Inference for a General Class of Models for Recurrent Events with application to cancer data

Juan R. González Ruiz
Center for Genomic Regulation and Catalan Institute of Oncology

Academic Dissertation for the Degree of PhD advised by

Dr. Edsel A. Peña
The University of South Carolina at Columbia
and
Dr. Pedro Delicado Useros
Universitat Politècnica de Catalunya

PhD thesis submitted to
Universitat Politècnica de Catalunya

Barcelona, April 2006

To Manoli,
Iker and Jan

To my parents

Acknowledgments

I specially wish to express my deepest gratitude to Professor Edsel Peña for his guidance and support all through the years of working on this PhD thesis. In spite of having a lot of work he spent a lot of his time reading hundreds of e-mails I sent him. He always had kind words about the work I did and encouraged me to continue working hard on my research. I am indebted to him, not only for everything I have previously mentioned, but also for being an excellent mentor, introducing me as a scientist, encouraging me to travel and bringing me into contact with many important researchers.

I also want to thank Professor Pedro Delicado for his guidance and for co-supervising this thesis with such great interest. He has supported me with advices and ideas, and he was always welcome to discuss my work and future directions with me.

A very special thanks goes out to Dr. Mercè Peris, head of Cancer Prevention and Control Unit from Catalan Institute of Oncology. She encouraged me to work on recurrent events settings applied in cancer data. She was very kind allowing me to make my work compatible in her unit with this PhD thesis.

I also would like to thank the two anonymous reviewers for the careful reading of the manuscript and for providing me lots of invaluable comments, constructive criticisms, and very interesting suggestions. Their helpful and thoughtful comments have improved many aspects of this thesis dissertation.

I like to thank Dr. Elizabeth Slate for her helpful feed-back on some parts of this dissertation. I would specially like to acknowledge her comments about the dynamic cancer model. I am also very grateful to her because she sent me some very useful `R` functions to summarize simulations

results showed in Chapter 5.

We thank Dr Octavi Servitge and Dr Cristina Muniesa for providing the non-Hodgkin's lymphoma data set and for their special effort in obtaining the treatment response information.

I must also acknowledge Professor Narcís Nabona for his helpful conversations on maximization problems, particularly on the topic about penalized approach. He was very kind to supply me with some Fortran routines to solve optimization problems.

Very special thanks I address to Dr. Virginie Rondeau with whom I shared ideas about penalized likelihood estimation for correlated survival data. She very kindly sent me the Marquart algorithm programmed in Fortran.

I would also like to express my gratitude to all people at the Cancer Prevention and Control Unit from Catalan Institute of Oncology. I would specially like to mention Mercè Margalef who always encouraged me to continue on my graduate career.

US NSF Grant DMS 02080301 is acknowledge because it allowed me to attend the International Conference on Reliability and Survival Analysis 2003 and met Professor Peña.

Grant to attend the seminars on *Empirical processes: theory and statistical applications* by UNESCO, held in EMS Summer School of Laredo is also acknowledge.

Es bastante difícil poder agradecer con un único párrafo la ayuda que Xavier Solé me ha prestado a lo largo de estos últimos años. Quizás su ayuda con el inglés haya sido una mera anécdota con el resto de cosas que he aprendido con él. Recuerdo que empezó ayudandome con el "maravillos" mundo de la creación de *dll*s hace más de cuatro años. Perdimos muchas horas hasta que conseguimos llamar por primera vez a Fortran desde `R`. También he recibido mucha ayuda a la hora de programar algunas cosas pues sus conocimientos en informática (a parte de otros campos de la ciencia) no parecen tener límite. Pero sin duda, su mejor ayuda a esta tesis, ha sido el estar disponible en cualquier momento. He podido hablar con él cuando necesitaba a alguien y siempre me ha dado excelentes consejos sobre todo cuando mi ánimo decaía. Encara que ja no estiguem treballant junts, ja saps que mai no deixarem de veure'ns. Persones com tu són les que realment

fan que la paraula AMIC tingui sentit i no sigui més que una definició al diccionari. Espero que la vida et torni tot allò que tu dones i que no canviïs mai. No te n'oblidis que tenim una cita anual al Xinés!.

También quisiera agradecer especialmente la ayuda que he recibido de Elisabet (*Looking*). Ya he perdido la cuenta de los párrafos en inglés que me ha corregido. También me gustaría agradecerle su amistad i así como las numerosas charlas que hemos tenido. Entre ellas, ha tenido que soportar algunos de mis rollos estadísticos.

A Klaus Langohr quiero agradecer su enorme paciencia que ha tenido al ayudarme a escribir esta tesis en LaTeX. Realmente ha sido como un manual on-line, en el buen sentido de la palabra. Siempre ha estado dispuesto a solucionar cualquier duda que me ha surgido. De hecho, todo lo poco que sé de LaTeX me lo ha enseñado él.

A pesar que sólo llevo unos pocos meses trabajando en el Centro de Regulación Genómica, también hay personas trabajando conmingo que merecen estar en estos agradecimientos. Lluís Armengol y Mónica Gratacós quizás sean las personas que más me han soportado durante los últimos meses en la realización de esta tesis. Este periodo final coincide sin duda con los meses más duros pues es cuando hay que preparar el manuscrito final y su defensa. Ambos me han ayudado a superar esta parte final de la tesis animándome y sobretodo dándome excelentes consejos.

Finalmente, y no por ser lo último es lo que menos ganas tengo de agradecer, quisiera mostrar mi mas sincero agradecimiento a mis padres. Desde que mi memoria puede recordar sólo veo el esfuerzo que realizaron para que todos sus hijos pudieran tener una educación que ellos no pudieron recibir. Mis ojos se llenan de lágrimas al recordar a mi padre queriendo que estudiara ya fuera verano o invierno, hubieran o no vacaciones. Este momento es el que culmina su propósito.

Esta tesis doctoral está dedicada completamente a mi mujer, Manoli, que me ha apoyado durante estos últimos años en todo momento y que siempre me ha animado a realizar estos estudios. Su amor, cariño, y ternura han hecho que este periodo de tiempo sea como un paseo por un precioso bosque lleno de bonitos lugares para tomar aliento. Nuestros hijos, Iker y Jan, han sido fruto del lugar más bello, sin duda.

Abstract

Survival analysis arises when we are interested in studying statistical properties of a variable which describes the time to a single event. This type of analysis occurs commonly in two areas: biomedicine and engineering. In biomedicine research it is known as survival analysis and refers often to the time from the beginning of the treatment to the occurrence of a particular condition or death. In some situations, we may observe that the event of interest occurs repeatedly in the same individual, such as when a patient diagnosed with cancer tends to relapse over time or when a person is repeatedly readmitted in a hospital. In this case we speak about survival analysis with recurrent events.

Recurrent nature of events makes necessary to use other techniques from those used when we analyze survival times from one single event. In this dissertation we deal with this type of analysis mainly motivated by two studies on cancer research that were created specially for this research. One of them belongs to a study on hospital readmissions in patients diagnosed with colorectal cancer, while the other one deals with patients diagnosed with non-Hodgkin's lymphoma. This last study is mainly relevant since we include information about the effect of treatment after relapses and some authors have stated the needed of developing an specific model for relapsing patients in cancer settings (Montoto et al., 2002). These two data sets together with two other existing examples and that have been extensively analyzed, have been used to illustrate the statistical methodology proposed in this work.

In this dissertation, we address two different types of statistical analysis similar to those one may carry out when we deal with survival analysis for a single event. This two types of analysis are known in the biomedical literature as univariate or multivariate analysis. Univariate analysis studies how one variable may modify the probability of observing a new recurrence. Our contribution to this problem is to propose a method to construct confidence intervals for

the median survival time in the case of recurrent event settings. Two different approaches are developed. One of them is based on asymptotic variances derived from two existing estimators of survival function (Peña et al., 2001 and Wang and Chang, 1999) while the other one uses bootstrap techniques. This last approach is useful since one of the estimators proposed by Peña et al. (2001) does not have any closed form for its variance yet. The new contribution to this work is the examination of the question of how to do bootstrapping in the presence of recurrent event data arising from a sum-quota accrual scheme and informativeness of right-censoring mechanism. Weak convergence is proved and asymptotic confidence intervals are built to according this result.

On the other hand, multivariate analysis addresses the problem of how incorporate more than one covariate in the analysis. In recurrent event settings, we also need to take into account that apart from covariates, the heterogeneity, the number of occurrences or specially, the effect of interventions after re-occurrences may modify the probability of observing a new event in a patient. This last point is a very important one since it has not been taken into consideration in biomedical studies yet. To address this problem, we base our work on this new model for recurrent events. Our contribution to this topic is to accommodate the situation of cancer relapses adopting the Peña and Hollander's model in which the effect of interventions is represented by an effective age process acting on the baseline hazard function. We call this model dynamic `cancer` model.

We also address the problem of estimating parameters of the general class of models for recurrent events proposed by Peña and Hollander (2004), where the dynamic cancer model may be seen as an special case of this general model. Two general approaches are developed. First approach is based on semiparametric inference, where a baseline hazard function is nonparametrically specified. The second one is a penalized likelihood approach. For the semiparametric inference we take two different strategies, depending on whether a frailty model is fitted or not. When frailties are included in the model, an EM algorithm is developed. Regarding penalized likelihood approach, two different strategies are also adopted. One of them was proposed in the shared frailty model context by Therneau et al. (2003). Their idea is based on penalizing the partial likelihood where the penalization bears on a regression coefficient. The second penalized approach, also applied in the shared frailty model, was proposed by Rondeau et al. (2003). Their method of estimation is based on the penalized full likelihood, and it gives a non-parametric estimation of the baseline hazard function using a continuous estimator. The solution is then approximated using splines. The main advantage of this method is that we can easily obtain smooth estimates of the hazard

function and an estimation of the variance of frailty variance, while in the other approaches this is not possible. In addition, this last approach has a quite less computational cost than the other ones. Simulations performed under different scenarios and sample sizes show the good properties of the proposed estimator. In addition, the results obtained using dynamic cancer model in real data sets, indicate that the flexibility of this method provides a safeguard for analyzing data where patients relapse over time and interventions are performed after tumoral reoccurrences.

Computational issue is another important contributions of this work to recurrent event settings. We have developed three `R` packages called `survrec`, `gcmrec`, and `frailtypack` that are available at CRAN, `http://www.r-project.org/`. These packages allow users to compute median survival time and their confidence intervals, to estimate the parameters involved in the Peña and Hollander's model (in particular in the dynamic cancer model) using EM algorithm, and to estimate this parameters using penalized approach, respectively.

Resumen

La necesidad del análisis de supervivencia aparece cuando necesitamos estudiar las propiedades estadísticas de una variable que describe el tiempo hasta que ocurre un evento único. Este tipo de análisis suele plantearse normalmente en dos áreas: biomedicina e ingeniería. En investigación biomédica, se conoce como análisis de supervivencia, y usualmente hace referencia, al tiempo desde el inicio del tratamiento hasta la ocurrencia de una condición en particular o la muerte. En algunas ocasiones, podemos observar que el evento de interes ocurre repetidamente en un mismo individuo, como puede ser el caso de un paciente diagnosticado de cáncer que recae a lo largo del tiempo o cuando una persona es reingresada repetidas veces en un hospital. En este caso hablamos de análisis de supervivencia con eventos recurrentes.

La naturaleza recurrente de los eventos hace necesario el uso de otras técnicas distintas a aquellas que utilizamos cuando analizamos tiempos de supervivencia para un evento único. En esta tesis, tratamos este tipo de ánalisis principalmente motivados por dos estudios en investigación en cáncer que fueron creados especialmente para este trabajo. Uno de ellos hace referencia a un estudio sobre readmisiones hospitalarias en pacientes diagnosticados con cáncer colorectal, mientras que el otro hace referencia a pacientes diagnosticados con linfomas no Hodgkinianos. Este último estudio es especialmente relevante ya que incluímos informacion sobre el efecto del tratamiento después de las recaídas y algunos autores han mostrado la necesidad de desarrollar un modelo específico para pacientes que presentan este tipo de enfermedades (Montoto et al., 2002). Estos dos conjuntos de datos, junto a otros dos existentes en la literatura biomédica y que han sido ampliamente analizados, han sido utilizados para ilustrar la metodología estadística que se propone en este trabajo.

En esta tesis, tratamos dos tipos diferentes de análisis estadísticos similares a los que se llevan a cabo cuando tratamos datos de supervivencia simples. Estos dos tipos de análisis son conocidos

en la literatura biomédica como análisis univariante y multivariante. El análisis univariante estudia cómo una variable puede modificar la probabilidad de observar una nueva ocurrencia. Nuestra contribución a este problema es proponer un método para construir intervalos de confianza para la mediana de supervivencia en el caso de eventos recurrentes. Para ello, hemos utilizado dos aproximaciones. Una de ellas se basa en las varianzas asintóticas derivadas de dos estimadores de la función de supervivencia (Peña et al., 2001 y Wang y Chang, 1999), mientras que el otro utiliza técnicas de remuestreo. Esta última aproximación es útil ya que uno de los estimadores propuestos por Peña et al. (2001) todavía no tiene una forma cerrada para su varianza. La nueva contribución de este trabajo es el estudio de cómo hacer remuestreo en la presencia de datos con eventos recurrentes que aparecen de un esquema conocido como "sum-quota accrual" y la informatividad del mecanismo de censura por la derecha que presentan este tipo de datos. Demostramos la convergencia débil y los intervalos de confianza asintóticos se construyen utilizando dicho resultado.

Por otro lado, el análisis multivariante trata el problema de cómo incorporar más de una covariable en el análisis. En problemas con eventos recurrentes, también necesitamos tener en cuenta que además de las covariables, la hetereogeneidad, el número de ocurrencias, o especialmente, el efecto de las intervenciones después de las reocurrencias pueden modificar la probabilidad de observar un nuevo evento en un paciente. Este último punto es muy importante ya que todavía no se ha tenido en cuenta en estudios biomédicos. Para tratar este problema, hemos basado nuestro trabajo en un modelo propuesto por Peña y Hollander (2004). Nuestra contribución a este punto es la adaptación de las recaídas en cáncer utilizando este nuevo modelo para eventos recurrentes en el que el efecto de las intervenciones se representa mediante un proceso llamado "edad efectiva" que actua sobre la función de riesgo basal. Hemos llamado a este modelo *modelo dinámico de cáncer* ("dynamic cancer model").

También tratamos el problema de la estimación de parámetros de la clase general de modelos para enventos recurrentes propuesta por Peña y Hollander (2004) donde el modelo dinámico de cáncer se puede ver como un caso especial de este modelo general. Hemos desarrollado dos aproximaciones. La primera se basa en inferencia semiparamétrica, donde la función de riesgo basal se especifica de forma no paramétrica. La segunda es una aproximación basada en verosimilitud penalizada. Para la inferencia semiparamétrica adoptamos dos estrategias diferentes dependiendo si ajustamos un modelo con fragilidad ("frailty") o no. El algoritmo EM se utiliza cuando la frag-

ilidad se incluye en el modelo. En cuanto a la aproximación mediante verosimilitud penalizada, de nuevo adoptamos dos estrategias diferentes. Una de ellas fue porpuesta por Therneau et al. (2003) en el contexto de un modelo de fragilidad compartida ("shared frailty model"). Su idea se basa en penalizar la verosimilitud parcial donde la penalización recae en los coeficientes de regresión. La segunda aproximación basada en penalización, también aplicada en este mismo modelo, fue propuesta por Rondeau et al. (2003). Su método de estimación se basa en penalizar la verosimilitud completa y da una estimación no paramétrica de la función de riesgo basal utilizando un estimador continuo. La solución se aproxima utilizando *splines*. La principal ventaja de este método es que podemos obtener fácilmente una estimación suave de la función de riesgo así como una estimación de la varianza de la varianza de la fragilidad, mientras que con las otras aproximaciones esto no es posible. Además este último método presenta un coste computacional bastante más bajo que los otros. Las simulaciones llevadas a cabo bajo diferentes escenarios y tamaños muestrales, han mostrado buenas propiedades de los estimadores propuestos. Además, los resultados obtenidos con datos reales, indican que la flexibilidad de este modelo es una garantía para analizar datos de pacientes que recaen a lo largo del tiempo y que son intervenidos depués de las recaídas tumorales.

El aspecto computacional es otra de las contribuciones importantes de esta tesis al campo de los eventos recurrentes. Hemos desarrollado tres paquete de `R` llamados `survrec`, `gcmrec` y `frailtypack` que están accesibles en CRAN, `http://www.r-project.org/`. Estos paquetes permiten al usuario calcular la mediana de supervivencia y sus intervalos de confianza, estimar los parámetros del modelo de Peña y Hollander (en particular el modelo dinámico de cáncer) utilizando el algoritmo EM y la verosimilitud penalizada, respectivamente.

Resum

La necessitat de l'anàlisi de supervivència apareix quan ens cal estudiar propietats estadístiques d'una variable que descriu el temps fins que succeeix un esdeveniment únic. Aquest tipus d'anàlisi se sol plantejar normalment en dues àrees: la biomedicina i l'enginyeria. En investigació biomèdica, es coneix com a anàlisi de supervivència, i usualment fa esment, al temps des de l'inici del tractament fins a l'esdeveniment d'una condició particular o la mort. En algunes ocasions, podem observar que l'esdeveniment d'interès se succeeix de manera recurrent en un mateix individu, com pot ser el cas d'un pacient diagnosticat de càncer que recau al llarg del temps o quan una persona és reingressada repetides vegades en un hospital. En aquest cas, parlem d'anàlisi de supervivència amb esdeveniments recurrents.

La naturalesa recurrent dels esdeveniments fa necessària la utilització de tècniques estadístiques distintes a les que emprem quan analitzem temps de supervivència en el cas d'un esdeveniment únic. En aquesta tesi, tractem aquest tipus d'anàlisi motivats per dos estudis de recerca en càncer que varen ser dissenyats per dur a terme aquest treball. Un d'ells gira en torn de readmissions hospitalàries en pacients diagnosticats amb càncer colorectal, per bé que l'altre fa referència a pacients diagnosticats amb limfomes no Hogkinians. Aquest darrer estudi és especialment rellevant ja que vàrem incloure informació sobre l'efecte del tractament després de les recaigudes, i alguns autors han mostrat la necessitat de desenvolupar un model específic per pacients que presenten aquest tipus de malalties (Montoto et al., 2002). Aquest dos conjunts de dades, juntament amb d'altres existents a la literatura biomèdica i que han estat àmpliament analitzats, han estat emprats per ilůlustrar la metodologia estadística que es proposa en aquest treball.

En aquesta tesi, hem tractat dos tipus diferents d'anàlisi estadístic similars als que es duen a terme quan es tracten dades de supervivència simples. Aquests dos tipus d'anàlisi són coneguts

en la literatura biomèdica com anàlisi univariant i multivariant. L'anàlisi univariant estudia com una variable pot modificar la probabilitat d'observar una nova ocurrència. La nostra contribució a aquest problema ha estat proposar un mètode per construir intervals de confiança per a la mediana de supervivència en el cas d'esdeveniments recurrents. És per això que hem utilitzat dues aproximacions. Una d'elles es basa en les variànces asimptòtiques derivades de dos estimadors de la funció de supervivència (Peña et al., 2001 y Wang y Chang, 1999), per bé que l'altre utilitza tècniques de remostratge. Aquesta darrera aproximació es útil ja que un dels estimadors proposats per Peña et al. (2001) encara no té una forma tancada per la seva variància. L'originalitat d'aquest treball rau en el disseny del remostratge quan es tracten dades amb events recurrents, que apareixen amb un esquema conegut com "sum-quota accrual" i la informativitat del mecanisme de censura per la dreta que presenten aquest tipus de dades. Demostrem la convergència dèbil i construïm els intervals de confiança assimptòtics utilitzant aquest resultat.

D'altra banda, l'anàlisi multivariant fa esment al problema de com incorporar més d'una covariable en l'anàlisi. En problemes amb esdeveniments recurrents també necessitem tenir en compte que, a més de les covariables, l'heterogeneïtat, el nombre d'ocurrències, o especialment, l'efecte de les intervencions després de les reocurrències poden modificar la probabilitat d'observar un nou esdeveniment en un pacient. Per abordar aquest problema, hem basat el nostre treball en un model proposat per Peña i Hollander (2004). La nostra contribució a aquest punt ha estat l'adaptació de les recaigudes en càncer utilitzant aquest nou model per a esdeveniments recurrents en què l'efecte de les intervencions es presenta mitjançant un procés anomenat "edat efectiva", que actua sobre la funció de risc basal. Hem anomenat a aquest model *model dinàmic de càncer* ("dynamic cancer model").

També tractem el problema de l'estimació de paràmetres de la classe general de models per esdeveniments recurrents proposat per Peña i Hollander (2004), on el model dinàmic de càncer es pot veure com un cas especial d'aquest model general. Hem desenvolupat dues aproximacions: la primera es basa en inferència semiparamètrica, on la funció de risc basal s'especifica de forma no paramètrica; la segona és una aproximació basada en versemblança penalitzada. En el cas d'inferència semiparamètrica adoptem dues estratègies diferents depenent de si ajustem un model amb fragilitat ("frailty") o de si no ho fem. L'algoritme EM és utilitzat quan la fragilitat és inclosa en el model. Pel que fa a l'aproximació mitjançant versemblança penalitzada, també adoptem dues estratègies diferents. Una d'elles va ser proposada per Therneau et al. (2003) en el context

d'un model de fragilitat compartida ("shared frailty model"). La seva idea es basa en penalitzar la versemblança parcial en què la penalització recau en els coeficients de regressió. La segona aproximació basada en penalització, també aplicada en aquest mateix model, va ser proposada per Rondeau et al. (2003). El seu mètode d'estimació es basa en penalitzar la versemblança completa i proporciona una estimació no paramètrica de la funció de risc basal utilitzant un estimador continu. La solució s'aproxima mitjançant splines. L'avantatge més important d'aquest mètode és que podem obtenir fàcilment una estimació suau de la funció de risc, així com una estimació de la variància de la variància de la fragilitat, ja que amb les altres aproximacions això no és possible. A més, aquest darrer mètode té un cost computacional bastant més baix que els altres. Les simulacions dutes a terme sota diferents escenaris i mides mostrals, han mostrat bones propietats dels estimadors proposats. A més, els resultats obtinguts amb dades reals indiquen que la flexibilitat d'aquest model es una garantia per analitzar dades de pacients que recauen al llarg del temps i que són intervinguts després de les recaigudes tumorals.

L'aspecte computacional és una altra de les contribucions importants d'aquesta tesi al camp dels esdeveniments recurrents. Hem desenvolupat tres paquets de `R` anomenats `survrec`, `gcmrec` y `frailtypack` que són accessibles a través de CRAN, `http://www.r-project.org/`. Aquest paquets permeten a l'usuari calcular la mediana de supervivència i els seus intervals de confiança, estimar els paràmetres del model de Peña i Hollander (en particular el model dinàmic de càncer) mitjançant l'algoritme EM y la versemblança penalitzada, respectivament.

Contents

List of Tables

List of Figures

Chapter 1

Introduction: State-of-the-art

1.1 Goals

The present thesis was primarily motivated by a study on hospital readmissions in patients diagnosed with colorectal cancer. Our aim was to examine whether there were differences between some clinical variables in the time until rehospitalization from the date of cancer surgery. We used a nonparametric estimation of survival function for this purpose including all observed times of hospital readmission related to the disease. Given the fact that the time between patients can be correlated it was necessary to use models more sophisticated than those normally employed in medical literature such as Kaplan-Meier or Cox model. We realized that we could not compute confidence intervals for median survival time when some of the existing estimators are used. Thus, our aim was to investigate bootstrapping schemes for estimating the sampling distribution of estimators of median survival time distribution in the presence of recurrent event data. Another important goal was to determine how to compute pointwise confidence intervals for median survival using asymptotic results from these estimators.

After dealing with these kind of data, we continued working on models for recurrent event data trying to extend some existing models to cancer settings. Thus, we deal with lymphoma data set which is mainly relevant since we include information about the effect of treatment after relapses. So far, this particular information has not been addressed in scientific publications and cannot be handled in a straightforward manner by statistical packages. In fact, some doctors asked for the needed to create "a model designed specifically for relapsing patients", as MacLaughlin

(2002) argued. For this reason, one of the main aims of this dissertation has been to investigate how to incorporate information about interventions in patients who relapse over time. To do so, we used a very flexible model designed for analyzing recurrent event data proposed by Peña and Hollander (2004). In particular, we used the response to the treatment after cancer relapses to model the effect of interventions after reoccurrences. We called this model dynamic *cancer* model. After that, another important goal was to develop statistical procedures for estimating model parameters involved in this new model such as penalized likelihood methods. Finally, we focused on implementing these methods in a widely used statistical software such as `R`, to help physicians to carry out their data analysis.

In the following sections we outline a survey on statistical methods for the analysis of recurrent event data. In particular, we discuss two important aspects in this type of studies: the *heterogeneity* and the *event dependence*. Finally, we illustrate how some models, mainly used in reliability settings, are useful to take into account another important topic which appears in these data: *the effect of interventions after re-occurrences*. The very flexible model proposed by Peña and Hollander (2004) which is the basis of our work is described at the end of this chapter.

1.2 Survival Analysis with Recurrent Events

Survival analysis arises when we are interested in studying statistical properties of the variable T, which describes the time to a single event. This type of analysis occurs commonly in two areas. In medical research it is known as survival analysis and refers often to the time from the beginning of the treatment to the occurrence of a particular condition or death. In engineering it is concerned with reliability and the analysis of failure times. That is, how long a component can be used until it fails. Counting process formulation and martingale theory have become the most used tool in the modern theory of these type of data. In the Appendix A we outline the main points in the use of counting processes as applied in survival analysis. Therneau and Grambsch (2000) can also be looked up for an overview from a very intuitive point of view.

However, in many other situations, we observe that the event of interest occurs repeatedly in the same subject such as when a patient diagnosed with cancer tend to relapse over time or when a person is repeatedly readmitted in a hospital. In that case we speak about survival analysis for recurrent events. Repeated events are prevalent in a vast variety of disciplines. These include

biomedicine, psychiatry, engineering, or sociology among others. Recurrent nature of events makes necessary to use other techniques from those used when we analyze survival times from one single event. In this case, random observation periods for each subject were allowed. Then, we observe the event of interest (always the same) at different times during the follow-up time. Finally, the last event is not observed since it would be observed after the end of study. Thus, censored data appears. This type of scheme which generates the data is known as *sum-quota accrual scheme* in the literature. This chapter starts by giving a representation of the sum-quota accrual scheme and introducing some notation of this scheme induced by the multiple occurrences and the randomness of follow-up period. The special nature of the data invalidates the direct use of martingales methods. To solve this problem, Section 1.2.2 illustrates another formulation, called *doubly-indexed process* that was first proposed by Gill (1981) and Sellke (1988) and extended by Peña et al. (2001). This technique will be useful to derive some theory for recurrent events. In Section 1.3 we give a survey of the statistical methodology for recurrent event data, focusing on nonparametric methods. Section 1.4 deals with two important topics which appear in studies with repeated measures: the heterogeneity and the event dependence. Both problems arise when the independent assumption is violated. This section also describes two methods to solve these problems: variance-corrected and frailty models. Finally, Section 1.4.3 describes conditional and marginal models which allow us to incorporate the effect of concomitant covariates, to control the event dependence, and the heterogeneity.

1.2.1 Sum-quota accrual scheme

Figure 1.1 illustrates a pictorial representation of our setting, known as *sum-quota accrual scheme*. We consider a patient diagnosed with cancer (e.g., an observational unit) which is being monitored for the occurrence of a recurrent event over a study period $[0, \tau]$, where τ may represent an administrative time, time of study termination, or some other right-censoring variable (see τ in Figure 1.1). The time τ could be a random time governed by an unknown probability distribution function $G(t) = \Pr(\tau \leq t)$. In recurrent event data, times are indexed in two scales, calendar and interoccurrence or gap times. Calendar times are defined by the sequence $S_0 \equiv 0 < S_1 < S_2 < S_3 < \ldots$ and correspond to the successive calendar times of event recurrences (see S in Figure 1.1). Interoccurrence times will be denoted by $T_1, T_2, T_3, \ldots$ and correspond to the time between successive event occurrences (see T in Figure 1.1). Thus, for $i = 1, 2, 3, \ldots$, $T_i = S_i - S_{i-1}$ and

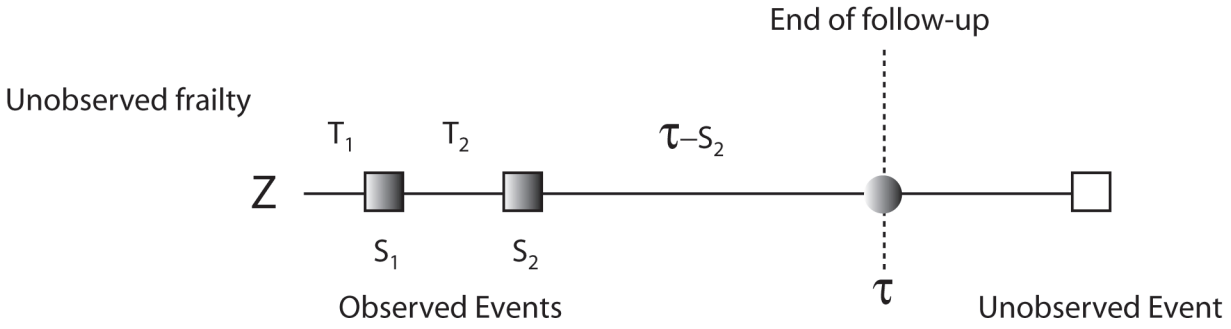

Figure 1.1: Graphical representation of the *sum-quota accrual scheme* for an individual.

$S_i = T_1 + T_2 + \ldots + T_i$. Over the observation period $[0, \tau]$, the number of event occurrences is $K = \max\{k \in \{0, 1, 2, \ldots\} : \ S_k \leq \tau\}$, which is a random variable whose distribution depends on the distributional properties of the inter-occurrence times T_is and the distribution $G(w) = P\{\tau_i \leq w\}$ of τ. As such, K is *informative* with regards to the distributional properties of event occurrences (in the example, $K = 2$). The fact of following-up patient for a fixed time, can lead to some event not being completely observed. Thus, Figure 1.1 illustrates that third event is not observed since it will appear after the end of study. So, T_3 is not observed completely: we only observe the censored time $\tau - S_2$. We notice that the interocurrence times are affected by unobserved variable called frailty (see Z in Figure 1.1). These frailties might make some patients have more recurrences than others depending on their values.

In the Appendix A we show that the risk indicator is one of the main components to use in counting processes and martingale theory. We illustrate that, in recurrent event situations, two different time scale are possible. Thus, we have two possible formulations for risk intervals: gap time or total time. Figure 1.2 shows the two types of risk intervals for two hypothetical subjects. For the first patient the interoccurrence times are 5 and 20 months and the third event is not observed after a follow-up of 15 months from the last event (15 is the censored time). These interoccurrence times correspond to calendar times 5, 25 and 40 respectively. For the second patient no events are observed after 30 months. Gap time represents the time from the prior event, e.g. the interoccurrence time. Thus, the gap time formulation (right bottom panel Figure 1.2) for our example indicates that the first subject is at risk of the first event during the interval $(0, 5]$, and for the second and third events during $(0, 20]$ and $(0, 15]$ respectively. On the other hand, total time is the time from a selected point. There exist several examples for this point, among others the date of diagnosis, the time of start of treatment, or the date of birth. In the

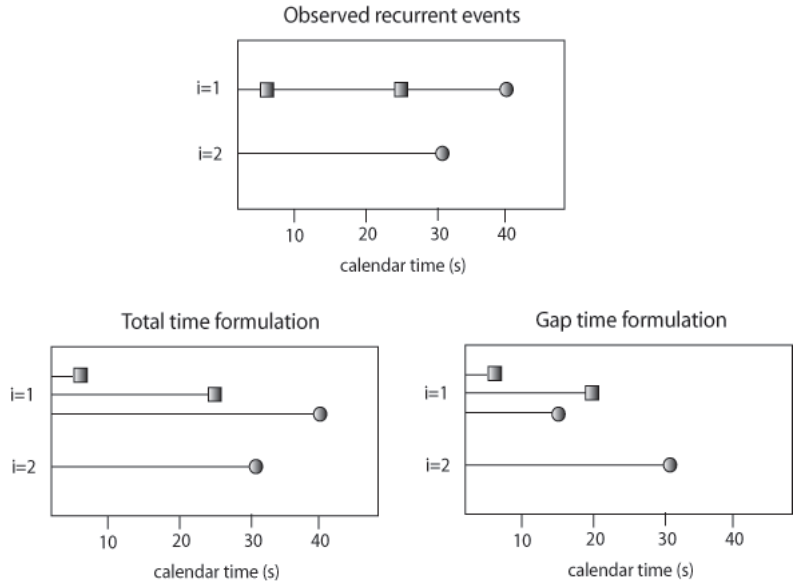

Figure 1.2: Illustrations of the risk interval formulation using an hypothetical data, where ■ corresponds to and event and • is censoring

example (bottom left panel Figure 1.2) , the first subject is at risk for the first, second and third events during the intervals $(0, 5]$, $(0, 25]$, and $(0, 40]$, respectively. Second subject is at risk during $(0, 30]$.

1.2.2 Doubly-indexed processes

Counting processes is a powerful tool in survival analysis. However, as Sellke (1988) observed, when dealing with recurrent event data, one should not only consider a calendar time formulation. We also need to consider gap time formulation. These idea originated from Gill (1981) and has been extended in Peña et al. (2001).

We begin by defining the following processes, which consider calendar time only. For $i = 1, \ldots, n$ and $s \geq 0$ let

$$N_i^\dagger(s) = \sum_{j=1}^{\infty} I(S_{ij} \leq s, S_{ij} \leq \tau_i) \tag{1.1}$$

be processes which count the number of failures for unit i at time s which have not been censored, and

$$Y_i^\dagger(s) = I(\tau_i \geq s), \tag{1.2}$$

which indicates if unit i has been censored at time s. Now, we define a filtration $\mathbf{G} = \{\mathcal{G}_s : s \geq 0\}$ such that $\{(N_i^\dagger(s), Y_i^\dagger(s)) : s \geq 0\}$ is $\mathbf{G}$-adapted. Moreover, let $A_i^\dagger(s) = \int_0^s Y_i^\dagger(v)\lambda(v - S_{iN_i^\dagger(v-)})dv$

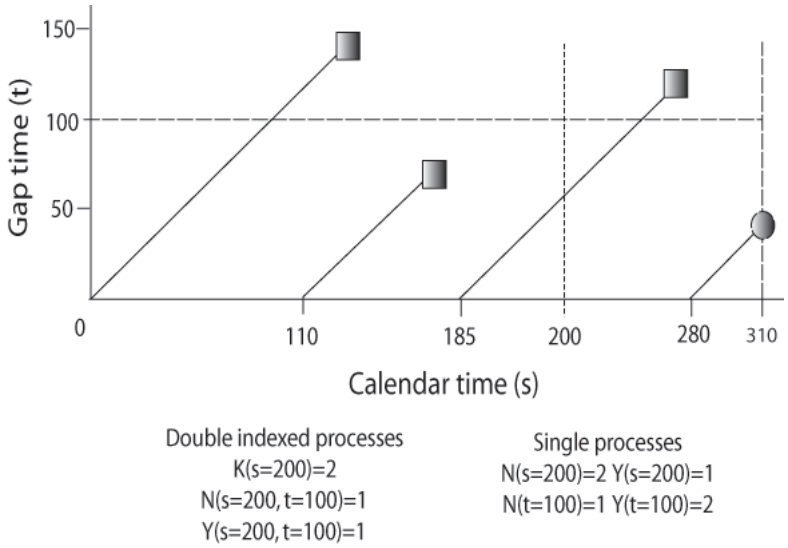

Figure 1.3: Doubly-indexed processes illustration for an hypothetical case

which makes $M_i^\dagger(s) = N_i^\dagger(s) - A_i^\dagger(s)$, $s \geq 0$ to be a local square-integrable **G**-martingale with predictable quadratic covariation process $\langle M_i^\dagger, N_i^\dagger \rangle(s) = A_i^\dagger(s) I\{i = i'\}$ (an introduction to these counting processes can be found in Peña et al., 2001).

Now, we have the main ingredients to define similar processes to those defined for single event case (see Appendix A) but for recurrent event situations. We introduce appropriate processes that are indexed by calendar time s and gap time t (doubly indexed processes) as follows. This processes are the basic ones considered in Peña et al. (2001) and Sellke (1988) and they provide the connection between the gap time formulation and that based on calendar time. For $i = 1, 2, \ldots, n$, let $Z_i(s,t) = I\{s - S_{iN_i^\dagger(s-)} \leq t\}$ be the indicator that for calendar time s at most t time units have elapsed since the time of the last event. For $s, t \geq 0$ we define

$$\begin{aligned} N_i(s,t) &= \int_0^s Z_i(v,t) dN_i^\dagger(v); \\ A_i(s,t) &= \int_0^s Z_i(v,t) dA_i^\dagger(v); \\ M_i(s,t) &= \int_0^s Z_i(v,t) M_i^\dagger(dv) = N_i(s,t) - A_i(s,t); \end{aligned} \tag{1.3}$$

$$Y_i(s,t) = \sum_{j=1}^{N_i^\dagger((s\wedge\tau_i)-)} I\{T_{ij} \geq t\} + I\{(s \wedge \tau_i) - S_{iN_i^\dagger((s\wedge\tau_i)-)} \geq t\}, i = 1, 2, \ldots, n.$$

We notice that $N(s,t)$ counts the number of observed events occurring over the calendar period $[0, s]$ whose interoccurrence times were at most t. On the other hand, $Y_i(s,t)$ counts the number

of observed events on calendar period $[0, s]$ whose interoccurrence times were at least t. Figure 1.3 shows a hypothetical case followed during 310 months. This patient presents three recurrences at months 110, 185, and 280 from the beginning of study. This fact implies that interoccurrence times are 110, 75, 55, and the censored time correspond to 30 months. Let us assume that we are interested in computing the single processes, $N(t)$ and $Y(t)$ for a selected interoccurrence time $t = 100$. In this case $N(t = 100) = 1$ and $Y(t = 100) = 2$. For the calendar time scale, $s = 200$, we have $N(s = 200) = 2$ and $Y(t = 200) = 1$. Now, let us assume that we would like to know double-indexed processes for both selected interoccurrence and calendar times. Using both time scales we observe that $N(s = 200, t = 100) = 1$ and $Y(s = 200, t = 100) = 1$.

1.3 Estimation of the survival function

The aim of this section is to give a survey of the statistical methodology for recurrent event data, focusing on nonparametric methods. We will also show how existing reliability models can be applied in biomedical or public health settings. Previously, we need to define a new process called *effective age* which is outlined in Section 1.3.1. Two estimators for the case of correlated interoccurrence times are described in Sections 1.3.3 and 1.3.4. The methodology described for nonparametric methods in reliability can be found in Hollander and Sethuraman (2002). The estimators for the correlated case are also described and compared in Peña et al. (2001) or in González and Peña (2004).

Statistical inference in the presence of recurrent event data has been considered by several authors such as Gill (1981), Vardi (1982a,b), McClean and Devine (1995), Soon and Woodroofe (1996), Wang and Chang (1999) (WC), and Peña, Strawderman, and Hollander (2001) (PSH). A main aspect with this type of data is the sum-quota accrual scheme which leads to an informative stopping rule as well as an informative censoring mechanism (see Section 1.2). Except in PSH (2001), most papers have used restrictive data accrual and censoring schemes for recurrent event data. In PSH (2001) it is assumed that the interoccurrence times represent independent and identically distributed (i.i.d.) observations from an unknown continuous distribution F, and that each subject is observed for a possibly random period of time. As a consequence, the number of event occurrences for a subject or unit is a random variable whose distribution depends on F, hence is informative about it. Moreover, the last observation for each subject is always right-

censored, with the censoring variable depending on the length of the observation period and on the previous interoccurrence times for that subject, rendering the censoring mechanism to become informative.

1.3.1 Effective age process

Reliability works with models which deal with repairable systems. These models incorporate the idea of *effective age*, also called *virtual age*, which describes the effect of repair. Kijima et al. (1988) introduce the notion of the virtual age as an improvement to minimal repair models. Maybe these models are more easily introduced in reliability settings than in biomedical problems due to the ability to monitor the effective age. Perhaps the main reason is that, upon failure, we can decide if the system is replaced by a new identical (perfectly repaired) or if it is restored to its state just before failure (minimally repaired). However, when we are dealing with patients, firstly, we cannot decide the degree of their "repair": we always want to eliminate all manifestation of the disease (e.g., "perfect" repair); and secondly, we always do not achieve the same response: we can obtain a complete remission of the disease, partial remissions or no response.

Although it seems natural to take into account the performed interventions, none of models which are mostly used in biomedical settings, described in Section 1.4.3, incorporate this effect. Before starting to describe how to adopt some existing reliability models in cancer data we focus more on the effect of interventions. For example, people with coronary heart disease are advised to alter their lifestyle by reducing stress level, quitting smoking habit, or doing regular and moderate physical activity. These advises try to modify the probability of presenting a new heart attack. Another example arises from patients with epileptic seizures, where recommendations in order to reduce the number of new seizures include sleeping 8 hours daily, avoiding exposition to flickering lights or reducing alcohol use. Considering cancer settings, in particular some indolent tumors, medical doctors may perform some prophylactic or curative interventions such as chemotherapy, radiotherapy or bone marrow transplantation to improve patient's disease status. Finally, in reliability the types of interventions are in general simplest since after having a break in the system the piece is repaired or replaced by a new one. There exist several reliability models which incorporate the effect of performed interventions. However, as we have pointed out above, these models have not been applied neither in biomedical nor in public health settings. Next sections give an overview of these models.

Another justification to use effective age based models is that most of repeated events survival models studied in medicine and public health problems are based on extensions of Cox model (see Section 1.4.3) which assumes that the effect after each intervention is always the same. For example, models which employ total time formulation assumes that all interventions produce a minimal improvement in the patient, e.g. disease continues in a stable manner. Models based on gap time formulation assume that all interventions lead to perfect recovery of the patient, e.g. disease disappears or a complete remission is achieved. However, it is not the case that the effect of treatment intervention will always be the same. For example, patients with a recurrent tumor usually are treated after each relapse and they may obtain a different response after each treatment. Patients sometimes achieve a complete remission, others minimal and others between none and complete remission, e.g. patients suffer a little improvement but the disease still remains (in reliability terms an intervention between perfect and minimal repair). Thus, as we will illustrate in Chapter 5, it can be very important to monitor the effective age process when we are dealing with some biomedical data. Although obtaining information about the effective age in health problems may be complicated, in cancer settings there exist the possibility to get this information using oncological terminology. The adoption of effective age to biomedical data, and in particular to some indolent cancers such as some lymphomas, has been an important part of this thesis.

The effective age for the i-th unit is defined as an observable processes $\{\mathcal{E}_i(s) : 0 \leq s \leq s^*\}$, $i = 1, 2, \ldots, n$, satisfying the following conditions: (I) $\mathcal{E}_i(0) = e_{i0}$, almost surely (a.s.), where $e_{i0}, i = 1, 2, \ldots, n$, are nonnegative real numbers; (II) $\mathcal{E}_i(s) \geq 0, i = 1, 2, \ldots, n$; and (III) On $[S_{ik-1}, S_{ik})$, $\mathcal{E}_i(s)$ is monotone and almost surely differentiable with a positive derivative $\mathcal{E}_i'(s)$.

To demonstrate it, Figure 1.3.1 shows the effective age for a unit. This process between 0 (or S_0) and S_2 is $\mathcal{E}(s) = s$. At the first event S_1, the unit is "minimally repaired". Thus, in this case patients have no improvement after first occurrence. Repair concept can be translated to biomedical context as "effect of treatment" because we cannot decide if a patient will receive either a "perfect" or a "minimal" repair. Physicians try to do the best for the patient and after the treatment we can observe if the intervention was "perfect" (e.g, patient achieves a complete remission) or "minimal" (e.g., disease is still active). A "perfect repair" occurs to the unit at the second failure time S_2. For medical people the treatment is also "perfect" since the disease disappears, although it is still in the patient but not active. After that, the effective age may be

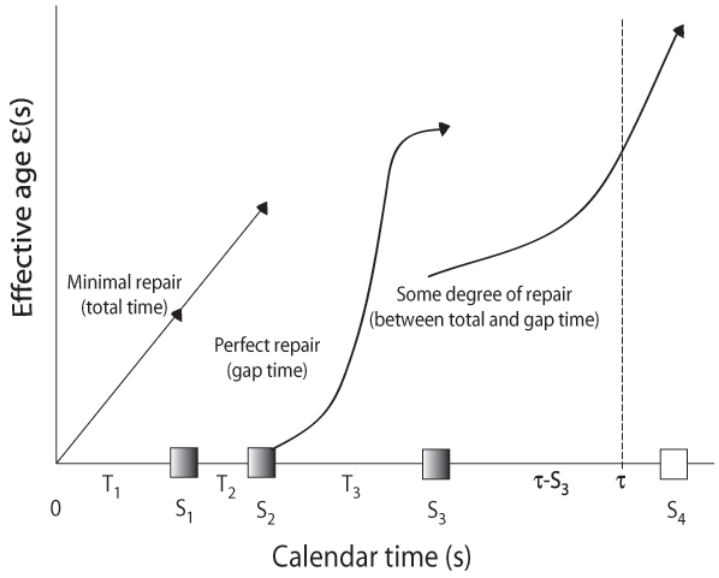

Figure 1.4: Effective age process, $\mathcal{E}(s)$, for an hypothetical case.

represented, in general, by a nonlinear function between S_2 and S_3 and also between S_3 and S_4. The unit is repaired after the third event between a minimal and perfect repair. In medicine we can say that the patient after third relapse suffers a little improvement produced by the treatment. The fourth failure, S_4, is not observed since the observational period for this unit is less than the calendar time. In consequence T_4 is censored at τ.

1.3.2 Reliability models

Next, we describe different existing models which take into account the effect of performed interventions. Most of them are models from reliability or engineering settings. We try to illustrate, using some real examples, how these models can be applied in biomedical or public health problems.

Minimal Repair Model

In this model, the repair restores the system to its state just before failure. Under this model, the hazard of event occurrence is identical to the intensity just prior to the event occurrence. In this case the effective age corresponds to calendar time, e.g., $\mathcal{E}(s) = s$ (in biomedical setting

this model is called total time model). Thus, in reliability terms one says that the subject is "minimally repaired" through the intervention. This model has been considered in Brown and Proschan (1983) and Lawless (1987).

In biomedical or public health settings we can say that treatment (intervention) is not effective, e.g. patient suffers a minimal or null disease improvement and the disease continues in a stable manner. Some examples can be found involving medical data. For example, let us suppose that patients are involved in a clinical trial. Let us also suppose that we are interested in estimating the time until a headache and that patients are treated with a new drug or placebo after each occurrence. The effective age for patients who received placebo must be calendar time (or total time) since the time until next headache will have the same probability distribution than the last headache occurrence. This is true because patients in the placebo arm are not intervened upon relapse since they do not receive any drug (of course, we omit placebo effect).

