 [0.10, 0.14] | 0.15 [0.13, 0.18] | 0.16 [0.13, 0.18] | |
| $36-40$ years | 0.01 [0.01, 0.01] | 0.04 [0.04, 0.05] | 0.13 [0.11, 0.16] | 0.16 [0.13, 0.20] | 0.17 [0.14, 0.20] | |
| $> 40$ years | 0 [0, 0.01] | 0.02 [0.02, 0.03] | 0.06 [0.05, 0.08] | 0.08 [0.06, 0.10] | 0.08 [0.06, 0.10] | |
| **Start cohort** | | | | | | $< 0.0001$ |
| [01/10/1990 - 30/09/1997] | 0.01 [0.01, 0.01] | 0.06 [0.05, 0.07] | 0.20 [0.17, 0.22] | 0.26 [0.23, 0.29] | 0.27 [0.24, 0.30] | |
| 01/10/1997 - 30/09/2004 | 0.01 [0.01, 0.01] | 0.08 [0.07, 0.09] | 0.27 [0.24, 0.31] | 0.35 [0.31, 0.39] | 0.36 [0.32, 0.40] | |
| 01/10/2004 - 30/09/2009 | 0.01 [0.01, 0.01] | 0.08 [0.07, 0.09] | 0.31 [0.27, 0.36] | 0.45 [0.39, 0.51] | 0.48 [0.42, 0.55] | |

**Table C.1:** The combined cumulative incidence function estimates and associated 95% pointwise confidence intervals for each category of covariate values and outcome Ph.D.-attainment - imputation method B.

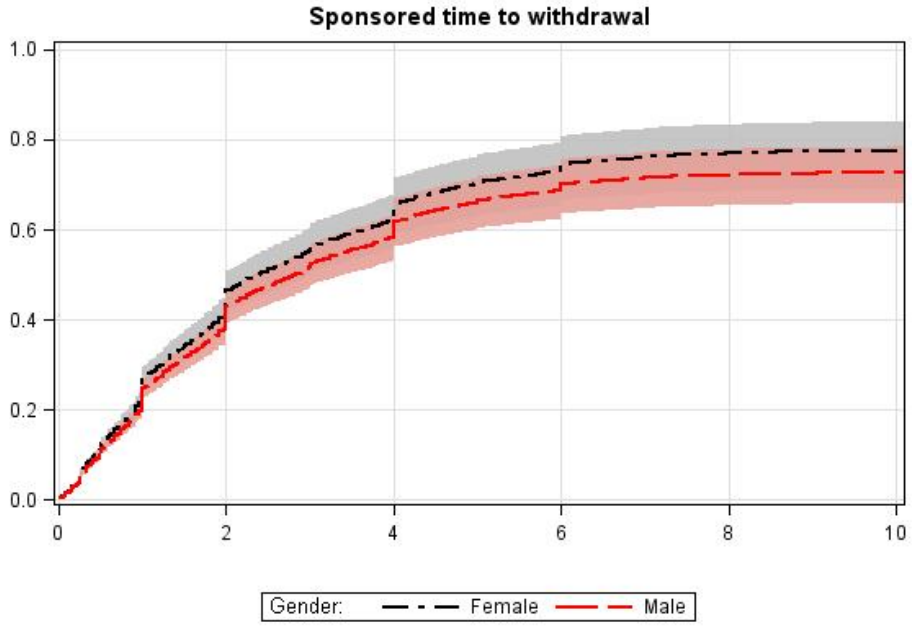


**Figure C.1:** The combined cumulative incidence function estimates and associated 95% pointwise confidence intervals for covariate 'gender' and outcome withdrawal - imputation method B.

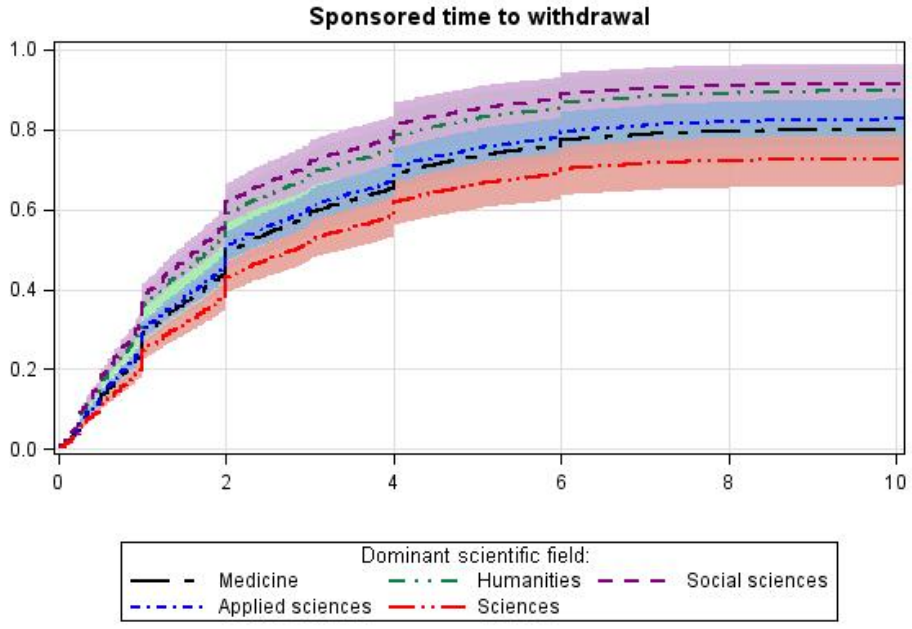


**Figure C.2:** The combined cumulative incidence function estimates and associated 95% pointwise confidence intervals for covariate 'dominant scientific field' and outcome withdrawal - imputation method B.

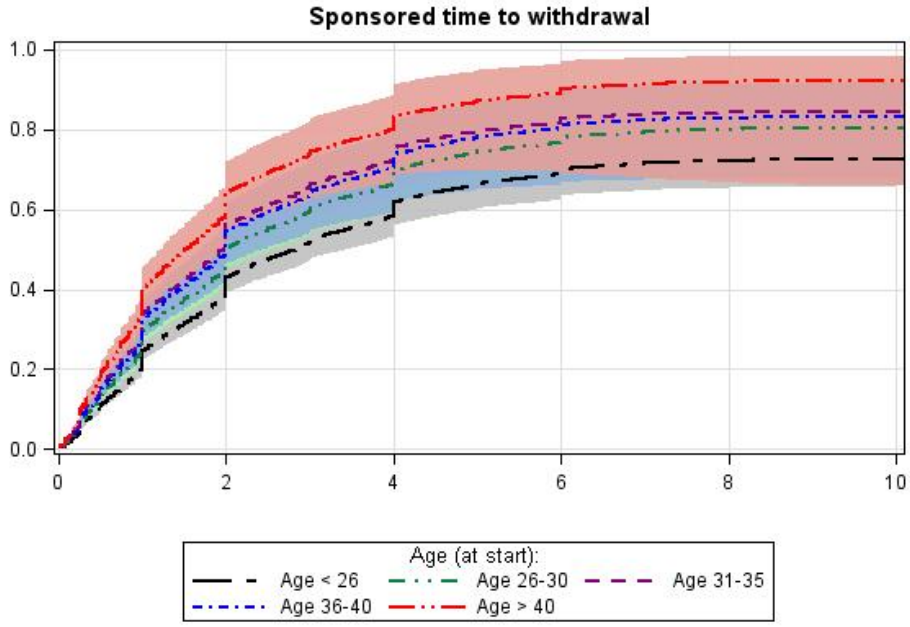


**Figure C.3:** The combined cumulative incidence function estimates and associated 95% pointwise confidence intervals for covariate 'age (at start)' and outcome withdrawal - imputation method B.