Perfect Repair Model

Upon failure, the failed system is replaced by a new one stochastically identical to the original so that $T_1, T_2, \ldots$ are i.i.d. according to F. These model can be seen as a gap time model where $\mathcal{E}(s) = s - S_{N^\dagger(s)}$ (where $s - S_{N^\dagger}(s)$ represents the elapsed time since the last event occurrence). In medicine one can find some examples when intervention completely treats the disease. A good example is recurrence in some cancers. For example, patients diagnosed to superficial bladder cancer can obtain a complete remission of their disease by means of a surgery intervention after each relapse. It is possible to get a complete remission upon each relapse because we are dealing with an indolent disease which can be controlled easily by means of the surgery. In this case, the perfect repair can be understood as some markers for the disease disappearance. At this moment the disease is not active, but disease is still in the patient since probably disease appears again as a new relapse or recurrence.

Brown and Proschan (1983) (BP) generalized the minimal repair model by allowing two types of repairs. After each failure, the system can be perfectly repaired with probability p and, with probability $1 - p$, a minimal repair is performed. If the success probability p is made to depend on the time of event occurrence, the Block et al. (1985) (BBS) model is obtained. Hollander et al. (1992) and Presnell, Hollander, and Sethuraman (1994) gives more details of this model. It is difficult to find a biomedical example where the effective age follows this model.

Kijima's Models

Models previously outlined are very useful when we deal with reliability data since we can easily control when a perfect or minimal repair is made. However, dealing with humans to assess the result of performed treatment interventions is often more complicated than if we deal with repairable systems. Thus, in biomedicine or public health setting, it is possible that interventions produce an improvement but not sufficient to say that the disease disappear completely. In reliability terms, one says that the degree of repair is between minimal and perfect.

Under this model, in medical terms, the improvement observed in patient can be classified according to a "degree-of-improvement". Kijima (1989) introduced models that allow improvements better than minimal but not necessarily as good as perfect in the reliability case. Kijima's reliability model restores the repaired item to an effective age that depends on its age just before failure as well as on "degree-of-repair" random variables. We let A_{j+1} denote the effective age of the system after the jth repair with $A_1 = 0$ (by definition). Let D_j, $j \geq 1$ denote the degree of repair random variables. They are assumed to be independently distributed on $[0, 1]$ and independent of other processes. Let $\bar{F} = 1 - F$ be the distribution function of interoccurrence times, T, then:

In Kijima's model I

$$P(T_j > x \mid T_1, \ldots, T_{j-1}, D_1, \ldots, D_{j-1}) = \frac{\bar{F}(x + A_j)}{\bar{F}(A_j)}, \tag{1.4}$$

where

$$A_j = \sum_{i=1}^{j-1} D_i T_i, \ j > 1$$

In Kijima's model II, $P(T_j > x \mid T_1, \ldots, T_{j-1}, D_1, \ldots, D_{j-1})$ is the same than in (1.4) but with the specification

$$A_j = \sum_{k=1}^{j-1} \left(\prod_{i=k}^{j-1} D_i \right) T_k, \ j > 1.$$

The effective age for this model is $\mathcal{E}(s) = A_{N^\dagger(s-)} + s - S_{N^\dagger(s-)}$.

Kijima's model I assumes that repairs served only to remove damage created in the last failure. However, model II assumes that the repair could remove all damage accumulated up to that point in time. Thus, if we are thinking on applying this model to cancer data, maybe model II is more appropriate than model I. For example, a patient diagnosed with a tumor relapses twice and

at each relapse chemotherapy just improves slightly the disease (e.g., patient achieves a partial remission). Then, the patient relapses again and is treated. In this case the chemotherapy makes that patient achieves a complete remission. Obviously, this complete remission means that the disease disappears completely (all damage accumulated) and not only the damage produced by the last recurrence. This model is used to connect reliability and biomedical models. It will be further discussed in Chapter 5.

Other reliability models

There exists other reliability models such a those proposed by Dorado et al. (1997), Last and Szekli (1998), or Kvam and Peña (2003). These models allows for other characteristics in reliability problems that in our opinion are difficult to be applied in health settings. In particular, Dorado et al. (1997) define a general repair model that contains many models previously mentioned and introduces new models as well. To do so, the authors introduce the term "life supplements" which could be viewed as improvement effects attributable to the performed interventions.

All previous reliability models are useful when independent assumption can be assumed. However, in the biomedical context it is somewhat restrictive because in biomedical settings the interoccurrence times may be correlated. To solve this problem, Wang and Chang (1999) and Peña et al. (2001) (PSH) propose two estimators under the case where the within-subject interoccurrence times are not independent. In particular, PSH describe an estimator which assumes that the interoccurrence times follow a gamma frailty model. We next describe these estimators.

1.3.3 Peña-Strawderman-Hollander estimator

Before going through the estimator for the correlated case we first outline the one proposed for the independent case. Peña et al. (2001) developed a nonparametric maximum likelihood estimator of the inter-event time survivor function under the assumption of i.i.d. model. This generalizes the product-limit estimator to the situation where the event is recurrent. This also generalizes Gill's estimator by allowing each process to be observed over a random time where the times are i.i.d. according to a distribution G. To describe this estimator, we first need to introduce some notation. For a given calendar time s and a gap time t, we define

$$K_i(s) = \sum_{j=1}^{\infty} I\{S_{ij} \leq s\},$$

and $N(s,t)$ and $Y(s,t)$ as in Section 1.2.2. The PSH (2001) generalized product-limit estimator of the common survivor function $\bar{F}$ of the event interoccurrence times is given by

$$\hat{\bar{F}}(s,t) = \prod_{w \le t} \left[1 - \frac{N(s, \Delta w)}{Y(s,w)} \right]. \tag{1.5}$$

The authors showed that the variance of this estimator is given by

$$\mathbf{V}\{\hat{\bar{F}}(s,t)\} = \bar{F}(s,t)^2 \sigma^2_{PSH}(s,t) \tag{1.6}$$

where $\sigma^2_{PSH}(s,t)$ is defined in Peña et al. (2001). An estimate of the variance is

$$\widehat{\mathbf{V}\{\hat{\bar{F}}(s,t)\}} = \hat{\bar{F}}(s,t)^2 \hat{\sigma}^2_{PSH}(s,t)$$

where

$$\hat{\sigma}^2_{PSH}(s,t) = \int_0^t \frac{N(s,dw)}{Y(s,w)[Y(s,w) - N(s,\Delta w)]} \tag{1.7}$$

This estimator is identical in form to the variance for the usual product-limit estimators for right-censored data. However, it is important to recognize that the at-risk processes are necessarily more complex. In this case, the computational form of $\hat{\sigma}^2(s,t)$ when there are no tied interoccurrence times is

$$\hat{\sigma}^2(s,t) = \sum_{i=1}^{n} \sum_{j=1}^{N_i^\dagger(s-)} \frac{I\{T_{ij} \le t\}}{Y(s,T_{ij})[Y(s,T_{ij}) - 1]},$$

where $Y(s,T_{ij}) = \sum_{i=1}^{n} \{ \sum_{k=1}^{N_i^\dagger(s-)} I\{T_{ik} \ge T_{ij}\} + I\{(s \wedge \tau_i) - S_{iN_i^\dagger(s-)} \ge T_{ij}\} \}$.

Following Peña et al. (2001) and their results from Section 2.2, an estimator indexed in the gap times, t, may be defined as follows:

$$\hat{\bar{F}}(t) = \lim_{s \to \infty} \hat{\bar{F}}(s,t) = \prod_{w \le t} \left[1 - \frac{N(\Delta w)}{Y(w)} \right]. \tag{1.8}$$

This estimator has variance given by

$$\mathbf{V}\{\hat{\bar{F}}(t)\} = \bar{F}(t)^2 \sigma^2_{PSH}(t), \tag{1.9}$$

where $\sigma^2_{PSH}(t) = \lim_{s \to \infty} \sigma^2_{PSH}(s,t)$ and may be estimated by

$$\hat{\sigma}^2_{PSH}(t) = \int_0^t \frac{N(dw)}{Y(w)[Y(w) - N(\Delta w)]} \tag{1.10}$$

Peña et al. (2001) also propose an estimator referred to as FRMLE (FRailty Maximum Likelihood Estimator) in their paper, of the common marginal distribution of the interoccurrence time distribution in the case of correlated interoccurrence times induced by a gamma frailty model. That model will be discussed later.

1.3.4 Wang-Chang estimator

Wang and Chang (1999) (WC) propose an estimator of the common marginal survivor function in the case where the within-unit interoccurrence times are correlated. They consider a correlation structure which is quite general, and includes as special cases both the i.i.d. and gamma frailty models. Setting all their weights to be equal to 1, their estimator is described below. For the ith unit, define

$$K_i^* = I\{K_i = 0\} + K_i I\{K_i > 0\}$$

and define the processes

$$d^*(t) = \sum_{i=1}^n \frac{1}{K_i^*} \sum_{j=1}^{K_i} I\{T_{ij} = t\};$$
$$R^*(t) = \sum_{i=1}^n \frac{1}{K_i^*} \left[\sum_{j=1}^{K_i} I\{T_{ij} \geq t\} + I\{\tau_i - S_{iK_i} \geq t\} I\{K_i = 0\}\right],$$

and with $\mathcal{T}$ denoting the set of distinct observed complete interoccurrence times for the n units. The WC estimator of $\bar{F}$ is given by

$$\hat{S}(t) = \prod_{\{T_k \in \mathcal{T};\ T_k \leq t\}} \left[1 - \frac{d^*(T_k)}{R^*(T_k)}\right]. \tag{1.11}$$

This estimator possesses less bias than the generalized product-limit estimator when interoccurrence times are correlated within subjects. For more discussions concerning these estimators and the comparisons of their properties, refer to PSH (2001). The variance of WC estimator is given by

$$\mathbf{V}\{\hat{S}(t)\} = S(t)^2 \sigma_{WC}^2(t), \tag{1.12}$$

where $\sigma_{WC}^2(t)$ is defined in Wang and Chang (1999) (see ϕ in their paper). This variance may be estimated by

$$\widehat{\mathbf{V}\{\hat{S}(t)\}} = \hat{S}(t)^2 \hat{\sigma}_{WC}^2(t)$$

where

$$\hat{\sigma}_{WC}^2(t) = \sum_{i=1}^n \frac{1}{K_i^*} \sum_{j=1}^n \frac{d^*(T_j)}{R^{*^2}(T_j)}. \tag{1.13}$$

1.4 Within-subject correlation

The Cox model is a widely used model for survival analysis. However, this semi-parametric model assumes that events occur independently, i.e., that the timing and occurrence of repeated events

is unrelated to the initial and subsequent occurrences. In particular, two features of repeated events processes violate the independence assumption: *heterogeneity across individuals* and *event dependence*.

Heterogeneity is produced because some subjects have a higher (or lower) event rate than other subjects due to unknown, unmeasured, or unmeasurable effects. For example, some patients have a genetic susceptibility to develop some disease, experiencing their first, second, third, etc., relapse more quickly than the rest of population. Perhaps, investigators do not know how to measure this susceptibility that are believed to be relevant for relapses. On the other hand, once again in biomedical setting, the occurrence of one event may make further relapses more or less likely. This event dependence may be produced by a learning process or by biologically weakening/strengthening the body and implies that the occurrence of a relapse itself may raise (or lower) the subsequent event rate. This dependence violates the independence assumption of the Cox model.

Any correlation among events (produced by heterogeneity, event dependence or jointly) has two important consequences. First, estimates are inefficient leading to incorrect estimates of standard errors. Kelly and Lim (2000) point out that under heterogeneity, the standard errors are too small. These problems may lead to incorrect conclusions regarding statistical significance in treatment effects. Second, violation of the independence assumption may induce biases in estimated effects. Aalen (1988) shows that unobserved heterogeneity produces attenuated estimated of treatment effects. Further, event dependence implies an event dependent baseline hazard rate, and possibly event dependence covariate effects.

Thus, variations of the Cox model have been proposed for estimation under recurrent events. In particular, many of *variance-corrected* and *frailty/random effects* models have been developed to account for correlations in event times that result from unknown sources of heterogeneity. Some of these models also attempt to control for event dependence by allowing baseline hazard rates to vary by event number (stratified models).

1.4.1 Variance-corrected models

Robust variance models are used to account for unobserved heterogeneity or event dependence. They are fit as though the data consist of independent observations, and then the variance is "fixed". Robust standard errors are based on the idea that observations are independent across

groups or clusters but not necessarily within groups. The robust variance estimator is then based on a "sandwich" estimate:

$$V = I^{-1}BI^{-1}$$

where I^{-1} is the usual variance estimate of a Cox model (the inverse of the information matrix I) and B is a correction factor. Therneau and Hamilton (1997) suggests that a natural correction, in survival settings, is to use the jackknife estimate of variance. We notice that the estimates of the variance-corrected standard errors are almost always larger than those from a "naive" estimates based on I^{-1} due to the unobserved intra-case correlations that are generally positive.

1.4.2 Frailty models

As we have mentioned previously, in the analysis of survival data it is frequently assumed that the history for the subjects under study are all statistically independent (at least conditionally on observed time-fixed covariates). In other words, the interoccurrence times appear in an independent manner. However, in many occasions, some patients are intrinsically more or less susceptible to experiencing the event of interest than are others. We may describe this fact as follows. Let us assume that an event occurs in a subject through a hazard function $\lambda(s)$, then another subject with frailty Z has hazard $Z\lambda(s)$. Thus, if the frailty is less than 1, then the subject tends to experience the event of interest at an later time, whereas the opposite occurs if Z is greater than 1. Other authors such as Vaupel et al. (1979), Hougaard (1984), Vaupel and Yashin (1985a,b), Hougaard (1987) interpret the frailty as modelling the effect of unobserved covariates which leads to some patients having more events than others. Frailty models assume that the distribution of these individual effects can be known, or at least approximated as Clayton (1978), Oakes (1982), Clayton and Cuzick (1985), (Hougaard, 1986a,b), Andersen et al. (1993, Chapter IX), and Hougaard (2000) suggest.

A specific type of model that results in correlated within-subject interoccurrence times is a multiplicative shared frailty model (see Andersen et al., 1993, Chapter IX; or Murphy, 1995 for the shared frailty model without covariates). A shared frailty model can be considered as a random effect model. In this model it is postulated that there exists for each subject an unobservable positive-valued frailty Z_i such that, conditionally on $Z_i = z_i$, the interoccurrence times $T_{i1}, T_{i2}, \ldots$

are i.i.d. with common conditional survivor function

$$\bar{F}(t \mid Z_i = z) = [\bar{F}_0(t)]^z = \exp\left(-z \int_0^t \lambda_0(u)du\right),$$

where $\lambda_0(\cdot)$ is the hazard function associated with a baseline survivor function $\bar{F}_0(\cdot)$. The frailties $Z_1, Z_2, \ldots, Z_n$ are assumed to be i.i.d. from an unknown distribution function H. In general, the Z's are not observed, so we are interested in estimating the marginal survivor of T_{ij}, which under this model is given by

$$\bar{F}(t) = \mathbf{E}\left\{\exp(-Z_1\Lambda_0(t))\right\} \tag{1.14}$$

where $\Lambda_0(t) = -\log[\bar{F}_0(t)]$ is the cumulative hazard function of $\bar{F}_0$.

A common choice of the unknown frailty distribution H is a gamma distribution with shape and scale parameters both equal to an unknown parameter α in order to guarantee identifiability. In this case, the common marginal survivor function $\bar{F}$ in (1.14) becomes

$$\bar{F}(t) = \left[\frac{\alpha}{\alpha + \Lambda_0(t)}\right]^{\alpha}. \tag{1.15}$$

The parameter α controls the degree of association between interoccurrence times within subject. In particular, as α increases (decreases), association between interoccurrence times decreases (increases). Letting $\alpha \longrightarrow \infty$, we obtain a model with independent interoccurrence times in which the T_{ij} has a common survivor function of $\bar{F}_0$.

Peña et al. (2001) showed that the estimation of α and Λ_0 of (1.15) can be obtained via the maximization of the marginal likelihood function of α and $\Lambda_0(\cdot)$ and with an implementation of the expectation-maximization (EM) algorithm (see, for details, Peña et al., 2001). This estimator of (1.15) is of form

$$\tilde{\bar{F}}(s,t) = \left[\frac{\hat{\alpha}}{\hat{\alpha} + \hat{\Lambda}_0(s,t)}\right]^{\hat{\alpha}}$$

where $\hat{\Lambda}_0(s,t)$ is an estimator of the marginal cumulative hazard function $\Lambda_0(t)$.

As pointed out in Peña et al. (2001), model parameter estimation can be carried out using the ideas in Nielsen et al. (1992) who suggest that the EM algorithm proposed by Dempster et al. (1977) can be used to obtain the maximizing values (further details in how to apply the EM algorithm when we deal with frailty models may be found in subsection 1.4.2).

The shared frailty model can be extended to a model with covariates by means of a multiplicative regression in which frailties act multiplicatively on hazard like in the Cox model. Thus,

proportional hazard model for subject i is written as:

$$\lambda_i(s \mid Z_i, X_i) = Z_i \lambda_0(s) \exp\{\beta' X_i(s)\} \tag{1.16}$$

where Z_i is the frailty for the ith subject.

In applied work, the most widely parametric distribution assumed for frailties is the gamma distribution (Vaupel et al., 1979; Clayton and Cuzick, 1985; Klein, 1992; and Andersen et al., 1993). However, another distributions have been employed (see for instance Hougaard, 1994).

Maximum likelihood estimation in the semiparametric frailty model (with gamma-distributed frailties) may be performed using EM algorithm (Dempster et al., 1977) as suggested Gill (1985). This method was further discussed by Nielsen et al. (1992), Klein (1992), and Guo and Rodriguez (1992). Assuming a parametric model, another possibility is to make direct use of the observed data (partial) likelihood (Aalen, 1988). McGilchrist and Aisbett (1991) and McGilchrist (1993) use partial likelihood procedures assuming that frailties follow a log-normal distribution. Penalized likelihood methods has also been studied by several authors: Verwerj and Houwelingen (1994), Therneau et al. (2003), Therneau and Grambsch (2000, section 9.6), Ripatti and Palmgren (2000). Next sections illustrate EM and penalized approaches to make inference for the frailty model.

EM algorithm for frailty models

We assume that the Z's are i.i.d. from a distribution $H(\cdot \mid \xi)$. A common choice of H is the gamma distribution with mean 1 and variance $1/\xi$, e.g. $H = \Gamma(\xi, \xi)$. Imposing this restriction is needed to have identifiability as we have pointed out in Subsection 1.4.2. Thus, if Z_i are known, the complete log-likelihood is given by

$$\sum_{i=1}^{n} \left(\int_0^\infty Y_i(t)[log(\lambda_0(t)) + log(Z_i) + \beta' X_i(t)] dN_i(t) - \int_0^\infty Y_i(t) Z_i \exp\{\beta' X_i(t)\} \lambda_0(t) dt + \log f(Z; \theta) \right).$$

As Z can be viewed as missing data, the problem can be approached using the EM algorithm implemented by Nielsen et al. (1992) in counting process frailty models. The main ingredients of this algorithm are two steps: E (expectation) and M (maximization). We outline briefly both steps, see Andersen et al. (1993) or Nielsen et al. (1992) for further details. Given $(\Lambda_0(\cdot), \beta)$, the observed times, and the covariates, the conditional expectation of Z_i is

$$\text{E-step: } \mathbf{E}\{Z_i \mid \Lambda_0(\cdot), \beta)\} = \frac{\xi + N_i(s^*)}{\xi + \int_0^{s^*} Y_i(v) \exp\{\beta' X_i(v)\} \lambda_0(v) dv},$$

where s^* denotes an upper limit of observation times. In the M-step $\Lambda_0(\cdot), \alpha, \beta$ can be estimated using procedures describe in Appendix A because after the E-step Z's are set to $\hat{Z}_i$. Thus, for example,

$$\text{M-step: } \hat{\Lambda}(t) = \int_0^t \frac{dN(u)}{\sum_{i=1}^n \hat{Z}_i Y_i(u) \exp\{\beta' X_i(u)\}}.$$

To estimate ξ Andersen et al. (1993, Subsection IX.4.2) propose to maximize the marginal profile likelihood for ξ.

Penalized likelihood estimation

The penalized regression formulation for the frailty model is easily developed by making the change $Z_j = \exp(z_j)$. Thus, equation (1.16) converts to

$$\lambda_i(s \mid z, X_i) = \lambda_0(s) \exp\{\beta' X_i(s) + z' M_i\},$$

where z' is a vector of frailties and M is a matrix of n indicator variables such that $M_{ij} = 1$ when observation i is a reoccurrence of individual j and 0 otherwise.

The penalized likelihood method was introduced by Good and Gaskins (1971) in the context of nonparametric probability density estimation. Its use in Cox regression model estimation was proposed by several authors: Zucker and Karr (1990), McGilchrist and Aisbett (1991), Verwerj and Houwelingen (1994), Therneau et al. (2003), and Ripatti and Palmgren (2000). The idea of this method is to maximize penalized partial likelihood equation

$$PPL = l(\beta, z) - g(z \mid \theta) \tag{1.17}$$

over both β and z. In this equation $l(\beta, z)$ is the log of the Cox partial likelihood given by

$$l(\beta, z) = \sum_{i=1}^n \int_0^\infty \left[Y_i(t)(\beta' X_i(t) + z' M_i) - \log \left(\sum_{j=1}^n Y_j(t) \exp\{\beta' X_j(t) + z' M_j\} \right) \right] dN_i(t),$$

and g is a penalty function chosen by the investigator to restrict the values of z. The parameter θ is a tuning constant which may be pre-specified or adapted to the data to control the amount of shrinkage. Typically, we are interested in choosing the penalty function to "shrink" z toward zero.

Using Newton-Raphson method we can estimate β and z solving the score equations. The penalty function does not involve β, so we can compute $\partial PPL/\partial\beta$ using $\partial PL/\partial\beta$ which is the

usual partial likelihood for the Cox model. Therefore, the score equation for β are the same as those for Cox model but incorporating z as an offset term (see Therneau and Grambsch, 2000 for further details).

In addition to the score vectors, the maximization algorithm requires the Hessian of the penalized partial log-likelihood which is given by

$$H = \mathbf{I} + \begin{pmatrix} 0 & 0 \\ 0 & g'' \end{pmatrix}$$

where $\mathbf{I} = \mathbf{I}(\beta, z)$ is the second derivative matrix of partial likelihood, also called information matrix, and g'' denote the second derivative of g.

1.4.3 Cox extension models

Many survival models based on Cox proportional hazards have been proposed that handle multiple event data (see Therneau and Grambsch, 2000, Chapter 8; Therneau and Hamilton, 1997; Kelly and Lim, 2000; or Barai and Teoh, 1997 for excellent reviews of these models; Barceló, 2002 has a review published in a Spanish journal). In general, there exists two different approaches that extend the Cox model: Wei, Lin, and Weissfeld (1989) (WLW) and Lee, Wei, and Amato (1992) (LWA) **marginal**, and Prentice, Williams, and Petersen (1981) (PWP) **conditional** models. On the other hand, frailty models can also be seen as a Cox-based model.

Before starting to illustrate both models we mention that there exists another approximation known as the AG model (Andersen and Gill, 1982). The coefficient estimates using AG apporach are exactly the same as those obtained using Cox model, only the standard errors are different. The AG model is the most simple variance corrected model, incorporating robust variance estimators. However, this model requires the strongest assumptions. The main hypothesis of AG model is that repeated events within-subject are independent (given the covariates). This assumption is called "independent increment", e.g., one event is not affected by previous events. This restriction means that event dependence cannot be estimated with this model, e.g, the model assumes that events do not change the subject and that the subject does not "learn" from previous events. In addition, AG model does not allow one to investigate effects that might change based on event specific covariate effects. However, we have noticed that there exists the possibility to incorporate event dependence via time-dependent covariates. Given these limitations, AG model is recommended when there is no event dependence and no covariate/event effects. The hazard function

of an individual i for the k event is given by

$$\lambda_{ik}(s \mid X_{ik}) = \lambda_0(s) \exp\{\beta' X_{ik}(s)\}.$$

In this case the set of subjects at risk, (e.g. the risk indicator) is given by

$$Y_{ik}(s) = I\{S_{i,k-1} < s \leq S_{ik}\}. \tag{1.18}$$

Marginal models

Wei, Lin, and Weissfeld (1989) (WLW) illustrated the marginal model with bladder cancer data set with multiple relapses per patient (see Chapter 2 for a more detailed description of these data). Their method model the marginal distribution of each failure time and no particular structure of dependence among distinct failure times on each subject is imposed. Each recurrence is modelled as a different strata. Data are used in each strata as marginal data, and as Therneau and Hamilton (1997) pointed out, "*what would result if the data recorder ignored all information except the given event (type)*". This model is marginal with respect to the risk set since each patient is at risk from the beginning of study and can be at risk for several events simultaneously. The intensity or hazard function for the kth event for the ith subject is

$$\lambda_{ik}(s \mid X_{ik}) = \lambda_{0k}(s) \exp\{\beta' X_{ik}(s)\},$$

For this model, the set of subjects at risk just prior to time s is with respect to the kth event. So, $Y_{ik}(s)$ can be given by

$$Y_{ik}(s) = I\{S_{ik} \geq s\},$$

which corresponds to total time formulation. The estimates of WLW can be either event-specific or overall. The overall estimate proposed by WLW is the weighted average of the event-specific estimates, $\hat{\beta}_1, \ldots, \hat{\beta}_k$, such that the corresponding weighted average of the robust variance is the smallest possible (Wei et al., 1989).

The WLW model presents important disadvantages. Cook and Lawless (1997) pointed out that WLW model is valid only under independent censoring. This disadvantage, however, does not have problems in practice, except if a recurrence is terminal. That is, if the end of study is related to interoccurence times. This model also requires that the data has a maximum number of events. This is a limitation if event-specific estimates become unreliable.

Conditional models

The WLW limitations can be solved using a conditional approach. In contrast to AG model, Prentice, Williams, and Petersen (1981) (PWP) propose a conditional model which allow for event dependence via stratification by event number; different events can have different baseline hazards. The main difference with marginal models is that in conditional models a subject cannot be at risk for the n-th event until the $(n-1)$-th event occurs, hence the name conditional model. Oakes (1991) argues for the conditional approach, and states that the marginal method is inefficient. The conditional model is another variance corrected model and as we have illustrated, it has intuitive appeal because it preserves the order of sequential events in the creation of the risk set and therefore incorporates events dependence.

PWP models can be estimated with the data organized in elapsed time (PWP-TT) (i.e., total time risk set or time from each unit's entry into the observation set) or interocurrence/gap time (PWP-GT) (i.e., gap time risk set or time since the previous event). Thus, the hazard function only differs in the risk intervals formulation. For PWP-TP model the hazard function is given by:

$$\lambda_{ik}(s \mid X_{ik}) = \lambda_{0k}(s) \exp\{\beta' X_{ik}(s)\},$$

and for PWP-GT model by

$$\lambda_{ik}(s \mid X_{ik}) = \lambda_{0k}(s - s_{k-1}) \exp\{\beta' X_{ik}(s)\}.$$

In consequence, $Y_{ik}(s)$ is different for each formulation. In the PWP-TT model $Y_{ik}(s)$ corresponds to $Y_{ik}(s) = I\{S_{ik} \geq s\}$, and for PWP-GT model it is given by

$$Y_{ik}(s) = I\{T_{ik} > s\}.$$

The choice between PWP-TT or PWP-GT depends on whether we are interested in the time that has elapsed since the patient entered the study or since the last recurrence. We notice that PWP model is a stratified AG model.

Some of these previous models have been compared using real and simulated data, giving different results as it is illustrated in Wei and Glidden (1997), Gao and Zhou (1997), Clayton (1994), Lin (1994), Therneau and Hamilton (1997), Barai and Teoh (1997), and Therneau and Grambsch (2000). The AG model is maybe the most used because of its efficiency as Therneau and Grambsch (2000, pag 229) conclude from their hidden covariate Monte Carlo simulation. In

addition, in the case of variance corrected models, the AG model gives the most reliable estimates of the overall effect. However, as Kelly and Lim (2000) pointed out, it remains unclear which models are suitable for recurrent event data, as well as the differences between existing models. They do not recommend LWA model because it allows a subject to be at risk several times for the same event. The WLW model overestimates treatment effect and they do not recommend it. Finally, the authors propose to use PWP-GT for analyzing recurrent event data. However, when there exists within-subject correlation, they recommend to use methods different from those which are based on a robust variance estimation, like random effects models.

Software

All Cox extension models, outlined previously, can be fitted using both S-plus (MathSoft, 1997) and R (R Development Core Team, 2005, Ihaka and Gentleman, 1996) functions creating an appropriate data set. After obtaining these data sets, we can fit the models mentioned above using `coxph` function to fit Cox models and the functions `strata`, `cluster`, or `frailty`. Therneau and Grambsch (2000, Chapter 8) or Therneau and Hamilton (1997) explain how to fit any of the models described in the previous sections. An excellent review of the different software packages used for analyzing correlated survival data has been recently written by Kelly (2004).

1.5 General class of models

As we have previously illustrated, there are currently several models and methods of analysis used for recurrent event data (see for instance Hougaard, 2000; or Therneau and Grambsch, 2000, Chapter 8). However, as Peña and Hollander (2004) point out, there is still a need for a general and flexible class of models that *simultaneously* incorporates the effects of covariates or concomitant variables, the impact on the unit of accumulating event occurrences, the effect of latent or unobserved variables which, for each unit, endow correlation among the inter-event times, as well as the effect of performed interventions after each event occurrence.

Most existing extensions of the Cox model deal with the majority of these effects. However, as we have pointed out in Section 1.3.1 these models assume that the effect after each intervention is always the same (minimal or perfect intervention if risk interval is total or gap time, respectively). As we have also illustrated, in many biomedical settings not always the effect of intervention is

the same. So, we need to use a more general model that allow us to incorporate different effects that affect to event occurrences. Peña and Hollander (2004) proposed a new class of models which generalize most of existing reliability and Cox-based models. This new model will allow us to connect reliability models, which models the effect of intervention via the effective age, with biomedical models, which incorporates the effect of concomitant covariates and the correlation among interoccurrence times. This model is briefly described in the next section.

1.5.1 Peña and Hollander model

Let $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$ be a vector of independent and identically distributed (i.i.d.) positive-valued random variables from a parametric distribution $H(z;\xi) = \Pr(Z \leq z|\xi)$ where ξ is a finite-dimensional parameter taking values in $\Xi \subseteq \Re^r$. These variables are unobservable random factors affecting the event occurrences for the subjects. Also, let $\mathbf{F} = \{\mathcal{F}_s : 0 \leq s \leq s^*\}$ be a filtration or history on some probability space $(\Omega, \mathcal{F}, \mathcal{P})$ such that $\mathbf{X}_i$s and $Y_i^\dagger$s are predictable and such that $N_i^\dagger$s are counting processes with respect to $\mathbf{F}$. Finally, let $\{\mathcal{E}_i(s) : 0 \leq s \leq s^*\},\ i = 1, 2, \ldots, n$, the effective age processes satisfying the conditions described in Section 1.3.1.

The class of models is obtained as follows. Conditionally on $\mathbf{Z}$, the $\mathbf{F}$-compensator of $N_i^\dagger$ is $\{A_i^\dagger(s|\mathbf{Z}, \mathbf{X}_i) : 0 \leq s \leq s^*\}$ with

$$A_i^\dagger(s|\mathbf{Z}, \mathbf{X}_i) = \int_0^s Y_i^\dagger(v)\, \lambda_i(v|\mathbf{Z}, \mathbf{X}_i)\, dv, \tag{1.19}$$

where

$$\lambda_i(s|\mathbf{Z}, \mathbf{X}_i) = Z_i\, \lambda_0[\mathcal{E}_i(s)]\, \rho[N_i^\dagger(s-); \alpha]\, \psi[\beta' \mathbf{X}_i(s)]. \tag{1.20}$$

This means that the process $M_i^\dagger(s|\mathbf{Z}, \mathbf{X}_i) = N_i^\dagger(s) - A_i^\dagger(s|\mathbf{Z}, \mathbf{X}_i)$ is a square-integrable $\mathbf{F}$-martingale. In (1.20), $\lambda_0(\cdot)$ is an unknown baseline hazard rate function. The effects of accumulating event occurrences are encoded in $\rho(\cdot;\alpha) : \mathbf{Z}_+ \equiv \{0, 1, 2, \ldots\} \to \Re_+$ which has a known functional form with $\rho(0;\alpha) = 1$ and with $\alpha \in \mathcal{A} \subseteq \Re^p$. The effect of covariates are considered in $\psi(\cdot)$ which is a nonnegative link function of known functional form with $\beta \in \mathcal{B} \subseteq \Re^q$. The dependence between interoccurrence times is modelled with Z_i which are unobserved frailties. The model also incorporates the effect of performed interventions, $\mathcal{E}_i(s)$, via the baseline hazard function. Thus, the unknown model parameters are $(\lambda_0(\cdot), \alpha, \beta, \xi)$, where $\lambda_0(\cdot)$ is non-parametrically specified, and α, β, and ξ are finite-dimensional parameters.

The statistical identifiability of this class of models without frailties has been established in Theorem 1 of Peña and Hollander (2004). The authors also showed that this class of models subsumes many existing models in the literature (see Peña and Hollander, 2004). In particular, some of models used in biomedical settings are special cases of this general model as we will illustrate in Chapter 5.

1.6 Thesis Overview

In the following chapters, we present the work that we have been developing during last few years. In the Chapter 2 we describe four data sets analyzed for the elaboration of this dissertation. We would like to emphasize that two of them were obtained from the Institution where I was working until July 2005. These data sets were created specially for this PhD thesis. The other two come from medical literature and have been widely used when researchers propose new methods to analyze recurrent event data. I must say that my motivation was to find real examples, which were suitable to be analyzed with the statistical methods I was working on. Perhaps the most relevant analysis was the one about lymphoma relapses, in which we included the information about the treatment after each relapse by means of the effective age, which had never been done before.

Then, in Chapter 3, we develop some procedures for estimating confidence intervals for median survival time or, in general, for some quantile. There, we propose some asymptotic confidence intervals which are based on asymptotic variances from existing estimators for survival function when we deal with recurrent events. We also propose how to estimate these confidence intervals using bootstrap techniques. The main contribution of this chapter has been the examination of the question of how to do bootstrapping in the presence of recurrent event data arising from a sum-quota data accrual scheme and informativeness of right-censoring mechanism. We show how to get bootstrap samples from the observed data, as many people normally do, is not correct when data are correlated.

Chapter 4 deals with procedures for estimating the parameters for the general model for recurrent events proposed by Peña and Hollander (2004). One possibility is to use the EM algorithm as in a joint work with Professors Edsel A. Peña and Elizabeth E. Slate showed (paper currently in first revision in Journal of Statistical Planning and Inference and published as a

technical report in Peña, Slate, and González, 2003). This work is presented in the Appendix B since my contribution was the programming of all the procedures exposed, as well as proposing alternative maximization methods. A result of that is the `gcmrec` package and this is why it is include in the thesis. These functions allowed us to perform simulation studies, and analyze the examples shown on the publication where I also participated. We have included this work in the appendix of the thesis because Chapter 4 is based on the notation and the results we get in it.

It is well-known that the EM algorith method have general drawbacks such as neither estimates of the variance of parameter nor frailty are directly estimated. Thus, still in Chapter 4, we proposed two alternative approaches, based on penalizing the likelihood, to fit Peña and Hollander model. One of them follows Therneau *et al.*'s 2003 work. They proposed to penalize regression coefficients. This approach still continues to have problems because the convergence can be slow and the variance of frailty cannot be directly estimated. Then, we propose to adopt another method of penalization described in Rondeau et al. (2003). Their idea is to penalize the full likelihood, instead of the partial likelihood as Therneau et al. (2003) proposed, and to obtain smooth estimates of the hazard function. This method has the main advantage of giving an estimate of variance of the frailty variance.

Finally, Chapter 5 address the problem of how to incorporate the effective age process in biomedical settings. So far, this concept has only be used in reliability problems. Our main contribution in this chapter has been to illustrate how to use the information regarding the effects of treatments or interventions after cancer relapses for modelling the effective age. Our motivation was firstly due to the fact that by analyzing some data sets, and carrying-out some simulations, we showed that one can obtain different results by using different effective age formulations. In addition, as some physicians pointed out *"it is necessary a model designed specifically for relapsing patients"* (MacLaughlin, 2002).

A section reproducing the on-line documentation of `R` packages developed for this dissertation that also can be obtained at `http://www.r-project.org/` are described in the Appendix D. We end the chapters by showing how to analyze the data presented in each chapter using these `R` functions.

Chapter 2

Studies with Recurrent Event Data

Herein, we present three studies dealing with recurrent events data which belong to cancer settings. In addition, we have also included another example from a study concerning small bowel motility that analyze the time of the migrating motor complex during fasting. This data set is an interesting example for illustrating how to analyze data when the interoccurrence times are independent within patients. We notice that the first two data sets are obtained from the institution where I was working until July 2005. The first one, concerning hospital readmission, appeared when we were evaluating the consume of medical resources in patients with cancer. On the other hand, the second data set, which is about cancer relapses, was created in order to illustrate the importance of monitoring the effective age in biomedical problems.

2.1 Hospital Readmission Times in Colorectal Cancer

The study took place in the Hospital de Bellvitge, a 960-bed public Universitary hospital in the metropolitan area of Barcelona, Spain. Between January of 1996 and December 1998, a total of 523 patients with incident colorectal cancer were identified. This study is based on 403 patients who were operated and gave written informed consent to participate. Other 120 (23%) patients were excluded because they died or were released before they were approached (n=74), refused to participate in the study (n=13), had incomplete information or interviews (n=27), or lived at 100 Km. or more from the hospital (n=6).

The outcome variable in this study was readmission, considering it as a potential recurrent event (colorectal cancer patients may have several readmissions after discharge). The date of

surgery was taken as the beginning of the observational period. Patients were actively followed up until June 2002. Consequently, the length of follow up can differ for each patient, depending on its surgery date. Some premature censoring might also occur due to death, migration or change of hospital. The first readmission time has been considered as the time between the date of the surgical procedure and the first re-hospitalization related to colorectal cancer. Following readmission times were considered as the difference between the last discharge date and the current hospitalization date. Totally, 1125 readmission events were recorded. Since co-morbidity may influence the likelihood of hospital readmission, only readmissions related to colorectal cancer have been considered. This information was obtained from the discharge diagnosis registered in the minimum basic data set maintained by the Department of Clinical Documentation (see Gonzalez et al. (2005) for further details). Two hundred sixty four re-hospitalizations were excluded because the main diagnostic or procedures were not related to colorectal cancer. Thus, the final data set consisted on 861 re-hospitalizations recorded on the 403 patients included in the study.

2.1.1 Variables of the data set

The main independent variable was sex, and other variables considered as potential confounders were age ($<$ 60, 60-74, 75 years), tumor site (rectum, colon), tumor stage (Dukes classification: A-B, C, or D), type of treatment (chemotherapy, radiotherapy), distance from living place to hospital (30 km, $>$30 km.), educational level (less than primary, primary, secondary, university). Given that radiotherapy is an exclusive treatment for patients with rectal cancer, to analyze both variables in multivariate models we have created a variable that combines both radiotherapy and tumor site (colon, rectum treated with radiotherapy and rectum treated without radiotherapy). In addition, to adjust the risk of readmissions for comorbidity, we have calculated Charlson index modified by Librero et al. (1999) that incorporates the information from the ICD-9-CM.

2.1.2 Descriptive analysis of the data set

González et al. (2005) analyzed these data both in frequency and time elapse between the readmissions. They used a graphical method to confirm the correlation between the times of re-hospitalization for each patient. After confirming that, they decided to model the data using a proportional hazard model including a random effect (frailty) to account for the within subject correlation between events. The main aim of the investigators was to study social-demographic

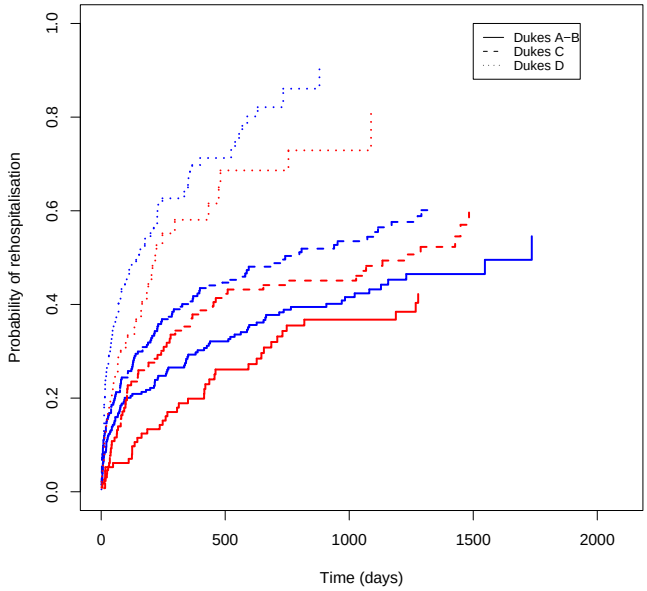

Figure 2.1: Probability of hospital readmission depending on Dukes stage estimated using frailty model (FRMLE). Blue lines represents males, while red lines are for females.

and clinical inequalities in hospital readmission among patients. The authors' main finding was that women with colorectal cancer are less likely than men to be readmitted to the hospital, after controlling for well-established predictors, such as tumor characteristics and comorbidity. Thus, authors compared patients' characteristics by sex which can be shown in table 2.1. No significant differences between males and females were observed in any of the variables analysed, though males tend to be older, with less advanced tumour stage, received less often chemotherapy and had a lower Charlson co-morbidity index.

The distribution of hospital readmissions is shown in Table 2.2. Most of the patients (70.7% of men and 82.3% of women) had none or one readmission and only about 5% of subjects had more than 5 readmissions. Male patients had, on average, more readmissions than women did (2.3 vs. 1.9, p=0.06). A higher number of hospitalizations was associated with more advanced tumor stages and treatment with chemotherapy ($p< 0.001$). However, patients with rectal cancer receiving radiotherapy have less readmissions (p=0.02). The number of hospitalization is greater for patients with university educational level and decreases with age, though the associations of these two variables were not significant. Figure 2.1 shows that the estimated probability of readmission is always higher for men than for women independently of Dukes tumor stage.

	Males n (%)	Females n (%)	p-value
Age (years)			
Average	65.4	63.8	0.123
<60	63 (26.4)	48 (29.3)	
60-74	119 (49.8)	75 (45.7)	
≥75	57 (23.8)	41 (25.0)	0.710
Tumour site			
Rectum	84 (35.1)	67 (40.9)	
Colon	155 (64.9)	97 (59.1)	0.290
Dukes stage			
A-B	115 (48.1)	65 (39.6)	
C	81 (33.9)	67 (40.9)	
D	43 (18.0)	32 (19.5)	0.226
Chemotherapy			
Yes	102 (42.7)	84 (51.2)	
No	137 (57.3)	80 (48.8)	0.112
Radiotherapy[a]			
Yes	45 (18.8)	34 (20.7)	
No	194 (81.2)	130 (79.3)	0.730
Distance			
Average (Km)	21.6	26.1	0.601
≤30 Km.	211 (88.3)	146 (89.6)	
>30 Km.	28 (11.7)	17 (10.4)	0.810
Hospitalisation			
Days	12.35	12.27	0.904
Educational Level			
Less than primary	104 (43.5)	72 (43.9)	
Primary	102 (42.7)	75 (45.7)	
Secondary	23 (9.6)	13 (7.9)	
University	10 (4.2)	4 (2.4)	0.711
Charlson Index[b]			
0	375 (68.3)	202 (64.7)	
1-2	36 (6.6)	10 (3.2)	
≥ 3	138 (25.1)	100 (32.1)	0.018
Follow-up			
Days	1393	1382	0.795

Table 2.1: Sex distribution of variables included in the hospital readmission for patients with colorectal cancer data set. Comparison uses a χ^2test with Yates' correction for categorical variables and t-test for continuous

[a]Results only for rectal cancer.

[b]Distribution for all readmission (time-dependent covariate).

	Number of hospital readmission							
	0	**1**	**2**	**3**	**4**	**≥5**	**mean**	**p-value**[a]
Sex								
Females	87 (53.0)	48 (29.3)	11 (6.7)	8 (4.9)	5 (3.0)	5 (3.0)	1.9	
Males	112 (46.9)	57 (23.8)	34 (14.2)	13 (5.4)	10 (4.2)	13 (5.4)	2.3	0.060
Age								
<60	47 (42.3)	32 (28.8)	11 (9.9)	7 (6.3)	8 (7.2)	6 (5.4)	2.4	
60-74	98 (50.5)	44 (22.7)	27 (13.9)	12 (6.2)	7 (3.6)	6 (3.1)	2.1	
≥ 75	54 (55.1)	29 (29.6)	7 (7.1)	2 (2.0)	0 (0.0)	6 (6.1)	1.8	0.072
Tumor site								
Colon	129 (51.2)	66 (26.2)	27 (10.7)	15 (6.0)	7 (2.8)	8 (3.2)	2.0	
Rectum	70 (46.4)	39 (25.8)	18 (11.9)	6 (4.0)	8 (5.3)	10 (6.6)	2.3	0.200
Dukes stage								
A-B	103 (57.2)	43 (23.9)	16 (8.9)	8 (4.4)	7 (3.9)	3 (1.7)	1.8	
C	67 (45.3)	40 (27.0)	20 (13.5)	7 (4.7)	6 (4.1)	8 (5.4)	2.2	
D	29 (38.7)	22 (29.3)	9 (12.0)	6 (8.0)	2 (2.7)	7 (9.3)	2.7	< 0.001
Chemotherapy								
Non	125 (57.6)	51 (23.5)	22 (10.1)	7 (3.2)	4 (1.8)	8 (3.7)	1.8	
Yes	74 (39.8)	54 (29.0)	23 (12.4)	14 (7.5)	11 (5.9)	10 (5.4)	2.5	< 0.001
Radiotherapy[b]								
Non	31 (40.3)	20 (26.0)	12 (15.6)	4 (5.2)	6 (7.8)	4 (5.2)	2.3	
Yes	39 (52.7)	19 (25.7)	6 (8.1)	2 (2.7)	2 (2.7)	6 (8.1)	2.0	0.022
Distance								
≤30 Km.	174 (48.7)	96 (26.9)	43 (12.0)	16 (4.5)	14 (3.9)	14 (3.9)	2.1	
>30 Km.	24 (53.3)	9 (20.0)	2 (4.4)	5 (11.1)	1 (2.2)	4 (8.9)	2.2	0.818
Education								
Less than primary	83 (47.2)	49 (27.8)	24 (13.6)	9 (5.1)	6 (3.4)	5 (2.8)	2.0	
Primary	91 (51.4)	45 (25.4)	16 (9.0)	8 (4.5)	7 (4.0)	10 (5.6)	2.2	
Secondary	21 (58.3)	8 (22.2)	2 (5.6)	1 (2.8)	2 (5.6)	2 (5.6)	2.0	
University	4 (28.6)	3 (21.4)	3 (21.4)	3 (21.4)	0 (0.0)	1 (7.1)	3.4	0.175

Table 2.2: Number (%) and mean of hospital readmission for variables analyzed in colorectal cancer data set.

[a] p value for Mann-Whitney U test or Kruskal-Wallis test

[b] Results only for rectal cancer

2.2 Non-Hodgkin's Lymphoma Cancer Relapses

The indolent non-Hodgkin's lymphomas (NHL) constitute a heterogeneous group of lymphoproliferative disorders. They encompass what were called low grade and some categories of intermediate grade NHL in the Working Formulation (Cheson et al., 1999). They are categorized based on pathologic and cytologic features. The indolent lymphomas include different subtypes of lymphomas such as Follicular Lymphomas, Small Lymphocytic Lymphoma, Lymphoma Marginal Zone, or Sezary Syndrome among others. Low grade lymphomas are associated with relatively prolonged survival. Because it is considered an indolent, but not curable, type of cancer, patients tend to relapse over time. Thus, patients are treated after each recurrence with intensive therapeutic approaches in an attempt to increase the time until next relapse (that is, to increase the disease-free survival). The treatments may produce different responses (e.g., complete response, CR, partial response PR, or null response NR) depending on disease status after therapy. It is well known that these responses may modify the probability of a subsequent relapse, and hence this intervention effect should be taken into account when modelling this type of data.

The data consist of the times to relapse, in months, for 63 patients with clinical, histopathological, and immunophenotypes of primary cutaneous marginal zone B-cell lymphoma (PCMZCL) as a particular subtype of indolent lymphoma. An analysis of a subset of these data based on 22 patients with a specific subtype of cutaneous lymphoma was presented in a recent paper Servitje et al. (2002). We use the date of first treatment as the beginning of the study. Relapsing times were considered as the difference between last relapse and the current one. Figure 2.2 shows a graphical representation of recurrences for this data set.

2.2.1 Summary of the data set

We have also obtained information about the response achieved after treatment upon relapses (CR, PR, or NR), depending on the disease status, for each relapse for each of the 63 patients. The total number of relapses among all patients is 112. The fraction of patients with no relapse is 57%, and only 7% have 3 or more events. The median follow-up time is 2.9 years (range 1 month to 13.5 years). Thirty eight of 49 (77.8%) responses to treatments administered after relapses are CRs, 9 (14.3%) are PRs, and 2 (4.5%) NRs. This information will be useful to model the effective age as we will illustrate along this thesis. As we can see in the Figure 2.3 the type of response to

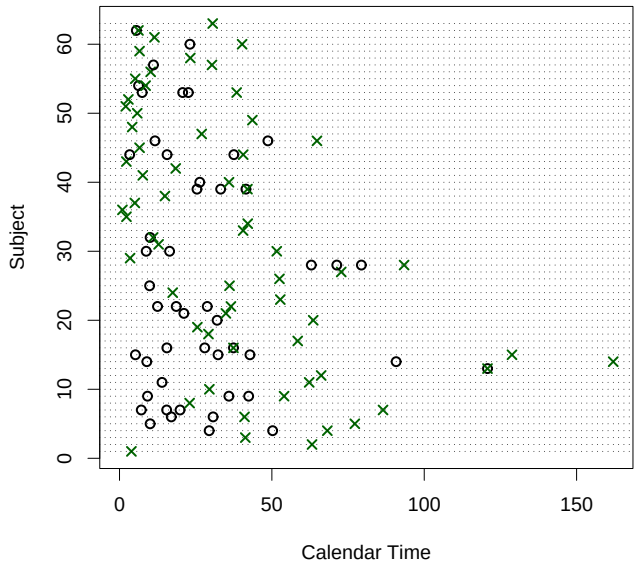

Figure 2.2: Graphical representation of the lymphoma data set. The graphic shows the times (∘) and censoring (×) of PCMZCL cancer recurrence for 63 patients.

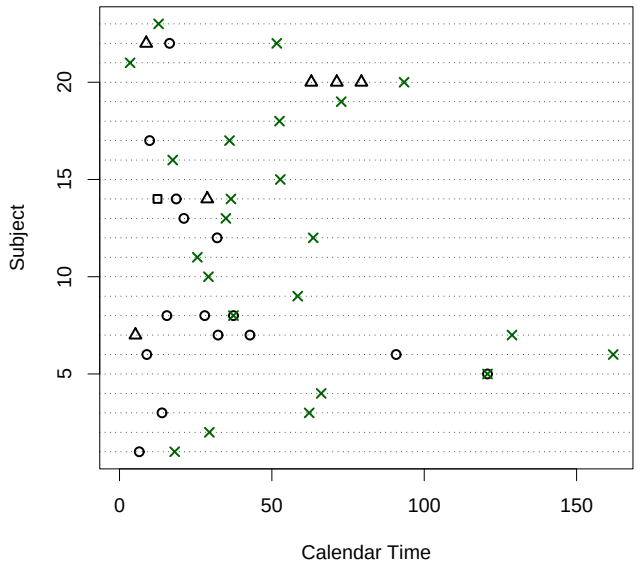

Figure 2.3: Graphical representation of the lymphoma data set including information about the response to treatment after relapses. The graphic shows the times (∘: complete remission, △: partial response, □: no response) and censoring (×) of non-Hodgkin's Lymphoma cancer recurrence only for 45 selected patients for for improving the presentation of the graphic.

the treatment is related to the time to next relapse.

We will also include in the analyses the covariates X_1: gender of patient (0=Male, 1=Female); X_2: delay between first symptom and date of first treatment as a continuous variable (in years); and X_3: lesions involved at diagnosis (0=Single, 1=Localized, 2=More than one nodal site, 3=Generalized), encoded as three indicator variables. The distribution of these covariates is as follows: 73 patients are males (65.2%) and 39 females (34.8%); the median time of the delay between first symptom and first treatment is 29.7 months (range 1 to 144 months); 28 patients (25.0%) presented single lesions at diagnosis, 43 localized lesions (38.4%), 35 more than one nodal site (31.2%), and 6 (5.4%) patients had generalized lesions.

2.3 Bladder Cancer Relapses

Last data set related to cancer is on recurrences of bladder tumor. These data have been used by many people to demonstrate methodology for recurrent event modelling and they can can be obtained from the `survival` package (Lumley and Therneau, 2003) in the `R` Library. Wei, Lin, and Weissfeld (1989) analyzed these data using marginal approach. These data provide the times to recurrence of bladder cancer for $n = 85$ subjects with superficial bladder tumors, which were removed when they entered the study. Forty seven of these patients were randomized into the placebo group, and 38 into the thiotepa group. Many patients have multiple recurrences of tumors in the study, and new tumors were removed at each visit. The data set contains the first four recurrences of the tumor for each patient, and each recurrence time was measured from the patient's entry time into the study. The total number of recurrences was 112. The covariates are X_1, the treatment indicator (1 = placebo; 2 = thiotepa); X_2, the size (in cm) of the largest initial tumor; and X_3, the number of initial tumors. Figure 2.4 shows a graphical representation of recurrences for this data set. No information about effective age is available for these data.

2.4 MMC data set

This data set belong to data from a study concerning small bowel motility (Husebye et al. (1990)). The aim of their analysis is to estimate the mean length of the Migratory Motor Complex (MMC) period (i.e., the mean interoccurrence time between two contractions in the small bowel during the

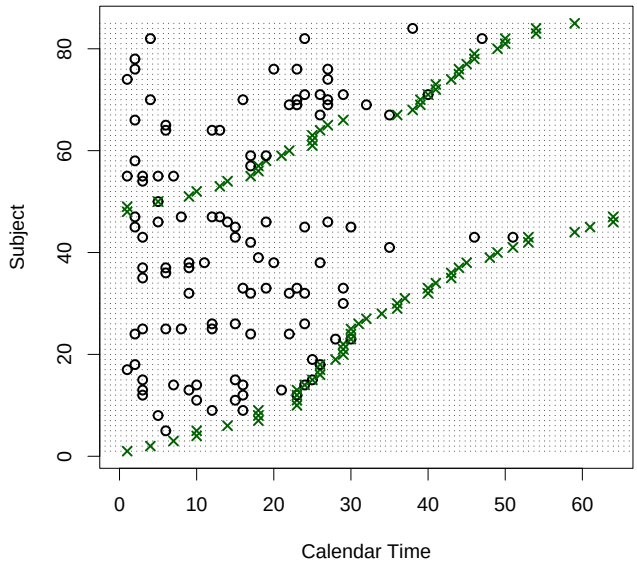

Figure 2.4: Graphical representation of the bladder data set used by Wei et al. (1989). The graphic shows the times (×) and censoring (○) of bladder cancer recurrence for 85 subjects. The 38 top subjects are treated with thiotepa and the upper 47 subjects are from placebo group.

interdigestive cycle). This data set was analyzed in Aalen and Husebye (1991) using a variance component model and an intensity-based formulation with a gamma frailty component using a parametric Weibull model. Then Peña et al. (2001) also analyze this data using two new estimators which are described in Sections 1.3.3 and 1.4.2 respectively. Aalen and Husebye (1991) stated that "the consecutive MMC periods for each individual appear (to be) approximate renewal process" but we need to verify this assumption. To do so, Peña et al. (2001) suggested that since formal statistical methods for checking this i.i.d. assumption are not yet available, a graphical method may be employed. We will illustrate it in Chapter 3.

Chapter 3

Confidence Intervals for Median Survival

This chapter addresses the problem about how to construct confidence intervals for the median survival time of a recurrent event. Two different approaches have been employed. One of them is based on asymptotic variance of Peña et al. (2001) and Wang and Chang (1999) estimators (see Sections 1.3.3 and 1.3.4 respectively) and some transformations. The other one uses bootstrap techniques. Two types of recurrent event models are considered: first is a model where the inter-event times are independent and identically distributed (see Section 1.3.3), and second is a model where the inter-event times are associated, with the association arising from a gamma frailty model (Section 1.4.2). Both bootstrap and asymptotic confidence intervals are studied through simulation. Weak convergence is proved and asymptotic confidence intervals are built according to these results. On the other hand, one of the major goals of this chapter is to study bootstrapping schemes for estimating the sampling distribution of estimators of the median survival time distribution in the presence of recurrent event data. Another important goal is to determine how to compute pointwise confidence intervals for median survival time using asymptotic theory for PSH and WC estimators.

We now discuss the motivations for the present Chapter. Peña et al. (2001) (under i.i.d. model) and Wang and Chang (1999) give asymptotic properties of their estimators. This fact allows us to approximate analytically the variance of survival function. Thus, we can use their estimators to compute confidence intervals for some quantile, ζ_p, in particular for median survival

time, $\zeta_{1/2}$. However, the estimator proposed by Peña et al. (2001) under correlated case does not have a closed form for its variance. If we can assume that our data have been generate under a gamma frailty model, we need an alternative method to estimate the variance of median survival time. As we have previously mentioned, we can use resampling techniques. Bootstrap procedures allow us to construct some confidence intervals for bootstrapped percentiles of ζ_p^*. However, before computing this variance, we need to determine how to obtain bootstrap samples by taking into account the true nature of this kind of data. We need to recognize that the informativeness of right-censoring mechanism, and the impact of the sum-quota accrual scheme, may influence the estimation of survival function, and consequently must be considered in generating bootstrap samples.

Section 3.1 provides some procedures to obtain asymptotic confidence intervals for median survival time (and other quantiles) using asymptotic results for PSH and WC estimators. Different bootstrapping schemes are described in section 3.2 for the i.i.d. model and for the correlated interoccurrence times model. In Section 3.3 a simulation is used to compare and discuss the statistical properties of median survival time estimated using these different bootstrap plans. In this section, we also compare the results obtained using improved bootstrap plans with those obtained using asymptotic theory. In Section 3.4 we apply these procedures to three real data sets described in Chapter 2. We illustrate the problems that one can have when analyze these kind of data. The first example belongs to data from small bowel motility study. This is a good example when interoccurrence times follow an i.i.d. model. The second example involves hospital readmissions in patients diagnosed with colorectal cancer. This example is an excellent example when interoccurrence times are correlated. The third example analyzes tumor recurrences in bladder cancer. This example shows how the results can be different depending on the model selection. Finally, Section 3.5 shows how to analyze data with recurrent events using `survrec` package.

3.1 Estimation of median survival time and other quantiles

After estimating the survival function of the interoccurrence times, we are usually interested in estimating some quantile of this function, $\zeta_p(x)$. For example, in biomedical settings, one may want to estimate the median survival interoccurrence time, $\zeta_{1/2}(x)$. After that, the main

concern is assessing the variability of the estimator, $\hat{\zeta}_p$, by estimating its standard deviation, and constructing confidence intervals for ζ_p based on $\hat{\zeta}_p$. Different approaches can be used to do so. Firstly, we can use asymptotic methods (see Burr and Doss, 1993 or Dabrowska and Doksum, 1987) taking into account that it is more difficult to apply this methodology in the case of $\zeta_p(x)$ than in other situations. Secondly, as several authors have stated, we can estimate the sampling variability of $\hat{\zeta}_p$ more accurately using resampling techniques (see for instance Efron, 1982; Efron, 1985b; Efron, 1985a; Bickel and Frieedman, 1981; Beran, 1982; and Singh, 1981). In particular, some authors have shown that these techniques are useful in the case of survival analysis under the Cox's model (see Burr, 1994, Hjort; 1985).

Let $0 < p < 1$, define

$$\zeta_p = F^{-1}(p) = \inf\{t : F(t) \geq p\} = \inf\{t : \bar{F}(t) \leq 1 - p\} \tag{3.1}$$

as the pth quantile of the interoccurrence distribution function, F. The quantiles can be estimated by taking the right-continuous inverse of the nonparametric estimated survival function $\hat{\bar{F}}$ or $\hat{S}$. Let $\hat{\bar{F}}^{(n)}$ be the nonparametric estimation of the interoccurrence times survival function using either $\hat{\bar{F}}$, or $\hat{S}$, estimators defined in (1.8) and (1.11) respectively. Thus,

$$\hat{\zeta}_p = \inf\{t : \hat{\bar{F}}^{(n)}(t) \leq 1 - p\}. \tag{3.2}$$

Proposition 1 *Let ζ_p and $\hat{\zeta}_p$ be defined in* (3.1) *and* (3.2) *respectively and let us assume the conditions stated in Theorem 1 from Peña et al. (2001) and in Theorem 1 from Wang and Chang (1999). For the PSH estimator, let also assume the condition $y(\infty, t^*) > 0$, for any $t^* > \zeta_p$ where $y(\infty, t^*) = \lim_{s\to\infty} \mathbf{E}\{Y(s,t)\}$. Then*

$$\sqrt{n}(\hat{\zeta}_p - \zeta_p) \xrightarrow{d} N\left(0, \frac{p^2\sigma^2(\zeta_p)}{f^2(\zeta_p)}\right)$$

where $f = -\bar{F}'$, and σ^2 corresponds to σ^2_{PSH} or σ^2_{WC} (defined in equations 1.9 and 1.12 respectively) depending on whether we estimate $\hat{\bar{F}}^{(n)}$ using either $\hat{\bar{F}}$ or $\hat{S}$ (respectively).

Proof: We demonstrate this proposition using the delta-method illustrated in Andersen et al. (1993, section IV.3.4). We write $\zeta_p = \phi(\bar{F})$ and $\hat{\zeta}_p = \phi(\hat{\bar{F}}^{(n)})$, where ϕ is the function defined by

$$\phi(G) = G^{-1}(p) = \inf\{x : G(x) \leq 1 - p\} \tag{3.3}$$

Using Proposition II.8.4 from Andersen et al. (1993) and $f(\zeta_p) > 0$, we have that ϕ is (tangentially) compactly differentiable at $\bar{F}$ with the derivative given by

$$\left(d\phi(\bar{F})\right)(p) \cdot h = \frac{h(\bar{F}^{-1}(p))}{f(\bar{F}^{-1}(p))}, \tag{3.4}$$

where h is continuous at $\bar{F}^{-1}(p)$.

Using functional delta-method we can approximate $\sqrt{n}(\phi(\hat{\bar{F}}^{(n)}) - \phi(\bar{F}))$ by $d\phi(\bar{F})\sqrt{n}(\hat{\bar{F}}^{(n)} - \bar{F})$, where $d\phi(\bar{F})$ is the derivative of ϕ at $\bar{F}$ and it acts on $\sqrt{n}(\hat{\bar{F}}^{(n)}(t) - \bar{F}(t))$ in a linear way. Now, using this approximation, and the Theorem IV.3.2 of Andersen et al. (1993), if we demonstrate that as $n \longrightarrow \infty$

$$\sqrt{n}(\hat{\bar{F}}^{(n)} - \bar{F}) \xrightarrow{d} Z \tag{3.5}$$

where $Z = \bar{F}U$, with U being a Gaussian martingale with $U(0) = 0$ and $cov(U(s_1), U(s_2)) = \sigma^2(s_1 \wedge s_2)$, we will be able to conclude that

$$\sqrt{n}(\phi(\hat{\bar{F}}^{(n)}) - \phi(\bar{F})) \xrightarrow{d} d\phi(\bar{F}) \cdot Z$$

and then,

$$\sqrt{n}(\hat{\zeta}_p - \zeta_p) \xrightarrow{d} \frac{pU(\bar{F}^{-1}(p))}{f(\bar{F}^{-1}(p))} \sim N\left(0, \frac{p^2\sigma^2(\zeta_p)}{f^2(\zeta_p)}\right)$$

Thus, to demonstrate (3.5) we need to prove that

(i) $\sqrt{n}(\hat{\bar{F}} - \bar{F}) \xrightarrow{d} Z_1$

(ii) $\sqrt{n}(\hat{S} - \bar{F}) \xrightarrow{d} Z_2$

where Z_1 and Z_2 must be a Gaussian martingale with $U(0) = 0$ and $cov(U(s_1), U(s_2)) = \sigma^2(s_1 \wedge s_2)$ as we have previously mentioned. The point (ii) is proved in Theorem 1 from Wang and Chang (1999) while point (i) may be proved using Theorem 2b from Peña et al. (2001) as follows. This Theorem states that for all fixed $s \in [0, \infty]$

$$\sqrt{n}(\hat{\bar{F}}(s,t) - \bar{F}(t)) \xrightarrow{d} Z.$$

Taking $s = \infty$ we have $\hat{\bar{F}}(\infty, t) \equiv \hat{\bar{F}}(t)$, and adding the condition $y(\infty, t^*) > 0$, where $y(\infty, t^*) = \lim_{s\to\infty} \mathbf{E}\{Y(s, t^*)\}$,

$$\sqrt{n}(\hat{\bar{F}}(t) - \bar{F}(t)) \xrightarrow{d} Z. \tag{3.6}$$

□

Let us notice that the condition that $y(\infty, t^*) > 0$ limits the interval in which we have convergence to be $[0, t^*]$ (see conditions of Theorem 1 from Peña et al., 2001). Let us also note that it will not work out if $t^* = \infty$. However this is an assumable situation at least in biomedical settings.

Using this Proposition, it is easy to estimate the asymptotic variance of the quantile inserting $\hat{\sigma}^2_{PSH}$ (1.7) or $\hat{\sigma}^2_{WC}$ (1.13) for $\sigma^2(\zeta_p)$ and $\hat{\zeta}_p$ (3.2) for ζ_p in the expression of Proposition 1. In order to estimate the density $f = -\bar{F}'$ we can use the kernel function estimator. Using an uniform kernel function, we can estimate f by

$$\hat{f}(t) = \frac{1}{2b}(\bar{F}^{(n)}(t-b) - \bar{F}^{(n)}(t+b)))$$

The problem, here, is to give some value for b (bandwidth parameter). A conservative choice is to take b to have 50% of observed times in the interval $(t-b, t+b)$. This criteria will be used in the examples.

It is well known that bandwidth selection is a difficult problem. For this reason, it is better to apply Brookmeyer and Crowley's 1982 procedure that do not need to estimate f (see Brookmeyer and Crowley, 1982). Thus, we take as an approximate 100(1-κ)% confidence interval for ζ_p all values ζ_p^0 which satisfies

$$\frac{\mid g(\bar{F}^{(n)}(\zeta_p^0)) - g(1-p) \mid}{\mid g'(\bar{F}^{(n)}(\zeta_p^0)) \mid g(\bar{F}^{(n)}(\zeta_p^0))\hat{\sigma}(\zeta_p^0)} \leq c_{\kappa/2}, \tag{3.7}$$

where $c_{\kappa/2}$ is the upper $\kappa/2$ quantile of the standard normal distribution. Brookmeyer and Crowley (1982) considered $g(x) = x$, but we can also consider some other transformations like $g(x) = \log(-\log(x))$ or $g(x) = \arcsin\sqrt{n}$ (see for instance, Kalbfleisch and Prentice, 1980 or Thomas and Grunkemeier, 1975). We may also estimate a confidence interval for ζ_p as follows:

Step 0. For a given p, estimate the pth quantile, $\hat{\zeta}_p$ using (3.2).

Step 1. Estimate the confidence interval of the survival function, $\hat{\bar{F}}^{(n)}$, at $t = \hat{\zeta}_p$ as follows:

1.1. If $g(x) = x$, take

$$\hat{\bar{F}}^{(n)}(t) \pm c_{\kappa/2}\hat{\sigma}(t)\hat{\bar{F}}^{(n)}(t) \tag{3.8}$$

1.2. If $g(x) = log(-log(x))$, take

$$\hat{\bar{F}}^{(n)}(t)^{exp\{\pm c_{\kappa/2}\hat{\sigma}(t)/log(\hat{\bar{F}}^{(n)}(t))\}} \tag{3.9}$$

1.3. If $g(x) = arcsin\sqrt{x}$, take

$$sin^2\left\{max\left(0, arcsin(\hat{\bar{F}}^{(n)}(t)^{1/2}) - \frac{1}{2}c_{\kappa/2}\hat{\sigma}(t)\left\{\frac{\hat{\bar{F}}^{(n)}(t)}{1-\hat{\bar{F}}^{(n)}(t)}\right\}^{1/2}\right)\right\} \leq$$

$$\leq \bar{F}^{(n)}(t) \leq$$

$$\leq sin^2\left\{min\left(\frac{\pi}{2}, arcsin(\hat{\bar{F}}^{(n)}(t)^{1/2}) + \frac{1}{2}c_{\kappa/2}\hat{\sigma}(t)\left\{\frac{\hat{\bar{F}}^{(n)}(t)}{1-\hat{\bar{F}}^{(n)}(t)}\right\}^{1/2}\right)\right\} \tag{3.10}$$

Step 2. Obtain the confidence interval for $\hat{\zeta}_p$ using the survival times corresponding to the confidence values obtained in the step 1 from the lower, $\hat{\zeta}_L$, and upper, $\hat{\zeta}_U$, pointwise confidence limits from $\bar{F}^{(n)}(t)$, e.g. we need to compute $\bar{F}^{(n)^{-1}}(\hat{\zeta}_L)$ and $\bar{F}^{(n)^{-1}}(\hat{\zeta}_U)$, respectively.

3.2 Bootstrapping ζ_p

In this section we will describe several plans to estimate the sampling distribution of estimators of the median survival, $\zeta^*_{1/2}$. If we are interested in obtaining a confidence interval for another quantile we can use the same schemes replacing $\hat{\zeta}^*_{1/2}$ by $\hat{\zeta}^*_p$ in the last step of each method. The new contribution of the present section is the examination of the question of how to do bootstrapping in the presence of recurrent event data arising from a sum-quota data accrual scheme and informativeness of right-censoring mechanism. In the schemes below, the number of bootstrap replications is denoted by B.

Method 1: (Bootstrapping the observed data)

Obtain B i.i.d. samples of form

$$\{(K^*_i, \tau^*_i, T^*_{i1}, T^*_{i2}, \ldots, T^*_{iK_i}, \tau^*_i - S^*_{iK_i}), i = 1, \ldots, n\},$$

with replacement, from the observed sample

$$\{(K_i, \tau_i, T_{i1}, T_{i2}, \ldots, T_{iK_i}, \tau_i - S_{iK_i}), i = 1, \ldots, n\}.$$

For each sample, compute PSH estimator $\hat{\bar{F}}$ of $\bar{F}$, and compute the resulting estimator of the median, i.e., $\zeta_{1/2}$ replacing $\hat{\bar{F}}^{(n)}$ by $\hat{\bar{F}}$ in (3.2).

Method 2: (Nonparametric bootstrap)

For $i = 1, \ldots, n$, a bootstrap sample is generated as follows:

Step 1. Take $\tau_i^* = \tau_i$;

Step 2. From the distribution $\hat{\bar{F}}^{(n)}$, continue generating an i.i.d sequence of T_{ij}^*'s until K_i^* where

$$\sum_{j=1}^{K_i^*} T_{ij}^* \leq \tau_i^* < \sum_{j=1}^{K_i^*+1} T_{ij}^*.$$

Step 3. The bootstrap sample for the ith unit is

$$(K_i^*, \tau_i^*, T_{i1}^*, T_{i2}^*, \ldots, T_{iK_i^*}^*, \tau_i^* - S_{iK_i^*}^*) \tag{3.11}$$

where $S_{ij}^* = \sum_{l=1}^{K_i^*} T_{il}^*$.

Step 4. For this bootstrap sample, compute $\hat{\bar{F}}^{(n)}$ and estimate the associated median estimate, $\hat{\zeta}_{1/2}^*$ using (3.2).

Method 3: (Semiparametric bootstrap)

Let $\tilde{\bar{F}}$ be the frailty estimator (FRMLE in Section 1.4.2) estimator of $\bar{F}$.

Step 1. Given the data, estimate $\hat{\alpha}$, the frailty parameter, and $\hat{\Lambda}_0$, the cumulative hazard function associated with $\bar{F}_0(t)$. Then, estimate the $\bar{F}_0$ distribution using

$$\hat{\bar{F}}_0(t) = \prod_{\{j:\ t_j \leq t\}} \left[1 - \Delta\hat{\Lambda}_0(t_j)\right]. \tag{3.12}$$

Step 2. Generate $Z_1^*, \ldots, Z_n^*$ according to a $Gamma(\hat{\alpha}, \hat{\alpha})$

For $i = 1, \ldots, n$, a bootstrap sample is generated as follows:

Step 3. Take $\tau_i^* = \tau_i$;

Step 4. From $\hat{\bar{F}}_0^{Z_i^*}$, continue generating an i.i.d sequence of T_{ij}^*'s until K_i^* where

$$\sum_{j=1}^{K_i^*} T_{ij}^* \leq \tau_i^* < \sum_{j=1}^{K_i^*+1} T_{ij}^*.$$

Step 5. The bootstrap sample for the ith unit is

$$(K_i^*, \tau_i^*, T_{i1}^*, T_{i2}^*, \ldots, T_{iK_i^*}^*, \tau_i^* - S_{iK_i^*}^*)$$

where $S_{ij}^* = \sum_{l=1}^{K_i^*} T_{il}^*$.

Step 6. For this bootstrap sample, compute FRMLE $\tilde{\bar{F}}$ of $\bar{F}$, and compute $\hat{\zeta}_{1/2}^*$ using (3.2).

We notice that Method 2 provide two different sampling distributions for the median survival time depending on the estimator selected in the Step 2. One of them is obtained replacing $\hat{\bar{F}}^{(n)}$ by $\hat{\bar{F}}$ and another one is obtained using $\hat{S}$ instead of $\hat{\bar{F}}$. In each case, in the Step 4 we have to compute $\hat{\zeta}^*_{1/2}$ replacing $\hat{\bar{F}}^{(n)}$ by $\hat{\bar{F}}$ or by $\hat{S}$ in (3.2) respectively.

There is another important question to investigate: how do we take into account the censoring mechanism? Except in Method 1, we generate times until their sum is bigger than the length of the period observed for this unit. Another way of obtaining τ_i^* is bootstrapping from the estimated empirical distribution of τ_i, $\hat{G}_n$. As we need that τ_i^* and T_{ij}^* be mutually independent, we can generate first τ_i^* from G_n and then obtain T_{ij}^* using Method 2 or Method 3. In the case that G depends on some covariates, we can extend this algorithm to that case easily.

Definitely, we have seven plans to compare. Plan I is Method 1. Plan II is nonparametric bootstrap (Method 2) when we estimate $\bar{F}$ using $\hat{\bar{F}}$. Plan III is the same as Plan II, except that for each bootstrap sample $\tau_i^*, i = 1, 2, \ldots, n$, is an i.i.d. sample from the empirical distribution G_n. Plan IV is a parametric bootstrap when we estimate $\bar{F}$ using $\hat{S}$. Plan V is the same as Plan IV, except that for each bootstrap sample $\tau_i^*, i = 1, 2, \ldots, n$, is an i.i.d. sample from the empirical distribution G_n. Plan VI is semiparametric plan, and finally, Plan VII is the same as plan VI, except that for each bootstrap sample $\tau_i^*, i = 1, 2, \ldots, n$, is an i.i.d. sample from the empirical distribution G_n.

After obtaining bootstrap samples of ζ_p we have to decide the method to form the bootstrap confidence interval because this may affect the results. There exists a vast number of ways to construct bootstrap confidence intervals such as that based on normality, the percentile, and Efron's BCa among others (see for instance Martin, 1990). We consider the percentile method because even though the theoretical justification for this method is weakest (see Efron and Tibshirani, 1993, or Singh, 1988), these intervals are the simplest to use, explain, and are the most frequently used in practice.

3.3 Simulation Study

3.3.1 Simulation Design

To assess the finite-sample performance of the proposed bootstrap schemes, and asymptotic pointwise confidence intervals, a simulation was performed. The data were generated under two sce-

narios: i.i.d. and gamma frailty models. To simulate the samples under the i.i.d. model, we first generate the monitoring time of each subject, τ_i, using $G(t|\nu) = 1 - \exp(-t/\nu)$, and then we simulate the interoccurrence times, T_{ij}, through $F(t|\theta) = 1 - \exp(-t/\theta)$. To simulate the samples under a gamma frailty model we also generate the monitoring times using the same G distribution and $F_0(t|\theta) = 1 - \exp(-t/\theta)$.

For each sample, median survival time has been estimated as we have described for each of the bootstrap plans. The true median survival time under the i.i.d. model is $-\theta \log(0.5)$ and under the gamma frailty model is

$$\frac{\theta\alpha(1 - 0.5^{1/\alpha})}{0.5^{1/\alpha}}.$$

We have simulated 2,000 samples and 500 bootstrap replicates (B=500). For each sample, the mean square error (MSE) and the 95% bootstrap percentile confidence interval (BPCI) have been calculated. In addition, for each BPCI the empirical coverage percentage was estimated by the proportion of times the BPCI covered the true median survival time in the 2,000 samples. Mean, median, and variance of the length of the BPCI bootstrap intervals have also been calculated. In order to compare asymptotic confidence intervals (AsyCI), we have also computed their empirical coverage, mean, median and variance of the length using different procedures mentioned in section 3.1. For the i.i.d. case we have also compared these results with those obtained from the best bootstrap method that has been found in the independent case (Method 2 or non parametric boostrap).

Samples were generated using $n \in \{15, 50, 80\}, \theta \in \{1/3, 1/6\}$ and $\nu = 1$, and for the correlated case $\alpha \in \{6, 2\}$. The simulation was carried out with a Fortran90 code. DRNUN subroutine from numerical libraries has been used as a random number generator.

3.3.2 Simulation Results

The results of the simulation for bootstrap methods are summarized in Tables 3.1 and 3.2 and in Tables 3.3 and 3.4 and in Figure 3.1. Tables 3.1 and 3.2 give the results for the i.i.d. model except for the plans VI and VII, because the results for these schemes showed poor coverages (less than 80%) and large biases (around 30% of the MSE). Tables 3.3 and 3.4 show the results for the correlated case except for the plans I, II and III, since these plans also present large biases (around 20%) and poor coverages (less than 80%). Figure 3.1 shows the observed distribution of the median survival time under an i.i.d. model and under a gamma frailty model, respectively.

In all simulations, as the sample size increases we obtain better coverage, less bias and less MSE, as is intuitively expected. From Tables 3.1 and 3.2 we see that, in terms of MSE, the best schemes for the i.i.d. case are plans I, II and III. However, plan I has a poorer coverage than both plans II and III. Regarding the length of the BPCI, the three plans show similar average size, but both plans II and III have the smallest variance. These conclusions are the same for all sample sizes and for both values of θ. When we examine the observed distribution of the median survival under the i.i.d. model (Figure 3.1, bottom panels), we immediately notice that plans I and III have less variance than plan V. We can also see that the three plans obtain a sample distribution centered at the true median survival. Similar results are obtained for sample sizes set equal to 15 and 80.

From Tables 3.3 and 3.4 we see that the best schemes for the correlated case in terms of MSE are both semiparametric bootstrap schemes (plans VI and VII). These plans have also the shortest BPCIs and smallest variances. Evidently, the performance of all plans degrades as the level of association among the within-unit interoccurrence times increases. These conclusions are the same for all sample sizes and for both values of θ. Figure 3.1 (top panels) shows the observed distribution of the median survival under a gamma frailty model. Examining these graphs, we see that resampling plan III outperforms plan V in in the i.i.d. model, whereas plan VII is best under the gamma frailty model. The performance of the resampling plan using the WC estimator seems intermediate between those based on the PSH and the FRMLE under the i.i.d. and the gamma frailty model, so in a sense this scheme may provide a robust procedure when uncertain about the model that generated the data. And this robustness property was the intent of Wang and Chang's (1999) proposing this estimator.

Tables 3.5 and 3.6 show the comparison among the different AsyCI with the most adequate bootstrap method under i.i.d. model. We can observe that the results are comparable with varying sample size, except for n=15 that AsyCI based on log-log transformation outperforms BCPI. Anyway, this comparison help us to determine that bootstrap procedures are correct and it will allow us to use them for the correlated case (gamma frailty model) in which asymptotic variances are not yet available.

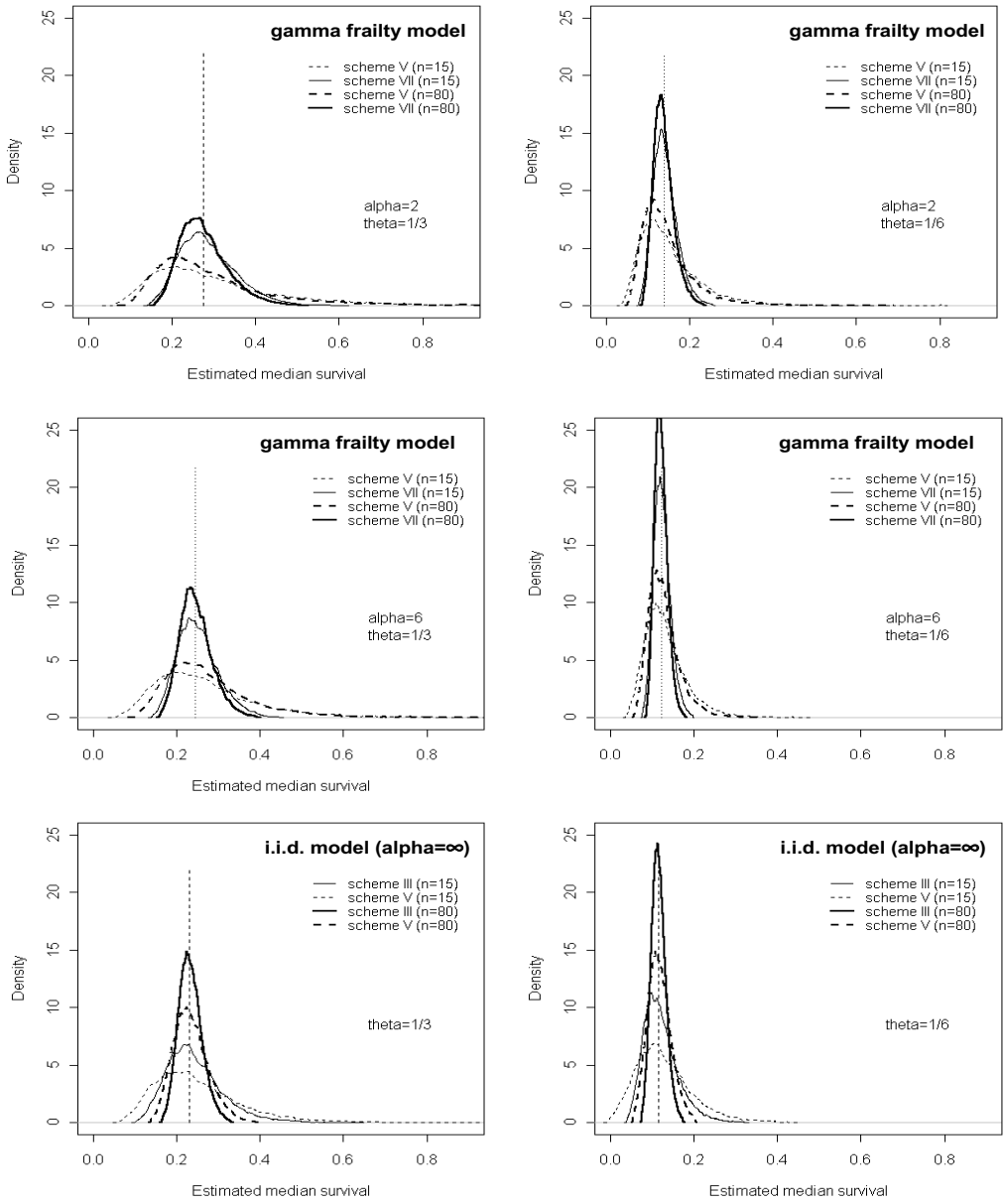

Figure 3.1: Observed distribution of the median survival estimator for an i.i.d. model and a gamma frailty model in 1,000,000 replications, for selected bootstrap plans. Each panel shows the observed distribution for all combinations of θ and α that we have simulated. Vertical lines represent the true median survival time.

		95% BPCI			
	MSE (bias)	**EC**	$\hat{\mu}$	$\tilde{\mu}$	$\hat{\sigma}^2$
n=15					
Plan I	283 (8.1)	88.4	0.19	0.17	1,034
Plan II	291 (11.3)	94.3	0.21	0.19	1,024
Plan III	291 (11.9)	95.2	0.22	0.20	1,119
Plan IV	703 (10.3)	93.6	0.32	0.28	3,291
Plan V	687 (10.4)	94.2	0.32	0.28	3,421
n=50					
Plan I	66 (3.1)	93.3	0.10	0.10	75
Plan II	66 (3.7)	94.6	0.10	0.10	63
Plan III	66 (3.7)	94.8	0.10	0.10	65
Plan IV	141 (3.2)	94.9	0.15	0.15	202
Plan V	142 (3.2)	94.5	0.15	0.15	198
n=80					
Plan I	39(2.1)	94.1	0.08	0.08	34
Plan II	38(2.5)	95.4	0.08	0.08	28
Plan III	38(2.6)	95.2	0.08	0.08	28
Plan IV	84(1.9)	95.3	0.12	0.12	88
Plan V	84(1.9)	95.4	0.12	0.12	91

Table 3.1: Simulation results for 2,000 samples and 500 bootstrap replicates under the i.i.d. model. Mean square error (MSE) ($\times 10^5$) and proportion of MSE due to bias. Empirical Coverage (EC) and mean ($\hat{\mu}$), median ($\tilde{\mu}$) and variance ($\hat{\sigma}^2$) of the length ($\times 10^5$) of 95% bootstrap percentile confidence intervals (BPCI). Results for the first five bootstrap schemes, varying sample sizes, $\theta = 1/3$ and ν=1.

		95% BPCI			
	MSE (bias)	**EC**	$\hat{\mu}$	$\tilde{\mu}$	$\hat{\sigma}^2$
n=15					
Plan I	140 (6.2)	90.4	0.16	0.15	926
Plan II	155 (9.8)	95.3	0.19	0.17	915
Plan III	156 (9.9)	95.1	0.20	0.18	965
Plan IV	321 (8.9)	94.6	0.25	0.24	1,562
Plan V	335 (9.1)	95.1	0.25	0.24	1,635
n=50					
Plan I	42 (2.8)	94.3	0.07	0.07	52
Plan II	43 (3.0)	95.6	0.07	0.07	50
Plan III	40 (3.0)	94.9	0.07	0.07	52
Plan IV	111 (2.9)	94.3	0.11	0.11	128
Plan V	115 (3.0)	94.7	0.11	0.11	135
n=80					
Plan I	25(2.0)	94.5	0.05	0.05	26
Plan II	26(2.1)	95.4	0.05	0.05	20
Plan III	28(2.2)	95.6	0.05	0.05	21
Plan IV	44(2.0)	95.1	0.10	0.10	56
Plan V	45(2.0)	94.8	0.10	0.10	59

Table 3.2: Simulation results for 2,000 samples and 500 bootstrap replicates under the i.i.d. model. Mean square error (MSE) ($\times 10^5$) and proportion of MSE due to bias. Empirical Coverage (EC) and mean ($\hat{\mu}$), median ($\tilde{\mu}$) and variance ($\hat{\sigma}^2$) of the length ($\times 10^5$) of 95% bootstrap percentile confidence intervals (BPCI). Results for the first five bootstrap schemes, varying sample sizes, $\theta = 1/6$ and ν=1.

			95% BPCI			
		MSE (bias)	**EC**	$\hat{\mu}$	$\tilde{\mu}$	$\hat{\sigma}^2$
$\boldsymbol{\alpha}$=2						
n=15	**Plan IV**	2,576 (15.2)	93.4	0.57	0.45	16,985
	Plan V	2,362 (15.6)	93.8	0.58	0.46	16,935
	Plan VI	1,876 (7.2)	92.1	0.45	0.34	13,115
	Plan VII	1,885 (7.3)	92.3	0.45	0.34	13,506
n=50	**Plan IV**	456 (5.4)	94.2	0.26	0.24	15,422
	Plan V	456 (5.5)	94.2	0.26	0.23	16,195
	Plan VI	258 (0.1)	93.8	0.20	0.19	518
	Plan VII	256 (0.1)	94.1	0.20	0.19	526
n=80	**Plan IV**	265 (5.6)	94.8	0.20	0.19	426
	Plan V	267 (5.8)	95.1	0.20	0.19	434
	Plan VI	167 (0.2)	94.4	0.16	0.15	206
	Plan VII	168 (0.2)	94.8	0.16	0.15	200
$\boldsymbol{\alpha}$=6						
n=15	**Plan IV**	1,203 (12.7)	92.9	0.40	0.33	7,160
	Plan V	1,154 (12.9)	93.1	0.41	0.33	7,085
	Plan VI	927 (16.7)	92.2	0.31	0.26	3,831
	Plan VII	921 (16.9)	92.7	0.31	0.26	4,667
n=50	**Plan IV**	209 (4.0)	93.7	0.18	0.17	319
	Plan V	208 (4.1)	93.9	0.18	0.17	352
	Plan VI	127 (2.0)	93.7	0.14	0.13	153
	Plan VII	127 (2.0)	93.5	0.14	0.13	151
n=80	**Plan IV**	120 (3.7)	95.3	0.14	0.14	157
	Plan V	121 (3.9)	95.4	0.14	0.14	158
	Plan VI	72 (1.0)	95.3	0.11	0.10	78
	Plan VII	72 (1.0)	95.2	0.11	0.10	77

Table 3.3: Simulation results for 2,000 samples and 500 bootstrap replicates under a gamma frailty model with shape and scale parameter set equal to α=2 and α=6. Mean square error (MSE) ($\times 10^5$) and proportion of MSE due to bias. Empirical Coverage (EC) and mean ($\hat{\mu}$), median ($\tilde{\mu}$), and variance ($\hat{\sigma}^2$) of the length ($\times 10^5$) of 95% bootstrap percentile confidence intervals (BPCI). Results for the last four bootstrap plans, varying sample sizes, $\theta = 1/3$ and ν=1.

			95% BPCI			
		MSE (bias)	**EC**	$\hat{\mu}$	$\tilde{\mu}$	$\hat{\sigma}^2$
$\boldsymbol{\alpha}$=2						
n=15	**Plan IV**	1,562 (12.2)	94.4	0.30	0.20	8,965
	Plan V	1,234 (12.1)	93.8	0.32	0.21	8,435
	Plan VI	976 (6.4)	93.1	0.25	0.24	7,215
	Plan VII	985 (6.3)	93.3	0.25	0.24	7,560
n=50	**Plan IV**	206 (4.6)	94.4	0.14	0.14	7,922
	Plan V	207 (4.2)	94.5	0.24	0.13	7,690
	Plan VI	152 (1.4)	94.6	0.10	0.09	218
	Plan VII	153 (1.1)	94.5	0.10	0.09	261
n=80	**Plan IV**	151 (4.7)	94.5	0.10	0.09	261
	Plan V	157 (5.0)	94.8	0.10	0.09	243
	Plan VI	107 (1.2)	95.2	0.08	0.07	106
	Plan VII	108 (0.9)	95.1	0.08	0.07	101
$\boldsymbol{\alpha}$=6						
n=15	**Plan IV**	842 (9.7)	93.9	0.22	0.19	3,262
	Plan V	758 (9.5)	93.8	0.23	0.20	3,381
	Plan VI	469 (10.2)	93.1	0.16	0.11	1,831
	Plan VII	450 (10.8)	93.4	0.17	0.11	2,005
n=50	**Plan IV**	112 (3.1)	94.7	0.08	0.07	157
	Plan V	111 (3.2)	94.2	0.08	0.07	145
	Plan VI	75 (1.6)	94.5	0.06	0.06	79
	Plan VII	77 (1.5)	94.0	0.06	0.05	65
n=80	**Plan IV**	70 (3.1)	95.1	0.06	0.06	89
	Plan V	69 (2.9)	95.2	0.06	0.06	82
	Plan VI	38 (1.0)	94.8	0.04	0.04	35
	Plan VII	38 (1.1)	94.3	0.04	0.04	26

Table 3.4: Simulation results for 2,000 samples and 500 bootstrap replicates under a gamma frailty model with shape and scale parameter set equal to α=2 and α=6. Mean square error (MSE) ($\times 10^5$) and proportion of MSE due to bias. Empirical Coverage (EC) and mean ($\hat{\mu}$), median ($\tilde{\mu}$), and variance ($\hat{\sigma}^2$) of the length ($\times 10^5$) of 95% bootstrap percentile confidence intervals (BPCI). Results for the last four bootstrap plans, varying sample sizes, $\theta = 1/6$ and ν=1.

	PSH estimator				WC estimator			
	EC	$\hat{\mu}$	$\tilde{\mu}$	$\hat{\sigma}^2$	**EC**	$\hat{\mu}$	$\tilde{\mu}$	$\hat{\sigma}^2$
n=15								
Estimating $f(t)$[a]	93.6	0.19	0.17	479	88.5	0.27	0.23	2,458
B&C no-trans.	99.9	0.40	0.38	1,685	99.8	0.63	0.58	6,922
B&C log-log-trans.	94.9	0.18	0.17	427	90.5	0.26	0.24	1,282
B&C arcsin-trans.	94.8	0.18	0.17	408	90.8	0.25	0.23	1,255
Bootstrap (Method 2)	94.3	0.21	0.19	1,024	93.6	0.32	0.28	3,291
n=50								
Estimating $f(t)$	93.7	0.10	0.10	28	94.2	0.15	0.14	123
B&C no-trans.	99.9	0.20	0.20	146	99.9	0.32	0.31	574
B&C log-log-trans.	94.6	0.10	0.10	54	94.0	0.15	0.14	186
B&C arcsin-trans.	94.4	0.10	0.10	53	94.3	0.15	0.14	183
Bootstrap (Method 2)	94.6	0.10	0.10	63	94.9	0.15	0.15	202
n=80								
Estimating $f(t)$	94.2	0.08	0.08	11	93.2	0.11	0.11	48
B&C no-trans.	99.9	0.16	0.16	65	99.9	0.24	0.24	245
B&C log-log-trans.	95.1	0.08	0.08	25	94.0	0.12	0.11	81
B&C arcsin-trans.	94.8	0.08	0.08	25	94.0	0.12	0.11	79
Bootstrap (Method 2)	95.4	0.08	0.08	28	95.3	0.12	0.12	88

[a] estimating $f(t)$ in the expression of asymptotic variance given in Proposition 1

Table 3.5: Simulation results for 2,000 samples under the i.i.d. model. Empirical Coverage (EC) and mean ($\hat{\mu}$), median ($\tilde{\mu}$) and variance ($\hat{\sigma}^2$) of the length ($\times 10^5$) of 95% different pointwise Brookmeyer and Crowley's (B&C) confidence intervals based on asymptotic variance of both PSH and WC estimators. Results for selected sample sizes, $\theta = 1/3$ and ν=1.