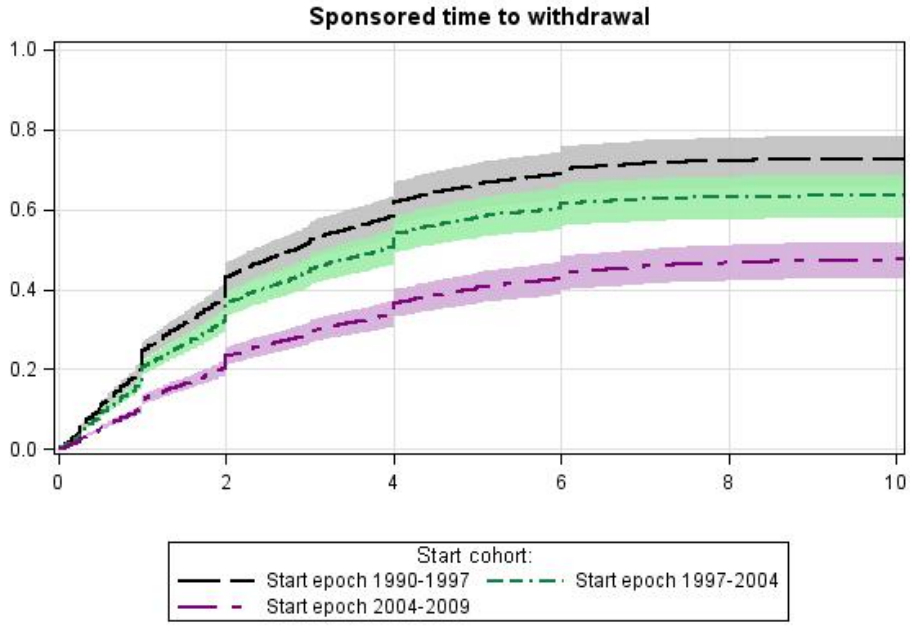


**Figure C.4:** The combined cumulative incidence function estimates and associated 95% pointwise confidence intervals for covariate 'start cluster' and outcome withdrawal - imputation method B.

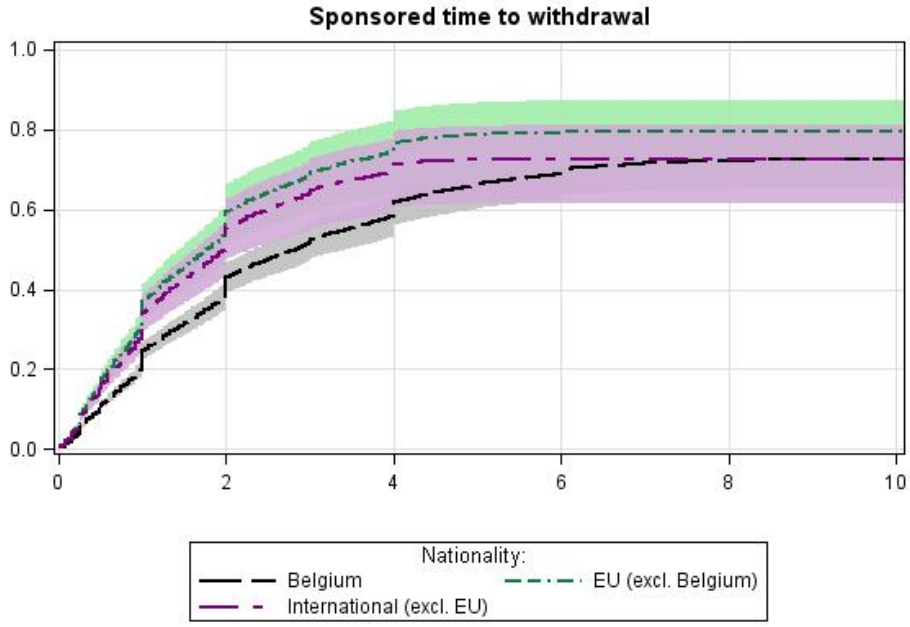


**Figure C.5:** The combined cumulative incidence function estimates and associated 95% pointwise confidence intervals for covariate 'nationality' and outcome withdrawal - imputation method B.