	PSH estimator				WC estimator			
	EC	$\hat{\mu}$	$\tilde{\mu}$	$\hat{\sigma}^2$	**EC**	$\hat{\mu}$	$\tilde{\mu}$	$\hat{\sigma}^2$
n=15								
Estimating $f(t)$[a]	93.6	0.19	0.17	479	88.5	0.27	0.23	2,458
B&C no-trans.	99.9	0.40	0.38	1,685	99.8	0.63	0.58	6,922
B&C log-log-trans.	94.9	0.18	0.17	427	90.5	0.26	0.24	1,282
B&C arcsin-trans.	94.8	0.18	0.17	408	90.8	0.25	0.23	1,255
Bootstrap (Method 2)	95.3	0.19	0.17	915	94.6	0.25	0.24	1,562
n=50								
Estimating $f(t)$	93.7	0.10	0.10	28	94.2	0.15	0.14	123
B&C no-trans.	99.9	0.20	0.20	146	99.9	0.32	0.31	574
B&C log-log-trans.	94.6	0.10	0.10	54	94.0	0.15	0.14	186
B&C arcsin-trans.	94.4	0.10	0.10	53	94.3	0.15	0.14	183
Bootstrap (Method 2)	95.3	0.19	0.17	63	94.3	0.11	0.11	128
n=80								
Estimating $f(t)$	94.2	0.08	0.08	11	93.2	0.11	0.11	48
B&C no-trans.	99.9	0.16	0.16	65	99.9	0.24	0.24	245
B&C log-log-trans.	95.1	0.08	0.08	25	94.0	0.12	0.11	81
B&C arcsin-trans.	94.8	0.08	0.08	25	94.0	0.12	0.11	79
Bootstrap (Method 2)	95.4	0.08	0.08	28	95.1	0.10	0.10	56

[a] estimating $f(t)$ in the expression of asymptotic variance given in Proposition 1

Table 3.6: Simulation results for 2,000 samples under the i.i.d. model. Empirical Coverage (EC) and mean ($\hat{\mu}$), median ($\tilde{\mu}$) and variance ($\hat{\sigma}^2$) of the length ($\times 10^5$) of 95% different pointwise Brookmeyer and Crowley's (B&C) confidence intervals based on asymptotic variance of both PSH and WC estimators. Results for selected sample sizes, $\theta = 1/6$ and ν=1.

3.4 Examples

In this section we will compute the confidence intervals for median survival time and its confidence interval using MMC, readmission and bladder data sets. We illustrate how to estimate PSH and WC using the `survrec` package for fitting these data in the Section 3.5.

3.4.1 MMC data set

The first data set pertains to data from the study concerning small bowel motility studied in Husebye et al. (1990) (see Section 2.4 for further description). The aim of their analysis is to estimate the mean length of the Migratory Motor Complex (MMC) period (i.e., the mean interoccurrence time). This data set was also analyzed in Aalen and Husebye (1991) using a variance component model and an intensity-based formulation with a gamma frailty component using a parametric Weibull model. Then Peña et al. (2001) analyze this data using the estimators described in section 1.3. Although Aalen and Husebye (1991) stated that "the consecutive MMC periods for each individual appear (to be) approximate renewal process" we need to verify this assumption. To do so, Peña et al. (2001) suggested that since formal statistical methods for checking this i.i.d. assumption are not yet available, a graphical method may be employed to assess the viability of the i.i.d. model by comparing the agreement among the PSH, WC, and FRMLE estimators in Peña et al. (2001). The resulting estimates of the interoccurrence time survivor function are presented in Figure 3.2. A close agreement among these three estimates is evident. Thus, this agrement provides support for Aalen and Husebye's assumption that the independence assumption is true.

After showing that we can assume that the data follow an IID model, we then estimate survival function using either PSH, WC or FRMLE estimators. We can also estimate pointwise confidence intervals of the survival function which can be seen in the Figure 3.3.

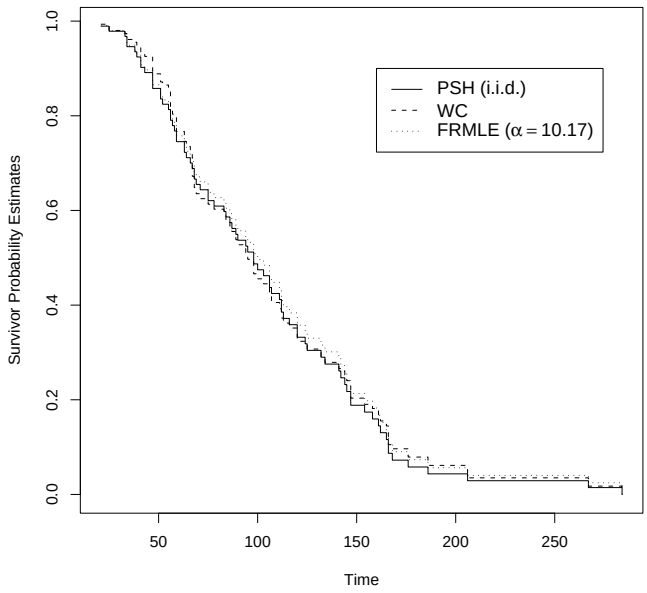

Figure 3.2: Survival function of interoccurrence times estimated using PSH, WC and FRMLE estimators for the MMC data set. This example corresponds to the IID model.

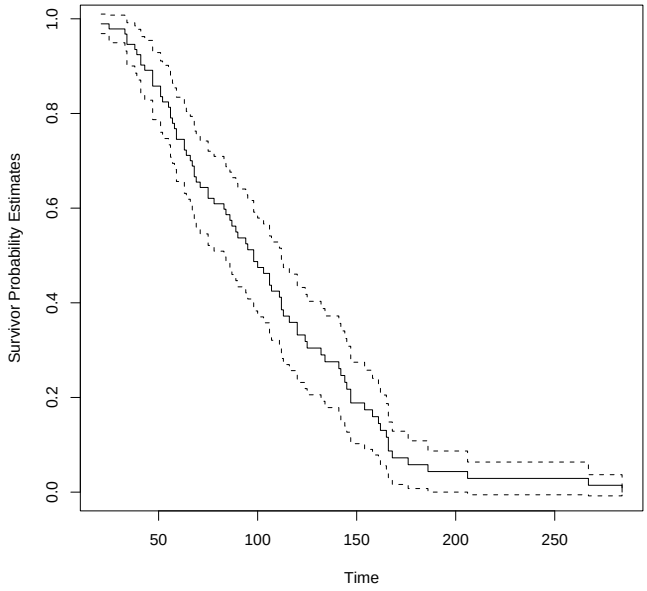

Figure 3.3: Survival function of interoccurrence times estimated using PSH estimators and their pointwise 95% confidence interval for MMC data set using log-log transformation.

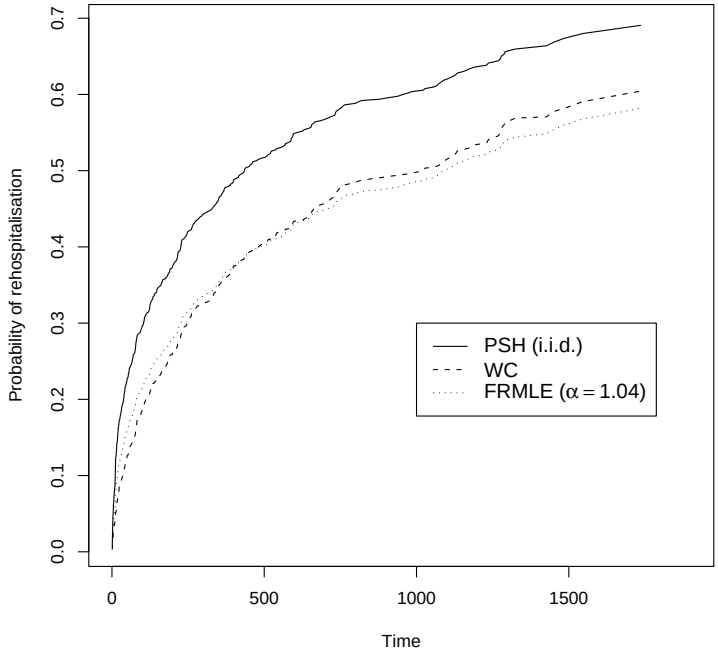

Figure 3.4: Probability distribution function of interoccurrence times estimated using PSH, WC and FRMLE estimators for the hospital readmission in colorectal cancer data set. This example corresponds to a frailty model with frailties following a gamma distribution.

3.4.2 Colorectal cancer rehospitalizations

The next data set concerns rehospitalization of patients diagnosed with colorectal cancer analyzed in González et al. (2005). A description of this data set can be found in Section 2.1. As in the previous example we need to determine if the i.i.d. model is viable. The resulting estimates of the readmission time distribution are presented in Figure 3.4. We have displayed the estimates of the distribution function instead of the survival function because in this study, the investigator is interested in analyzing the probability of readmission instead of the probability of not visiting the hospital.

A considerable difference between these three estimates is evident. The difference is clear between PSH and both WC and FRMLE estimators. Thus, basing on Peña, Strawderman and Hollander's argument, we may conclude that the i.i.d. model is not appropriate for this readmission data set. In making practical conclusions, it behooves therefore to use the inference obtained from the gamma frailty model. In addition, the estimate of the frailty parameter, α, is near 0

indicating a high correlation in the interoccurrence times. Thus, following these results, it was decided to analyze data using a Cox gamma shared frailty model (see González et al., 2005).

Median survival times and probabilities to be readmitted at one and three years using FRMLE model can be found in Table 3.7. We realize that the one-year probability of rehospitalisation was 0.26 in Dukes stages A-B, 0.38 for patients with stage C, and 0.64 for those with stage D. This indicates that the probability of being readmitted strongly depends on advanced tumor stages. The three-year probability shows a similar difference. Males, chemotherapy use, high educational level, mortality, and high co-morbidity were also associated with smaller times between readmissions.

After that, in order to verify if these observed differences in Dukes stage are statistically significant we compute their confidence intervals. Table 3.8 shows the confidence interval obtained using asymptotic variance and bootstrap procedures described in Section 3.1. As a general comment we can say that, and as we expected, that median survival time differs depending on the model. Under i.i.d assumption median survival is underestimated. Regarding Dukes stage, we can conclude that there are differences between median survival time for patients with stage D and both patients with stages A-B and C. These differences are statistically significant since their confidence intervals does not overlap. This conclusion is the same assuming both independent and correlated model. We also observe that the α estimates are very small in all cases (Table 3.8). So, it seems reasonable to use a frailty model for making these comparisons.

Regarding differences in confidence intervals and their width, we should mention that these results agree with those observed in the simulation study. We first consider the confidence intervals computed using asymptotic variances. Table 3.8 shows that the narrowest confidence intervals are those obtained using the log-log transformation (joint with arcsinus in some occasions) as simulation studies showed (Tables 3.5 and 3.6). On the other hand, when bootstrap method is used, confidence intervals are close to those obtained using log-log transformation although they are slightly wider, as the simulations also indicated.

		Readmission Probability		Median
	n (%)	At one year	At three years	readmission time (days)
Sex				
Females	164 (40.7)	0.32	0.46	1427
Males	239 (59.3)	0.39	0.53	799
Age				
<60	111 (27.5)	0.39	0.54	799
60-74	194 (48.1)	0.36	0.48	1230
≥75	98 (24.3)	0.33	0.49	1188
Tumor site				
Colon	252 (62.5)	0.34	0.49	1116
Rectum	151 (37.5)	0.39	0.51	1022
Dukes stage				
A-B	180 (44.7)	0.26	0.41	2175
C	148 (36.7)	0.38	0.50	1073
D	75 (18.6)	0.64	0.89	199
Chemotherapy				
No	217 (53.8)	0.31	0.45	1427
Yes	186 (46.2)	0.41	0.55	734
Radiotherapy[1]				
No	77 (51.0)	0.34	0.48	1188
Yes	74 (49.0)	0.41	0.58	589
Distance				
≤30 Km.	381 (94.8)	0.35	0.50	1073
>30 Km.	21 (5.2)	0.37	0.43	1128
Educational Level				
Less than primary	176 (43.7)	0.36	0.53	819
Primary	177 (43.9)	0.36	0.49	1188
Secondary	36 (8.9)	0.31	0.44	NA
University	14 (3.5)	0.55	0.67	227

Table 3.7: Readmission probability at one and three years and median survival time for variables included in the analysis of readmissions of colorectal cancer data set using FRMLE estimator.

	Dukes stage		
	A-B	**C**	**D**
Asymptotic CI (PSH)			
Estimating $f(t)$[a]	1157 (741.7,1572.3)	398(257.5,538.4)	107 (65.2,148.8)
B&C no-trans.	1157 (521.0,1736.0)	398 (202.0,1104.0)	107 (40.0,223.0)
B&C log-log-trans.	1157 (710.0,1547.0)	398 (285.0,654.0)	107 (67.0,165.0)
B&C arcsin-trans.	1157 (521.0,1736.0)	398 (280.0,654.0)	107 (67.0,165.0)
Asymptotic CI (WC)			
Estimating $f(t)$	1736 (1446.1,2025.8)	1028(589.3,1466.7)	199 (116.9,281.1)
B&C no-trans.	1736 (655.0,∞)	1028 (276.0,1483.0)	199 (79.0,474.0)
B&C log-log-trans.	1736 (1157.0,∞)	1028 (489.0,1291.0)	199 (158.0,297.0)
B&C arcsin-trans.	1736 (1157.0,∞)	1028 (462.0,1291.0)	199 (158.0,297.0)
Bootstrap CI			
Plan II (PSH)	1157 (718.0,1736.0)	398 (290.0,733.0)	107 (70.0,176.0)
Plan IV (WC)	1736 (1188.0,∞)	1028 (489.0,1325.0)	199 (161.0,350.0)
Semiparametric	2175 (1188.0,∞)	1073 (450.0,1288.0)	199 (109.0,297.0)
$\hat{\alpha}$	1.11	1.46	2.19

[a] estimating $f(t)$ in the expression of asymptotic variance given in Proposition 1

Table 3.8: Median survival time and confidence intervals (CI) for hospital readmission data set, using PSH or WC asymptotic variance with Brookmeyer and Crowley (B&C) procedure with no transformation, log-log-transformation and arcsin-transformation. Table also shows bootstrap percentile confidence intervals for selected plans.

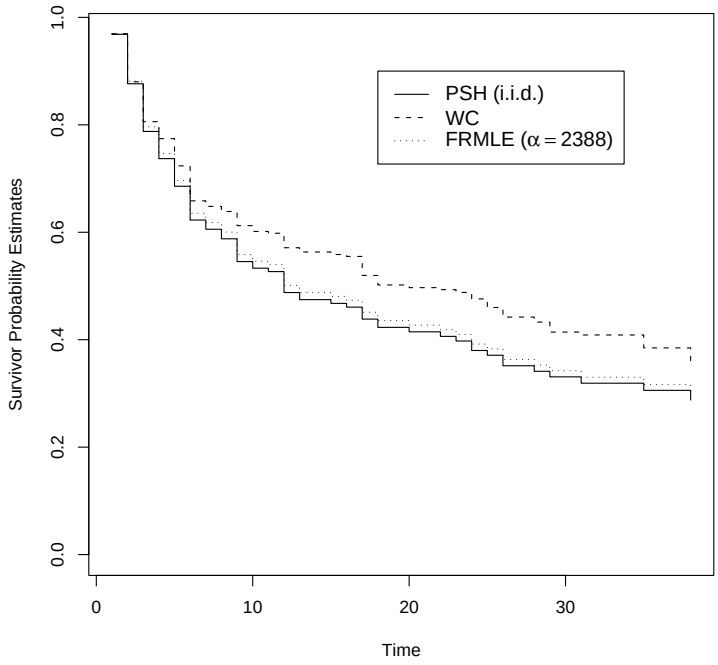

Figure 3.5: Survival function of interoccurrence times estimated using PSH, WC and FRMLE estimators for the bladder cancer data set. This example corresponds to a frailty model when the frailties does not follow a gamma distribution.

3.4.3 Bladder cancer data

Finally, the last example is an application to recurrences on bladder cancer analyzed in Wei et al. (1989). A description of this data set can be found in Section 2.3. Using again a graphical test based on Peña et al. (2001), although both FRMLE and PSH estimators agree (the estimation of α parameter also confirms this assumption), we cannot assume that interoccurrence times are i.i.d. because WC clearly differs from PSH estimator (see Figure 3.5). This indicates the need of using WC estimator or FRMLE with another distribution for the frailties.

Table 3.9 shows median survival time and their asymptotic and bootstrap confidence intervals depending on treatment. As in the previous example, regarding the width of the confidence intervals, we may also state that these results completely agree with those obtained in the simulation studies. The log-log and arcsinus transformations showed the narrower confidence intervals. Similar confidence intervals are obtained using the bootstrap method. Regarding the biomedical interpretation, we can see that patients who receive thiotepa have greater survival than those

	Treatment	
	placebo	**thiotepa**
Asymptotic CI (PSH)		
Estimating $f(t)^a$	10 (6.9,13.1)	20(8.8,31.2)
B&C no-trans.	10 (5.0,29.0)	20 (3.0,∞)
B&C log-log-trans.	10 (7.0,15.0)	20 (7.0,24.0)
B&C arcsin-trans.	10 (7.0,15.0)	20 (7.0,24.0)
Asymptotic CI (WC)		
Estimating $f(t)$	13 (4.3,21.7)	26(-3.0,55.0)
B&C no-trans.	13 (4.0,35.0)	26 (2.0,∞)
B&C log-log-trans.	13 (7.0,28.0)	26 (12.0,∞)
B&C arcsin-trans.	13 (7.0,28.0)	26 (12.0,∞)
Bootstrap CI		
Plan II (PSH)	10 (8.0,16.0)	20 (11.0,∞)
Plan IV (WC)	13 (8.0,29.0)	26 (12.0,∞)
Semiparametric	10 (9.0,18.0)	24 (12.0,∞)
$\hat{\alpha}$	∞	2.99

[a] estimating $f(t)$ in the expression of asymptotic variance given in Proposition 1

Table 3.9: Median survival time and confidence intervals (CI) for bladder data set, using PSH or WC asymptotic variance with Brookmeyer and Crowley (B&C) procedure with no transformation, log-log-transformation and arcsin-transformation. Table also shows bootstrap percentile confidence intervals for selected plans.

who did not receive any drug. However, looking at confidence intervals we cannot say that this difference was statistically significant. The same conclusions are obtained using another approach such as AG model. We observe that patients who received thiotepa have a 19.2% less probability to relapse than those who only received placebo. As we concluded by comparing median survival time, this difference was not statistically significant (p-value for Likelihood ratio test 0.277).

3.5 R instructions for `survrec` package

MMC example

The resulting estimates of the interoccurrence time survivor function for MMC data set showed in Figure 3.2 can be obtained by executing

```
> fit.PSH<-survfitr(Survr(id,time,event)~1,data=MMC,type="p")
> plot(fit.PSH,conf.int=FALSE,cex.lab=2)
> fit.WC<-survfitr(Survr(id,time,event)~1,data=MMC,type="w")
> lines(fit.WC,lty=2)
> fit.FRMLE<-survfitr(Survr(id,time,event)~1,data=MMC,type="M")

Needs to Determine a Seed Value for Alpha
 Seed Alpha:  20.02853

 Alpha estimate= 10.17623

> lines(fit.FRMLE,lty=3)
>
> legend(160,0.9,expression(paste("PSH (i.i.d.)"),paste("WC"),
+        paste("FRMLE (",alpha==10.17,")")),lty=c(1:3),cex=1.2)
> text(25,0.1,"a) i.i.d. case",cex=1.8,adj=0)
                     # produces Figure 3.2
```

Note that when we fit FRMLE estimator using `survfitr` function, appears a message indicating that the algorithm needs a seed value for α. Then, the program calls to another subroutine which computes an initial value for α in order to obtain good convergence in the EM algorithm. It is carried out by the maximization of the profile likelihood for alpha using golden search method.

The pointwise confidence intervals of the survival function can be fitted by writing

```
> fit.PSH<-survfitr(Survr(id,time,event)~1,data=MMC,type="p")
> plot(fit.PSH,conf.int=TRUE) # Figure 3.3
```

In order to compute mean (and median) survival time we can use generic `print` function. We obtain the following results

```
> print(fit.PSH,digits=c(4,2))
Survival for recurrent event data
   n events  mean se(mean) median recurrences: min max median
  19     80 104.1    5.869     98                1   9      4

> print(fit.WC,digits=c(4,2))
```

```
Survival for recurrent event data
   n events  mean se(mean) median recurrences: min max median
  19     99 106.0     12.7     95                2  10      5

> print(fit.FRMLE,digits=c(4,2))
Survival for recurrent event data
   n events  mean se(mean) median recurrences: min max median
  19     80 108.1    6.697    100                1   9      4
```

As we expect, we observe agrement between the three estimators due to the independence between interoccurrence times.

Hospital Readmission example

The resulting estimates of the probability distribution function for hospital readmission data set showed in Figure 3.4 can be obtained as previously but modifying the `prob` argument.

```
> fit.PSH<-survfitr(Survr(id,time,event)~1,data=readmission,type="p")
> fit.WC<-survfitr(Survr(id,time,event)~1,data=readmission,type="w")
> fit.FRMLE<-survfitr(Survr(id,time,event)~1,data=readmission,type="M")

Needs to Determine a Seed Value for Alpha
 Seed Alpha:  0.5

 Alpha estimate= 1.046656
> plot(fit.PSH,xlim=c(0,2000),prob=TRUE,conf.int = FALSE)
> lines(fit.WC,prob=TRUE,lty=2)
> lines(fit.FRMLE,prob=TRUE,lty=3)
>
> legend(1000,0.3,expression(paste("PSH (i.i.d.)"),paste("WC"),
+          paste("FRMLE (",alpha==1.04,")")),lty=c(1:3),cex=1.2)
> text(250,0.1,"b) correlated case",cex=1.8,adj=0)
          # Figure 3.4
```

Median survival times and probabilities of being readmitted showed in Table 3.7 can be obtained by writing

```
> fit.FRMLE<-survfitr(Survr(id,time,event)~as.factor(dukes),
+              data=readmission,type="M")

Needs to Determine a Seed Value for Alpha
 Seed Alpha:  18.18003

 Alpha estimate= 1.113895
```

```
Needs to Determine a Seed Value for Alpha
 Seed Alpha:  12.55364

 Alpha estimate= 1.460141

Needs to Determine a Seed Value for Alpha
 Seed Alpha:  9.342046

 Alpha estimate= 2.193977



# Median survival estimates
> print(fit.FRMLE,digits=c(4,2))
Survival for recurrent event data. Group= as.factor(dukes)
    n events   mean se(mean) median recurrences: min max median
1 180    144 1350.0    67.50   2175                0   6      0
2 148    183  841.5    46.82   1073                0  16      1
3  75    131  363.2    45.96    199                0  22      1

# Probability at one and three years
> for (i in 1:length(fit.FRMLE))
+   { +     cat(i,"\n") +     cat("at one year:
",1-fit.FRMLE[[i]]$surv[sum(fit.FRMLE[[i]]$time<365)],"\n")
+     cat("at three years:
",1-fit.FRMLE[[i]]$surv[sum(fit.FRMLE[[i]]$time<1095)],"\n")
+   }

1
at one year:  0.2604853
at three years:  0.4072520
2
at one year:  0.3814841
at three years:  0.5045216
3
at one year:0.6418163
at three years:  0.8867255
```

Asymptotic confidence intervals, for WC estimator, described in Section 3.1 can be obtained in R as follows:

```
> fit.WC<-survfitr(Survr(id,time,event)~as.factor(dukes),data=readmission,t="w")
> cat("dukes=1 \n")
dukes=1
```

```
> print(Brook.Crowley(fit.WC$"1",0.5))
$percentile [1] 1736

$ci95.asymptotic $ci95.asymptotic$bandwith [1] 1148
$ci95.asymptotic$ci95 [1] 1446.15 2025.85

$ci95.id [1]  655 1736          # formula 3.8
$ci95.log.log [1] 1157 1736     # formula 3.9
$ci95.arcsin [1] 1157 1736      # formula 3.1

> cat("dukes=2 \n")
dukes=2
> print(Brook.Crowley(fit.WC$"2",0.5))
$percentile [1] 1028

$ci95.asymptotic $ci95.asymptotic$bandwith [1] 735
$ci95.asymptotic$ci95 [1]  589.27 1466.73

$ci95.id [1]  276 1483          # formula 3.8
$ci95.log.log [1]  489 1291     # formula 3.9
$ci95.arcsin [1]  462 1291      # formula 3.1

> cat("dukes=3 \n")
dukes=3
> print(Brook.Crowley(fit.WC$"3",0.5))
$percentile [1] 199

$ci95.asymptotic $ci95.asymptotic$bandwith [1] 132

$ci95.asymptotic$ci95 [1] 116.87 281.13


$ci95.id [1]  79 474          # formula 3.8
$ci95.log.log [1] 158 297     # formula 3.9
$ci95.arcsin [1] 158 297      # formula 3.1
```

where `Brook.Crowley` function has been created because it is not included in the `survrec` package (see Appendix E)

On the other hand, the same confidence intervals using bootstrap procedures described in Section 3.2 can be fitted using `survdiffr` function. Let us compute bootstrap Plan II, that is, nonparametric bootstrap estimating $\hat{F}$ using WC estimator, and simulating G from its empirical distribution.

```
> fit<-survdiffr(Survr(id,time,event)~as.factor(dukes),data=readmission,
+                   q=0.5,boot.F="WC",boot.G="empirical",B=999)
> print(fit)
$"1" CASE RESAMPLING BOOTSTRAP FOR CENSORED DATA

Call: survdiffr(formula = Survr(id, time, event) ~
as.factor(dukes),
    data = readmission, q = 0.5, B = 999, boot.F = "WC", boot.G = "empirical")

Bootstrap Statistics :
    original   bias    std. error
t1*     1736 -459.972    810.8057

$"2" CASE RESAMPLING BOOTSTRAP FOR CENSORED DATA

Call: survdiffr(formula = Survr(id, time, event) ~
as.factor(dukes),
    data = readmission, q = 0.5, B = 999, boot.F = "WC", boot.G = "empirical")

Bootstrap Statistics :
    original    bias    std. error
t1*     1028 -54.57057    251.8067

$"3" CASE RESAMPLING BOOTSTRAP FOR CENSORED DATA

Call: survdiffr(formula = Survr(id, time, event) ~
as.factor(dukes),
    data = readmission, q = 0.5, B = 999, boot.F = "WC", boot.G = "empirical")

Bootstrap Statistics :
    original   bias    std. error
t1*      199 14.52452    46.15239
```

We notice that `survdiffr` function returns an object of class `"boot"`. Thus, we can use `boot` package for summarizing the object `fit` using generic `print` function. Then, bootstrap confidence intervals showed in Table 3.8 can be fitted using `boot.ci`. This function is also included in `boot` package.

```
> # Dukes stage A-B
> boot.ci(fit$"1")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS Based on 999 bootstrap
replicates
```

```
CALL : boot.ci(boot.out = fit$"1")

Intervals : Level      Normal             Basic
95%   (1148, 2347 )   (1297, 2315 )

Level     Percentile            BCa
95%   (1157, 2175 )   (1188, 2175 )
Calculations and Intervals on Original Scale


> # Dukes stage C
> boot.ci(fit$"2")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS Based on 999 bootstrap
replicates

CALL : boot.ci(boot.out = fit$"2")

Intervals : Level      Normal             Basic
95%   ( 608, 1566 )   ( 731, 1567 )

Level     Percentile            BCa
95%   ( 489, 1325 )   ( 453, 1325 )
Calculations and Intervals on Original Scale


> # Dukes stage D
> boot.ci(fit$"3")

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS Based on 999 bootstrap
replicates

CALL : boot.ci(boot.out = fit$"3")

Intervals : Level      Normal             Basic
95%   ( 95.3, 275.3 )   ( 48.0, 240.0 )

Level     Percentile            BCa
95%   (158, 350 )   (113, 247 )
Calculations and Intervals on Original Scale Some BCa intervals
may be unstable
```

Chapter 4

Inference for the General Class of Models

This chapter gives procedures for estimating the parameters of the general class of models for recurrent events proposed by Peña and Hollander (2004). An approach based on semiparametric inference (Peña, Slate, and González, 2003) has been developed. When model without frailties is estimated, a generalization of the partial likelihood for the Cox model is obtained. On the other hand, when frailties are included in the model, an EM algorithm is developed. As we have mentioned in Section 1.6, this work was not only carried out by myself, so we have included it in the Appendix B. We encourage the reading of this appendix before going through this chapter because the present chapter is based on the notation and the results we get in it.

Herein, we present another methodology based on a penalized likelihood approach. Two different strategies are adopted. One of them was developed in the shared frailty model context by Therneau et al. (2003). Their idea is based on penalizing the partial likelihood where the penalization bears on a regression coefficient (see Section 1.4.2 for further details). The second penalized approach, also applied in the shared frailty model, was proposed by Rondeau et al. (2003). Their method of estimation is based on the penalized full likelihood, and it gives a non-parametric estimation of the baseline hazard function using a continuous estimator. The solution is then approximated using splines.

The motivations for the present chapter are mainly due to general drawbacks which appears when EM algorithm is used. In particular, direct estimates of the variance of parameters are

not provided. Thus, to solve this problem, we first propose to estimate parameters involved in Peña and Hollander's model using the ideas described in Therneau et al. (2003). However, this approach still continue having problems such as the convergence can be slow and the variance of frailty variance cannot be estimated directly. To improve this approach, an alternative method of penalization proposed by Rondeau et al. (2003) is adopted. The main advantage of this method is that we can easily obtain smooth estimates of the hazard function and an estimation of the variance of frailty variance parameter.

In Section 4.1 we begin by giving how to fit Peña and Hollander model using EM algorithm. We also compare the resulting estimates using this approach with those obtained using existing methods such as AG, WLW and PWP models described in Section 1.4.3. Section 4.2 deals with both methods of estimation based on penalization above outlined. In Section 4.3, based on Nielsen et al.'s Nielsen et al. work, we give some procedures to test whether the frailty is necessary. Some computational issues are discussed in Section 4.4. In Section 4.5 the estimation procedures are illustrated using both readmission and bladder cancer data sets. This last example is also used to compare the results obtained using Peña and Hollander's model to those obtained using three methods of analyzing recurrent event data mentioned in Chapter 1. Bladder cancer data set is also used to illustrate which is the impact of miss-specifying effective age process. Finally, Section 4.6 shows how to fit the general class of models using `gcmrec` and `frailtypack` packages.

4.1 Semiparametric inference: EM algorithm

Before going through the penalized likelihood inference we focus on presenting how to fit Peña and Hollander model using EM algorithm described in Appendix B. To do this, we consider the data set belonging to patients with colorectal cancer described in Section 2.1. In this analysis we consider only the variables tumor stage (Dukes classification: A-B, C or D) and gender. This example will also be useful to describe how to fit and interpret the results obtained using the general class of models as well as compare them to the typical Cox models for recurrent event data described in Section 1.4.3.

This data set was first analyzed in Gonzalez et al. (2005) using a shared gamma frailty model since interoccurrence times were correlated. Thus, instead of using a marginal models such as those described in 1.4.3 we have fitted both AG and PWP models including a frailty term. Table

4.1 shows the resulting estimates using these approaches together with those obtained using the general class of models. First of all, we observe that there is a significant random effect since ξ (frailty precision, see Appendix B) is quite small in all cases. When the effective age corresponds to perfect repair, the resulting estimates from Peña and Hollander model are close to those obtained with the PWP conditional method. On the other hand, when a minimal repair effective age formulation is used, the results are close to those obtained using the AG model although the hazard risk are little different.

The differences in hazard estimates are probably due to the impact of accumulating event readmissions. This effect is incorporated in the Peña and Hollander model via the ρ function. Using, $\rho(k;\alpha) = \alpha^k$ we model different scenarios. As an example, if α is less than unity, the increasing number of rehospitalization has a beneficial effect. In our case, the probability of being rehospitalized would decrease with the number of hospitalization. Looking at our results (Table 4.1) we observe that α is greater than unity, indicating that each hospitalization increases the risk of further hospitalization. Using other words, α parameter greater than one suggests that there is different risk of being hospitalized depending on the number of rehospitalizations. We may further analyze this fact looking at the cumulative hazard functions for the time since last event estimated using the PWP stratified model. The resulting plot, shown in Figure 4.1, shows that the probability of being readmitted increases as the number of hospitalization increases (red lines). On the other hand, the cumulative hazard estimates using Peña and Hollander model is common for all events, and approximates to the average of all marginal cumulative hazards. For this model, the *alpha* parameter allows changing the baseline hazard and it play the same role as the stratification in the PWP model.

Regarding risk estimate, as PWP model is an stratified model, we can fit separate coefficients for both sex and dukes variables to each stratum. Table 4.2 contains these coefficients where the rehospitalizations greater than fourth are combined. We can observe that the differences between males and females are only statistically significant in the probability of being readmitted for the second rehospitalization. This fact is difficult to be explained as the differences observed between males and females using a simple fit model (Table 4.1) as Gonzalez et al. (2005) pointed out. On the other hand, patients with advanced tumoral stage (dukes D) have and approximately 4.5-fold first recurrence rate as compared to those with early tumoral stage (dukes A-B). This difference decrease when posterior rehospitalizations are analyzed to be approximately 2-fold risk. Patients

	Peña and Hollander model		Shared Gamma Frailty model	
	perfect repairs[a]	minimal repairs[b]	PWP model	AG model
Covariate	HR (CI95%)	HR (CI95%)	HR (CI95%)	HR (CI95%)
Gender				
Female	1	1	1	1
Male	1.56 (1.20-2.03)	1.69 (1.19-2.40)	1.47 (1.15-1.88)	1.85 (1.34-2.54)
Dukes stage				
A-B	1	1	1	1
C	1.49 (1.08-2.05)	1.57 (1.04-2.36)	1.46 (1.12-1.91)	1.63 (1.15-2.31)
D	3.06 (2.04-4.58)	4.05 (2.25-7.29)	3.45 (2.47-4.81)	5.03 (3.35-7.56)
Frailty ξ	2.50	1.03	3.09	0.73
$\log N(s-)\ \alpha$	1.08 (0.97-1.19)	1.11 (0.61-1.60)		

[a]Effective Age is backward recurrence time ($\mathcal{E}(s) = s - S_{N^\dagger(s-)}$)
[b]Effective Age is calendar time ($\mathcal{E}(s) = s$).

Table 4.1: Hazard ratios and 95% confidence intervals for the probability of rehospitalization to the readmission data set. Estimates using Andersen-Gill (AG) and Prentice, Williams and Peterson (PWP) methods with a frailty term, together the estimates obtained using Peña and Hollander model using two different effective age formulations.

with dukes C only differ from patients with early tumoral stages (dukes A-B) in the probability of having the first hospitalization (HR:1.66, CI95%: 1.19-2.75).

4.2 Penalized Likelihood Inference

Herein, we propose a different method of parameter estimation that solves problems which appear when EM algorithm is used. In Section 4.2.1, we begin by showing how to adapt Therneau *et al.*'s 2003 approach in our case while in Section 4.2.2 we develop another different penalized approach that is able to give an estimation of the variance of frailty variance parameter. This gives us a possibility making a formal test to see whether data are correlated or not.

Let us assume that $\mathbf{X}_i$ is time-independent, an increasing number of event occurrences is of form $\rho(N_i^\dagger(s-);\alpha) = \alpha^{N_i^\dagger(s-)}$ (that is, α^k, where k is the number of occurrence), and $\psi(x) = \exp(x)$. Assuming this situation, the Peña and Hollander model (1.20) can be also written as follows:

$$\lambda_i(s|Z_i, \mathbf{X}_i) = Z_i\, \lambda_0[\mathcal{E}_i(s)]\, \alpha^{N_i^\dagger(s-)} \exp[\beta' \mathbf{X}_i], \tag{4.1}$$

As in Peña et al. (2003), we also assume that the frailties $Z_1, \ldots, Z_n$ are i.i.d. from a gamma distribution. However, in our case we consider the parameterization $\nu = 1/\xi$. Therefore, we

Covariate	HR	(CI95%)	p value
Male			
1st event	1.26	(0.93-1.70)	0.130
2nd event	1.80	(1.16-2.79)	0.009
3rd event	0.92	(0.49-1.72)	0.790
≥ 4	1.60	(0.78-3.28)	0.200
Dukes C			
1st event	1.66	(1.19-2.31)	0.003
2nd event	1.36	(0.83-2.24)	0.220
3rd event	0.82	(0.39-1.72)	0.600
≥ 4	1.20	(0.53-2.75)	0.660
Dukes D			
1st event	4.54	(3.12-6.59)	<0.001
2nd event	2.15	(1.22-3.78)	0.008
3rd event	2.16	(1.02-4.56)	0.043
≥ 4	2.03	(0.82-5.03)	0.130

Table 4.2: Hazard ratios and 95% confidence intervals per event for the probability of rehospitalization to the readmission data set. Estimates using stratified Prentice, Williams and Peterson (PWP) model.

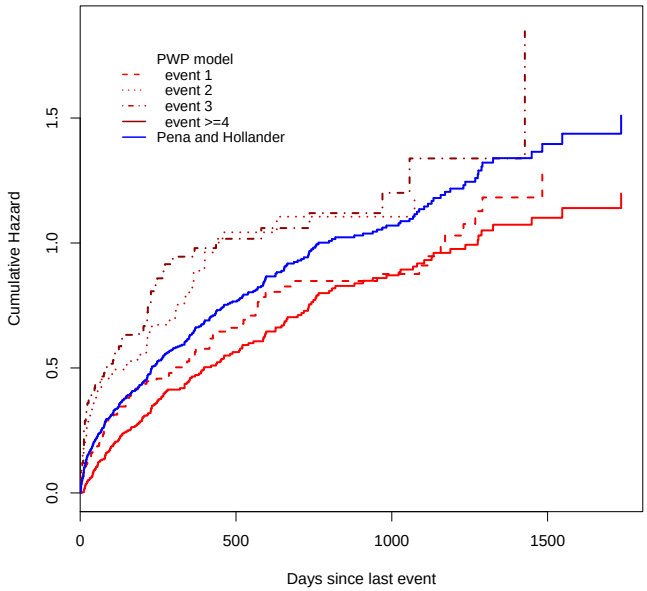

Figure 4.1: Baseline cumulative hazards for each event number using Prentice, Williams and Peterson (PWP) conditional method and Peña and Hollander model to readmission data set.

estimate frailty variance instead of frailty precision. This will be discussed later in Section 4.3.

Using previous notation, the complete log-likelihood process, which is obtained from (B.15), is:

$$\sum_{i=1}^{n}\left\{\int_0^{s^*}\log\left[Z_iY_i^{\dagger}(v)\lambda_0[\mathcal{E}_i(v)]\alpha^{N_i^{\dagger}(v-)}\exp(\beta'\mathbf{X}_i)\right]N_i^{\dagger}(dv)\right.$$
$$\left.-\int_0^{s^*}Z_iY_i^{\dagger}(v)\lambda_0[\mathcal{E}(v)]\alpha^{N_i^{\dagger}(s-)}\exp(\beta'\mathbf{X}_i)dv+\log f(Z_i;\nu)\right\}, \tag{4.2}$$

where $f(Z_i;\nu)$ is the density function of a Gamma distribution. We notice that the frailties, Z, can be viewed as missing data, so expectation-maximization (EM) algorithm can be used for solving the problem of parameters estimation as we have illustrated in Section B.2.

Let $\hat{A}_i \equiv \hat{A}_i(s;\lambda_0,\alpha,\beta) = \sum_{i=1}^{n} Y_i(s;\alpha,\beta)\lambda_0(s)ds$, where the sum is over the distinct jump times. Then, the log-likelihood for the observed data (i.e., full log-likelihood process which is just the logarithm of equation B.16)

$$l_F(s|\lambda_0,\alpha,\beta,\nu) = \sum_{i=1}^{n}\left\{\log\left[\frac{\Gamma(\nu+N_i^{\dagger}(s^*))}{\Gamma(\nu)}\right]-(\nu+N_i^{\dagger}(s^*))\,(\nu+\hat{A}_i)\right.$$
$$\left.+\nu log(\nu)+\sum_{j=1}^{N_i^{\dagger}(s^*)}\log[\lambda_0(\mathcal{E}_i(s))\,\alpha^{N_i^{\dagger}(s-)}\,\exp(\beta'\mathbf{X}_i)]\right\}, \tag{4.3}$$

is found by integrating the distribution of Z. Thus, the estimation of ν can be done by maximizing the profile log-likelihood of this function.

$$l_F(s^*|\nu) = l_F(s^*|\hat{\alpha},\hat{\beta},\hat{\lambda}_0;\nu) \tag{4.4}$$

as we have previously mentioned.

4.2.1 Penalized partial likelihood

The general class of models for recurrent event data (4.1) can be written as a penalized partial likelihood formulation following the approach proposed by Therneau et al. (2003) in the case of the shared frailty model. The idea is to introduce the reparametrization $Z_j = \exp(z_j)$ and to consider them as additional regression coefficients which are constrained by a penalty function. The equation (4.1) can be written as

$$\lambda_i(s|z_i,\mathbf{X}_i) = \lambda_0[\mathcal{E}_i(s)]\,\alpha^{N_i^{\dagger}(s-)}\exp[\beta'\mathbf{X}_i+z'\mathbf{M}_i]. \tag{4.5}$$

where M is a matrix of n indicator variables such that $M_{ij} = 1$ when observation i is a reoccurrence of individual j and 0 otherwise. Therneau et al. (2003) proposed to estimate the parameters involved in this model by maximizing a penalized partial log-likelihood

$$pl_P(s|\alpha,\beta,z) = l_P(s|\alpha,\beta,z) - g(z;\nu) \tag{4.6}$$

over α, β, and z. In this equation, g is a penalty function chosen to restrict the values of z. The parameter ν is a constant which can be given by the user or adapted to the data. For example, one possibility is to choose the penalty function to "shrink" z toward 0 and use ν to control the amount of shrinkage. The function l_P denotes the logarithm of the profile likelihood

$$l_P(s|\alpha,\beta,w) = \sum_{i=1}^{n} \int_0^s \Big\{ N_i^\dagger(v-)\log(\alpha) + \beta'\mathbf{X}_i + z'\mathbf{M}_i - \log \tilde{S}_0(s,\mathcal{E}_i(v)|\alpha,\beta,w) \Big\} N_i^\dagger(dv),$$

where, $\tilde{S}_0$ is a similar process to that described in the Appendix B but including z as an another parameter. That is, $\tilde{S}_0(s,\mathcal{E}_i(v)|\alpha,\beta,z) = \sum_{i=1}^n Y_i(s,t|\alpha,\beta,z)$ and $Y(\cdot,\cdot|\alpha,\beta,w)$ corresponds to the at-risk process defined in Proposition 2 from Peña et al. (2003) by considering that in this case the equation (B.11) becomes:

$$\tilde{\varphi}_{ij}(s|\alpha,\beta,z) \equiv \frac{\alpha^{N_i^\dagger(s-)} e^{\beta'\mathbf{X}_i + z'\mathbf{M}_i}}{\mathcal{E}'_{ij}(s)}. \tag{4.7}$$

To estimate α, β, and z, we solve the score equations. Because the penalty function does not involve neither α, nor β, then $\partial pl_P/\partial\alpha = \partial l_P/\partial\alpha$, and $\partial pl_P/\partial\beta = \partial l_P/\partial\beta$. That is,

$$\frac{\partial pl_P}{\partial\alpha} = \sum_{i=1}^{n} \int_0^{s^*} \left[\frac{\frac{\partial}{\partial\alpha}\alpha^{N_i^\dagger(v-)}}{\alpha^{N_i^\dagger(v-)}} - \frac{\frac{\partial}{\partial\alpha}\tilde{S}_0(s,\mathcal{E}_i(v)|\alpha,\beta,z)}{\tilde{S}_0(s,\mathcal{E}_i(v)|\alpha,\beta,z)} \right] N_i^\dagger(dv) = \mathbf{0}; \tag{4.8}$$

$$\frac{\partial pl_P}{\partial\beta} = \sum_{i=1}^{n} \int_0^{s^*} \left[\frac{\frac{\partial}{\partial\beta} e^{\beta'\mathbf{X}_i + z'\mathbf{Z}_i}}{e^{\beta'\mathbf{X}_i + z'\mathbf{M}_i}} - \frac{\frac{\partial}{\partial\beta}\tilde{S}_0(s,\mathcal{E}_i(v)|\alpha,\beta,z)}{\tilde{S}_0(s,\mathcal{E}_i(v)|\alpha,\beta,z)} \right] N_i^\dagger(dv) = \mathbf{0}. \tag{4.9}$$

Therefore, the score equations for α, and β are the same to those for general class of models treating $z'\mathbf{M}$ as an offset term. On the other hand, the score equation for z is

$$\frac{\partial pl_P}{\partial z} = \sum_{i=1}^{n} \int_0^{s^*} \left[\frac{\frac{\partial}{\partial z_i} e^{\beta'\mathbf{X}_i + z'\mathbf{M}_i}}{e^{\beta'\mathbf{X}_i + z'\mathbf{M}_i}} - \frac{\frac{\partial}{\partial z_i}\tilde{S}_0(s,\mathcal{E}_i(v)|\alpha,\beta,z)}{\tilde{S}_0(s,\mathcal{E}_i(v)|\alpha,\beta,z)} - \frac{\partial g(z;\nu)}{\partial z_i} \right] N_i^\dagger(dv) = \mathbf{0}. \tag{4.10}$$

We also recall that since $N_i^\dagger(\cdot)$ is a step process with a finite number of jumps, then previous estimating equations can be written as finite sums with respect to the event interoccurrence

times S_{ij}s. As in the case without frailties described in Section B.2, we may better understand the estimation equations (4.8), (4.9), and (4.10) using similar notation. For $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, N_i^\dagger(s)$, and using (4.7), define

$$\tilde{Q}_{ij}(s,t|\alpha,\beta,z) = I_{[\mathcal{E}_{ij-1}(S_{ij-1}),\ \mathcal{E}_{ij-1}(S_{ij})]}(t)\,\tilde{\varphi}_{ij-1}\left(\mathcal{E}_{ij-1}^{-1}(w);\alpha,\beta,z\right), \tag{4.11}$$

and

$$\tilde{R}_i(s,t|\alpha,\beta,z) = I_{[\mathcal{E}_{iN_i^\dagger(s-)}(S_{iN_i^\dagger(s-)}),\ \mathcal{E}_{iN_i^\dagger(s-)}(\min(s,\tau_i))]}(t)\,\tilde{\varphi}_{iN_i^\dagger(s-)}\left(\mathcal{E}_{iN_i^\dagger(s-)}^{-1}(t);\alpha,\beta,z\right) \tag{4.12}$$

Some algebra (see Section B.1) shows that the estimating equations (4.8), (4.9), and (4.10) become

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{N_i^\dagger(s^*)}\left[\frac{j-1}{\alpha} - \tilde{\mathbf{A}}(s^*,\mathcal{E}_{ij-1}(S_{ij})|\alpha,\beta,z)\right]\Delta N_i^\dagger(S_{ij}) &= \mathbf{0},\\
\sum_{i=1}^{n}\sum_{j=1}^{N_i^\dagger(s^*)}\left[\mathbf{X}_i - \tilde{\mathbf{B}}(s^*,\mathcal{E}_{ij-1}(S_{ij})|\alpha,\beta,z)\right]\Delta N_i^\dagger(S_{ij}) &= \mathbf{0},\\
\sum_{i=1}^{n}\sum_{j=1}^{N_i^\dagger(s^*)}\left[M_{ij}\delta_i - \tilde{\mathbf{C}}(s^*,\mathcal{E}_{ij-1}(S_{ij})|\alpha,\beta,z) - \frac{\partial g(w;\nu)}{\partial w_i}\right]\Delta N_i^\dagger(S_{ij}) &= \mathbf{0},
\end{aligned}$$

where

$$\tilde{\mathbf{A}}(s,t|\alpha,\beta,z) = \frac{1}{\alpha}\frac{\sum_{i=1}^{n}\left\{\sum_{j=1}^{N_i^\dagger(s)}(j-1)\,\tilde{Q}_{ij}(s,t|\alpha,\beta,z) + N_i^\dagger(s-)\,\tilde{R}_i(s,t|\alpha,\beta,z)\right\}}{\sum_{i=1}^{n}\left\{\sum_{j=1}^{N_i^\dagger(s)}\tilde{Q}_{ij}(s,t|\alpha,\beta,z) + \tilde{R}_i(s,t|\alpha,\beta,z)\right\}},$$

$$\tilde{\mathbf{B}}(s,t|\alpha,\beta,z) = \frac{\sum_{i=1}^{n}\mathbf{X}_i\left\{\sum_{j=1}^{N_i^\dagger(s)}\tilde{Q}_{ij}(s,t|\alpha,\beta,z) + \tilde{R}_i(s,t|\alpha,\beta,z)\right\}}{\sum_{i=1}^{n}\left\{\sum_{j=1}^{N_i^\dagger(s)}\tilde{Q}_{ij}(s,t|\alpha,\beta,z) + \tilde{R}_i(s,t|\alpha,\beta,z)\right\}},$$

and

$$\tilde{\mathbf{C}}(s,t|\alpha,\beta,z) = \frac{\sum_{i=1}^{n}\mathbf{M}_i\left\{\sum_{j=1}^{N_i^\dagger(s)}\tilde{Q}_{ij}(s,t|\alpha,\beta,z) + \tilde{R}_i(s,t|\alpha,\beta,z)\right\}}{\sum_{i=1}^{n}\left\{\sum_{j=1}^{N_i^\dagger(s)}\tilde{Q}_{ij}(s,t|\alpha,\beta,z) + \tilde{R}_i(s,t|\alpha,\beta,z)\right\}}.$$

Upon obtaining the estimators $\hat{\alpha}$, $\hat{\beta}$, and $\hat{z}$, the estimator of underlying cumulative hazard based on the realizations over $[0, s^*]$ is

$$\hat{\Lambda}_0(s^*,t) = \int_0^t \left\{\frac{I\{\tilde{S}_0(s^*,u|\hat{\alpha},\hat{\beta},\hat{z}) > 0\}}{\tilde{S}_0(s^*,u|\hat{\alpha},\hat{\beta},\hat{z})}\right\}\left\{\sum_{i=1}^{n}N_i(s^*,du)\right\}. \tag{4.13}$$

which is just the equation (B.18) of M-step with different notation.

Thus, after having defined analytic forms of estimating equations for α, β and z, we may estimate them using a Newton-Raphson procedure, where the Hessian of penalized likelihood can be written as follows:

$$H = H(\alpha, \beta, w) = \mathcal{I} + \begin{pmatrix} 0 & 0 \\ 0 & g'' \end{pmatrix},$$

were $\mathcal{I} = \mathcal{I}(\alpha, \beta, w)$ is the information matrix of general class of models. That is, the second derivative matrix of 4.7 with respect to α, β, and w.

The algorithm for fitting parameters consists of an inner and outer loop. For a fixed ν, Newton-Raphson iteration is used to solve the penalized model in a few steps (usually 5-6), and return the corresponding value of the pl_P. This first step have some computational problems since the information matrix may have many parameters. Section 4.4 deals with this problem. The outer loop chooses ν to maximize the profile likelihood showed in equation (4.2) that can be easily done using one dimensional procedures.

4.2.2 Penalized full likelihood

Another possibility of penalizing the likelihood is to penalize the full log-likelihood as Rondeau et al. (2003) proposed for the shared frailty model. Now, the idea is to penalize the baseline hazard by a term which takes large values for rough estimations of the function. Thus, the penalized log-likelihood can be defined as follows:

$$pl_F(\alpha, \beta, \Lambda_0(\cdot), z) = l_F(\alpha, \beta, \Lambda_0(\cdot), z) - \kappa \int_0^\infty (\lambda_0^{''})^2(t)dt \tag{4.14}$$

where $l_F(\alpha, \beta, \Lambda_0(\cdot), z)$ is the full log-likelihood for the general class of models defined in (4.3), and $\kappa \geq 0$, is a positive smoothing parameter which controls the trade-off between the data fit and the smoothness of the functions. Next, we briefly outline the main aspects of their approach developed for the shared frailty model that may easily accommodate to our case. Further details can be found in Rondeau et al. (2003) or in Rondeau and Gonzalez (2005).

The main problem in estimating parameter involved in (4.14) arise from the method of maximization because we estimate $\hat{\lambda}(\cdot)$ approximating it by a linear combination of m cubic M-splines $\tilde{\lambda}(.) = \sum_{i=1}^{m} \eta_j M_j(.)$ (see Ramsay, 1988 for further details). Thus, we need to estimate η as well as α and β parameters. In that case, neither the score nor the Hessian of log-likelihood have a simple

analytical form. One possibility, used in Rondeau et al. (2003), is to compute numerically these quantities by using finite differences. That procedure is integrated out in the Marquardt algorithm which is useful for high-dimensional problems such as our case. This algorithm is a combination between a Newton-Raphson algorithm and a steepest descent algorithm (see Section 4.4 for further details). After estimating both η, we use this vector to get the cumulative hazard function with I-splines (integrated M-splines). After that, we may obtain approximate Bayesian pointwise 95% confidence bands for the hazard function by using : $\tilde{\lambda}_0(t) \pm 1.96\sqrt{\mathbf{M}'(\mathbf{t})[\widehat{\text{var}}(\hat{\eta})]\mathbf{M}(\mathbf{t})}$ where $\mathbf{M}'(\mathbf{t}) = (M_1(t), ..., M_m(t))$ is the spline vector in t. As we have mentioned previously, further details of these procedures can be found in Rondeau et al. (2003).

Another important point which has been taken into account when we are dealing with non-parametric methods, is how to chose the smoothing parameter, κ. In practice, it is sometimes sufficient to choose it heuristically, by plotting several curves and choosing that which seems more realistic. Furthermore, Rondeau et al. (2003) also proposed two other approaches to determine the smoothing parameter. One of them is based on an approximate cross-validation score as in Joly et al. (1999) (see also O'Sullivan, 1988). While another one introduces a priori knowledge by fixing the number of degrees of freedom to estimate the hazard function as Gray (1987) proposed.

In some cases, the search for the smoothing parameter may not be reliable because of local extrema. Thus the estimate of the smoothing parameter is not optimal. This can be examined by taking different starting points. Moreover, it seems that the cross-validation score tends to undersmooth, especially for small samples, so in this case the smoothing parameter may be fixed a priori. We have to mention that we implemented the cross-validation procedure for model (4.1) when $Z = 1$ and $\mathcal{E}_i(s) = s - S_{N_i^\dagger(s-)}$ and that a method for other models will be part of our future research.

4.3 Statistical Inference

As Nielsen et al. (1992) stated, a very important point when we are dealing with frailty models, is to test whether the frailty is necessary. In our case and using Peña et al.'s (2003) parameterization, when $\xi \to \infty$ the frailty variance tends to zero and the model becomes frailtyless (all Z's are identically equal to one). Using the parameterization $\nu = 1/\xi$ we may also check the need of frailty by testing the null hypothesis $H_0 : \nu = 0$ (e.g., $H_0 : \xi = \infty$). Let us notice that using this

parameterization we directly estimate frailty variance instead of frailty precision as Peña et al. (2003) propose. Nielsen et al. (1992), in the context of shared frailty model, showed that the value $\nu = 0$ of frailty variance is not on the parameter boundary, so standard likelihood inference methods can be used to test the null hypothesis. Here, using a similar argument, we show that this argument may also be used in the Peña and Hollander model.

Peña et al. (2003) stated that conditional on the data from the time interval $[0, s^*)$, the expectation of Z_i is (see Appendix B)

$$\frac{\xi + N_i^\dagger(s^*)}{\xi + \int_0^{s^*} Z_i\, Y_i^\dagger(v)\, \lambda_0[\mathcal{E}_i(v)]\, \alpha^{N_i^\dagger(v-)} \exp[\beta' \mathbf{X}_i(v)]\, dv}.$$

They also showed that conditionally on $\mathbf{Z}$, the $\mathbf{F}$-compensator of $N_i^\dagger$ is $\{A_i^\dagger(s|\mathbf{Z}, \mathbf{X_i}) : 0 \leq s \leq s^*\}$ with components

$$\frac{\xi + N_i^\dagger(s^*)}{\xi + \int_0^{s^*} Z_i\, Y_i^\dagger(v)\, \lambda_0[\mathcal{E}_i(v)]\, \alpha^{N_i^\dagger(v-)} \exp[\beta' \mathbf{X}_i(v)]\, dv} Y_i^\dagger(v)\, \lambda_0[\mathcal{E}_i(v)]\, \alpha^{N_i^\dagger(v-)} \exp[\beta' \mathbf{X}_i(v)].$$

Now, dividing the numerator and the denominator of the first term by ξ and using the parameterization $\nu = 1/\xi$, previous equation my be written as follows:

$$\frac{1 + \nu N_i^\dagger(s^*)}{1 + \nu \int_0^{s^*} Z_i\, Y_i^\dagger(v)\, \lambda_0[\mathcal{E}_i(v)]\, \alpha^{N_i^\dagger(v-)}\, \exp[\beta' \mathbf{X}_i(v)]\, dv} Y_i^\dagger(v)\, \lambda_0[\mathcal{E}_i(v)]\, \alpha^{N_i^\dagger(v-)}\, \exp[\beta' \mathbf{X}_i(v)]$$

We notice that the general class of models makes sense for all ν (including $\nu = 0$) when the first term is non-negative. This happens for all ν larger than ν^*, the maximum of the quantities $-1/N_i^\dagger(s^*)$ and $-1/\int_0^{s^*} Z_i\, Y_i^\dagger(v)\, \lambda_0[\mathcal{E}_i(v)]\, \alpha^{N_i^\dagger(v-)} \exp[\beta' \mathbf{X}_i(v)]\, dv$.

Let us observe that $\nu^* < 0$, and then $\nu = 0$, is not in the boundary of the parameter space $[\nu^*, \infty)$. Thus, the null hypothesis $H_0 : \nu = 0$ of independence between re-occurrences may be tested using either likelihood ratio test statistic or the Wald test. In the first case, the test for the frailty is twice the difference between the log partial-likelihood with the frailty term integrated out, and the loglikelihood of a no frailty model. We can test this hypothesis at α-level using the χ^2 on one degree of freedom. On the other hand, as in the case of penalized full likelihood approach we obtain an estimation of variance of frailty variance a Wald test type statistic (i.e., $\hat{\nu}/\mathrm{SE}(\hat{\nu})$) may also be used to test $H_0 : \nu = 0$. In this case we refer to a standard normal.

4.4 Computational Issues

The estimation procedures for the general class of models proposed by Peña and Hollander (2004) have been implemented in an `R` package which is available at the CRAN project (Ihaka and Gentleman, 1996, R Development Core Team, 2005, URL:`http://www.R-project.org`). The function *gcmrec* performs estimation for model (1.20) with or without frailties, with $\rho(j;\alpha) = 1$ or $\rho(j;\alpha) = \alpha^j$, and with $\psi(\cdot) = \exp(\cdot)$. These procedures are implemented using the combination of `R` (flexible, high-level statistical language) with `Fortran` (high execution speed of iterative procedures). Thus, we have implemented a *dynamic link library (dll)* in `Fortran 77` that is called by *gcmrec* `R` function. For efficiency, the numerically intensive steps of the algorithm are coded in `Fortran 77`, loaded as a dynamic linked library into `R`, and invoked from the top-level `R` routine (also named *gcmrec*).

A major problem in the parameter estimation is the maximization procedures of the likelihood. We comment briefly on some aspects of methods used in the programming. Estimation of α and β requires maximizing the logarithm of the profile likelihood in the model without frailty, and, additionally in the model with frailty, obtaining the conditional expectation of $\{Z_i\}$. Newton-Raphson (N-R) algorithm can be used to obtain solutions as well as to obtain the inverse of the approximate Hessian. The algorithm continues until convergence, or until a pre-set maximum number of iterations is reached. This algorithm may diverge if the Hessian is not positive definite. To circumvent this, we can modify the Hessian by adding a large enough constant to its diagonal, making the matrix positive definite but perturbing the Hessian as little as possible. This modification requires computing the Choleski factorization in some steps of the N-R procedure. Both the N-R and modified Hessian procedures are implemented in *gcmrec* function.

When we perform estimation with frailty models by penalizing partial likelihood, memory and time-consuming considerations may become an important issue. For instance, let us assume that we have 403 patients such as in the hospital readmissions data set and that we fit the model with frailties. Let us also assume that we are including 6 other variables. Then, the full information matrix has $409^2 = 167281$ elements. One possible solution to this problem is to use quasi-Newton methods that can be used when the Hessian matrix is difficult or time-consuming to evaluate. In this case, instead of obtaining an estimate of the Hessian matrix at a single point, these methods gradually build up an approximate Hessian matrix by using gradient information from some or

all of the previous iterations. BFGS algorithm is one of the widely used quasi-Newton methods. Another possibility is to use a modification of this algorithm, which is known as L-BFGS-B that is a limited-memory quasi-Newton code for large-scale bound-constrained or unconstrained optimization (see Byrd and Nocedal, 1995 for further details). The main disadvantage of these algorithms is that as the inverse Hessian is estimated at each step, we cannot use it for estimating the variance of the parameters. Thus, in our programs, we adopted another approach that was proposed by Therneau et al. (2003) in the case of shared frailty model. Their idea was to partition the inverse of the Hessian according to the rows of X and M (that makes reference to covariates and frailties, respectively) as follows:

$$I = \begin{pmatrix} I_{MM} & I_{MX} \\ I_{XM} & I_{XX} \end{pmatrix}.$$

Then, they propose to use a sparse computation option, where only the diagonal of I_{MM} is retained. They indicate that this method has not a large impact on the estimation procedure because neither the score vector nor the likelihood are changed. Thus, the solution is identical to the one obtained in the non-sparse case, but the speed of the algorithm increased dramatically (see Therneau et al., 2003 for further details).

Finally, when penalized full likelihood approach is used, both the first derivative (the score) and the second derivative (the Hessian) of the log-likelihood themselves do not have a simple analytical form, so another different method from N-R must be used. To solve this problem, we have chosen the Marquardt algorithm which computes numerically these quantities using finite differences. This algorithm was first published by Levenberg (1944). Then, it was rediscovered by Marquardt (1963) who applied it to statistical problems. The Levenberg-Marquardt algorithm (LMA) provides a numerical solution to the problem of maximizing a sum of square of several, generally nonlinear functions that depend on a common set of parameters. This method is applied in our case by maximizing the sum of squares of the partial derivatives to find their solutions. The LMA is a combination between the N-R algorithm and the method of gradient descent and is more robust than the N-R, which means that in many cases it finds a solution even if it starts very far off the final minimum.

In our problem, to be sure of having a positive function at all stages of the algorithm, we restrict all the spline coefficients η_j to be positive for all j. We imposed a constraint of positivity for the parameter ν, so we did not consider a negative dependence in the model, which obviously

do not have a frailty interpretation (although negative values make sense in the intensity as we have illustrated in Section 4.3).

In our approach we use a modified Marquardt algorithm like the Newton-Raphson procedure. Here, we outline the main points of this algorithm also used in Rondeau et al. (2003). Let θ be the parameters to be estimated (in our case $(\eta, \alpha, \beta, \nu)$). If necessary the diagonal of the Hessian at iteration k, $H^{(k)}$, is inflated to obtain a positive definite matrix $H^{*(k)} = H_{ij}^{*(k)}$ with

$$H_{ii}^{*(k)} = H_{ii}^{(k)} + \varrho\left[(1-\delta)\,|H_{ii}^{k}| + \delta\,\mathrm{trace}(H)\right]$$

where ϱ and δ are parameters with initial values set equal to 0.01. They are reduced when H^* is positive definite and increased if not, and

$$H_{ij}^{*(k)} = H_{ij}^{(k)}.$$

The estimates $\theta^{(k)}$ are then updated to $\theta^{(k+1)}$ using both the current modified hessian, $H^{*(k)}$, and the current gradient of the parameters according to the formula:

$$\theta^{(k+1)} = \theta^{(k)} - \phi\frac{\nabla(\theta^{(k)})}{H^{*(k)}}$$

where ∇ denotes the gradient and if necessary, ϕ is modified to ensure that the log likelihood is improved at each iteration.

One-dimensional maximization

In the model with frailties, we further need to maximize the marginal likelihood with respect to only one parameter, ν. The *gcmrec* package provides for two options here: the Newton-Raphson method, and Brent's algorithm (Brent, 1973) which has a faster linear rate of convergence than golden section search. This method is as a one-dimensional maximization without derivatives. First we bracket the maximizing value, and then we obtain it using Brent's method in one dimension (see Brent, 1973 for further details). In both cases, optimization is performed using the reparameterization to $\log(\nu)$. Another possibility is to use $\nu^{1/2}$ as Therneau and Grambsch (2000) used in the context of shared frailty model.

Test of `R` functions

Therneau and Grambsch (2000, Appendix E) gave a set of test data with known answers. These data sets were mainly created to illustrate the computations of statistical algorithms and to ensure

the accuracy and quality of software programs (applied in the survival analysis settings). As the Cox model and shared frailty model are particular cases of the Peña and Hollander model, we have used these data sets to test `gcmrec` function. Using our function, we reproduce the exact results that are obtained using `survival` library (that used in Therneau and Grambsch, 2000). For the Cox model we set up $Z = 1$, $\rho^{N_i^\dagger}$=1 and $\mathcal{E}_i(s) = s - S_{iN_i^\dagger(s-)}$ in the `gcmrec` function and we obtain the same results as using `coxph` function. Similarly, the shared frailty model (using `coxph` and `frailty` functions in `survival` package) showed the same results as using `gcmrec` function when $\rho^{N_i^\dagger}$=1 and $\mathcal{E}_i(s) = s - S_{iN_i^\dagger(s-)}$.

4.5 Hospital Readmission and Bladder Cancer Data Sets Revisited

In this section we apply the estimation procedures developed in preceding sections to two real data sets previously analyzed in Chapter 3 using a more complex model.

4.5.1 Hospital Readmission Study

In this example we consider only the variables tumor stage (Dukes classification: A-B, C or D) and gender corresponding to the data about rehospitalization in patients with colorectal cancer described in Section 2.1. Since in this example we do not have information about the effective age, we assume the backward recurrence time, $\mathcal{E}(s) = s - S_{N^\dagger(s-)}$. We fitted the general model (with frailties), taking $s^* = 2060$, the maximum follow-up time.

After 31 iterations in the EM algorithm (see Section 4.6), the estimate of the frailty precision is $\xi = 1/\nu$ is quite small ($\hat{\xi} = 2.50$), so we may conclude that the frailty component of the model is important for these data. We also saw this fact in Section 3.4.2 by using a graphical method. However, we cannot conclude its statistical signification since the EM algorithm does not provide any estimation of frailty variance. On the other hand, we may also say that among these covariates, the advanced tumor stages (C or D) and males are associated with an elevated risk of rehospitalization. Furthermore, since the estimate of α is larger than unity, there is an indication that each hospitalization increases the risk of further hospitalization, as one could expect.

We may use another method of estimation such as penalized likelihood from the two different points of view we have outlined in Section 4.2. Table 4.3 shows the hazard ratios and its confidence

	Penalized approach		EM approach
	Partial Likelihood	Full Likelihood	Jackknife
Covariate	HR (95%CI)	HR (95%CI)	HR (95%CI95)
Gender			
Female	1	1	1
Male	1.56 (1.17-2.09)	1.63 (1.25-2.13)	1.56 (1.20-2.03)
Dukes stage			
A-B	1	1	1
C	1.49 (1.11-2.00)	1.51 (1.14-2.01)	1.49 (1.08-2.05)
D	3.06 (2.14-4.39)	3.41 (2.39-4.86)	3.06 (2.04-4.58)
log N(s-) α	1.08 (0.97-1.19)	1.14 (1.03-1.24)	1.08 (0.81-1.35)
Frailty ν	0.40	0.58	0.40
(SE ν)	(NA)	(0.15)	(NA)
ξ	2.50	1.72	2.50
κ		3.36×10^{11}	

Table 4.3: Hazard ratios (HR) and 95% confidence intervals for the probability of rehospitalization for the colorectal data set. Estimates using both penalized and EM approach assuming effective age gap time formulation, $\mathcal{E}(s) = s - S_{N^\dagger(s-)}$. Standard errors are computed using $H^{-1}IH^{-1}$ for penalized approach and using Jackknife for the EM approach.

intervals using the EM approach and both methods of penalization. First of all we deal with the importance of frailty. As we have previously mention, we have used a graphical method to check independence assumption. Now, using the penalized full likelihood approach, standard error of the frailty variance can be estimated. So, we can then verify independence assumption using one-side Wald statistic (see Nielsen et al., 1992 or Self and Liang, 1987 and also Section 4.3). In the model $\hat{\nu}/\mathrm{SE}(\hat{\nu}) = 0.58/0.15 = 3.87$ and `1-pnorm(3.87)=5.441768e-05`. Another possibility to test the null hypothesis $H_0 : \nu = 0$ is to use a likelihood ratio test as Therneau and Grambsch (2000) or Nielsen et al. (1992) indicated. As we have mentioned in Section 4.3, this likelihood ratio test may be performed as twice the difference between the log-partial-likelihood with the frailty terms integrated out, and the log-likelihood of a model without frailties. In our case, the test for significance of the frailty is -2751.2 vs. -2719.7, which gives a chi-square statistic of 31.5 on one degree of freedom for a p-value of `2.11e-15`. So, one can conclude that there is heterogeneity among the interoccurrence times. We notice that that EM algorithm and penalized partial likelihood approach should be identical as Therneau and Grambsch (2000) proved in the case of shared frailty model. The prove of this results for the Peña and Hollander model is beyond the scope. We also realize that there is some little differences between the risks estimates using EM algorithm and penalized full likelihood, probably due to differences in the maximization procedures and the selection of number of knots and bandwidth. However, in all case we can conclude that males and patients with advance tumoral stage have more probability to be readmitted and that these risks are statistically significant. On the other hand, we also can see as the width of confidence intervals for hazard ratios are very different depending on the method of estimation used.

Having observed these differences, it is of interest to compare the estimates of variance of α and β parameters using the three different approaches. To do so, we have carried out a little simulation study following Peña et al.'s 2003 procedure (this simulation study is also described in Section 5.3). Table 4.4 summarizes the standard deviation estimates using EM algorithm and both penalized likelihood approaches: partial (PPL) and full (PFL). From this table we note that EM algorithm and PPL approach overestimate the empirical standard deviation of parameters associated with the covariates, β, while the variance of α estimates is quite well estimated. On the other hand, PPL approach clearly underestimates the empirical standard deviations, specially for the α parameter. Thus, after our results, we can say that the conclusions obtained after analyzing

γ	ξ	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\alpha}}^{\text{JACK}}$	$\hat{\sigma}_{\hat{\alpha}}^{\text{PPL}}$	$\hat{\sigma}_{\hat{\alpha}}^{\text{PFL}}$	$\hat{\sigma}_{\hat{\beta}_1}$	$\hat{\sigma}_{\hat{\beta}_1}^{\text{JACK}}$	$\hat{\sigma}_{\hat{\beta}_1}^{\text{PPL}}$	$\hat{\sigma}_{\hat{\beta}_1}^{\text{PFL}}$	$\hat{\sigma}_{\hat{\beta}_2}$	$\hat{\sigma}_{\hat{\beta}_2}^{\text{JACK}}$	$\hat{\sigma}_{\hat{\beta}_2}^{\text{PPL}}$	$\hat{\sigma}_{\hat{\beta}_2}^{\text{PFL}}$
0.9	2	0.007	0.008	0.007	0.003	0.387	0.562	0.483	0.331	0.228	0.312	0.293	0.196
2	2	0.006	0.007	0.007	0.002	0.352	0.498	0.475	0.326	0.180	0.271	0.261	0.166
0.9	6	0.007	0.007	0.006	0.002	0.262	0.403	0.382	0.244	0.162	0.241	0.238	0.149
2	6	0.006	0.007	0.008	0.002	0.221	0.331	0.341	0.204	0.129	0.196	0.185	0.101
0.9	∞	0.007	0.007	0.008	0.002	0.163	0.221	0.214	0.159	0.108	0.183	0.168	0.081
2	∞	0.005	0.006	0.006	0.001	0.128	0.202	0.196	0.131	0.074	0.158	0.146	0.056

Table 4.4: Summary of empirical standard deviations of the estimators of α ($\hat{\sigma}_{\hat{\alpha}}$), and β ($\hat{\sigma}_{\hat{\beta}_1}$ and $\hat{\sigma}_{\hat{\beta}_2}$) using EM algorithm and penalized likelihood approaches (JACK: jackniffe estimates, PPL: penalized partial likelihood, PFL: penalized full likelihood). This corresponds to the simulation study performed in Peña et al. (2003) for the case where the α parameter is 1.05 and the sample size $n = 30$. The true value of β is $(1, -1)$, and 1000 replications were run for each parameter combination. The other parameter are: γ the Weibull shape and ξ the frailty precision.

readmission data set (Table 4.1) are in general correct, except for the statistically significance of α parameter stem from PFL approach.

The last analysis performed in this data set was to compare the baseline survivor function estimates using EM algorithm and PFL approach. Figure 4.2 shows a graphical comparison between the estimation of baseline survivor function using both approaches. We can see that the curve provided by PFL method is completely smooth, while EM algorithm gives a function which jumps at each observed readmission time.

4.5.2 Bladder Cancer Study

We analyze the covariates: X_1, the treatment indicator (1 = placebo; 2 = thiotepa); X_2, the size (in cm) of the largest initial tumor; and X_3, the number of initial tumors. First, we fit the Peña and Hollander model using the gap time formulation, $\mathcal{E}(s) = s - S_{N^\dagger(s-)}$, as effective age. With $s^* = 64$, the maximum observation period, the general model *without* frailties estimates: $\hat{\alpha} = 0.9826$ and $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (-0.3188, -0.0154, 0.1353)$. Same results are also obtained using model *with* frailties since in that case $\hat{\xi} = 5432999$ ($\hat{\nu} \approx 0$). Thus, using the approximate inverse of the partial likelihood information matrix from fitting the model without frailties, the associated estimated standard errors are .0736 for $\hat{\alpha}$ and $(0.2051, 0.0695, 0.0511)$ for $\hat{\beta}$ (Table 4.5).

The effective age for these data is not known, so we also fitted the general model with frailties assuming a calendar time for effective age, $\mathcal{E}(s) = s$. For this case, the estimates are $\hat{\alpha} = .789$, $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (-.5743, -.0315, .2220)$, and $\hat{\xi} = .974$ ($\hat{\nu} = 1.03$). Here we could say that the frailty

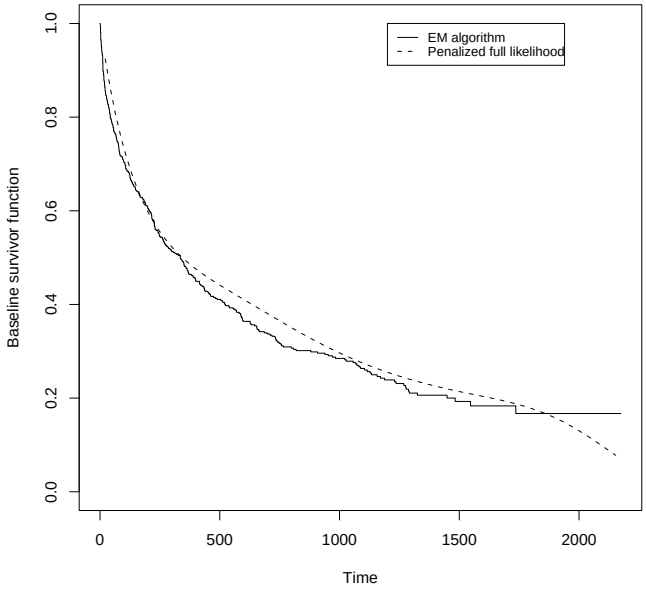

Figure 4.2: Baseline survivor function estimated using both EM algorithm and penalized full likelihood approach for time until next rehospitalization corresponding to readmission data set.

parameter is important. In order to compare both models, the estimates of the survivor functions for the two effective age specifications are presented in Figure 4.3. The lower curves (blue lines), corresponding to the placebo group, are obtained by setting $X_1 = 1$ in the expression given by

$$\{\hat{\bar{F}}_0(t)\}^{\exp\{\hat{\beta}_1 X_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3\}},$$

while the upper curves (dark green lines) are for the thiotepa group obtained by setting $X_1 = 2$. The observed means were $\bar{X}_2 = 2.01$ and $\bar{X}_3 = 2.11$. The solid curves are for the backward recurrence time effective age, while the dashed curves are for $\mathcal{E}(s) = s$. We observe that thiotepa group shows a higher survival rate than the placebo group, although the statistical significance of this difference depends on which effective age process was used.

Then, we compare these results with those obtained using the three existing methods of analysis described in Therneau and Hamilton (1997) and Therneau and Grambsch (2000). Table 4.5 summarizes the estimates from, AG, WLW, and PWP methods, together with the estimates obtained from the Peña and Hollander model using two specifications of the effective age process, $\mathcal{E}(s) = s - S_{N^\dagger(s-)}$ and $\mathcal{E}(s) = s$. The authors analyzed the data set by using models described in Section 1.4.3 (AG,PWP and WLW). These models are marginal models, while the general class

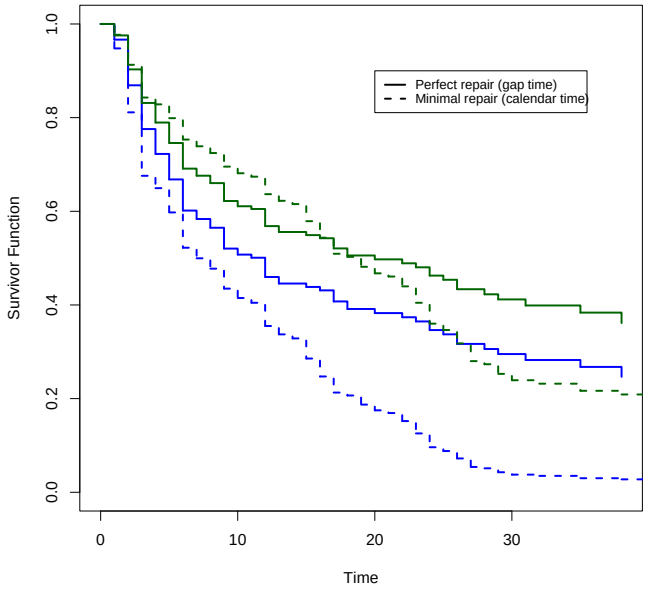

Figure 4.3: Estimates of the survivor function for bladder data set when the Peña and Hollander model is fitted. The blue curve corresponds to the placebo group, while the dark green curve is for the thiotepa group, both evaluated at the mean values of size of initial tumor and mean number of initial tumors. The solid curves show effective age $\mathcal{E}(s) = s - S_{N^\dagger(s-)}$ (perfect repair), while the dashed curves are when $\mathcal{E}(s) = s$ (minimal repair).

of models is a frailty model. As we have mentioned in Section 1.4.2, the estimates from a frailty model have a subject-specific interpretation, while marginal models have a population-average interpretation. Hence, the estimates are not directly comparable to those presented in Therneau and Hamilton (1997) or in Therneau and Grambsch (2000). To make the results comparable we fitted both AG and PWP models including a frailty term (Table 4.5).

First of all, we highlight the role that the effective age process plays in data analysis. We also note that the results obtained using some of existing models are similar to those obtained using the general model. As an example, when 'perfect repairs' is assumed, the results obtained are close to those obtained using PWP model. In both cases the frailty term was not important. On the other hand, the results obtained using 'minimal repairs' are close to those obtained from both AG with frailties and WLW method. We see that in this case it is necessary to include a frailty term to model the association among the inter-event times for each patient. This similitude is what one would expect since the time scale acting in the hazard for the PWP model is gap time, $\mathcal{E}(s) = s - S_{N^\dagger(s-)}$, while in the case of WLW method is time scale formulation, $\mathcal{E}(s) = s$. We realize that the parameter estimates using the AG model without covariates lie between WLW and Peña and Hollander model (estimates from Therneau and Grambsch, 2000) indicating the need of using a frailty model.

The observed differences between models indicate the importance of the effective age and the need to monitor this information. This example along with the results we are showing in the next section, are the basis to further study how to incorporate the notion of effective age in biomedical setting. This was our goal in Chapter 5 where we illustrate how to incorporate information about intervention after relapses in patients with cancer.

4.5.3 Miss-specification of effective age

Because of the importance of the effective age process as demonstrated by previous application to the bladder cancer data, we examined further through a simulation study the impact of miss-specifying the effective age process. We again consider the simulation study carried out in Peña et al. (2003). We examine the impact of two types of effective age process miss-specification: that the interventions following event occurrences are all minimal repair, or that they are all perfect repair. Tables 4.6 and 4.7 show the results for a given simulation (when $\alpha = .9$).

The results indicate an interesting interplay between the nature of the baseline survivor func-

Covariate	Parameter	WLW	AG	PWP	Peña and Hollander Model	
		Marginal	Frailty	Frailty	Perfect[a]	Minimal[b]
$\log N(s-)$	α	-	-	-	0.98 (.07)	0.79
Frailty	ξ	-	0.92	∞	∞	0.97
	ν	-	1.08	5×10^{-7}	5×10^{-7}	1.03
rx	β_1	-0.58 (.20)	-0.61 (.22)	-0.33 (.22)	-0.32 (.21)	-0.57
Size	β_2	-0.05 (.07)	-0.02 (.07)	-0.01 (.07)	-0.02 (.07)	-0.03
Number	β_3	0.21 (.05)	0.24 (.06)	0.12 (.05)	0.14 (.05)	0.22

[a]Effective Age is backward recurrence time ($\mathcal{E}(s) = s - S_{N^\dagger(s-)}$).
[b]Effective Age is calendar time ($\mathcal{E}(s) = s$).

Table 4.5: Summary of estimates for the bladder data set from the Wei, Lin and Weissfeld (WLW), and both Andersen-Gill (AG), and Prentice, Williams and Peterson (PWP) methods including a frailty term to bladder cancer data set, together with the estimates obtained from the Peña and Hollander model using two effective ages corresponding to 'perfect repairs' and 'minimal repairs.'

α	γ	ξ	n	NC	$\hat{\mu}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\mu}_{\hat{\beta}_1}$	$\hat{\sigma}_{\hat{\beta}_1}$	$\hat{\mu}_{\hat{\beta}_2}$	$\hat{\sigma}_{\hat{\beta}_2}$
0.90	0.9	2	10	18	0.915	0.207	0.590	6.786	-1.044	0.669
0.90	0.9	2	30	0	0.909	0.059	1.036	0.414	-1.034	0.270
0.90	0.9	6	10	10	0.877	0.180	1.001	3.418	-1.084	0.482
0.90	0.9	6	30	0	0.907	0.051	1.043	0.327	-1.042	0.220
0.90	0.9	Inf	10	11	0.864	0.145	1.015	4.627	-1.226	0.450
0.90	0.9	Inf	30	0	0.900	0.039	1.057	0.235	-1.060	0.166
0.90	2.0	2	10	34	0.847	0.140	0.777	0.928	-0.802	0.712
0.90	2.0	2	30	0	0.881	0.054	0.721	0.283	-0.716	0.179
0.90	2.0	6	10	22	0.825	0.122	0.856	0.685	-0.820	0.483
0.90	2.0	6	30	0	0.869	0.051	0.740	0.228	-0.741	0.142
0.90	2.0	Inf	10	5	0.805	0.099	0.597	5.694	-0.906	0.280
0.90	2.0	Inf	30	0	0.852	0.038	0.799	0.172	-0.801	0.120

Table 4.6: Results of simulation runs when minimal repair is always assumed after each event occurrence when the actual effective age process is a general minimal repair with perfect repair probability of 0.6.

α	γ	ξ	n	NC	$\hat{\mu}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\mu}_{\hat{\beta}_1}$	$\hat{\sigma}_{\hat{\beta}_1}$	$\hat{\mu}_{\hat{\beta}_2}$	$\hat{\sigma}_{\hat{\beta}_2}$
0.90	0.9	2	10	30	0.876	0.127	1.958	31.850	-1.033	0.628
0.90	0.9	2	30	0	0.894	0.030	1.016	0.390	-1.016	0.247
0.90	0.9	6	10	11	0.861	0.129	1.661	22.382	-1.080	0.449
0.90	0.9	6	30	0	0.893	0.028	1.032	0.305	-1.030	0.192
0.90	0.9	Inf	10	14	0.868	0.067	0.941	4.588	-1.156	0.367
0.90	0.9	Inf	30	0	0.889	0.026	1.045	0.219	-1.053	0.145
0.90	2.0	2	10	2	0.934	0.084	0.433	5.692	-0.761	0.345
0.90	2.0	2	30	0	0.939	0.015	0.755	0.274	-0.748	0.151
0.90	2.0	6	10	3	0.931	0.051	0.802	0.419	-0.771	0.255
0.90	2.0	6	30	0	0.938	0.014	0.736	0.194	-0.744	0.116
0.90	2.0	Inf	10	1	0.929	0.025	0.805	0.297	-0.798	0.186
0.90	2.0	Inf	30	0	0.932	0.013	0.767	0.145	-0.765	0.083

Table 4.7: Results of simulation runs when perfect repair is always assumed after each event occurrence when the actual effective age process is a general minimal repair with perfect repair probability of 0.6.

tion, $\bar{F}_0$, (Decreasing Failure Rate (DFR), $\gamma = 0.9$ or Increase Failure Rate (IFR), $\gamma = 2$) and the behavior of $\hat{\alpha}$. We observed that under the minimal repair miss-specification (Table 4.6) , when $\bar{F}_0$ is DFR, $\hat{\alpha}$ is positively biased. Additionally for this miss-specification, when $\bar{F}_0$ is IFR, $\hat{\alpha}$ is negatively biased. On the other hand, when the miss-specification is perfect repair, an underlying baseline DFR (IFR) is associated with negative (positive) bias in $\hat{\alpha}$. Peña et al. (2003) explain this interplay between $\rho(\cdot;\cdot)$ and $\epsilon(\cdot)$ functions as follows: "When the model mistakenly assumes minimal repair after each reoccurrence, it tends to overestimate the effective age of subjects. Hence, in the case of DFR, the model anticipates longer interoccurrence times than are realized in the data, creating the negative bias, especially for larger interoccurrence times, in the estimates of the baseline survivor function (where the effective age acts). In the case of IFR, the minimal repair miss-specification leads to longer interoccurrence times in the data than are anticipated by the model, creating a positive bias in the estimated baseline survivor function." We can also explain the behavior observed in the case of perfect repair using a similar reasoning. We notice that this behavior induces biases in both $\hat{\alpha}$, and β estimates. As we have previously mentioned, these simulation results further indicate the importance of monitoring the effective age process and it was one of the basis of our research in the next Chapter.

4.6 R instructions for gcmrec package

Hospital readmission example

The resulting estimates of the parameters can be obtained using gcmrec package as follows:

```
> mod.per<-gcmrec(Survr(id,time,event)~as.factor(dukes)+sex,data=readmission,
+   s=2060,typeEffage="perfect",Frailty=TRUE,rhoFunc="alpha to k")
> print(mod.per)
Call: gcmrec(formula = Survr(id, time, event) ~ as.factor(dukes) +
    sex, data = readmission, s = 2060, Frailty = TRUE)


                   coef exp(coef) se(coef)  z  p
as.factor(dukes)2 0.400      1.49       NA NA NA
as.factor(dukes)3 1.119      3.06       NA NA NA
sex               0.446      1.56       NA NA NA

  General class model parameter estimates
    rho function:  Alpha to k
      alpha (s.e.): 1.08 (NA)
    Frailty parameter, Xi (s.e. Jacknife):  2.51  ( NA )

  Marginal log-likelihood= -2747.06
  n= 403
  n times= 861
  number of iterations:  31  EM steps
```

The typeEffage argument indicates which type of effective age is used. Backward recurrence time or gap time formulation corresponds to "perfect" and calendar time to "minimal". Many of the labels in this output are self-explanatory. Some may need some clarification as Marginal log-likelihood. This value corresponds to marginal likelihood in Equation (B.16) or (4.3), evaluated at maximum likelihood estimators of $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\Lambda}$. We also can see that, when model with frailties is fitted, the output does not provide any estimation for the variances. One possibility, before using other approaches such as penalized likelihood inference, is to provide the jackknife estimates of variance parameters. This procedure is very time consuming. After 56 minutes (Pentium III) and adding se="Jack" we obtain

```
> mod.per.Jack<-gcmrec(Survr(id,time,event)~as.factor(dukes)+sex,data=readmission,
+   s=2060,typeEffage="perfect",Frailty=TRUE,rhoFunc="alpha to
```

```
k",se="Jack")
> print(mod.per.Jack)
Call: gcmrec(formula = Survr(id, time, event) ~ as.factor(dukes) +
    sex, data = readmission, s = 2060, Frailty = TRUE, rhoFunc = "alpha to k",
    typeEffage = "perfect", se = "Jack")


                   coef exp(coef) se(coef) Jacknife    z       p
as.factor(dukes)2 0.400      1.49             0.163 2.44 1.5e-02
as.factor(dukes)3 1.119      3.06             0.206 5.42 5.9e-08
sex               0.446      1.56             0.133 3.36 7.8e-04

  General class model parameter estimates
    rho function:  Alpha to k
      alpha (s.e. Jacknife): 1.08 (0.139)
    Frailty parameter, Xi (s.e. Jacknife):  2.51

  Marginal log-likelihood= -2747.06
  n= 403
  n times= 861
  number of iterations:  31  EM steps
```

Next, we will adopt another method of estimation such as penalized full likelihood. The case of penalizing full likelihood can be obtained using `gcmrecPenal` function with `typePen="full"`.

```
# # Penalized full likelihood #
>fit.PenFull<-gcmrecPenal(Survr(id,time,event)~as.factor(dukes)+as.factor(sex),
+ data=readmission,typeEffage="perfect",typePen="full",n.knots=4,
+ kappa1=100000,cross.validation=TRUE,rhoFunc = "alpha to k", +
s=2060)
>print(fitPenFull)
Call:
gcmrecPenal(Survr(id,time,event)~as.factor(dukes)+as.factor(sex),
+ data=readmission,typeEffage="perfect",typePen="full",n.knots=4,
+ kappa1=100000,cross.validation=TRUE,rhoFunc = "alpha to k", +
s=2060)

General class model parameter estimates using a Penalized
Likelihood on the hazard function
                   coef    exp(coef) SE coef (H) SE coef (HIH)    z      p
as.factor(dukes)2  0.415   1.51        0.1447      0.1447     2.867 4.1e-03
as.factor(dukes)3  1.226   3.40        0.1809      0.1809     6.777 1.2e-11
sex                0.490   1.63        0.1353      0.1353     3.617 3.0e-04

    rho function:  Alpha to k
```

```
  alpha (s.e.): 1.14 (0.052)
Frailty parameter, Xi: 1.72 (SE (H): 0.153 ) (SE (HIH): 0.153 )
penalized marginal log-likelihood = -3272.98
n=861
n groups=403
number of iterations:  8
Exact number of knots used:  4
Best smoothing parameter estimated by
  an approximated Cross validation: 3.36E+11
```

This output is also self-explanatory but again we need to clarify some labels. In that case `"Number of iterations"` makes reference to the Marquardt algorithm. The smoothing parameter is estimated assuming a Cox model where the seed is `kappa1=100000`. On the other hand, if we are interesting in penalizing partial likelihood we can fit

```
# # Penalized full likelihood #
>fit.PenPart<-gcmrecPenal(Survr(id,time,event)~as.factor(dukes)+as.factor(sex),
+
data=readmission,typeEffage="perfect",typePen="part",rhoFunc="alpha
to k",s=2060)
>print(fitPenPart)
Call:
fit.PenPart<-gcmrecPenal(Survr(id,time,event)~as.factor(dukes)+as.factor(sex),
+
data=readmission,typeEffage="perfect",typePen="part",rhoFunc="alpha
to k",s=2060)

General class model parameter estimates using Penalized Partial
Likelihood
                   coef    exp(coef) SE coef (H) SE coef (HIH)    z      p
as.factor(dukes)2  0.400   1.49        0.1497      0.1501     2.665 7.7e-03
as.factor(dukes)3  1.119   3.06        0.1838      0.1837     6.091 1.1e-03
sex                0.446   1.56        0.1425      0.1492     2.989 2.7e-03

    rho function:  Alpha to k
      alpha (s.e.): 1.08 (0.054)
    Frailty parameter, Xi: 0.475
    penalized marginal log-likelihood = -2747.06
    n=861
    n groups=403
    number of iterations:  12
```

Chapter 5

Dynamic Cancer Model for Tumor Relapses

In this chapter, we address the problem of how to monitor the effective age process for some biomedical problems such as cancer relapses. We illustrate how to use the information regarding the effects of treatments or interventions for this purpose. Thus, we adopt the general model for recurrent events proposed by Peña and Hollander (2004), in which the effect of interventions is represented by an effective age process acting on the baseline hazard rate function. To accommodate the situation of cancer relapses, we propose an effective age function that encodes three possible therapeutic responses: complete remission, partial remission, and null response.

The motivation of this chapter is mainly due to the importance of the effective age process in biomedical settings as we have demonstrated by analyzing bladder data set and simulations under mis-specification in previous chapter. In addition, and regarding some indolent tumors, MacLaughlin (2002) points out that "*it is necessary [that] a model designed specifically for relapsing patients*" be utilized. The author justifies this by arguing that it is well-known that the impact of therapy after each relapse is a significant prognostic factor for the occurrence of the next one (see Montoto et al., 2002 or Spinolo et al., 1992).

The specification, and consequent analysis, of cancer prognostic models either for overall (time until death) or for disease-free survival (time until relapse or progression) are very useful in making adequate patient management. In particular, there are some indolent type of cancers (e.g. patients with long survival but that tend to relapse over time) where the study of factors

related to the time until progression is important because most patients die from causes related to the disease (see Lister, 1991 or Romaguera et al., 1991). Thus, the estimation of the risk of recurrence would allow for better planning of follow-up after diagnosis or first treatment, and would permit clinicians to consider new therapeutic approaches depending on the patient's risk of relapse.