| Variable | $\hat{I}_w(t\|\mathbf{z}^*)$ and associated 95% confidence interval | | | | | |
|---|---|---|---|---|---|---|
| | $t=2$ | $t=4$ | $t=6$ | $t=8$ | $t=10$ | $p$-value |
| **Gender** | | | | | | $< 0.0001$ |
| [Male] | 0.41 [0.37, 0.44] | 0.62 [0.56, 0.67] | 0.69 [0.63, 0.75] | 0.72 [0.65, 0.78] | 0.73 [0.66, 0.78] | |
| Female | 0.44 [0.40, 0.48] | 0.66 [0.59, 0.72] | 0.74 [0.66, 0.80] | 0.77 [0.69, 0.83] | 0.78 [0.69, 0.84] | |
| **Dominant scientific field** | | | | | | $< 0.0001$ |
| [sciences] | 0.41 [0.37, 0.44] | 0.62 [0.56, 0.67] | 0.69 [0.63, 0.75] | 0.72 [0.65, 0.78] | 0.73 [0.66, 0.78] | |
| medicine | 0.47 [0.42, 0.51] | 0.69 [0.62, 0.74] | 0.77 [0.69, 0.82] | 0.80 [0.72, 0.85] | 0.80 [0.72, 0.86] | |
| humanities | 0.56 [0.51, 0.61] | 0.78 [0.70, 0.85] | 0.86 [0.76, 0.92] | 0.89 [0.77, 0.95] | 0.90 [0.78, 0.95] | |
| social | 0.59 [0.54, 0.63] | 0.81 [0.74, 0.87] | 0.88 [0.79, 0.93] | 0.91 [0.80, 0.96] | 0.92 [0.81, 0.96] | |
| applied | 0.48 [0.45, 0.52] | 0.71 [0.65, 0.75] | 0.79 [0.73, 0.83] | 0.82 [0.75, 0.87] | 0.83 [0.76, 0.88] | |
| **Nationality** | | | | | | $< 0.0001$ |
| [Belgian] | 0.41 [0.37, 0.44] | 0.62 [0.56, 0.67] | 0.69 [0.63, 0.75] | 0.72 [0.65, 0.78] | 0.73 [0.66, 0.78] | |
| European Union (excl. Belgium) | 0.57 [0.49, 0.63] | 0.77 [0.66, 0.85] | 0.79 [0.67, 0.87] | 0.79 [0.67, 0.87] | 0.79 [0.67, 0.87] | |
| Other | 0.53 [0.45, 0.60] | 0.71 [0.60, 0.80] | 0.73 [0.61, 0.81] | 0.73 [0.61, 0.81] | 0.73 [0.61, 0.81] | |
| **Dominant statute classification** | | | | | | $< 0.0001$ |
| Assistant lectureship | 0.16 [0.15, 0.18] | 0.28 [0.25, 0.31] | 0.34 [0.30, 0.37] | 0.36 [0.33, 0.40] | 0.37 [0.34, 0.40] | |
| Compet. scholarship (Flanders) | 0.06 [0.05, 0.07] | 0.10 [0.09, 0.11] | 0.11 [0.10, 0.13] | 0.12 [0.10, 0.13] | 0.12 [0.10, 0.13] | |
| Compet. scholarship (own university) | 0.10 [0.08, 0.12] | 0.17 [0.13, 0.21] | 0.19 [0.15, 0.23] | 0.19 [0.15, 0.23] | 0.19 [0.15, 0.23] | |
| Project funding (FWO, BOF, IUAP) | 0.16 [0.15, 0.18] | 0.28 [0.25, 0.30] | 0.32 [0.29, 0.35] | 0.33 [0.30, 0.36] | 0.33 [0.30, 0.37] | |
| [Project funding (other)] | 0.41 [0.37, 0.44] | 0.62 [0.56, 0.67] | 0.69 [0.63, 0.75] | 0.72 [0.65, 0.78] | 0.73 [0.66, 0.78] | |
| **Age (at start)** | | | | | | $< 0.0001$ |
| [$\leq 25$ years] | 0.41 [0.37, 0.44] | 0.62 [0.56, 0.67] | 0.69 [0.63, 0.75] | 0.72 [0.65, 0.78] | 0.73 [0.66, 0.78] | |
| $26-30$ years | 0.48 [0.43, 0.52] | 0.70 [0.63, 0.76] | 0.77 [0.69, 0.84] | 0.80 [0.71, 0.86] | 0.80 [0.72, 0.87] | |
| $31-35$ years | 0.53 [0.47, 0.59] | 0.76 [0.66, 0.83] | 0.82 [0.71, 0.89] | 0.84 [0.73, 0.91] | 0.85 [0.73, 0.92] | |
| $36-40$ years | 0.52 [0.44, 0.58] | 0.74 [0.62, 0.83] | 0.81 [0.66, 0.89] | 0.83 [0.68, 0.91] | 0.83 [0.68, 0.92] | |
| $> 40$ years | 0.61 [0.53, 0.69] | 0.83 [0.69, 0.91] | 0.90 [0.69, 0.97] | 0.92 [0.67, 0.98] | 0.92 [0.66, 0.98] | |
| **Start cohort** | | | | | | $< 0.0001$ |
| [01/10/1990 - 30/09/1997] | 0.41 [0.37, 0.44] | 0.62 [0.56, 0.67] | 0.69 [0.63, 0.75] | 0.72 [0.65, 0.78] | 0.73 [0.66, 0.78] | |
| 01/10/1997 - 30/09/2004 | 0.34 [0.31, 0.37] | 0.54 [0.49, 0.58] | 0.61 [0.55, 0.66] | 0.63 [0.57, 0.68] | 0.63 [0.58, 0.69] | |
| 01/10/2004 - 30/09/2009 | 0.22 [0.20, 0.24] | 0.37 [0.33, 0.40] | 0.43 [0.39, 0.47] | 0.47 [0.42, 0.51] | 0.47 [0.43, 0.52] | |