In Chapter 1 we have shown that there exists many survival models that handle recurrent event data. However, several prognostic studies in major cancer and epidemiologic journals estimate the risk of relapse only using information about the time until first occurrence (follicular lymphomas Lombardo et al. 2002, Lopez-Guillermo et al., 2000, acute leukemias Godder et al., 2004, colorectal cancer Schwandner et al., 2000, or breast cancer Fredriksson et al., 2002, among others). This approach ignores the information of subsequent relapses, hence statistical inference will tend to be inefficient. To avoid this problem, other cancer studies, such us mammary tumors for rats (see Gail et al., 1980) or patients with superficial bladder cancer (see Cheuvart, 1988 or Byar, 1980) rely on the Cox's proportional hazards models and its variants which handle both intra-subject correlation and event dependence.

As we have illustrated in Section 1.3.1, another aspect that may modify the event occurrence intensity arises from the interventions performed on the subject after each event occurrence. In cancer settings, patients with the disease are treated after observing a progression of the tumor. In the particular case of indolent lymphomas, as we are dealing with non-curable disease, the therapy aims to increase as much as possible the time until the next relapse. After giving some therapy, the patient is monitored and then we observe if cancer or some disease-related symptoms disappear. Thus, patients whose disease completely disappears will have less probability to relapse than those where little or no response is observed. Hence such interventions can be viewed as improving the patient. In the reliability literature this is referred to as adjusting the *effective age* of the system. This is the basis of work we developed in present chapter.

We briefly outline the contents of this chapter. We begin by providing some additional notation for the general class of models in Section 5.1. This section also outlines how this class of models subsumes some existing models for recurrent event data which have been used in biomedical settings. Section 5.2 presents a description of the effective age process in cancer settings. A simulation study is performed to study the model behavior under effective age mis-specification. The simulation results and design are discussed in Section 5.3. Section 5.4 illustrates the use of

the model with a real data set from low grade lymphomas. Finally, Section 5.5 shows how to fit the dynamic cancer model using `gcmrec` package.

5.1 The Peña and Hollander Model Revisited

For a patient, as in the general case, in cancer settings we can observe a, possibly time-varying, q-dimensional vector of covariates such as gender, age, race, disease status, beta-2 microglobulin level, treatment regimen, etc. We denote this covariate process by

$$\{\mathbf{X}(s) = (X_1(s), X_2(s), \ldots, X_q(s))' : 0 \leq s \leq \tau\}.$$

In addition, in our case, after treatments or interventions are administered upon relapses, information about patient status may be obtained. Examples of interventions are chemotherapy, radiotherapy, and bone marrow transplant, among others. We denote this information by a vector

$$\psi = (\psi_1, \psi_2, \ldots, \psi_K)',$$

where ψ_j signifies a certain type of response to the intervention after the jth relapse. This will be explained in more detail later. Consequently, if in the study there are n patients, over the period $[0, \tau^*]$ where $\tau^* \equiv \max_{i \leq n} \tau_i$, we will have the following data:

$$\mathbf{D}(\tau^*) \equiv \{[(\mathbf{X}_i(s) : 0 \leq s \leq \tau), \psi_i, K_i, \tau, S_{i1}, S_{i2}, \ldots, S_{iK_i}, \tau_i - S_{iK_i}], i = 1, 2, \ldots, n\}.$$

We recall that the conditional intensity function given in 4.1 is

$$\lambda_i(s|\mathbf{Z}, \mathbf{X}_i) = Z_i \, \lambda_0[\mathcal{E}_i(s)] \, \alpha^{N_i^\dagger(s-)} \, \exp[\beta^t \mathbf{X}_i(s)].$$

This model incorporates the effect of performed interventions through the effective age, $\mathcal{E}_i(s)$, which serves as the argument to the baseline hazard rate function Peña and Hollander (2004). Here, we briefly show as some of existing models that have been used in recurrent events problems are particular cases of this general class of models. The simplest model regarding the effective age is when $Z = 1$, $\rho[N_i^\dagger(s-); \alpha] = 1$, and the effective age is always the same type and is either of the following possibilities: First, a patient can achieve a perfect response to the treatment. This means that the patient recovers perfectly and the status is the same as at the beginning of the study (if there is no time-dependent covariate). This model has been considered by Prentice

et al. (1981), Lawless (1987), and Aalen and Husebye (1991). Second, the therapy does not have an effect on the patient, so the state of the patient is the same as just before the relapse. This is known in reliability literature as "minimal repair" and it has been studied Prentice et al. (1981), Brown and Proschan (1983), and Lawless (1987).

We realize that some of the models mentioned above can be formulated using two different expressions for effective age function. As an example, some of the conditional models examined Prentice et al. (1981) (also called PWP in the introduction) can be formulated by organizing the data in calendar time (PWP-TT) (i.e., total time risk set or time from each unit's entry into the observation set) or interoccurrence/gap time (PWP-GT) (i.e., gap time risk set or time since the previous event). Thus, the hazard function only differs in the at-risk process formulation. In the case of PWP-TT the effective age corresponds to $\mathcal{E}_i(s) = s$ and in the PWP-GP formulation to $\mathcal{E}_i(s) = s - S_{iN_i^\dagger(s-)}$. The choice between PWP-TT or PWP-GT depends on whether we are interested in the time that has elapsed since a patient entered the study or since the last relapse. Models which employ calendar time formulation assume that all interventions produce a minimal or no improvement in the patient. In medical terms, the disease is continuing in a stable manner. Models based on gap-time formulation assume that all interventions lead to perfect recovery for the patient, e.g. disease disappears, which is known as complete remission in the cancer literature.

Other models that have been used in biomedical problems where neither $\rho[\cdot;\alpha]$ nor Z are both identically unity are the following. Gail et al. (1980) cancer occurrence model is obtained if we take, $Z = 1$, $\rho[N_i^\dagger(s-);\alpha] = \max\{\alpha - N_i^\dagger(s-), 0\}$, where α is some real number, and $\lambda_0(s) = \lambda_0$, where λ_0 is some positive constant. In this model α can be interpreted as an initial measure of the patient's susceptibility to events, which is becoming weaker as the relapses accumulate. Shared frailty model (see Oakes, 1991) arises from Peña and Hollander's model by taking $\rho[N_i^\dagger(s-);\alpha] = 1$, $\mathcal{E}_i(s) = s - S_{N_i^\dagger(s-)}$, and putting some parametric distribution to the frailty component Z, such as a gamma or a lognormal distribution.

In all examples previously mentioned, the expressions of $\mathcal{E}_i(s)$ do not depend on the number of relapses. In other words, the effect of treatment after each occurrence is always the same. However, we know that the effect of intervention after each relapse is not always the same. So, we need to define how to incorporate the effect of intervention upon relapses via a more general effective age. To do so, response to therapy will become crucial. The next section deals with this aspect.

5.2 Effective Age Process for Cancer Data

In Section 1.3.1 some examples about effective age in biomedical settings were discussed. Now, we will focus on cancer problem. Patients with indolent lymphomas, and in general with cancer, are monitored from the date of diagnosis to the time of death or loss to follow-up. At any evaluation, complete remission (CR) is defined as the disappearance of tumor masses and disease-related symptoms, as well as the normalization of the previous test and/or biopsies, lasting for at least one month. Partial remission (PR) is said to occur when measurable lesions have decreased by at least 50%. Patients not included in these categories are called non-responders (NR) (see Cheson et al., 1999). It is well-known that the response to the treatment is related to the time until the next relapse (see for instance Montoto et al., 2002, Weisdorf et al., 1992, or Davidge-Pitts et al., 1996), so it is reasonable to have a model which incorporates this information. However, neither AG, PWP, WLW, GSB, nor frailty models, which are mostly used in biomedical settings, have incorporated the effect of performed interventions upon event reoccurrences, though some of these methods could accommodate such information through the use of time-dependent covariates as will be illustrated in the example presented in Section 5.4.

We propose using the response to therapy, defined as CR, PR, or NR, to define a model for the effective age for relapsing patients as follows. Consider a single patient and let $\{A_j : j = 0, 1, 2, \ldots\}$ be a sequence satisfying

$$A_0 = 0, \quad A_j = A_{j-1} + \left(\prod_{k=1}^{j} [1 - \psi_k] \right) T_j, \quad j \geq 1, \tag{5.1}$$

where $\boldsymbol{\psi} = (\psi_1, \psi_2, \ldots, \psi_K)'$, with $\psi_j \in \{0, .5, 1\}$ and with the interpretation that $\psi_j = 0$ means that an NR (non-response) has occurred after the jth relapse, $\psi_j = 1$ means that a CR (perfect intervention) has occurred, while $\psi_j = .5$ means that a PR (partial remission) has transpired. The values of the ψ_js can be assessed by the clinician(s) monitoring the patient. Our proposed effective age process for cancer relapse is

$$\mathcal{E}(s) = A_{N^\dagger(s-)} + \left(s - S_{N^\dagger(s-)} \right). \tag{5.2}$$

The effective age (5.2) is a particular case of Kijima's 1989 model II (Section 1.3.2), where in his model the ψ_js are assumed to take any values in $[0, 1]$, whereas in our model we assume that they only take three possible values. Kijima's aim was to model the situation where after

each failure some repair is performed and the effectiveness of this repair could be quantified by a number that is between zero and one. We note that in cancer problems, we may also assess this "degree" of response according to a number in $[0, 1]$; however, in realistic settings clinicians only need to know if a CR, PR, or NR was achieved to make a good therapy determination, hence our restriction of the possible values of the ψ_js to the set $\{0, .5, 1\}$. Dorado et al. (1997) also studied effective age functions that encompass the form (5.2). Note that if all responses are CR, i.e. $\psi_i = 1, i = 1, 2, \ldots$, then the effective age corresponds to gap time formulation, $\mathcal{E}(s) = s - S_{N^\dagger(s-)}$ since all A_js in (5.1) become 0. Similarly, if all responses are NR, the effective age corresponds to a calendar time formulation, $\mathcal{E}(s) = s$.

To demonstrate the notion of an effective age in biomedical settings, Figure 5.1 shows the effective age for a patient in a cancer study. This process between 0 (or S_0) and S_2 corresponds to $\mathcal{E}(s) = s$ (calendar or elapsed time formulation). At the first event S_1, treatment or intervention did not improve the disease status. In medical parlance, the patient did not respond to treatment, i.e., NR is achieved. After the second event, which occurred at S_2, the patient responds perfectly to treatment, achieving a complete resolution of all clinical manifestation of the disease. It is considered a CR. In this case the effective age corresponds to $\mathcal{E}(s) = s - S_{N^\dagger(s-)}$ (backward recurrence time). However, after the third event at time S_3, the patient reverts to a state between a CR and a NR, that is, the patient experiences a little improvement or a PR to treatment. Finally, a progressive disease is observed for the fourth relapse at S_4, possibly due to some complications. We have decided to show a progressive possibility despite we do not include it in our analysis because physicians may observe this types of responses. Thus, we indicate how this information should be incorporated in the model. Finally, the fifth failure which would have happened at S_5 is not observed since the end of observational period τ for this hypothetical patient is less than S_5. Consequently, the gap time for the fifth event is right-censored by $\tau - S_4$. .

Next, we illustrate the effective age with a numerical example. Let us suppose that we observe a patient who receives an initial treatment at time 0, and this patient relapses four times at calendar times 30, 55, 100, and 150, and gets censored at calendar time 175. Thus, the gap times are 30, 25, 45, and 50, respectively. If we assume perfect repair model, e.g. the patient achieves CR after each intervention, the effective age at each recurrence will be 0, 30, 25, 45, and 50. On the other hand, if we assume minimal repair model, e.g. treatment is not effective at each recurrence or NR after each treatment, the effective age at each relapse will be 0, 30, 55, 100, and

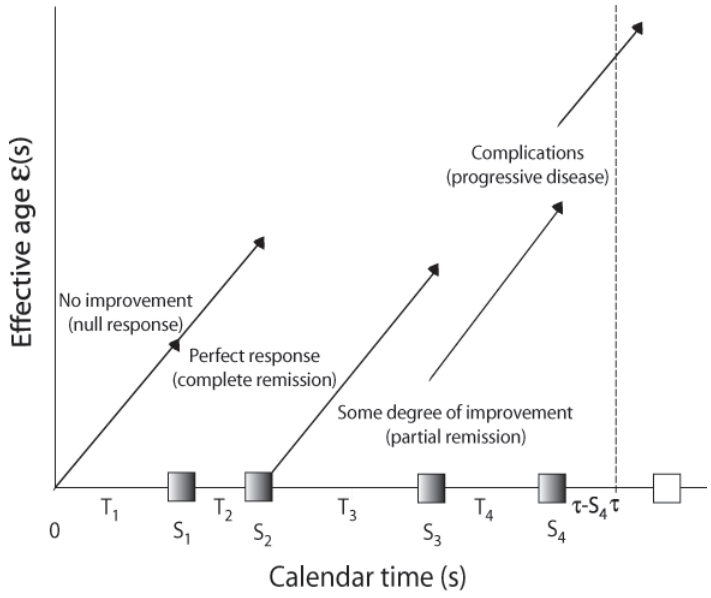

Figure 5.1: Pictorial representation of effective age vs. calendar time for a hypothetical unit in cancer settings.

150 (the same as calendar times). But the effect of the intervention at each recurrence need not be always the same. For example, let us suppose that the treatment after the first relapse does not improve the patient's health at all, so we observe an NR. Then after the second relapse the patient achieves a CR. A PR is observed after third recurrence, and finally an NR is achieved at the fourth occurrence. Thus, if the patient has no improvement after the first intervention, the effective age will be the same as if the patient had a minimal repair. After the second relapse the patient has a perfect intervention. In that case the effective age corresponds to gap time or backward recurrence time. After the third recurrence the patient acquires some, but not total, improvement, so the effective age will be between those observed in the perfect and the minimal repair situations, say halfway. Finally, in the last relapse patient does not get better, so the effective age will start at a higher value and proceed possibly in a linear fashion. Under this hypothetical situation, the effective age would take the values 0, 30, 55, 45, and 72.5, with the effective age at the time of censoring (calendar time 175) being higher than 97.5.

We will refer to Peña and Hollander model with effective age in 5.1 as a dynamic *cancer* model. Because this model is just a special case of the general recurrent event model considered in Chapter 4, the procedures for estimating the parameters of this general class of models therefore applies to this cancer model.

5.3 Simulation Study

5.3.1 Simulation Design

We have carried out simulation studies to examine empirically the properties of the parameter estimators described in Sections 5.1 and 5.2. In particular we study in the simulation:

- the effect of sample size (n)
- bias and variance of the estimators
- the performance of the estimator of the baseline survival function $\bar{F}_0(t) = \exp\{-\int_0^t \lambda_0(w)dw\}$, in terms of its bias function and root-mean-square-error (RMSE) function at specified time points
- the consequences of miss-specifying the effective age function

In the last item we consider the situation where the data have been generated by the model allowing for different responses after each intervention, but with the resulting data analyzed by assuming that the patients always achieve the same response (that is, either always CR or always NR) after each tumor reoccurrence.

We mimicked the simulation study performed by Peña et al. (2003). We point out that because the simulation was meant to cover general biomedical settings, some of the parameter values we considered may not be realistic for cancer settings, for instance, the mean number of events per patient, and effective age function.

For the simulation, we considered an effective age function corresponding to the cancer model. That is, a patient can achieve a complete, a partial, or a null response depending on the vector ψ. We have assumed three different scenarios according to the following probability functions for ψ which takes values in the set {1=CR, 0.5=PR, 0=NR}: $\{(.8, .1, .1), (.3, .5, .2), (.1, .2, .7)\}$. Thus, in the first case, we assume that patients achieve CR with a probability of 80%, and PR or NR 10% of the time, respectively. These three sets of distributions allow us to cover three different scenarios: the first assumes that in a large majority of cases, perfect response is achieved after each relapse; the third has minimal response predominating; and the second distribution is an in-between scenario. For notation in the sequel, when we write $p(\psi) = (p_1, p_2, p_3)$ to indicate that

the ψ values are chosen such that 1 (= CR) occurs with probability p_1, .5 (= PR) occurs with probability p_2, and 0 (= NR) occurs with probability p_3.

To examine the impact of sample size, we select two values of $n : n \in \{30, 50\}$. These values are sufficient to study the limitations of the model in cancer settings since it is usual to have at least these sample sizes in this context. The censoring variables τ_i, $i = 1, \ldots, n$, are generated according to same scheme as in previous simulation study. That is, a uniform distribution over $[0, B]$, where B is chosen such that under the assumption that patients always achieve a complete remission after relapses (i.e., $\mathcal{E}(s) = s - S_{N^\dagger(s-)}$) and with no accumulating events effect (i.e., $\alpha = 1$). On average, there are approximately 5 events per patient since in cancer problems, it is difficult to find situations with more than this number of reoccurrences. For the baseline hazard function λ_0, we choose a Weibull distribution, with unit scale parameter and shape parameter, γ, taking values in $\{.9, 2\}$. Thus, we are able to study two different situations: one for an increasing baseline hazard rate function and the other one for a decreasing case. The impact of the accumulating number of relapses is assumed to be of form $\rho(k, \alpha) = \alpha^k$ form. We have selected $\alpha \in \{0.9, 1, 1.05\}$. Thus, we are able to study the case where an increasing number of relapses increases the time of the next relapse (that is, beneficial effect) which is the case when $\alpha = .9$, as well as the case where there is no effect which is for $\alpha = 1$, and the case where there is an adverse effect which is when $\alpha = 1.05$, respectively. In order to take into account the effect of covariates, we have simulated a two-dimensional covariate vector (X_1, X_2). Then, to have both categorical and continuous covariates, X_1 has been simulated to have a Bernoulli distribution with success probability of .5 and X_2 was set to have a standard normal distribution. These covariates were generated to be stochastically independent. The regression coefficient vector (β_1, β_2) was set to $(1, -1)$. Finally, the frailty component was generated under a gamma distribution with unit mean and variance $1/\xi$. The parameter ξ took values in $\{2, 6, \infty\}$, with ∞ corresponding to the absence of frailties.

We performed 1,000 replications for each combination of simulation parameters. To create the bias and RMSE curves of the estimator of the baseline survivor function, we chose the time values that corresponded to the $[0 : (.01) : .99]$ quantiles of the true baseline distribution function. In order to study the miss-specification of effective age function, we have also estimated the parameters for the dynamic cancer model assuming a model with an effective age function that considers that patients always achieve a perfect response (CR) and another where patients

always achieve a minimal response (NR). Through these simulations we are able to highlight the importance of the effective age function in relation to either under- or over-estimation of the baseline survivor function. Given that we wanted to compare three different effective age functions (minimal, perfect, and cancer model) and given the three discrete distributions that specified the effective age process, two sample sizes, two hazard shapes, tree levels of event dependence, and three degrees of correlation, we conducted a total of 324 simulation experiments.

5.3.2 Simulation Results

In the discussion of the simulation results, we will focus on the consequences of analyzing data from the cancer model when analyzed using models which always assume the same response, either always perfect or always minimal. Regarding distributional properties of the estimators of α, β, and $\eta \equiv \xi/(1+\xi)$, and the estimator of the baseline survivor function when the correct model is utilized, we only focus on the results presented in the tables and figures. Further details of this simulation are described in Peña et al. (2003).

Results of the simulations are shown in the following tables. Table 5.1 summarizes the mean values and standard deviations of the sampling distributions of the estimators of α, β_1, β_2, and η for α, n, and ξ varying in the sets mentioned above and for $p(\psi) = (0.8, 0.1, 0.1)$. Figure 5.2 shows plots of the bias and RMSE curves for the non-parametric estimator of $\bar{F}_0$, where $\alpha = 0.9$, $p(\psi) = (0.8, 0.1, 0.1)$, and $n \in \{30, 50\}$. The lines of each plot represent the three different values of ξ: 2, 6, and ∞. Each plot frame contains the figure for two Weibull shape parameters, $\gamma = 0.9$, and $\gamma = 2.0$. Tables 5.2 and 5.3 present the summary of simulation results belonging to the effective age mis-specification analysis. Figure 5.3 shows the estimated baseline survivor function, the bias, and the RMSE curves calculated under effective age miss-specification, showing the effect of different effective age function chosen (always minimal or perfect response), the impact of different ξ values, and for the three different cancer models analyzed. The results are for $\alpha = 0.9$ and $n = 30$.

As one may expect, when there is no miss-specification and when the sample size increases, the performance of the estimators of the finite-dimensional parameters and the baseline survivor function improved, as can be seen by noting that both bias and standard error decrease. We also notice that when the sample size is small, there is considerable over-estimation of η (Table 5.1). Examining the curves in Figure 5.2 we observe that the estimator of baseline survivor function is

α	γ	ξ	η	n	NC	$\hat{\mu}_{Ev}$	$\hat{\mu}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\mu}_{\hat{\beta}_1}$	$\hat{\sigma}_{\hat{\beta}_1}$	$\hat{\mu}_{\hat{\beta}_2}$	$\hat{\sigma}_{\hat{\beta}_2}$	$\hat{\eta}$
0.9	0.9	2	0.67	30	0	2.87	0.896	0.035	1.033	0.415	−1.012	0.241	0.714
0.9	0.9	2	0.67	50	0	3.98	0.896	0.022	1.021	0.325	−1.024	0.174	0.69
0.9	0.9	6	0.86	30	0	6.13	0.895	0.032	1.02	0.327	−1.027	0.198	0.895
0.9	0.9	6	0.86	50	0	4.5	0.897	0.022	1.005	0.23	−1.015	0.141	0.877
0.9	0.9	∞		30	0	5.4	0.894	0.026	1.022	0.237	−1.022	0.152	
0.9	0.9	∞		50	0	4.1	0.896	0.019	1.03	0.169	−1.022	0.1	
0.9	2	2	0.67	30	0	7.13	0.905	0.019	0.981	0.302	−0.994	0.168	0.763
0.9	2	2	0.67	50	0	8.32	0.904	0.014	1.014	0.222	−0.985	0.121	0.731
0.9	2	6	0.86	30	0	6.8	0.903	0.017	0.983	0.206	−0.987	0.127	0.896
0.9	2	6	0.86	50	0	7.9	0.903	0.013	0.997	0.154	−1.001	0.09	0.883
0.9	2	∞		30	0	7.5	0.899	0.017	1.036	0.161	−1.02	0.099	
0.9	2	∞		50	0	8.2	0.898	0.012	1.014	0.113	−1.016	0.067	
1	0.9	2	0.67	30	10	2.87	0.994	0.032	1.003	0.444	−1.001	0.251	0.728
1	0.9	2	0.67	50	1	4	0.999	0.015	1.006	0.321	−0.99	0.192	0.713
1	0.9	6	0.86	30	5	5.2	0.995	0.029	1.033	0.353	−1.019	0.199	0.903
1	0.9	6	0.86	50	0	4.72	0.998	0.014	1.015	0.264	−1.009	0.155	0.884
1	0.9	∞		30	0	2.3	0.995	0.024	1.006	0.254	−1.031	0.165	
1	0.9	∞		50	0	4.8	0.997	0.014	1.025	0.198	−1.019	0.121	
1	2	2	0.67	30	0	5.07	1.009	0.027	0.988	0.327	−0.973	0.183	0.744
1	2	2	0.67	50	0	4.86	1.008	0.019	0.99	0.25	−0.981	0.137	0.73
1	2	6	0.86	30	1	4.97	1.006	0.026	1	0.262	−0.988	0.144	0.889
1	2	6	0.86	50	0	4.06	1.007	0.017	0.993	0.19	−0.991	0.11	0.867
1	2	∞		30	0	7.23	0.997	0.023	1.031	0.202	−1.023	0.118	
1	2	∞		50	0	3.92	0.999	0.016	1.024	0.149	−1.014	0.094	
1.05	0.9	2	0.67	30	5	6.53	1.05	0.017	1.021	0.421	−0.997	0.239	0.742
1.05	0.9	2	0.67	50	0	6.42	1.051	0.008	0.984	0.315	−0.991	0.184	0.707
1.05	0.9	6	0.86	30	6	2.03	1.051	0.016	0.989	0.317	−1.003	0.205	0.894
1.05	0.9	6	0.86	50	3	4.96	1.051	0.008	1.002	0.231	−0.994	0.154	0.889
1.05	0.9	∞		30	0	6.2	1.05	0.014	1.029	0.229	−1.036	0.155	
1.05	0.9	∞		50	0	7.34	1.051	0.008	1.022	0.166	−1.01	0.106	
1.05	2	2	0.67	30	1	5.77	1.055	0.022	0.988	0.334	−0.98	0.187	0.758
1.05	2	2	0.67	50	0	7.26	1.051	0.012	1.01	0.237	−1.006	0.138	0.72
1.05	2	6	0.86	30	0	8.63	1.053	0.02	1.008	0.246	−1.011	0.145	0.889
1.05	2	6	0.86	50	0	7.54	1.052	0.013	1.005	0.192	−1.008	0.105	0.872
1.05	2	∞		30	0	6.57	1.049	0.018	1.034	0.187	−1.012	0.119	
1.05	2	∞		50	0	7.2	1.048	0.013	1.03	0.141	−1.023	0.082	

Table 5.1: Summary of simulated means and standard deviations of the estimators of α, β, and $\eta = \xi/(\xi+1)$. The true value of β is $(1, -1)$. Results correspond to the case of $p(\psi) = (0.8, 0.1, 0.1)$, and 1000 replications were done for each parameter combination. The others columns are: γ Weibull shape parameter, n sample size, NC number of replicates in which there was no model convergence; $\hat{\mu}_{Ev}$ mean number per patient in all the simulation replications.

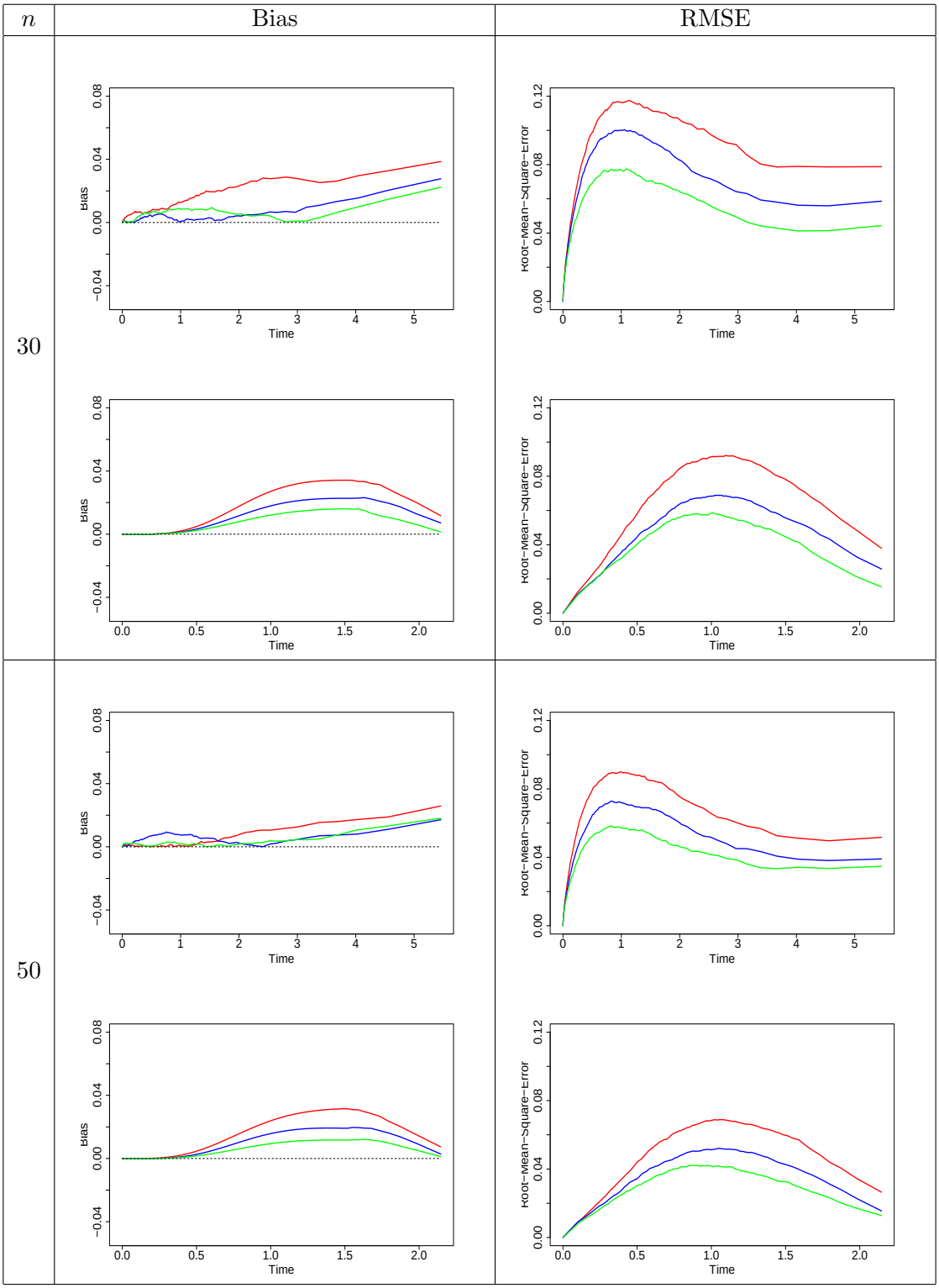

Figure 5.2: Bias and root mean squared error curves for the estimator of the baseline survivor function as the frailty parameter ξ varies ($\xi = 2$ red line; $\xi = 6$ blue line; $\xi = \infty$ green line) for the two sample sizes. This is for the case where $\alpha = .90$ and $p(\psi) = (0.8, 0.1, 0.1)$. The upper plot frame in each cell is for Weibull shape parameter of 0.90, while the lower plot frame is for shape parameter of 2.0.

α	γ	ξ	η	n	NC	$\hat{\mu}_{Ev}$	$\hat{\mu}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\mu}_{\hat{\beta}_1}$	$\hat{\sigma}_{\hat{\beta}_1}$	$\hat{\mu}_{\hat{\beta}_2}$	$\hat{\sigma}_{\hat{\beta}_2}$	$\hat{\eta}$
0.9	0.9	2	0.67	30	0	2.43	0.911	0.105	1.02	0.516	−1.017	0.301	0.737
0.9	0.9	2	0.67	50	0	2.82	0.907	0.075	1.004	0.36	−1.024	0.225	0.697
0.9	0.9	6	0.86	30	2	3.53	0.905	0.093	1.03	0.415	−1.031	0.252	0.965
0.9	0.9	6	0.86	50	1	2.8	0.911	0.065	1.021	0.289	−1.016	0.183	0.912
0.9	0.9	∞		30	0	3.07	0.882	0.07	1.082	0.301	−1.068	0.196	
0.9	0.9	∞		50	0	2.2	0.897	0.046	1.022	0.216	−1.033	0.146	
0.9	2	2	0.67	30	1	5.43	0.886	0.057	0.808	0.358	−0.808	0.209	0.797
0.9	2	2	0.67	50	0	5.26	0.885	0.038	0.809	0.255	−0.806	0.151	0.777
0.9	2	6	0.86	30	1	5.97	0.879	0.054	0.803	0.252	−0.82	0.177	0.931
0.9	2	6	0.86	50	0	4.66	0.884	0.037	0.819	0.193	−0.812	0.131	0.917
0.9	2	∞		30	0	6.53	0.878	0.042	0.828	0.212	−0.828	0.142	
0.9	2	∞		50	0	6.08	0.883	0.032	0.798	0.154	−0.811	0.103	

Table 5.2: Summary of simulated means and standard deviations of the estimators of α, β, $\eta = \xi/(\xi+1)$ when minimal response is always assumed after each event relapse when the true effective age is a cancer model with probability of complete response 0.3, partial response 0.5, and minimal response 0.2, $p(\psi) = (0.3, 0.5, 0.2)$. The true value of β is $(1, -1)$, and 1000 replications were done for each parameter combination. The others columns are: γ Weibull shape parameter, n sample size, NC number of replicates in which there was no model convergence; $\hat{\mu}_{Ev}$ mean number per patient in all the simulation replications.

positively biased, with larger bias and RMSE in the middle portion of the survivor function.

On the other hand, under effective age mis-specification, the estimators of the finite-dimensional model parameters and the baseline survivor function are highly biased (see Tables 5.2 and 5.3 and Figure 5.3). The parameter that controls event dependence, as well as the parameters associated with the covariates are more biased than when the correct model is used. In particular, this observed bias is highly evident when the baseline hazard function is increasing, that is, when $\gamma = 2$.

Regarding the survivor curves in Figure 5.3 we observe that under a cancer model with probabilities $p(\psi) = (0.8, 0.1, 0.1)$, if we consider a model where always minimal response is achieved, we get an extremely negatively biased estimator of survival function. In contrast, to consider a perfect response after each relapse produces a less biased estimator. Obviously, this happens because the data were generated assuming that most of the interventions after relapses achieve a complete remission ($p_1 = 0.8$), so a model which assumes that always a CR is achieved is closer to this scenario than one that always assumes NR. In the case where $p(\psi) = (0.1, 0.2, 0.7)$, the preceding statement remains valid when we change perfect to minimal. That is, in that case the minimal model is better than the perfect one because the data are generated assuming

α	γ	ξ	η	n	NC	$\hat{\mu}_{Ev}$	$\hat{\mu}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\mu}_{\hat{\beta}_1}$	$\hat{\sigma}_{\hat{\beta}_1}$	$\hat{\mu}_{\hat{\beta}_2}$	$\hat{\sigma}_{\hat{\beta}_2}$	$\hat{\eta}$
0.9	0.9	2	0.67	30	1	3.23	0.876	0.058	1.039	0.49	−1.033	0.283	0.732
0.9	0.9	2	0.67	50	0	3	0.882	0.039	1.012	0.369	−1.031	0.2	0.681
0.9	0.9	6	0.86	30	0	1.6	0.876	0.06	1.037	0.395	−1.042	0.249	0.961
0.9	0.9	6	0.86	50	0	2.62	0.879	0.04	1.047	0.292	−1.043	0.16	0.885
0.9	0.9	∞		30	0	2.87	0.868	0.049	1.08	0.298	−1.07	0.184	
0.9	0.9	∞		50	0	2.88	0.879	0.032	1.038	0.216	−1.049	0.135	
0.9	2	2	0.67	30	0	5.57	0.996	0.025	0.605	0.257	−0.599	0.147	0.917
0.9	2	2	0.67	50	0	5.52	0.992	0.018	0.604	0.189	−0.604	0.119	0.895
0.9	2	6	0.86	30	0	7.3	0.991	0.022	0.598	0.199	−0.609	0.113	0.964
0.9	2	6	0.86	50	1	4.96	0.988	0.014	0.609	0.155	−0.6	0.092	0.935
0.9	2	∞		30	0	7.47	0.984	0.019	0.635	0.166	−0.634	0.098	
0.9	2	∞		50	0	5.24	0.983	0.014	0.62	0.128	−0.618	0.069	

Table 5.3: Summary of simulated means and standard deviations of the estimators of α, β, $\eta = \xi/(\xi+1)$ when minimal response is always assumed after each event relapse when the true effective age is a cancer model with probability of complete response 0.3, partial response 0.5, and minimal response 0.2, $p(\psi) = (0.3, 0.5, 0.2)$. The true value of β is $(1, -1)$, and 1000 replications were done for each parameter combination. The others columns are: γ Weibull shape parameter, n sample size, NC number of replicates in which there was no model convergence; $\hat{\mu}_{Ev}$ mean number per patient in all the simulation replications.

that most of interventions have none response ($p_3 = 0.7$). However, under a non-extreme case, e.g., $p(\psi) = (0.3, 0.5, 0.2)$ using a minimal response as well as perfect response model lead to unacceptable results as the estimators become highly biased.

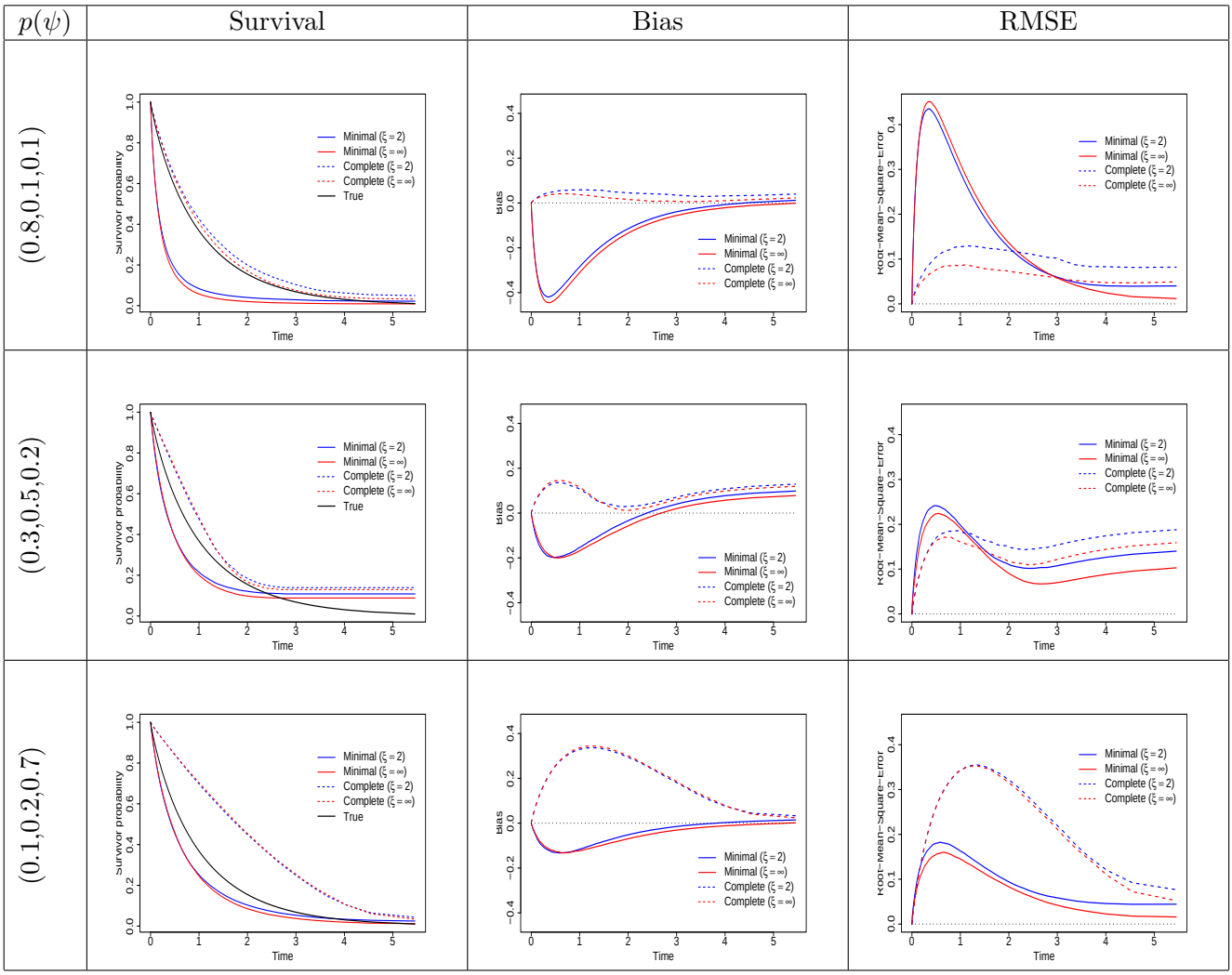

Figure 5.3: Estimated baseline survivor function, bias and root mean squared error curves for the estimator of the baseline survivor function as both the effective age $\mathcal{E}_i(s)$ (always minimal response, $\mathcal{E}_i(s) = s$, and always perfect response, $\mathcal{E}_i(s) = s - S_{N_i^\dagger(s-)}$) and the frailty parameter ξ varies ($\xi = 2$ and $\xi = \infty$). This case corresponds to $\alpha = 0.9$ and $n = 30$.

5.4 The non-Hodgkin's lymphoma study

Herein, we analyze the data belonging to the times to relapse for patients diagnosed with non-Hodgkin's lymphoma (see Section 2.2 for a description). The main importance of this data set is that it has recorded information about the effective age by using the disease status after each reoccurrence. We first examine the effect of assumptions concerning the effective age function. To do so, we fit some simple models that include only the lesion at diagnosis (X_3) as a covariate. We compare the results obtained using the cancer model with those obtained from Peña and Hollander's model assuming always NR or always CR for the effective age. Then, we also compare these results with the AG model including response to treatment as a time-dependent covariate. We denote by β the length-three coefficient associated with X_3 coded as a dummy variable. Figure 5.4 gives the estimated disease-free survival curves for three different effective age specifications (always NR, always CR, and cancer model) for patients with single and with more than 1 site affected. When CR is assumed at each relapse, the survival probability tends to be underestimated for short times and overestimated for longer times, relative to using the cancer model incorporating information about the intervention effect. But when NR is assumed at each relapse, the survival probability tends to be overestimated for short times and underestimated for longer times. Intuitively, the assumption of a constant intervention effect, when in fact it varies, leads to an incorrect time scale in the hazard rate function, thus inducing bias in the estimators.

Regarding the parameter estimates, the three assumed forms of the effective age give rise to differences mainly in the frailty parameter, as shown in Table 5.4. If we use the minimal repair effective age $\mathcal{E}(s) = s$ (always NR), we obtain a small value of frailty precision, $\hat{\xi} = 2.24$ ($\nu = 0.45$), indicating the need to include a frailty component. On the other hand, if we assume $\mathcal{E}(s) = s - S_{N^\dagger(s-)}$ as effective age (always CR) we obtain $\hat{\xi} = 11145048$ ($\nu =$`8.97e-8`), a very large value that indicates that there is no need for the frailty component. Finally, if we use the cancer model formulation $\mathcal{E}(s) = A_{N^\dagger(s-)} + \left(s - S_{N^\dagger(s-)}\right)$ as effective age (different responses can be achieved) the resulting estimates again indicate the importance of the frailty with $\hat{\xi} = 1.36$ ($\nu = 0.73$). We can test the significance of the frailty to verify these statements. A likelihood ratio test for the frailty can be computed as twice the difference between the log-partial-likelihood with the frailty terms integrated out, and the log-likelihood of a model without frailties (Section 4.3. These values for cancer model are -181.46 and -176.27, respectively. That is, the likelihood

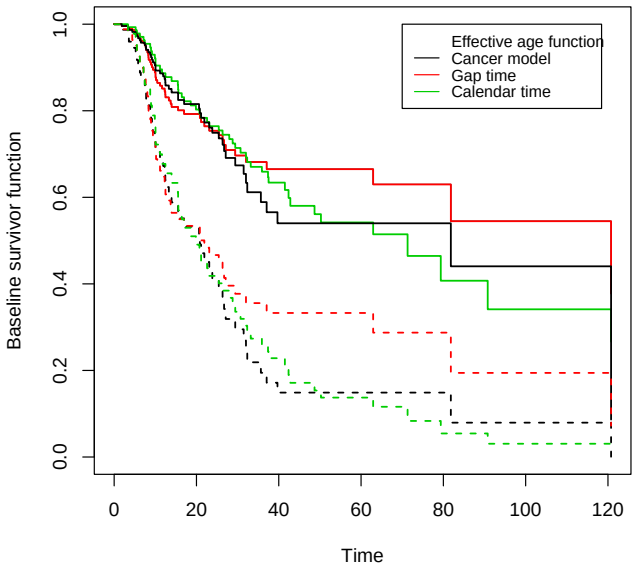

Figure 5.4: Estimates of survivor function (with frailties set to one) for multiple events for non-Hodgkin's lymphoma data set by lesions involved at diagnosis using three different formulation for effective age function.

ratio test can be computed as chi-square statistic of $2(181.46 - 176.27)$ on one degree of freedom leading a p-value $= 0.0013$. The same procedure for the minimal and perfect repair models yields p-values of 0.999 and 0.0201, respectively. These results partly confirm that the need for the frailty term depends on the form of the effective age function.

The hazard ratios (HR) for the risk of relapse associated with X_3 vary little for the three forms of effective age. In all models, patients with localized lesions and with generalized lesions at diagnosis have a higher risk of relapse compared to those with single lesions, showing a similar HR for each model (Table 5.4). This risk is also high for patients with more than one nodal site being statistically significant in all models. However, we observe some differences in their magnitude. Finally, we notice that none of models provide a confidence interval for α (based on approximate normality) that excludes 1.

Now, continuing to use only the lesions at diagnosis, X_3, as covariate, we compare the estimates resulting from the cancer model with other approaches. A simple way to incorporate the response to treatment in the AG model is by considering the disease status after relapses as a time-dependent covariate (see Therneau and Grambsch, 2000 or Therneau and Hamilton, 1997

	Minimal[a]	**Perfect**[b]	**Cancer**[c]
α	0.72 (.37)	0.90 (.15)	0.68 (0.21)
Frailty ξ	2.24	∞	1.36
ν	0.45	8.97×10^{-8}	0.73
Lesions			
Single	1	1	1
Localized	2.48 (0.93-6.60)	2.42 (0.90-6.50)	2.59 (0.88-7.62)
>1 nodal site	4.47 (1.64-12.15)	3.26 (1.20-8.87)	4.55 (1.22-16.93)
Generalized	3.24 (0.70-14.95)	2.69 (0.59-12.31)	3.09 (1.01-9.37)

[a]Effective Age is $\mathcal{E}(s) = s$.
[b]Effective Age is $\mathcal{E}(s) = s - S_{N^\dagger(s-)}$.
[c]Effective Age is $A_{N^\dagger(s-)} + \big(s - S_{N^\dagger(s-)}\big)$.

Table 5.4: Hazard ratios and confidence intervals at 95% (in parenthesis) for the probability of relapse depending on lesions involved at diagnosis for the non-Hodgkin's lymphoma data set. Estimates obtained from the general model using three different effective ages processes.

for further details). After preparing the data and including the treatment response as a dummy variable, the HR for variable lesions at diagnosis are: 2.46 (CI95% 1.06 to 5.77), 3.25 (CI95% 1.25 to 8.47), and 2.77 (CI95% 0.83 to 9.22), respectively. These results are similar to those obtained using the perfect repair model. Considering that the cancer model reveals that a frailty component is important, perhaps the AG model is not adequate since this model assumes that there is no heterogeneity among patients. Finally, we compare our results to those obtained using only time to first relapse in a Cox model. In that case, the HR are: 1.40 (CI95% 0.48 to 4.03), 2.76 (CI95% 0.95 to 8.03), and 3.24 (CI95% 0.28 to 37.50). Here, we ignore the information in the subsequent relapse times and we observe that this fact substantially affects the estimate of the coefficients and their statistical significance, especially in patients with localized lesions.

The heterogeneity of the risk of relapse may be explained by subject-specific factors other than lesions involved at diagnosis, such as gender or delay between first treatment and first symptom. Thus, we now include all three covariates and compare the estimates of the regression coefficients from the cancer model with those obtained using some of the currently-used models (AG, AG with time-dependent covariates, WLW, and shared gamma frailty model). Table 5.5 shows these resulting HRs for the PCMZCL data. After adjusting for gender and delay, the variance of frailty decreased to 0.11 (1/8.85) indicating that the frailty is not necessary (likelihood ratio test $2(180.07 - 179.40) = 1.34$, $p = 0.2476$). Similarly, although the estimate of α differs from 1, it is not statistically significant (based on assumed asymptotic normality), so it seems that the

prior number of event occurrences does not have an impact. The results also indicate that shared frailty model gives lower risk estimates than those obtained using cancer model, while WLW method gives higher risk estimates than cancer model. Only models based on AG approach show similar results to those obtained using cancer model. In that case gender differences is statistically significant only if the AG model is chosen, while cancer model is the only one that indicates an statistically significant increased risk for those patients with generalized lesions as compared to patients with single lesions.