**Table C.2:** The combined cumulative incidence function estimates and associated 95% pointwise confidence intervals for each category of covariate values and outcome withdrawal - imputation method B.

# Bibliography

[1] Andersen, P. K., Gill, R. D. (1982), "Cox's Regression Model for Counting Processes: A Large Sample Study," *The Annals of Statistics*, Vol. 10, No. 4, 1100-1120.

[2] Baert, K., Goetghebeur, E. (2010), *Time to Ph.D.-completion: Technical Document*, unpublished report, Consortium Stat-Gent, Ghent University.

[3] Barnard, J., Meng, X. L. (1999), "Applications of Multiple Imputation in Medical Studies: from AIDS to NHANES," *Statistical Methods in Medical Research*, **8**, 17-36.

[4] Bie, O., Borgan, Ø., Liestøl, K. (1987), "Confidence Intervals and Confidence Bands for the Cumulative Hazard Rate Function and Their Small Sample Properties," *Scandinavian Journal of Statistics*, Vol. 14, No. 3, 221-233.

[5] Cheng, S. C., Fine, J. P., Wei, L. J. (1998), "Prediction of Cumulative Incidence Function Under the Proportional Hazards Model," *Biometrics*, Vol. 54, No. 1, 219-228.

[6] Fine, J. P., Gray, R. J. (1999), "A Proportional Hazards Model for the Subdistribution of a Competing Risk," *Journal of the American Statistical Association*, Vol. 94, No. 446, 496-509.

[7] Goetghebeur, E. (2010), *Survival Analysis*, unpublished course, Ghent University.

[8] Link, C. L. (1984), "Confidence Intervals for the Survival Function Using Cox's Proportional-Hazard Model with Covariates," *Biometrics*, Vol. 40, No. 3, 601-609.

[9] Karatzas, I., Shreve, S. E. (1998), Second Edition, *Brownian Motion and Stochastic Calculus*, Springer.

[10] Klein, P. J., and Moeschberger, M. L. (1997), *Survival Analysis: Techniques for Censored and Truncated Data*, New York: Springer-Verlag.

[11] Little, R. J. A., and Rubin, D. B. (2002), Second Edition, *Statistical Analysis with Missing Data*, New York: John Wiley.

[12] Othus, M., Li, Y., Tiwari, R.C. (2009), "A Class of Semiparametric Mixture Cure Survival Models with Dependent Censoring," *Journal of the American Statistical Association*, Vol. 104, No. 487, 1241-1250.

[13] Rubin, D. B. (1996), "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association*, Vol. 91, No. 434, 473-489.

[14] Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London.

[15] Schafer, J. L. (1999), "Multiple Imputation: a Primer", *Statistical Methods in Medical Research*, **8**, 3-15.

[16] Sy, J. P., Taylor, J. M. G. (2000), "Estimation in a Cox Proportional Hazards Cure Model," *Biometrics*, **56**, 227-236.

[17] Tsiatis, A. A. (1981), "A Large Sample Study of Cox's Regression Model," *The Annals of Statistics*, Vol. 9, No. 1, 93-108.

[18] Tu, X. M., Meng, X. L., Pagano, M. (1993), "The AIDS Epidemic: Estimating Survival after AIDS Diagnosis from Surveillance Data," *Journal of the American Statistical Association*, Vol. 88, No. 421, 26-36.

[19] Visscher, A., Varewyck, M., Baert, K., Goetghebeur, E. (2010), *Time to Ph.D. Degree or Withdrawal: A Competing Risks Analysis*, unpublished report, Consortium Stat-Gent, Ghent University.