The estimates of the cumulative hazard functions for the multiple event data for cancer model and for AG model with treatment response as a time-dependent covariate are shown in Figure 5.5. The solid lines correspond to hazard of relapse for patients with single lesions at diagnosis, obtained via

$$\hat{\Lambda}_0(s)\exp(\hat{\beta}_2\bar{X}_2) \quad \text{and} \quad \hat{\Lambda}_0(s)\exp(\hat{\beta}_1+\hat{\beta}_2\bar{X}_2)$$

for males and females, respectively, where $\hat{\Lambda}_0(\cdot)$ is the estimated cumulative baseline hazard. The dotted lines in this figure are for patients with generalized lesions which correspond to

$$\hat{\Lambda}_0(s)\exp(\hat{\beta}_2\bar{X}_2+\hat{\beta}_5) \quad \text{and} \quad \hat{\Lambda}_0(s)\exp(\hat{\beta}_1+\hat{\beta}_2\bar{X}_2+\hat{\beta}_5)$$

for males and females, respectively. The observed means are $\bar{X}_2 = 2.4$ for males, and $\bar{X}_2 = 2.7$ for females. These plots indicate that different risks are associated with the number of lesions involved at diagnosis, as is clear from their associated HRs. The AG and cancer model estimates of the hazard rate functions differ more for patients with generalized lesions than with single lesions.

Covariate	AG Frailty	AG2 Frailty	WLW	Shared Frailty	Cancer[a]
α	-	-	-	-	.88 (.40)
Frailty ξ	24.51	∞	-	∞	8.85
ν	0.04	5×10^{-7}	-	5×10^{-7}	0.11
Gender					
Males	1	1	1	1	1
Females	2.01	2.01	1.83	1.73	1.84
	(1.01-4.00)	(1.02-3.92)	(0.81-4.14)	(0.88-3.40)	(0.82-4.10)
delay					
in years	0.99	1.01	1.02	1.04	0.99
	(0.89-1.12)	(0.89-1.12)	(0.88-1.18)	(0.77-1.40)	(0.80-1.23)
Lesions					
Single	1	1	1	1	1
Localized	3.70	3.83	5.23	3.24	3.57
	(1.18-1.16)	(1.23-11.9)	(1.71-15.96)	(1.05-9.96)	(1.17-10.89)
>1 nodal site	4.71	4.77	6.45	3.99	4.67
	(1.62-13.7)	(1.62-14.0)	(2.37-17.56)	(1.41-11.28)	(1.25-17.4)
Generalized	4.75	4.60	23.16	3.44	4.60
	(0.92-24.4)	(0.86-21.5)	(5.02-106.9)	(0.69-16.97)	(1.30-16.3)

[a]Effective Age is $A_{N^\dagger(s-)} + \big(s - S_{N^\dagger(s-)}\big)$.

Table 5.5: Hazard ratios and confidence intervals at 95% (in parenthesis) for the probability of relapse for the PCMZCL data set. Estimates from the Andersen-Gill (AG), and Andersen-Gill with response to treatment after relapse as time-dependent covariate (AG2) including a frailty term, together with the estimates obtained form Wei, Lin and Weissfeld (WLW), Shared Gamma Frailty model, and dynamic *cancer* model.

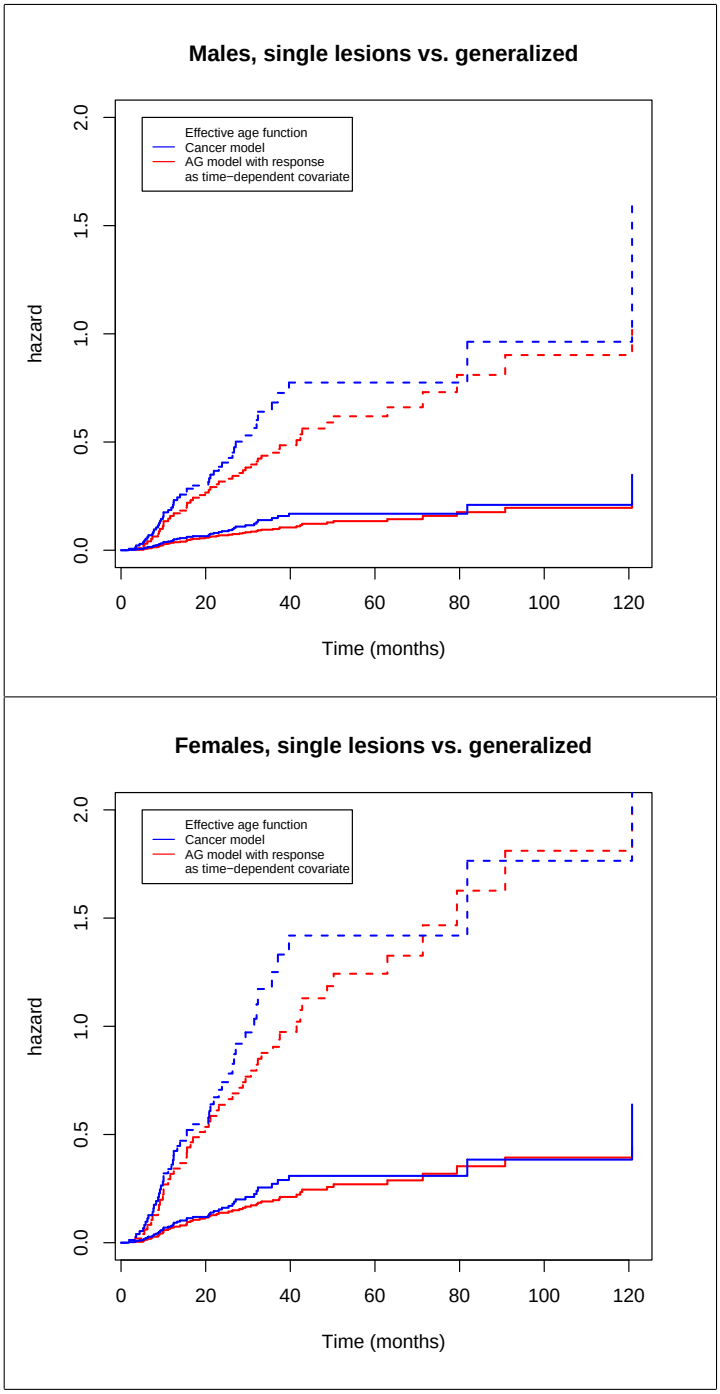

Figure 5.5: Estimates of cumulative hazard function for multiple events for non-Hodgkin's lymphoma data set by sex and lesions involved at diagnosis, all cases evaluated at the mean value of delay between first treatment and first symptom. The blue line shows the hazard assuming the cancer model and red lines correspond to AG model which includes the response to treatment as a time-dependent covariate. Solid lines are patients with single lesions and dotted lines are for patients with generalized lesions at diagnosis.

5.5 R instructions for gcmrec package

Parameter estimates for the dynamic cancer model can be obtained using `gcmrec` function and indicating where it is the information about effective age. In our case `cancer=lymphoma$effage`.

```
> mod.can<-gcmrec(Survr(id,time,event)~as.factor(distrib),
+ data=lymphoma, s=1000, Frailty=TRUE , se="Jacknife", +
cancer=lymphoma$effage)
> mod.can
Call: gcmrec(formula = Survr(id, time, event) ~
as.factor(distrib),
    data = lymphoma, s = 1000, Frailty = TRUE, se = "Jacknife",
    cancer = lymphoma$effage)


                     coef exp(coef) se(coef) Jacknife      z     p
as.factor(distrib)1 0.953      2.59             0.556 1.7146 0.086
as.factor(distrib)2 1.516      4.55             0.667 2.2738 0.023
as.factor(distrib)3 1.129      3.09             0.569 1.9842 0.047

  General class model parameter estimates
    rho function:  Alpha to k
      alpha (s.e. Jacknife): 0.683 (0.214)
    Frailty parameter, Xi (s.e. Jacknife):  1.36

  Marginal log-likelihood= -176.27
  n= 63
  n times= 112
  number of iterations:  74  EM steps
```

Chapter 6

Concluding Remarks and Future Research

6.1 Conclusions

The general class of model proposed by Peña and Hollander (2004) have been demonstrated to be very useful to deal with recurrent event data. In particular, the dynamic cancer model developed in this PhD thesis has demonstrate to be very useful in analyzing indolent diseases in which a relapsing pattern is observed. We have to mention that although some physicians are interested in analyzing this type of data (see MacLaughlin, 2002), these models have received no attention, so far. The model we propose for analyzing cancer data, which includes as special cases many well-known models in survival analysis, is also important because it takes into account the effect of interventions which are performed after each event occurrence through the notion of an effective age, the possible weakening (or strengthening) effect of accumulating event occurrences, the possible presence of unobserved frailties that could be inducing correlations among the inter-event times per unit, and the effect of observable covariates.

Regarding procedures for estimating the parameters of this model, three different approaches have been described. One of them is based on the EM algorithm, while the other two used penalized likelihood inference. Regarding the general behavior of the model proposed, our simulations suggested that an under-specification of the model, in the sense of analyzing a data generated from the model with frailties using procedures developed from the model without frailties, could

have unacceptable consequences in that the resulting estimators will have non-negligible systematic biases. On the other hand, it was found that over-specification of the model may provide a robust method of analysis with an acceptable loss in efficiency. The application of the procedures to the bladder cancer data set highlights the importance of monitoring the effective age process. The simulations for the miss-specification of effective age by not incorporating information about the intervention effect in cancer settings, but instead assuming that always a complete or null response is achieved, has undesirable consequences because the resulting estimators of the finite-dimensional parameters and the baseline survivor function are highly biased. Our application of the cancer model to an indolent lymphoma data set also highlights the need to incorporate information about the effect of intervention after each relapse.

The main advantage of using the general class of models proposed by Peña and Hollander (2004) with respect to analyze recurrent event data using other existing models, is the ability of incorporating information about the performed interventions after reoccurrences. In particular, we have shown that this model may be useful to analyze data set arising form cancer settings, in which the response to the treatment after each relapse is an important factor to predict new relapses. The dynamic cancer model may be used in a variety of applications when information about the response to intervention upon relapses can be obtained. The bladder cancer data set is an example. This data set, however, does not contain information about the effective age function and this leads to be the main limitation of this model. One possibility to solve this drawback is to use simple forms of the effective age, such as perfect or minimal repair formulations as it has been illustrated through this thesis. In this sense, the cancer model tries to model more complicated forms for the effective age function in which the response to the treatment after relapses are included in the model acting in the baseline hazard function.

On the other hand, this PhD thesis has also addressed the problem of how to calculate confidence intervals for median survival time. We have proposed two different methods. One of them is based on asymptotic variances in the case that interoccurrence times are i.i.d. and another used bootstrap techniques. We have also studied several bootstrapping schemes to estimate the sampling distribution of median survival time estimators in the presence of recurrent event data and in consideration of the sum-quota data accrual which induces informative stopping and censoring. We proposed several resampling plans under the i.i.d. model and a correlated interoccurrence times model. From the simulations studies we carried out, we may conclude that

the best bootstrapping scheme to estimate the median survival sample distribution under an i.i.d. model are just bootstrapping from observed data and non-parametric bootstraps (plans I, and II or III, respectively). For a correlated interoccurrence times (under a gamma frailty model), both semiparametric plans (VI and VII) are the best ones. Plan IV, which is anchored in using the WC (1999) estimator of the inter-event survivor function appears to offer a robust procedure when uncertain about the model that generated the data. Based on the simulation studies, it appears that bootstrapping from the empirical distribution of the monitoring times do not provide improvements.

6.2 Future Research

Maybe the main important future work regarding confidence interval for median survival time is to prove asymptotic convergence of bootstrap procedures. There are still many important questions that need to be examined with regards to the general model proposed by Peña and Hollander (2004). The first is the ascertainment of asymptotic properties of the estimators, such as their asymptotic normality or the weak convergence to a Gaussian process of a properly normed estimator of the baseline survivor function. The resolution of this asymptotic problem may require methods utilized in Murphy (1994, 1995) and Parner (1998). Some asymptotic results for the general class of models when baseline hazard function is parametrically specified can be found in Kvam and Peña (2003). Through such asymptotic analysis we will be able to obtain expressions for approximating analytically the standard errors of the estimators and compare them with those obtained using penalized approaches. Another issue of importance is whether bootstrapping methods could be utilized to obtain standard errors of estimators for the purpose of constructing confidence intervals and/or bands for the parameters. The problem of how to validate this class of models after it has been fitted to a specific data set is another open problem, and calls for suitable goodness-of-fit and model validation procedures.

Regarding the dynamic cancer model, new applications for the cancer model may require modifications of our formulation. One important consideration is the value of A_0 in the effective age (see equation (5.2)). Because all our lymphoma patients achieved CR at first treatment, we were able to use this date as study origin, assured that all patients had the same initial status with $A_0 = 0$. In other situations, however, CR may not be achievable at first treatment for all

subjects. In this case, A_0 is not uniformly zero for all subjects, and disease status after first treatment may be assessed to assign a positive value for A_0.

Another aspect of our work that requires consideration in future research is that the time between treatment and assessment of the response to treatment (the ψ_js) following each relapse may not be negligible, contrary to our context and earlier developments in reliability. The delay between application of treatment and evaluation of patient response is widely recognized. However, in cancer studies, at least in hematological diseases, this may not be a problem since all patients are routinely monitored for one month following administration of therapy.

Bibliography

O. Aalen. Heterogeneity in survival analysis. *Stat. Med.*, 7:1121–1137, 1988.

O. Aalen and E. Husebye. Statistical analysis of repeated events forming renewal processes. *Stat. Med.*, 10:1227–1240, 1991.

P. Andersen, O. Borgan, R. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer-Verlag, New York, 1993.

P. Andersen and R. Gill. Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, 10:1100–1120, 1982.

U. Barai and N. Teoh. Multiple statistics for multiple events, with application to repeated infectious in the growth factor studies. *Stat. Med.*, 16:941–949, 1997.

M. Barceló. Modelos marginales y condicionales en el análisis de supervivencia multivariante. *Gac. Sanit.*, 16 (Supl. 2):59–68, 2002.

R. Beran. Estimated sampling distributions: the bootstrap and competitors. *Annals of Statistics*, 10:212–225, 1982.

P.J. Bickel and D.A. Frieedman. Some asymptotic theory for the bootstrap. *Annals of Statistics*, 9:1196–1217, 1981.

H. Block, W. Borges, and T. Savits. Age-dependent minimal repair. *J. Appl. Prob.*, 22:51–57, 1985.

R. Brent. *Algorithms for Minimization without Derivatives*. Englewood Cliffs, NJ: Prentice-Hall, 1973.

N. Breslow. "discussion of professor cox's paper". *J. Royal Statist. Soc., B*, 34:216–217, 1972.

R. Brookmeyer and J.J. Crowley. A confidence interval for the median survial time. *Biometrics*, 38:29–41, 1982.

M. Brown and F. Proschan. Imperfect repair. *J. Appl. Prob.*, 20:851–859, 1983.

D. Burr. A comparison of certain bootstrap confidence intervals in the Cox model. *J. Amer. Statist. Assoc.*, 89:1290–1302, 1994.

D. Burr and H. Doss. Confidence bands for the median survival time as a function of the covariates in the cox model. *Journal of the American Statistical Association*, 88:1328–1340, 1993.

D. Byar. *The veterans administration study of chemoprophylaxis for recurrent stage I bladder tumors: Comparisons of Placebo, Piroxidine, and Topical Thiotepa,* in Bladder Tumors and Other Topics in Urological Oncology, eds. M. Pavone-Macaluso and P. H. Smith and F. Edsmyn. Plenum, New York, 1980.

R. H. Byrd and J. Nocedal. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific and Statistical Computing*, 16:1190–1208, 1995.

BD. Cheson, SJ. Horning, B. Coiffier, MA. Shipp, RI. Fisher, JM. Connors, TA. Lister, J. Vose, A. Grillo-Lopez, A. Hagenbeek, F. Cabanillas, D. Klippensten, W. Hiddemann, R. Castellino, NL. Harris, JO. Armitage, W. Carter, R. Hoppe, and GP. Canellos. Report of an international workshop to standardize response criteria for non-hodgkin's lymphomas. NCI sponsored international working group. *J Clin Oncol*, 17:1244–1253, 1999.

B. Cheuvart. A nonparametric model for multiple recurrences. *Appl Statist*, 37:157–168, 1988.

D. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65:141–151, 1978.

D. Clayton. Some approaches to the analysis of recurrent event data. *Stat. Meth. Med. Res.*, 3: 244–262, 1994.

D. Clayton and J. Cuzick. Multivariate generalizations of the proportional hazards model (with discussion). *J. Roy. Statist. Soc., A*, 148, 1985.

R. Cook and J. Lawless. Marginal analysis of recurrent events and a terminating event. *Stat. Med.*, 16:911–924, 1997.

D. Cox. Regression models and life tables (with discussion). *J. Royal Statist. Soc., B*, 34:187–220, 1972.

D.M. Dabrowska and K.A. Doksum. Estimates and confidence intervals for median and mean life in the proportional hazard model. *Biometrika*, 74:799–807, 1987.

M. Davidge-Pitts, R. Dansey, and W. Bezwoda. Prolonged survival in follicular non hodgkin's lymphoma is predicted by achievement of complete remission with initial treatment: results for a long term study with multivariate analysis of prognostic factors. *Leuk Lymphoma*, 24: 131–140, 1996.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood estimation from incomplete data via the em algorithm (with discussion). *J. Roy. Statist. Soc., B*, 39:1–38, 1977.

C. Dorado, M. Hollander, and J. Sethuraman. Nonparametric estimation for a general repair model. *Ann. Statist.*, 25:1140–1160, 1997.

B. Efron. The jacknife, the bootstrap, and other resampling plans. *SIAM-NSF, CMBS No. 38, Philadelphia*, 1982.

B. Efron. Better bootstrap confidence intervals. *Technical report No. 226, Dept of Statistics, Stanford University*, 1985a.

B. Efron. Bootstap confidence intervals for parametric problems. *Biometrika*, 72:45–58, 1985b.

B. Efron and R. Tibshirani. *An introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

T. Fleming and D. Harrington. *Counting Processes and Survival Analysis*. Wiley, New York, 1991.

I. Fredriksson, G. Liljegren, L. Arnesson, S. Emdin, M. Palm-Sjovall, T. Fornader, M. Holmqvist, and L. Holmberg. Local recurrence in the breast after conservative surgery - a study of prognosis and prognostic factors in 391 women. *Eur J Cancer*, 38:1860–1870, 2002.

M. Gail, T. Santner, and C. Brown. An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics*, 36:255–266, 1980.

S. Gao and X-H. Zhou. An empirical comparison of tow semi-parametric approaches for the estimation of covariate effects from multivariate failure time data. *Stat. Med.*, 16:2049–2062, 1997.

R. Gill. Testing with replacement and the product-limit estimator. *Ann. Statist.*, 9:853–860, 1981.

R. Gill. Discussion of the paper by d. clayton and j. cuzick. *J. Roy. Statist. Soc., A*, 148:108–109, 1985.

K. Godder, M. Eapen, JH. Laver, MJ. Zhang, BM.ç Camitta, AS. Wayne, RP. Gale, JJ. Doyle, LC. Yu, AR. Chen, JH Jr. Garvin, ES. Sandler, AM. Yeager, JR. Edwards, and MM Horowitz. Autologous hematopoietic stem-cell transplantation for children with acute myeloid leukemia in first or second complete remission: a prognostic factor analysis. *J Clin Oncol*, 22:3798–3804, 2004.

J. Gonzalez, Fernandez E. Moreno V., Ribes J. Peris M. Navarro M. Cambray M. and Borras J. Gender differences in hospital readmission among colorectal cancer patients. *J Epidemiol Community Health*, 59:506–511, 2005.

J. González and E. Peña. Estimación no paramétrica de la función de supervivencia para datos con eventos recurrentes. *Rev. Española de Salud Pública*, 78:211–220, 2004.

JR. González, E. Fernandez, V. Moreno, J. Ribes, M. Peris, M. Navarro, M. Cambray, and JM. Borras. Gender differences in hospital readmission among colorectal cancer patients. *J Epidemiol Comm Health*, 59:506–511, 2005.

I. Good and R. Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58:255–277, 1971.

R. Gray. Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *JASA*, 82:605–610, 1987.

G. Guo and G. Rodriguez. Estimating a multivariate proportional hazards model for clustered data using em algorithm, with an application to child survival in guatemala. *J. Amer. Statist. Assoc.*, 87:969–976, 1992.

N. Hjort. Bootstrapping Cox's regression model. Technical Report NSF-241, Department of Statistics, Stanford University, 1985.

M. Hollander, B. Presnell, and J. Sethuraman. Nonparametric methods for imperfect repair models. *Ann. Statist.*, 20:879–896, 1992.

M. Hollander and J. Sethuraman. Nonparametric inference for repair models. *Sankya, Series A*, 64:693–706, 2002.

P. Hougaard. Life table methods for heterogeneous populations: distributions describing the heterogeneity. *Biometrika*, 73:387–396, 1984.

P. Hougaard. A class of multivariate failure time distributions. *Biometrika*, 73:671–678, 1986a.

P. Hougaard. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73:387–396, 1986b.

P. Hougaard. Modelling multivariate survival. *Scand. J. Statist.*, 14:291–304, 1987.

P. Hougaard. Frailty models for survival data. *Lifetime Data Analysis*, 13:2427–2436, 1994.

P. Hougaard. *Analysis of Multivariate Survival Data.* Springer, New York, 2000.

E. Husebye, V. Skar, O. Aalen, and M. Osnes. Digital ambulatory manometry of the small intestine in healthy adults: estimation of the variation within and between individuals and statistical management of aborted mmc periods. *Digestive Diseases and Sciences*, 35:1057–1065, 1990.

R. Ihaka and R. Gentleman. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314, 1996.

J. Jacod. Multivariate point processes: Predictable projection, radon-nikodym derivatives, representation of martingales. *Z. Wahrsch. verw. Geb*, 31:235–253, 1975.

D. Joly, A. Letenneur, D. Alioum, and D. Commenges. Phmpl: a computer program for hazard estimation using a penalized likelihood method with interval-censored and left-truncated data. *Computer Methods and Programs in Biomedicine*, 60:225–231, 1999.

J D Kalbfleisch and R L Prentice. *The statistical analysis of failure time data*. Wiley, New York, 1980.

A. Kaplan and P. Meier. Non-parametric estimation for incomplete estimations. *J. Amer. Statist. Assoc.*, 53:457–481, 1958.

P. Kelly. A review of software packages for analyzing correlated survival data. *Am Stat*, 58: 337–342, 2004.

P. Kelly and L. Lim. Survival analysis for recurrent event data: an application to childhood infectious diseases. *Stat. Med.*, 19:13–33, 2000.

M. Kijima. Some results for repairable systems with general repair. *J. Appl. Prob.*, 26:89–102, 1989.

M. Kijima, H. Morimura, and Y. Suzuki. Periodical replacement problem without assuming minimal repair. *Europ. J. Operat. Res.*, 37:194–203, 1988.

J. Klein. Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, 48:795–806, 1992.

P. Kvam and E. Peña. Estimating load-sharing properties in a dynamic reliability system. *Currently under revision for JASA*, 2003.

G. Last and R. Szekli. Asymptotic and monotonicity properties of some repairable systems. *Adv. in Appl. Probab.*, 30:1089–1110, 1998.

J. Lawless. Regression methods for poisson process data. *J. Amer. Statist. Assoc.*, 82:808–815, 1987.

E. Lee, L. Wei, and D. Amato. *Cox-type regression analysis for large numbers of small groups of correlated failure time observations*, pages 237–247. Kluwer, Dordrecht, 1992.

K. Levenberg. A method for the solution of certain problems in least squares. *Quart. Appl. Math.*, 2:164–168, 1944.

J. Librero, S. Peiro, and R. Ordiñana. Chronic comorbidity and outcome of hospital care: lenght of stay, mortality, and readmission at 30 and 365 days. *J Clin Epidemiol*, 52:171–179, 1999.

D. Lin. Cox regression analysis of multivariate failure time data: the marginal approach. *Stat. Med.*, 13:2233–2247, 1994.

T. Lister. The management of follicular lymphoma. *Ann Oncol*, 2 (Suppl 2):131–135, 1991.

M. Lombardo, F. Morabito, F. Merli, S. Molica, L. Cavanna, S. Sacchi, C. Broglia, F. Angrilli, F. Ilariucci, C. Stelitano, D. Luisi, R. Berte, S. Luminari, M. Federico, and M Brugiatelli. Bleomycin, epidoxorubicin, cyclophosphamide, vincristine and prednisone (bacop) in patients with follicular non-hodgkin's lymphoma: results of a retrospective, multicenter study of the gruppo italiano per lo studio dei linfomi (GISL). *Leuk Lymphoma*, 43:1795–1801, 2002.

A. Lopez-Guillermo, F. Cabanillas, P. McLaughlin, T. Smith, F. Hagemeister, MA. Rodriguez, JE. Romaguera, A. Younes, AH. Sarris, HA. Preti, W. Pugh, and MS. Lee. Molecular response assessed by PCR is the most important factor predicting failure-free survival in indolent follicular lymphoma: update the MDACC series. *Ann Oncol*, 11:137–140, 2000.

T. Lumley and T. Therneau. *The survival Package*. The Comprehensive R Archive Network, http://cran.r-project.org, 2003.

P. MacLaughlin. Editorial on 'Survival after progression in patients with follicular lymphoma: analysis of prognostic factors'. *Ann. Oncol.*, 13:499–500, 2002.

D. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal of Applied Mathematics*, 11:431–441, 1963.

M.A. Martin. On bootstrap iteration for coverage correction in confidence intervals. *Journal of the American Statistical Association*, 85:1105–1118, 1990.

MathSoft. *S-PLUS 4 Guide to Statistics*. Data Analysis Products Division, Seattle, 1997.

S. McClean and C. Devine. A nonparametric maximum likelihood estimator for incomplete renewal data. *Biometrika*, 82(4):791–803, 1995.

C. McGilchrist. Reml estimation for survival models with frailty. *Biometrics*, 49:221–225, 1993.

C. McGilchrist and C. Aisbett. Regression with frailty in survival analysis. *Biometrics*, 47: 461–466, 1991.

S. Montoto, A. Lopez-Guillermo, A. Ferrer, M. Camos, A. Alvarez-Larran, F. Bosch, and et al. Survival after progression in patients with follicular lymphoma: analysis of prognostic factors. *Ann. Oncol.*, 13:523–530, 2002.

S. Murphy. Consistency in a proportional hazards model incorporating a random effect. *The Annals of Statistics*, 22:712–731, 1994.

S. Murphy. Asymptotic theory for the frailty model. *Ann. Statist.*, 23(1):182–198, 1995.

G. Nielsen, R. Gill, P. Andersen, and T. Sorensen. A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statist.*, 19:25–43, 1992.

D. Oakes. A model for association in bivariate *J. Roy. Statist. Soc., B*, 44:414–422, 1982.

D. Oakes. *Frailty models for multiple event times*, chapter In *Survival Analysis: State of the Art*, pages 371–379. Kluwer, Dordrecht, (J. P. Klein and P. K. Goel, (eds.) edition, 1991.

F. O'Sullivan. Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal of Applied Mathematics*, 9:363–379, 1988.

E. Parner. Asymptotic theory for the correlated gamma-frailty model. *The Annals of Statistics*, 26:183–214, 1998.

E. Peña and M. Hollander. Models for recurrent events in reliability and survival analysis. In *Mathematical Reliability* eds. T. Mazzuchi and N. Singpurwalla and R. Soyer, pages 105–123, 2004.

E. Peña, E. Slate, and J. González. Semiparametric inference for a general class of models for recurrent events. Technical Report 214, Department of Statistics, University of South Carolina, October 2003.

E. Peña, R. Strawderman, and M. Hollander. Nonparametric estimation with recurrent event data. *J. Amer. Statist. Assoc.*, 96(456):1299–1315, December 2001.

R. Prentice, B. Williams, and A. Petersen. On the regression analysis of multivariate failure time data. *Biometrika*, 68:373–379, 1981.

B. Presnell, M. Hollander, and J. Sethuraman. Testing the minimal repair assumption in an imperfect repair model. *J. Amer. Statist. Assoc.*, 89:289–297, 1994.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

J. Ramsay. Monotone regression splines in action. *Statistical Science*, 3:425–461, 1988.

S. Ripatti and J. Palmgren. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56:1016–1022, 2000.

JE. Romaguera, P. McLaughlin, L. North, D. Dixon, KB. Silvermintz, LA. Garnsey, WS. Velasquez, FB. Hagemeister, and F. Cabanillas. Multivariate analysis of prognostic factors in stage IV follicular low-grade lymphoma: a risk model. *J Clin Oncol*, 9:762–769, 1991.

V. Rondeau, D. Commenges, and P. Joly. Maximum penalized likelihood estimation in frailty models. *Lifetime Data Analysis*, 9:139–153, 2003.

V. Rondeau and J. R. Gonzalez. frailtypack: A computer program for the analysis of correlated failure time data using penalized likelihood estimation. *Comput Methods Programs Biomed*, 2005.

O. Schwandner, T. Schiedeck, H. Bruch, M. Duchrow, U. Windhoevel, and R. Broll. p53 and Bcl-2 as significant predictors of recurrence and survival in rectal cancer. *Eur J Cancer*, 94: 2813–2820, 2000.

S. Self and K. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *JASA*, 82:605–610, 1987.

T. Sellke. *Weak convergence of the Aalen estimator for a censored renewal process*, volume 2, pages 183–194. 1988.

O. Servitje, F. Gallardo, T. Estrach, RM. Pujol, A. Blanco, A. Fernandez-Sevilla, L. Petriz, J. Peyri, and V. Romagosa. Primary cutaneous marginal zone B-cell lymphoma: a clinical, histopathological, immunophenotypic and molecular genetic study of 22 cases. *Br J Dermatol*, 147:1147–1158, 2002.

K. Singh. On the accuracy of efron's bootstrap. *The Annals of Statistics*, 9:1187–1195, 1981.

K. Singh. Theoretical comparison of bootstrap confidence intervals (with discussion). *The Annals of Statistics*, 16:927–985, 1988.

G. Soon and M. Woodroofe. Nonparametric estimation and consistency for renewal processes. *Journal of Statistical Planning and Inference*, 53:171–195, 1996.

J. Spinolo, F. Cabanillas, D. Dixon, S. Khorana, P. McLaughlin, W. Velasquez, and et al. Therapy of relapsed or refractory low-grade follicular lymphomas: factors associated with complete remission, survival and time to treatment failure. *Ann. Oncol.*, 3:227–232, 1992.

T. Therneau and P. Grambsch. *Modeling Survival Data: Extending the Cox Model.* Springer, New York, 2000.

T. Therneau, P. Grambsch, and Pankratz V. Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, 1:156–175, 2003.

T. Therneau and S. Hamilton. rhDNase as an example of recurrent event analysis. *Stat. Med.*, 16:2029–2047, 1997.

D.R. Thomas and G.L. Grunkemeier. Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, 70:865–871, 1975.

Y. Vardi. Nonparametric estimation in the presence of length bias. *Ann. Statist.*, 10:616–620, 1982a.

Y. Vardi. Nonparametric estimation in renewal processes. *Ann. Statist.*, 10:772–785, 1982b.

J. Vaupel, K. Manton, and E. Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16:439–454, 1979.

J. Vaupel and A. Yashin. *The deviant dynamics of death in heterogeneous populations*, chapter In Tuma, N.B., editor,*Social Methodology*, pages 179–211. Josey-Bass, London, 1985a.

J. Vaupel and A. Yashin. Heterogeneity's ruses: Some surprising effects of selection on population dynamics. *Amer. Stat.*, 39:176–185, 1985b.

J. Verwerj and H. Houwelingen. Penalized likelihood in cox regression. *Stat. Med.*, 13:2427–2436, 1994.

M. C. Wang and S. H. Chang. Nonparametric estimation of a recurrent survival function. *J. Amer. Statist. Assoc.*, 94:146–153, 1999.

L. Wei and D. Glidden. An overview of statistical methods for multiple failure time data in clinical trials. *Stat. Med.*, 16:833–839, 1997.

L. Wei, D. Lin, and L. Weissfeld. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Amer. Statist. Assoc.*, 84:1065–1073, 1989.

D. Weisdorf, J. Andersen, J. Glick, and M Oken. Survival after relapse of low-grade non-hodgkin's lymphoma: implications for marrow transplantation. *J Clin Oncol*, 10:942–947, 1992.

D. Zucker and A. Karr. Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *Ann. Statist.*, 18:329–353, 1990.

Appendix A

Counting Processes in Survival Analysis

Survival analysis arises when we are interested in studying statistical properties of the variable T, which describes the time to a single event. This type of analysis occurs commonly in two areas. In medical research it is known as survival analysis and refers often to the time from the beginning of the treatment to the occurrence of a particular condition or death. In engineering it is concerned with reliability and the analysis of failure times. That is, how long a component can be used until it fails.

We can use some functions to describe T. Let $f(t)$ be the probability density function of the failure time. The survivor function, $\bar{F}(t)$ (also called $S(t)$), which is the probability of surviving to at least until time t, is given by

$$\bar{F}(t) = \int_t^\infty f(\tau)d\tau = 1 - F(t), \tag{A.1}$$

where $F(t)$ is the cumulative distribution function. The hazard function, $\lambda(t)$, is the instantaneous probability that the event occurs a time t given that the individual survived up to time t, and is given by

$$\lambda(t) = \frac{f(t)}{\bar{F}(t)} \tag{A.2}$$

Finally, the cumulative hazard rate is defined as

$$\Lambda(t) = \int_0^t \lambda(\tau)d\tau, \tag{A.3}$$

hence $\bar{F}(t) = \exp(-\Lambda(t))$.

If we are interested in estimating the previous functions we have to take into account an important aspect which makes survival analysis different from traditional statistical analysis. Let us assume that a researcher is studying the effectiveness of a new treatment for a generally terminal disease. The major variable of interest is the number of days that patients survive. In principle, one could use the standard parametric and nonparametric statistics to describe the average survival, and to compare the new treatment to traditional methods. However, at the end of the study there will be patients who survived over the entire study period, in particular those patients who entered the study in its final stage. Surely, one would not want to exclude all those patients from the study by declaring them to be missing data (since most of them are "survivors" and, therefore, they reflect the success of the new treatment method). Those observations, which contain only partial information are called censored observations (e.g., "patient A survived at least 5 months before he moved away and we lost touch"). Thus, the presence of censored data leads to complications in the analysis.

To denote that times can be censored (e.g., not observed completely) we use the following notation. We assume that for n individuals in a sample, we have $Y_1, ..., Y_n$ independent, identically distributed nonnegative random variables ("lifetimes") with common continuous distribution F. Suppose $C_1, ..., C_n$ are independent, identically distributed nonnegative random variables ("censoring sequence") with a common distribution function G. Assume also that both sets of variables are independent. Thus, in the setting of survival analysis data with random right censorship, we observe $(T_1, \delta_1), ..., (T_n, \delta_n)$, where

$$T_i = min\{Y_i, C_i\},$$
$$\delta_i = I\{Y_i \leq C_i\}.$$

This notation is known as the "traditional" description of time to event data. However, counting process notation has been established as the theoretical basis for analyzing survival data. In the next two sections we will outline this alternative notation and the procedures to estimate $\bar{F}(t)$ as well as $\Lambda(t)$ using it.

A.1 Counting processes approach

Counting processes formulation and martingale theory play a fundamental role in modern theory of survival analysis. Andersen et al. (1993) is an excellent book to understand this approach. However, it is necessary for the reader to assimilate some mathematical concepts that can be difficult for applied statisticians. Fortunately, Therneau and Grambsch (2000, Chapter 1) gives a very intuitive explanation to counting processes and martingales using practical examples. The martingale processes provide direct ways of studying large sample properties of estimators and significance tests for right censored failure time data, and provide tools for analyzing event history data more complicated than censored data as we illustrate in the next chapter. Next sections give a short overview of the mathematical details of counting processes from an applied statisticians point of view.

We begin by giving some notation. We define $f(t-)$ as shorthand for $lim_{\delta \downarrow 0} f(t - \delta t)$. $I\{\}$ denotes the indicator function so $I\{A\} = 1$ if A is true and $I\{A\} = 0$ otherwise. $\mathcal{P}\{\ldots\}$ is probability, $\mathbf{E}\{\ldots\}$ is expectation, and $\mathbf{V}\{\ldots\}$ denotes variance.

There are three key variables which are functions of time in the counting process approach: the *counting process* variable, the *risk indicator*, and the *intensity process*. The random variable $N(t)$ represents a *counting process* on $[0, \infty)$ if

1. $N(t)$ is a non-negative integer
2. $N(s) \leq N(t)$ for $s < t$
3. $dN(t) = N(t) - N(t-)$ is either 0 or 1
4. $\mathbf{E}\{N(t)\} < \infty$

The *risk indicator* is a dummy variable indicating whether observation i is "at risk" for the event of interest at time t,

$$Y_i(t) = I\{T_i \geq t\}$$

To define the intensity, first we need to introduce some nomenclature. The *filtration* $\mathcal{F}_t$ (often called *history*) of a counting process is all that is known at time t. In particular, the history includes the values of random variables known up to and including the time t. $\mathcal{F}_{t-}$ represents what is known up to but not including time t, i.e.: vital status, age, treatment received, hemoglobin

level, blood preasure, etc. Thus, a filtration is defined as an increasing family of σ-algebras defined in the sample space. Using data description mentioned in previous section we take $\mathcal{F}_t$ to mean the values of T_i and δ_i for all i such that $T_i \leq t$, otherwise just the filtration that $T_i > t$. For $\mathcal{F}_{t-}$ we have to change $\leq$ by $<$ and $>$ by $\geq$. A formal definition of filtration notion can be found in Andersen et al. (1993, Section II.2).

The probability (conditional on the filtration) of $dN(t) = 1$ at any time can be written in terms of an *intensity* $\alpha(t)$:

$$\mathcal{P}\{N(t+dt) - N(t-) = 1 \mid \mathcal{F}_{t-}\} \simeq \alpha(t)dt.$$

or, equivalently:

$$\mathcal{P}\{dN(t) = 1 \mid \mathcal{F}_{t-}\} = dA(t)$$

where

$$A(t) = \int_0^t \alpha(w)dw$$

is the *integrated intensity*. We notice that $A(t)$ is required to be *predictable* with respect to $\mathcal{F}_t$. That is, $A(t)$ is known given $\mathcal{F}_t$. In practice, this means that $A(t)$ has to be continuous. Finally, we say that the processes N is *adapted* (to the filtration) if $N(t)$ is $\mathcal{F}_{\sqcup}$ measurable for each t. The process N is called *cadlag* if its sample paths $(N(t) : t \in \mathcal{F})$, are right-continuous with left-hand limits.

Now we can show some elementary martingale theory. First of all we give the definition of a martingale. A *martingale* is a cadlag adapted process M which is integrable, i.e.,

$$\mathbf{E}\{\mid M(t) \mid\} < \infty \text{ for all } t \in$$

and satisfies the martingale property:

$$\mathbf{E}\{M(t) \mid \mathcal{F}_s\} = M(s) \text{ for all } s \leq t.$$

The process is a *submartingale* replacing previous equation by the inequality

$$\mathbf{E}\{M(t) \mid \mathcal{F}_s\} \geq M(s) \text{ for all } s \leq t. \tag{A.4}$$

When we have (A.4) with the inequality reversed, M is called a *supermartingale*. A martingale is called *square integrable* if

$$\sup_{t \in \mathcal{T}} \mathbf{E}\{M(t)^2\} < \infty,$$

where $\mathcal{T} = [0, \tau)$ where τ may be finite or infinite.

Thus, a martingale is a process without drift. Conditional on its past, the best prediction of any future value is its current value.

We also can define $M(t)$, a counting processes *martingale*, as

$$M(t) = N(t) - A(t)$$

and its expectation becomes

$$\mathbf{E}\{dM(t) \mid \mathcal{F}_{t-}\} = 0$$

Equivalently, for any $0 \leq s < t$,

$$\mathbf{E}\{M(t) \mid \mathcal{F}_s\} = M(s)$$

which implies,

$$\mathbf{E}\{M(t) \mid M(u); 0 \leq u \leq s\} = M(s).$$

Finally, the counting process $N(t)$ can be written using the *Doob-Meyer decomposition* theorem, as a unique sum of a predictable, right continuous process, called *compensator*, and a martingale:

$$N(t) = A(t) + M(t).$$

As it is mentioned in Therneau and Grambsch (2000), *"The decomposition: counting process = compensator + martingale is analogous to the statistical decomposition: data = model + noise or, more to the point since we are dealing with counts, observed count = expected count + error".*

A.2 Nonparametric methods

In survival analysis the counting process formulation replaces the pair of variables (T_i, δ_i) with the pair of counting processes $(N_i(t), Y_i(t))$, where $N_i(t)$ represents whether or not the event has happened by or at t for unit i, and $Y_i(t)$ is an indicator for being at risk:

$$N_i(t) = I\{T_i \leq t, \delta_i = 1\},$$
$$Y_i(t) = I\{T_i \geq t\}.$$

Note that $Y_i(t) = 1 - N_i(t-)$ for an uncensored individual and that $Y_i(t)$ is an example of a *predictable process*, since its value at time t is known infinitesimally before t (in other words, at time $t-$).

Now, we only need to define the third ingredient of counting process approach: the intensity. In survival analysis, the intensity, $\alpha(t)$, is equal to the hazard function, $\lambda(t)$, when the individual is at risk of the event and equal to zero when the event has happened. We express this by writing the intensity as

$$Y(t)\lambda(t)$$

A major focus of survival methods is the hazard function. It turns out to be much easier to estimate the cumulative or integrated hazard (A.3), than the hazard function (A.2). For the no-covariate case, the most common estimate of $\Lambda(t)$ is the Nelson-Aalen estimate. This estimator is based on the aggregated processes $Y_+ = \sum_{i=1}^n Y_i(t)$, $N_+ = \sum_{i=1}^n N_i(t)$, $M_+ = \sum_{i=1}^n M_i(t)$, and $\Lambda_+(t)$. Thus, $N_+(t)$ is a counting process, $A_+(t)$ its compensator and $M_+(t)$ its martingale:

$$N_+(t) = A_+(t) + M_+(t). \tag{A.5}$$

The compensator can be written in terms of the individual hazards, $\lambda_i(t)$, as

$$A_+(t) = \int_0^t \sum_{i=1}^n Y_i(u)\lambda_i(u)du. \tag{A.6}$$

However, if all individuals are exposed to the same hazard then the expression for the overall compensator reduces to

$$A_+(t) = \int_0^t Y_+(u)\lambda(u)du = \int_0^t Y_+(u)d\Lambda(u). \tag{A.7}$$

The idea of estimation of the integrated hazard is as follows. The decomposition (A.5) expressed in differentials and the formula (A.7) becomes

$$dN_+(t) = Y_+(t)d\Lambda(t) + dM_+(t),$$

where $N_+(t)$ and $Y_+(t)$ are the data. Conditional on the filtration $\mathcal{F}_{t-}$, $dM_+(t)$ has zero expectation. So an estimate of Λ can be obtained by setting $dM_+(t)$ equal to zero. Using previous formula we have that

$$d\hat{\Lambda}(t) = \frac{dN_+(t)}{Y_+(t)}. \tag{A.8}$$

Hence, the estimated integrated hazard from (A.8) is

$$\hat{\Lambda}(t) = \int_0^t \frac{dN_+(u)}{Y_+(u)}.$$

An equivalent representation of the Nelson-Aalen estimate is the sum

$$\hat{\Lambda}(t) = \sum_{j:t_j \leq t} \frac{dN_+(t_j)}{Y_+(t_j)}. \tag{A.9}$$

Now, let the n distinct, uncensored, event times from a set of n individuals be $a_1, \ldots, a_j, \ldots, a_n$ with $a_{j-1} < a_j$. Thus, we notice that the numerator of the integrand is zero unless $t = a_j$ for some j and the denominator is the number in the risk set at a_j, conventionally written as r_j. Using this notation, and if there are no ties, the Nelson-Aalen estimator is therefore:

$$\hat{\Lambda}(t) = \sum_{j:a_j \leq t} \frac{1}{r_j}.$$

If we have censored data between two failures a_{j-1} and a_j then the individual is counted in the risk sets up to and including the set at a_{j-1} but not in any subsequent ones. On the other hand, if an individual is censored at a failure time a_j then that individual is included in the risk set for a_j but not in any later ones.

Now, we can estimate the survival function, $S(t) = \exp[-\Lambda(t)]$ using two methods. The first one was proposed by Kaplan and Meier (1958). The second one was proposed by Breslow (1972) who suggested that $\hat{S}_B(t) = \exp[-\hat{\Lambda}(t)]$. Both estimators can be written as

$$\hat{S}_{KM}(t) = \prod_{j:a_j \leq t} [1 - d\hat{\Lambda}(a_j)], \tag{A.10}$$

and

$$\hat{S}_B(t) = \prod_{j:a_j \leq t} e^{-d\hat{\Lambda}(a_j)}, \tag{A.11}$$

respectively.

A.3 Cox proportional hazards model

The Cox proportional hazards model Cox (1972) has become the most used method for modelling the relationship of covariates to a survival data. Let $X_{ij}(t)$ be the jth covariate of the ith person. We use X_i to denote the covariate vector for subject i, e.g. the ith row of the matrix. We notice that we use the notation $X_i(t)$ to emphasize that the covariates can be time-dependent ones. Using this notation the Cox model relates the hazard function $\lambda_i(t)$ with the ith individual to the vector of explanatory covariates $X_i(t)$ by

$$\lambda_i(t \mid X_i) = \lambda_0(t) \exp\{\beta' X_i(t)\}, \tag{A.12}$$

where "′" denotes vector transpose, β is a vector of parameters, and λ_0 is the baseline hazard function. The essence of proportional hazards modelling is that β can be estimated without needing to estimate $\lambda_0(t)$ using the partial likelihood function introduced by Cox (1972). For untied failure time data we have

$$PL(\beta) = \prod_{i=1}^{n} \prod_{t \geq 0} \left(\frac{Y_i(t) \exp\{\beta' X_i(t)\}}{\sum_{j=1}^{n} Y_j(t) \exp\{\beta' X_j(t)\}} \right)^{dN_i(t)}. \tag{A.13}$$

The log partial likelihood is given by

$$l(\beta) = \log PL(\beta) = \sum_{i=1}^{n} \int_0^{\infty} \left[\beta' X_i(t) - \log \left(\sum_{j=1}^{n} Y_j(t) \exp\{\beta' X_j(t)\} \right) \right] dN_i(t). \tag{A.14}$$

The score for β, $U(\beta) = \frac{\partial}{\partial \beta} \log PL(\beta)$, is

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\infty} \left[X_i(t) - \frac{\sum_{j=1}^{n} Y_j(t) X_j(t) \exp\{\beta' X_j(t)\}}{\sum_{j=1}^{n} Y_j(t) \exp\{\beta' X_j(t)\}} \right] dN_i(t). \tag{A.15}$$

Maximum partial likelihood estimates β are found by solving the p simultaneous equations $U(\beta) = 0$, where p is the number of covariates. We notice that when data contain tied observation times the partial likelihood need to be changed (see Fleming and Harrington, 1991, Chapter 4, for further details).

There are occasions when an estimate of baseline cumulative hazard function $\Lambda_0(t) = \int_0^t \lambda_0(u) du$ is necessary. We can use a method very similar to that used to derive the Nelson-Aalen estimator. The compensator (A.6) in the proportional hazards formulation becomes

$$A_+(t) = \int_0^t \sum_{i=1}^{n} Y_i(u) \lambda_0(u) \exp\{\beta' X_i(u)\} du.$$

As in the derivation of the Nelson-Aalen estimator, we estimate $dA_+(t)$ by $dN_+(t)$, so we have

$$d\hat{A}_+(t) = \sum_{i=1}^{n} Y_i(u) \exp\{\hat{\beta}' X_i(u)\} d\Lambda(t) = dN_+(t),$$

where we have replaced β by its estimate, $\hat{\beta}$, obtained maximizing (A.14). Finally, rearranging previous equality and integrating, we obtain

$$\hat{\Lambda}(t) = \int_0^t \frac{dN_+(t)}{\sum_{i=1}^{n} Y_i(u) \exp\{\hat{\beta}' X_i(u)\}} \tag{A.16}$$

called Aalen-Breslow estimator.

A.3.1 Stratified Cox model

An extension of the Cox model allows for multiple strata. This model assumes that the strata divide the subjects in disjoint groups, each of which has a distinct baseline hazard function but common values for β. The hazard for an individual i, who belongs to stratum k is

$$\lambda_k(t) \exp\{\beta' X_i(t)\}$$

This model is very useful when we analyze multicenter studies because patients from different centers tend to have populations, with different patterns. The overall log-likelihood becomes the sum

$$\sum_{k=1}^{K} l_k(\beta)$$

where $l_k(\beta)$ is the equation (A.14), but we only sum the subjects in stratum k. This model is also useful when we deal with recurrent event data. In that case, each recurrence can be modelled using a different baseline hazard to control the event dependence.

Appendix B

Semiparametric Inference for Peña and Hollander model

B.1 Case without Frailties

Chapter 1 introduces the model proposed by Peña and Hollander (2004) to deal with recurrent events data (see section 1.5). Now, we address the problem of estimating the model parameters $\Lambda_0(s) = \int_0^s \lambda_0(w)dw$, α and β for the model without frailties. That is, equation (1.20) where it is assumed that $Z_i \equiv 1$. Thus, the model has intensity process

$$\lambda_i(s|\mathbf{X}_i) = \lambda_0[\mathcal{E}_i(s)]\,\rho[N_i^\dagger(s-);\alpha]\,\psi(\beta'\mathbf{X}_i(s)). \tag{B.1}$$

The observables for the n subjects including the effective age processes are

$$\left\{\left(\mathbf{X}_i(s), N_i^\dagger(s), Y_i^\dagger(s), \mathcal{E}_i(s)\right) : 0 \le s \le s^*\right\},\ i = 1, 2, \ldots, n,$$

where $N_i^\dagger(s)$ and $Y_i^\dagger(s)$ are defined in (1.1) and (1.2), respectively.

By letting

$$A_i^\dagger(s) = \int_0^s Y_i^\dagger(v)\,\lambda_0[\mathcal{E}_i(v)]\,\rho[N_i^\dagger(v-);\alpha]\,\psi(\beta'\mathbf{X}_i(v))\,dv,$$

then with respect to the filtration $\mathbf{F}$, the vector of processes

$$\mathbf{M}^\dagger = (M_1^\dagger, \ldots, M_n^\dagger) = \mathbf{N}^\dagger - \mathbf{A}^\dagger = (N_1^\dagger - A_1^\dagger, \ldots, N_n^\dagger - A_n^\dagger)$$

consists of orthogonal square-integrable martingales with predictable quadratic covariation processes $\langle M_{i_1}^\dagger, M_{i_2}^\dagger\rangle(s) = A_{i_1}^\dagger(s)\,I\{i_1 = i_2\}$. Peña et al. (2003) stated that the usual martingale

theory does not apply directly for the purpose of estimating $\Lambda_0(\cdot)$. They argued that that the $\lambda_0(\cdot)$ appearing in $A_i^\dagger(\cdot)$ is time-transformed by the observable predictable process $\mathcal{E}_i(\cdot)$. The authors propose to use the techniques used in Peña et al. (2001) which are based on defining the double-indexed processes, $N_i(s,t)$, $Y_i(s,t)$, $A_i(s,t)$, and $M_i(s,t)$ described in Section 1.2.2.

Using Proposition 1 from Peña et al. (2003) which gives an expression for $A_i(s,t)$ which involves $\lambda_0(t)$ directly, we have $M_i(s,t) = N_i(s,t) - \int_0^t Y_i(s,w)\Lambda_0(dw), i = 1,2,\ldots,n$, so that $\sum_{i=1}^n M_i(s,dw) = \sum_{i=1}^n N_i(s,dw) - S_0(s,w)\Lambda_0(dw)$, where

$$S_0(s,t) \equiv S_0(s,t|\alpha,\beta) = \sum_{i=1}^n Y_i(s,t|\alpha,\beta). \tag{B.2}$$

Since the mean of $\sum_{i=1}^n M_i(s,dw)$ is zero, a method-of-moments 'estimator' of $\Lambda_0(t)$, given (α,β) is therefore

$$\hat{\Lambda}_0(s,t;\alpha,\beta) = \int_0^t \left\{ \frac{J(s,w|\alpha,\beta)}{S_0(s,w|\alpha,\beta)} \right\} \left\{ \sum_{i=1}^n N_i(s,dw) \right\}, \tag{B.3}$$

with $J(s,w|\alpha,\beta) = I\{S_0(s,w|\alpha,\beta) > 0\}$ and with the convention that $0/0 = 0$.

After that Peña et al. (2003) develop the profile likelihood for (α,β) following Jacod (1975) as follows. Assuming that the distribution G of τ does not involve the model parameters, the likelihood process associated with the observables for the Peña and Hollander model without frailties is

$$L^\dagger(s|\lambda_0(\cdot),\alpha,\beta) = \left\{ \prod_{i=1}^n \prod_{v=0}^s \left[Y_i^\dagger(v)\, \rho[N_i^\dagger(v-);\alpha]\, \psi(\beta'\mathbf{X}_i(v))\, \lambda_0[\mathcal{E}_i(v)] \right]^{N_i^\dagger(\Delta v)} \right\} \times$$
$$\left\{ \exp\left[-\sum_{i=1}^n \int_0^s Y_i^\dagger(v)\, \rho[N_i^\dagger(v-);\alpha]\, \psi(\beta'\mathbf{X}_i(v))\, \lambda_0[\mathcal{E}_i(v)]\, dv \right] \right\}. \tag{B.4}$$

Using (B.3), we can estimate the cumulative hazard function using the expression

$$\hat{\Lambda}_0(s,dw|\alpha,\beta) = \frac{\sum_{i=1}^n N_i(s,dw)}{S_0(s,w|\alpha,\beta)}.$$

On the other hand, substituting $\hat{\Lambda}_0(s,w|\alpha,\beta)$ for $\Lambda_0(w)$ in the first term of (B.4), we obtain the relevant portion of the profile likelihood of (α,β) to be

$$L_p(s|\alpha,\beta) = \prod_{i=1}^n \prod_{j=1}^{N_i^\dagger(s)} \left[\frac{\rho(j-1;\alpha)\, \psi[\beta'\mathbf{X}_i(S_{ij})]}{S_0[s,\mathcal{E}_i(S_{ij})|\alpha,\beta]} \right]^{\Delta N_i^\dagger(S_{ij})}. \tag{B.5}$$

The logarithm of the profile likelihood may also be expressed as

$$l_P(s|\alpha,\beta) = \sum_{i=1}^n \int_0^s \left[\log \rho[N_i^\dagger(v-);\alpha] + \log \psi(\beta'\mathbf{X}_i(v)) - \log S_0(s,\mathcal{E}_i(v)|\alpha,\beta) \right] N_i^\dagger(dv) \tag{B.6}$$

Finally, the estimators of α and β may be computed using the next estimating equations:

$$\sum_{i=1}^{n}\int_0^{s^*}\left[\frac{\frac{\partial}{\partial\alpha}\rho[N_i^\dagger(v-);\alpha]}{\rho[N_i^\dagger(v-);\alpha]}-\frac{\frac{\partial}{\partial\alpha}S_0(s,\mathcal{E}_i(v)|\alpha,\beta)}{S_0(s,\mathcal{E}_i(v)|\alpha,\beta)}\right]N_i^\dagger(dv) \;=\; \mathbf{0}; \tag{B.7}$$

$$\sum_{i=1}^{n}\int_0^{s^*}\left[\frac{\frac{\partial}{\partial\beta}\psi(\beta'\mathbf{X}_i(v))}{\psi(\beta'\mathbf{X}_i(v))}-\frac{\frac{\partial}{\partial\beta}S_0(s,\mathcal{E}_i(v)|\alpha,\beta)}{S_0(s,\mathcal{E}_i(v)|\alpha,\beta)}\right]N_i^\dagger(dv) \;=\; \mathbf{0}. \tag{B.8}$$

After that, Newton-Raphson algorithm may be employed to obtain the estimates $\hat{\alpha}$ and $\hat{\beta}$ as we have described in Section 4.4.

Peña et al. (2003) proposed and alternative notation to better understand equations (B.7) and (B.8). For $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, N_i^\dagger(s)$, and recalling the definition of the function $\varphi_{ij}(\cdot;\alpha,\beta)$ in (B.11), the authors define

$$Q_{ij}(s,w|\alpha,\beta) = I_{(\mathcal{E}_{ij-1}(S_{ij-1}),\ \mathcal{E}_{ij-1}(S_{ij})]}(w)\,\varphi_{ij-1}\left(\mathcal{E}_{ij-1}^{-1}(w);\alpha,\beta\right) \tag{B.9}$$

$$R_i(s,w|\alpha,\beta) \;=\; I_{(\mathcal{E}_{iN_i^\dagger(s-)}(S_{iN_i^\dagger(s-)}),\ \mathcal{E}_{iN_i^\dagger(s-)}(\min(s,\tau_i))]}(w)\,\varphi_{iN_i^\dagger(s-)}\left(\mathcal{E}_{iN_i^\dagger(s-)}^{-1}(w);\alpha,\beta\right) \tag{B.10}$$

where

$$\varphi_{ij}(s;\alpha,\beta) \equiv \frac{\rho[N_i^\dagger(s-);\alpha]\psi[\beta'\mathbf{X}_i(s)]}{\mathcal{E}'_{ij}(s)}. \tag{B.11}$$

and $\mathcal{E}'_{ij}(s) = \frac{d}{ds}\mathcal{E}_{ij}(s)$.

Using these processes, $S_0(s,w|\alpha,\beta)$ could be re-expressed via

$$S_0(s,w|\alpha,\beta) = \sum_{i=1}^{n}\left\{\sum_{j=1}^{N_i^\dagger(s)} Q_{ij}(s,w|\alpha,\beta) + R_i(s,w|\alpha,\beta)\right\}. \tag{B.12}$$

The authors noticed that the Q_{ij}s can be interpreted as the contributions of the uncensored values, while the R_is are the contributions of the right-censored values. Introducing new notation: $\rho^{(\alpha)}(\cdot;\alpha) = \partial\rho(\cdot;\alpha)/\partial\alpha$, $\psi'(\cdot)$ be the derivative of $\psi(\cdot)$, and

$$\mathbf{V}(j;\alpha) = \frac{\rho^{(\alpha)}(j;\alpha)}{\rho(j;\alpha)} \quad \text{and} \quad \mathbf{W}(\mathbf{x};\beta) = \frac{\mathbf{x}\psi'(\beta'\mathbf{x})}{\psi(\beta'\mathbf{x})}.$$

the authors showed that, assuming $\rho(k;\alpha) = \alpha^k$, $\psi(w) = \exp(w)$, and that the covariate vector process is time-independent, the estimating equations in (B.7) and (B.8) become

$$\sum_{i=1}^{n}\sum_{j=1}^{N_i^\dagger(s^*)}\left[\frac{j-1}{\alpha} - A(s^*,\mathcal{E}_{ij-1}(S_{ij})|\alpha,\beta)\right]\Delta N_i^\dagger(S_{ij}) \;=\; 0;$$

$$\sum_{i=1}^{n}\sum_{j=1}^{N_i^\dagger(s^*)}\left[\mathbf{X}_i - \mathbf{B}(s^*,\mathcal{E}_{ij-1}(S_{ij})|\alpha,\beta)\right]\Delta N_i^\dagger(S_{ij}) \;=\; \mathbf{0}.$$

where

$$A(s,w|\alpha,\beta) = \frac{1}{\alpha}\frac{\sum_{i=1}^{n}\left\{\sum_{j=1}^{N_i^\dagger(s)}(j-1)\,Q_{ij}(s,w|\alpha,\beta) + N_i^\dagger(s-)\,R_i(s,w|\alpha,\beta)\right\}}{\sum_{i=1}^{n}\left\{\sum_{j=1}^{N_i^\dagger(s)}Q_{ij}(s,w|\alpha,\beta) + R_i(s,w|\alpha,\beta)\right\}};$$

$$\mathbf{B}(s,w|\alpha,\beta) = \frac{\sum_{i=1}^{n}\mathbf{X}_i\left\{\sum_{j=1}^{N_i^\dagger(s)}Q_{ij}(s,w|\alpha,\beta) + R_i(s,w|\alpha,\beta)\right\}}{\sum_{i=1}^{n}\left\{\sum_{j=1}^{N_i^\dagger(s)}Q_{ij}(s,w|\alpha,\beta) + R_i(s,w|\alpha,\beta)\right\}},$$

After obtaining the estimators $\hat{\alpha}$ and $\hat{\beta}$, the authors also showed that the estimator of $\Lambda_0(t)$ based on the realizations of the observables over $[0, s^*]$ is obtained by substituting $(\hat{\alpha}, \hat{\beta})$ for (α, β) in the expression of $\hat{\Lambda}_0(s^*, t|\alpha, \beta)$ given in (B.3). That is,

$$\hat{\Lambda}_0(s^*,t) = \int_0^t \left\{\frac{J(s^*,w|\hat{\alpha},\hat{\beta})}{S_0(s^*,w|\hat{\alpha},\hat{\beta})}\right\}\left\{\sum_{i=1}^{n}N_i(s^*,dw)\right\}. \tag{B.13}$$

Finally, for an estimator of the baseline survivor function associated with $\Lambda_0(\cdot)$ defined via $\bar{F}_0(t) = \exp\{-\Lambda_0(t)\}$, by the product-integral representation and the substitution principle, we obtain

$$\hat{\bar{F}}_0(s^*,t) = \prod_{w=0}^{t}\left[1-\hat{\Lambda}_0(s^*,dw)\right] = \prod_{w=0}^{t}\left[1-\frac{\sum_{i=1}^{n}N_i(s^*,dw)}{S_0(s^*,w|\hat{\alpha},\hat{\beta})}\right]. \tag{B.14}$$

B.2 Case with Frailties

In this section we consider the estimation of the parameters when the class of models includes frailties. It will be assumed that the frailties $Z_1, Z_2, \ldots, Z_n$ are IID from a distribution $H(\cdot|\xi)$ where $\xi \in \Xi \subseteq \Re^r$. A common choice for this H is the gamma distribution with unit mean and variance $1/\xi$, $H = \text{Gamma}(\xi, \xi)$ (see Section 1.4.2). The restriction that the gamma shape and scale parameters are identical is needed to have model identifiability. The intensity function given in (1.20) is:

$$\lambda_i(s|Z_i,\mathbf{X}_i) = Z_i\,\lambda_0[\mathcal{E}_i(s)]\,\rho[N_i^\dagger(s-);\alpha]\,\psi(\beta'\mathbf{X}_i(s)).$$

The complete likelihood process for the model parameters $(\lambda_0, \alpha, \beta, \xi)$ is

$$L_C^\dagger(s^*|\lambda_0(\cdot),\alpha,\beta,\xi,Z) = \prod_{i=1}^{n}\left[\frac{\xi^\xi}{\Gamma(\xi)}Z_i^{\xi-1}\exp -\xi Z_i\times\right.$$
$$\left\{\prod_{v=0}^{s^*}\left[Z_iY_i^\dagger(v)\,\lambda_0(\mathcal{E}_i(v))\,\rho[N_i^\dagger(v-);\alpha]\,\psi(\beta'\mathbf{X}_i(v))\right]^{N_i^\dagger(\Delta v)}\right\}\times$$
$$\left.\exp\left\{-\int_0^{s^*}Z_iY_i^\dagger(v)\,\lambda_0(\mathcal{E}_i(v))\,\rho[N_i^\dagger(v-);\alpha]\,\psi(\beta'\mathbf{X}_i(v))\,dv\right\}\right]. \tag{B.15}$$

By integrating out $\mathbf{Z}$ according to its joint (gamma) distribution in the previous equation, the full likelihood process is obtained as follows:

$$L_F(s^*|\lambda_0(\cdot),\alpha,\beta,\xi) = \prod_{i=1}^{n}\left\{\left[\frac{\Gamma(\xi+N_i^\dagger(s^*))}{\Gamma(\xi)}\right]\times \left[\frac{\xi}{\xi+\int_0^{s^*} Y_i^\dagger(v)\,\rho[N_i^\dagger(v-);\alpha]\,\psi(\beta'\mathbf{X}_i(v))\,\lambda_0[\mathcal{E}_i(v)]\,dv}\right]^{\xi+N_i^\dagger(s^*)}\times \left(\prod_{v=0}^{s^*}\left[\frac{Y_i^\dagger(v)\,\rho[N_i^\dagger(v-);\alpha]\,\psi(\beta'\mathbf{X}_i(v))\,\lambda_0[\mathcal{E}_i(v)]}{\xi}\right]^{N_i^\dagger(\Delta v)}\right)\right\}. \tag{B.16}$$

To estimate the model parameters ξ, $\Lambda_0(\cdot)$, α, and β, the authors generalize and extend the approach implemented in Peña, et al. (2001). They used the expectation-maximization (EM) algorithm introduced by Dempster, et al. (1977), and implemented in counting process frailty models by Nielsen, et al. (1992).

The two steps of the EM algorithm for obtaining the maximum likelihood estimator of α, β, and Λ_0 are:

1. E-step. Compute the frailties estimates, Z, as the expected value given the current values α, β, Λ_0, and the data using the formula

$$\hat{Z}_i = \frac{\hat{\xi}+N_i^\dagger(s^*)}{\hat{\xi}+\int_0^{s^*} Y_i^\dagger(v)\,\rho[N_i^\dagger(v-);\hat{\alpha}]\,\psi(\hat{\beta}'\mathbf{X}_i(v))\,\hat{\lambda}_0[\mathcal{E}_i(v)]\,dv} \tag{B.17}$$

2. M-step. Treating the estimates of Z as a fixed offset, we update α, β, and λ_0 as in the case without frailties. That is, solving the score functions for the profile likelihood. Given Z, α, β, and the data

$$\hat{\Lambda}_0(s^*,t|Z,\alpha,\beta) = \int_0^t \left\{\frac{J(s^*,u|w,\alpha,\beta)}{S_0(s^*,u|w,\alpha,\beta)}\right\}\left\{\sum_{i=1}^{n} N_i(s^*,du)\right\}, \tag{B.18}$$

where $J(s,u|w,\alpha,\beta) = I\{S_0(s,u|w,\alpha,\beta)>0\}$ with

$S_0(s,u|w,\alpha,\beta) = \sum_{i=1}^{n} Z_i Y_i(s,u|\alpha,\beta)$.

For a given $(\Lambda_0(\cdot),\alpha,\beta)$, we can obtain an estimation of frailty parameter ξ maximizing the marginal profile likelihood for ξ from (B.16). Since we are dealing with an one-dimensional problem other algorithms different from Newton-Raphson can be used (see Section 4.4).

Having obtained an estimator of the baseline hazard function $\Lambda_0(\cdot)$ given by $\hat{\Lambda}_0(s^*, \cdot)$, the semiparametric estimator of the baseline survivor function $\bar{F}_0(\cdot)$ for this model with frailty is obtained via

$$\hat{\bar{F}}_0(s^*, t) = \prod_{\{w:\ w \leq t\}} \left[1 - \hat{\Lambda}_0(s^*, dw)\right]. \tag{B.19}$$

Appendix C

Published work related to present thesis

C.1 Papers

I Juan R González, Edsel A Peña, Elizabeth H Slate. Modelling Intervention Effects after Cancer Relapses. (*Statistics in Medicine*, 2005;24:3959-75. DOI: 10.1002/sim.2394). This article addresses the problem of incorporating information regarding the effects of treatments or interventions into models for repeated cancer relapses. In contrast to many existing models, our approach permits the impact of interventions to differ after each relapse. We adopt the general model for recurrent events proposed by Peña and Hollander, in which the effect of interventions is represented by an effective age process acting on the baseline hazard rate function. To accommodate the situation of cancer relapse, we propose an effective age function that encodes three possible therapeutic responses: complete remission, partial remission, and null response. The proposed model also incorporates the effect of covariates, the impact of previous relapses, and heterogeneity among individuals. We use our model to analyze the times to relapse for 63 patients with a particular subtype of indolent lymphoma and compare the results to those obtained using existing methods.

II Edsel A Peña, Elizabeth H Slate, Juan R González. Semiparametric Inference for a General Class of Models for Recurrent Events. (in first revision in *Journal of*

Statistical Planning and Inference). Procedures for estimating the parameters of the general class of semiparametric models for recurrent events proposed by Peña and Hollander (2004) are developed. This class of models incorporates an effective age function which encodes the changes that occur after each event occurrence such as the impact of an intervention, it al- lows for the modeling of the impact of accumulating event occurrences on the unit, it admits a link function in which the effect of possibly time-dependent covariates are incorporated, and it allows the incorporation of unobservable frailty components which induce dependencies among the inter-event times for each unit. The estimation procedures are semiparametric in that a baseline hazard function is non-parametrically specified. The sampling distribution properties of the estimators are examined through a simulation study, and the consequences of miss-specifying the model are analyzed. The results indicate that the flexibility of this general class of models provides a safeguard for analyzing recurrent event data, even data possibly arising from a frailty-less mechanism. The estimation procedures are applied to real data sets arising in the biomedical and public health settings, as well as from reliability and engineering situations. In particular, the procedures are applied to a data set pertaining to times to recurrence of bladder cancer and the results of the analysis are compared to those obtained using three methods of analyzing recurrent event data.

III Virginie Rondeau, Juan R González. frailtypack: a computer program for the analysis of correlated failure time data using penalized likelihood estimation. (*Computer Methods and Programs in Biomedicine*, 2005;**80**: 154-164. DOI:10.1016/j.cmpb.2005.06.01

Correlated survival outcomes occur quite frequently in the biomedical research. Available softwares are limited, and especially if we want to obtain a smooth curve for the baseline hazard function and a random effects model for correlated data.

The main objective of this paper is to describe an `R` package, *frailtypack* that can be used to estimate the parameters in a shared gamma frailty model with potentially right censored, left truncated and stratified survival data, using maximum penalized likelihood estimation. Time-dependent structure for the explanatory variables and/or extension of the Cox regression model to recurrent events are also allowed. This program can also be used simply to obtain directly a smooth estimates of the baseline hazard function.

To illustrate the program we used two data sets, one with clustered survival times, the

other one with recurrent events, ie the rehospitalizations of patients diagnosed with colorectal cancer. With this example we show how to fit the recurrent events, with time-dependent variables using the Andersen-Gill approach.

IV Juan R González, Edsel A Peña. Nonparametric Estimation of Survival Function with Recurrent Event Data. (*Rev. Española de Salud Pública*, **78**, 211-220, 2004. In Spanish). Recurrent events when we deal with survival studies demand a different methodology from what is used in standard survival analysis. The main problem that we found when we make inference in these kind of studies is that the observations may not be independent. Thus, biased and inefficient estimators can be obtained if we do not take into account this fact. In the independent case, the interoccurrence survival function can be estimated by the generalization of the limit product estimator (Peña et al., 2001). However, if data are correlated, other models should be used such as frailty models or an estimator proposed by Wang and Chang (1999), that take into account the fact that interocurrence times were or not correlated. The aim of this paper has been the illustration of these approaches by using two real data sets.

V Juan R González, Esteve Fernández, Victor Moreno, *et al.*. Gender differences in hospital readmission among colorectal cancer patients (*Journal or Epidemiology and Community Health*, **59**, 506-11, 2005. DOI:10.1136/jech.2004.028902).

Background: While several studies have analyzed gender and socioeconomic differences in cancer incidence and mortality, gender differences in oncological health care have been seldom considered.

Objective: The aim of this study was to investigate gender-based inequalities in hospital readmission among patients diagnosed with colorectal cancer.

Design: Prospective cohort study. Setting: Hospital Universitary in L'Hospitalet (Barcelona, Spain).

Participants: Four hundred and three patients diagnosed with colorectal between January 1996 and December 1998 were actively followed up until 2002.

Main outcome measurements and methods: Hospital readmission times related to colorectal cancer after surgical procedure. Cox proportional model with random effect (frailty) was used to estimate hazard rate ratios and 95% confidence intervals of readmission time

for covariates analyzed.

Results: Crude hazard rate ratio of hospital readmission in males was 1.61 (95% confidence interval: 1.21-2.15). When other significant determinants of readmission were controlled for (Dukes' stage, mortality, and Charlson's index) a significant risk of readmission was still present for males (hazard rate ratio: 1.51, 95% confidence interval: 1.17-1.96).

Conclusions: In the case of colorectal cancer, women are less likely than men to be readmitted to the hospital, even after controlling for tumor characteristics, mortality, and comorbidity. New studies should investigate the role of other non-clinical variable such as differences in help-seeking behaviors or structural or personal gender bias in the attention given to patients.

C.2 R packages

VI Juan R González, Edsel A Peña, Robert L Strawderman, (2005). survrec: Survival analysis for recurrent event data. R package version 1.1-3. http://www.r-project.org. Estimation of survival function for recurrent event data using Peña-Strawderman-Hollander, Whang-Chang estimators and MLE estimation under a Gamma Frailty model.

VII Juan R González, Elizabeth H. Slate, Edsel A Peña, (2005). gcmrec: General class of models for recurrent event data. R package version 0.9-1. http://www.r-project.org. Parameters estimation of the general semiparametric model for recurrent events data proposed by Peña and Hollander.

VIII Juan R González, Virginie Rondeau, (2005). frailtypack: Frailty models using maximum penalized likelihood estimation. R package version 2.0-0. http://www.r-project.org. Fit a shared gamma frailty model and Cox proportional hazards model using a Penalized Likelihood on the hazard function. Left truncated, censored data and strata (max=2) are allowed. Clustered and recurrent survival times can be studied (the Andersen-Gill (1982) approach has been implemented for recurrent events). An automatic choice of the smoothing parameter is possible using an approximated cross-validation procedure.

C.3 Oral Contributions in Meetings

IX Juan R González, Edsel A Peña. Modelling treatment effect after cancer relapses. 25th Annual Conference of the International Society for Clinical Biostatistics Leiden, the Netherlands, August 15-19, 2004.

X Juan R González, Edsel A Peña. Bootstrapping median survival with recurrent event data. 9th Spanish Conference on Biometrics A Corunha, Spain, May 28-30th, 2003 (In Spanish). This oral contribution was awarded with the Young Biometric Researchers Award.

XI Juan R González, J Ribes, E Fernández, *et al.* Non-parametric estimation with recurrent events. Application to hospital readmission in patients with colorectal cancer. XX Scientific Meeting of Spanish Society of Epidemiology Barcelona, Spain, September 12-14, 2002 (In Spanish).

Appendix D

R Functions and Classes

In this appendix we reproduce the on-line documentation for function and classes that are used in the examples in the text. The documentation is also available at CRAN `http://www.r-project.org/`.

D.1 The survrec Package

This package deals with the estimation of survival function for recurrent event data using Peña-Strawderman-Hollander, Whang-Chang estimators and maximum likelihood estimation under a Gamma Frailty model. In addition, this package also estimates median survival time and their confidence intervals using both asymptotic or bootstrap variance as we have described in this chapter.

`MMC` *Migratory Motor Complex*

Description

This contains the Migratoty Motor Complex data

Usage

```
data(MMC)
```

Format

This data frame contains the following columns:

id ID of each subject. Repeated for each recurrence

time recurrence o censoring time

event censoring status. All event are 1 for each subject excepting last one that it is 0

group a factor with levels

Note: The group have been created (at random) to illustrate a group comparison

Source

Husebye E, Skar V, Aalen O and Osnes M (1990), Digestive Diseases and Sciences, p1057

`Survr` *Create a Survival recurrent object*

Description

Create a survival recurrent object, usually used as a response variable in a model formula

Usage

```
Survr(id, time, event)
is.Survr(x)
```

Arguments

`id`	Identifier of each subject. This value is the same for all recurrent times of each subject.
`time`	time of reccurence. For each subject the last time are censored.
`event`	The status indicator, 0=no recurrence 1=recurrence. Only these values are accepted.
`x`	any R object.

Value

An object of class Survr. Survr objects are implemented as a matrix of 3 columnns. No method for print.

In the case of is.Survr, a logical value T if x inherits from class "Survr", otherwise an F.

See Also

`survfitr`,`psh.fit`,`wc.fit`,`mlefrailty.fit`

Examples

```
data(MMC)
Survr(MMC$id,MMC$time,MMC$event)
```

---

surv.search *Calculate the survival in selected times*

Description

Auxiliary function called from pshPLE, wcPLE and MLEFrailty.

The estimation using PLE (e.g. Kaplan-Meier) is a decreasing constant piecewise function with jumps in the times with events. Thus, to estimate the survival at any time we take the time of the precious event.

Usage

```
surv.search(tvals,time,surv)
```

Arguments

`tvals`	vector of times where the survival function has to be estimated
`time`	vector of failures times (distinct)
`surv`	vector of survival of each time

Value

Returns the survival in each selected time (tvals) from a vector of survival values

Examples

```
# we have the times 4,7,9,15,21,67
time<-c(4,7,9,15,21,67)

# and its survival (note: in this example there may be more
#                         than one event in some times)
surv<-c(0.8,0.7,0.65,0.55,0.43,0.22)

# We want to calculated the survival at times 1, 10, 32,64
surv.search(c(1,10,32,74),time,surv)
```

---

colon *Rehospitalization colorectal cancer*

Description

This contains rehospitalization times after surgery in patients with colorectal cancer

Usage

```
data(colon)
```

Format

This data frame contains the following columns:

hc identificator of each subject. Repeated for each recurrence

time rehospitalization o censoring time

event censoring status. All event are 1 for each subject excepting last one that it is 0

chemoter Did patient receive chemotherapy? 1: No 2:Yes

dukes Dukes' tumoral stage: 1:A-B 2:C 3:D

distance distance from living place to hospital 1:<=30 Km. 2:>30 Km.

Source

González, JR., Fernandez, E., Moreno, V. et al. Gender differences in hospital readmission among colorectal cancer patients. J Epidem Community Health, 2005

`mlefrailty.fit` *Survival function estimator for correlated recurrence time data under a Gamma Frailty Model*

Description

Estimation of survival function for correlated recurrence time data under a Gamma Frailty model using the maximum likelihood criterion. The resulting object of class "survfitr" is plotted by 'plot.survfitr', before it is returned.

Usage

```
mlefrailty.fit(x,tvals, lambda=NULL, alpha=NULL, alpha.min, alpha.max,
tol=1e-07, maxiter=500,alpha.console=TRUE)
```

Arguments

`x`	a survival recurrent event object.
`tvals`	vector of times where the survival function can be estimated.
`lambda`	optional vector of baseline hazard probabilities at t (see details). Default is numdeaths/apply(AtRisk,2,sum).
`alpha`	optional parameter of shape and scale for the frailty distribution. If this parameter is unknown it is estimated via EM algorithm. In order to obtain the convergence of this algorithm a seed is calculated (see details).
`alpha.min`	optional left bound of the alpha parameter in order to obtain a seed to estimate alpha parameter. Default value is 0.5.
`alpha.max`	optional right bound of the alpha parameter in order to obtain a seed to estimate alpha parameter. Default value is the maximum of distinct times of events.

`tol`

`maxiter` optional maximum number of iterations of the EM algorithm used to estimate the alpha parameter. Default is 500.

`alpha.console`

if TRUE prints in the console the estimates initial value for alpha and the alpha estimate via the EM algorithm, if FALSE not.

Details

The product limit estimator developed by Peña, Strawderman and Hollander (2001) are valid when the interoccurrence times are assumed to represent an IID sample from some underlying distribution F. This assumption is clearly restrictive in biomedical applications, and one obvious generalization that allows association between interocurrence times is a frailty model.

A common and convenient choice of frailty distribution is a gamma distribution with shape and scale parameters set equal to an unknown parameter α. The common marginal survival function can be written as follows

$$\bar{F}(t) = \left[\frac{\alpha}{\alpha + \Lambda_0(t)}\right]^{\alpha}$$

.

The parameter α controls the degree of association between interoccurrence times within a unit. Peña, Strawderman and Hollander (2001) showed that the estimation of α and Λ_0 can be obtained via the maximisation of the marginal likelihood function and the expectation-maximisation (EM) algorithm. For details and the theory behind this estimator, please refer to Peña, Strawderman and Hollander (2001, JASA).

In order to obtain a good convergence, α is estimated previously. This estimation is used as initial value in the EM procedure and it carried out by the maximisation of the profile likelihood for α. In this case the arguments of mlefrailty.fit function called alpha.min and alpha.max are the boundaries of this maximisation. The maximum is obtained using the golden section search method.

Value

If the convergence of EM algorithm is not obtained, the initial value of alpha can be used as a alpha.min argument and recalculate.

`n`	number of unit or subjects observed.
`m`	vector of number of recurrences in each subject (length n).
`failed`	vector of number of recurrences in each subject (length n*m). Vector ordered (e.g. times of first unit, times of second unit , ..., times of n-unit).
`censored`	vector of times of censorship for each subject (length n).
`numdistinct`	number of distinct failures times.
`distinct`	vector of distinct failures times.
`status`	0 if the estimation is can be provided and 1 if not depending if alpha could be estimate or not.
`alpha`	parameter of Gamma Frailty Model.
`lambda`	Estimates of the hazard probabilities at distinct failures times.
`survfunc`	vector of survival estimated in distinct times.
`tvals`	copy of argument.
`MLEAttvals`	vector of survival estimated in tvals times.

References

Peña, E.A., Strawderman, R. and Hollander, M. (2001). Nonparametric Estimation with Recurrent Event Data. *J. Amer. Statist. Assoc*, 96, 1299-1315.

See Also

`survfitr` `Survr`

Examples

```
data(MMC)
fit<-mlefrailty.fit(Survr(MMC$id,MMC$time,MMC$event))
fit
```

```
plot(fit)

# compare with pena-straderman-hollander

fit<-psh.fit(Survr(MMC$id,MMC$time,MMC$event))
fit
lines(fit,lty=2)

# and with wang-chang

fit<-wc.fit(Survr(MMC$id,MMC$time,MMC$event))
fit
lines(fit,lty=3)
```

`plot.survfitr` *Plots estimated survival function from an object of class 'survrec'.*

Description

Additional plots can be added to the same of axes using 'lines.survrec'.

Usage

```
plot.survfitr(x, conf.int=TRUE, prob = FALSE, ...)
```

Arguments

`x`	Object of class `survrec` (output from calling `survrec` function).
`conf.int`	Print the pointwise confidence intervals of the probability or survival function if its value is TRUE or FALSE.
`prob`	Print of the probability or survival function if its value is TRUE or FALSE respectively.

`...`

additional arguments passed to the plot function.

Value

Print a plot of class `survrec`

See Also

`psh.fit` `mlefrailty.fit` `wc.fit`

`print.survfitr` *Print a Short Summary of a Survival Recurrent Curve*

Description

Print number of observations, number of events, the restricted mean survival and its standard error, the median survial and the minimum, maximum and median number of recurrences for each subject.

Usage

```
print.survfitr(x, scale=1, digits=max(options()$digits - 4, 3), ...)
```

Arguments

`x`	the result of a call to the survfit, psh.fit, wc.fil or mlefrailty.fit functions
`scale`	a numeric value to rescale the survival time, e.g., if the input data to survfit were in days, scale=365 would scale the printout to years
`digits`	number of digits to print
`...`	other unusued arguments

Details

The restricted mean and its standard error are based on a truncated estimator. If the last observation(s) is not a death, then the survival curve estimate does not got to zero and the mean survival time cannot be estimated. Instead, the quantity reported is the mean of survival restricted to the time before the last censoring. When the last censoring time is not random this quantity is occasionally of interest.

Any randomness in the last censoring time is not taken into account in computing the standard error of the restricted mean. The restricted mean is shown mainly for compatibility with S.

The median are defined by drawing a horizontal line at 0.5 on the plot of the survival curve.

Value

x, with the invisible flag set

See Also

`summary.survfitr`, `survfitr`

Examples

```
data(MMC)
fit<-survfitr(Survr(id,time,event)~group,data=MMC)
print(fit)
```

`psh.fit` *Survival function estimator for recurrence time data using the estimator developed by Peña, Strawderman and Hollander*

Description

Estimation of survival function for recurrence time data by means the generalized product limit estimator (PLE) method developed by Pe a, Strawderman and Hollander. The resulting object of class "survfitr" is plotted by 'plot.survfitr', before it is returned.

Usage

```
psh.fit(x,tvals)
```

Arguments

`x`	a survival recurrent event object
`tvals`	vector of times where the survival function can be estimated.

Details

The estimator computed by this object is the nonparametric estimator of the inter-event time survivor function under the assumption of a renewal or IID model. This generalizes the product-limit estimator to the situation where the event is recurrent. For details and the theory behind this estimator, please refer to Peña, Strawderman and Hollander (2001, JASA).

Value

`n`	number of unit or subjects observed.
`m`	vector of number of recurrences in each subject (length n).
`failed`	vector of number of recurrences in each subject (length n*m). Vector ordered (e.g. times of first unit, times of second unit, ..., times of n-unit)
`censored`	vector of times of censorship for each subject (length n).
`numdistinct`	number of distinct failures times
`distinct`	vector of distinct failures times
`AtRisk`	matrix of number of persons-at-risk at each distinct time and for each subject
`survfunc`	vector of survival estimated in distinct times
`tvals`	copy of argument
`PSHpleAttvals`	vector of survival estimated in tvals times

References

Peña, E.A., Strawderman, R. and Hollander, M. (2001). Nonparametric Estimation with Recurrent Event Data. *J. Amer. Statist. Assoc*, 96, 1299-1315.

See Also

`survfitr` `Survr`

Examples

```
data(MMC)
fit<-psh.fit(Survr(MMC$id,MMC$time,MMC$event))
fit
plot(fit,conf.int=FALSE)

# compare with MLE Frailty

fit<-mlefrailty.fit(Survr(MMC$id,MMC$time,MMC$event))
fit
lines(fit,lty=2)

# and with wang-chang
```

```
fit<-wc.fit(Survr(MMC$id,MMC$time,MMC$event))
fit
lines(fit,lty=3)
```

`q.search` *Calculate the survival time of a selected quantile*

Description

Auxiliary function called from survdiffr function. Given a survfitr object we obtain the quantile from a survival function

Usage

```
q.search(f, q = 0.5)
```

Arguments

`f`	survdifr object
`q`	quantile. Default is 0.5

Value

Returns the time in a selected quantile

Examples

```
data(MMC)
fit<-survfitr(Survr(id,time,event)~1,data=MMC)

# 75th percentile from the survival function
q.search(fit,q=0.75)
```

`summary.survfitr` *Summary of a Survival of Recurrences Curve*

Description

Returns a matrix containing the survival curve and other information. If there are multiple curves, returns a list that contains the previous matrix for each curve.

Usage

```
summary.survfitr(object,...)
```

Arguments

`object`	output from a call to survfitr, psh.fit, wc.fit or mlefrailty.fit.
`...`	other unused arguments.

Value

For one survival curve returns a matrix, and for multiple curves a list with the same matrix for each curve. This matrix contains the distinct failure times, and the number of events, at risk subjects, survival and standard error for each distinct time

See Also

`survfitr`

Examples

```
data(MMC)
summary(survfitr(Survr(id,time,event)~group,data=MMC))
```

`survdiffr` *Test median survival differences (or other quantile)*

Description

Obtain bootstrap replicates of the median survival time for different groups of subjects. We can compute confidence intervals using boot package.

Usage

```
survdiffr(formula, data, q, B = 500, boot.F = "WC", boot.G = "none", ...)
```

Arguments

`formula`	A formula object. If a formula object is supplied it must have a Survr object as the response on the left of the operator and a term on the right. For a single bootstrap median survival the " 1" part of the formula is required.
`data`	A data frame in which to interpret the variables named in the formula.
`q`	Quantile that we are interested in to obtain a bootstrap sample from survival function
`B`	Number of boostrap samples
`boot.F`	a character string specifying the boostrap procedure. Possible value are either "PSH" or "WC" for nonparametric bootstrap or "semiparametric" for semiparametric boostrap. The default is "WC". Only the first words are required, e.g "P","W","se"
`boot.G`	a character string specifying if we also resample form censored empirical distribution. Possible value are either "none" or "empirical". The default is "none". Only the first words are required, e.g "n","e"
`...`	additional arguments passed to the type of estimator.

Details

See reference. Some procedures can be slow

Value

A boot object. Bootstrap confidence intervals can be computed using boot.ci function from boot package

References

Gonzalez JR, Peña EA. Bootstraping median survival with recurrent event data. IX Conferencia Española de Biometría; 2003 May 28-30; A Coruña, España.

Paper available upon request to the mantainer

See Also

survfitr,boot.ci

Examples

```
data(colon)

#We will compare the median survival time for three dukes stages
fit<-survdiffr(Survr(hc,time,event)~as.factor(dukes),data=colon,q=0.5)
boot.ci(fit$"1")
boot.ci(fit$"2")
boot.ci(fit$"3")

# 75th quantile of survival function
fit<-survdiffr(Survr(hc,time,event)~as.factor(dukes),data=colon,q=0.75)
# bootstrap percentile confidence interval
quantile(fit$"1"$t,c(0.025,0.975))
quantile(fit$"2"$t,c(0.025,0.975))
quantile(fit$"3"$t,c(0.025,0.975))

# We can execute this if there is none Inf value
# boot.ci(fit$"1")
# boot.ci(fit$"2")
# boot.ci(fit$"3")

#We can modify the bootstrap procedure modifiying boot.F parameter
fit<-survdiffr(Survr(hc,time,event)~as.factor(dukes),data=colon,q=0.5,boot.F="PSH")
# bootstrap percentile confidence interval
quantile(fit$"1"$t,c(0.025,0.975))
quantile(fit$"2"$t,c(0.025,0.975))
quantile(fit$"3"$t,c(0.025,0.975))
```

`survfitr` *Compute a Survival Curve for Recurrent Event Data given a covariate*

Description

Computes an estimate of a survival curve for recurrent event data using either the Peña-Strawderman-Hollander, Wang-Chang or MLE Frailty estimators. It also computes the asymptotic standard errors. The resulting object of class "survfitr" is plotted by 'plot.survfitr', before it is returned.

Usage

```
survfitr(formula, data, type="MLEfrailty",...)
```

Arguments

`formula`	A formula object. If a formula object is supplied it must have a Survr object as the response on the left of the operator and a term on the right. For a single survival curve the " 1" part of the formula is required.
`data`	a data frame in wich to interpret the variables named in the formula.
`type`	a character string specifying the type of survival curve. Possible value are "pena-strawderman-hollander", "wang-chang" or "MLEfrailty". The default is "MLEfrailty". Only the first words are required, e.g "pe","wa","ML"
`...`	additional arguments passed to the type of estimator.

Details

See the help details of psh.fit, wc.fit or mlefrailty depending on the type chosen

Value

a survfitr object. Methods defined for survfitr objects are provided for print,plot,lines and summary.

Note

The mantainer wishes to thank Professors Chiung-Yu Huang and Shu-Hui Chang for their help for providing us with the Fortran code which computes standard errors of Wang and Chang's estimator.

References

1. Peña, E.A., Strawderman, R. and Hollander, M. (2001). Nonparametric Estimation with Recurrent Event Data. *J. Amer. Statist. Assoc*, 96, 1299-1315.

2. Wang, M.-C. and Chang, S.-H. (1999). Nonparametric Estimation of a Recurrent Survival Function. *J. Amer. Statist. Assoc*, 94, 146-153.

See Also

`print.survfitr`,`plot.survfitr`, `lines.survfitr`, `summary.survfitr`, `Survr`,`psh.fit`,`wc.fit`, `mlefrailty.fit`

Examples

```
data(colon)
# fit a pena-strawderman-hollander and plot it
fit<-survfitr(Survr(hc,time,event)~as.factor(dukes),data=colon,type="pena")
plot(fit,ylim=c(0,1),xlim=c(0,2000))
# print the survival estimators
fit
summary(fit)

# fit a MLE Frailty and plot it (in this case do not show s.e.)
fit<-survfitr(Survr(hc,time,event)~as.factor(dukes),data=colon,type="MLE")
plot(fit)
# print the survival estimators
fit
summary(fit)
```

`wc.fit` *Survival function estimator for recurrence time data using the estimator developed by Wang and Chang.*

Description

Estimation of survival function for correlated or i.i.d. recurrence time data by means of the product limit estimator (PLE) method developed by Wang and Chang. The resulting object of class "survfitr" is plotted by 'plot.survfitr', before it is returned.

Usage

```
wc.fit(x,tvals)
```

Arguments

`x`	a survival recurrent event object.
`tvals`	vector of times where the survival function can be estimated.

Details

Wang and Chang (1999) proposed an estimator of the common marginal survivor function in the case where within-unit interrocurence times are correlated. The correlation structure considered by Wang and Chang (1999) is quite general and contains, in particular, both the i.i.d. and multiplicative (hence gamma) frailty model as special cases.

This estimator removes the bias noted for the produc-limit estimator developed by Pena, Strawderman and Hollander (PSH, 2001) when interrocurence times are correlated within units. However, when applied to i.i.d. interocurrence times, this estimator is not expected to perform as well as the PSH estimator, especially with regard to efficiency.

Value

`n`	number of unit or subjects observed.
`m`	vector of number of recurrences in each subject (length n).
`failed`	vector of number of recurrences in each subject (length n*m). Vector ordered (e.g. times of first unit, times of second unit, ..., times of n-unit)
`censored`	vector of times of censorship for each subject (length n).
`numdistinct`	number of distinct failures times.
`distinct`	vector of distinct failures times.

`AtRisk` matrix of number of persons-at-risk at each distinct time and for each subject.

`survfunc` vector of survival estimated in distinct times.

`tvals` copy of argument.

`PSHpleAttvals`
vector of survival estimated in tvals times.

Note

The mantainer wishes to thank Professors Chiung-Yu Huang and Shu-Hui Chang for their help for providing us with the Fortran code which computes standard errors of Wang and Chang's estimator.

References

Wang, M.-C. and Chang, S.-H. (1999). Nonparametric Estimation of a Recurrent Survival Function. *J. Amer. Statist. Assoc*, 94, 146-153.

See Also

`survfitr` `Survr`

Examples

```
data(MMC)

fit<-wc.fit(Survr(MMC$id,MMC$time,MMC$event))
fit
plot(fit,conf.int=FALSE)

# compare with pena-straderman-hollander

fit<-psh.fit(Survr(MMC$id,MMC$time,MMC$event))
fit
lines(fit,lty=2)

# and with MLE frailty

fit<-mlefrailty.fit(Survr(MMC$id,MMC$time,MMC$event))
fit
lines(fit,lty=3)
```

D.2 The gcmrec Package

This package is a computational implementation of the procedures and algorithms to estimate parameters involved in the general class of models proposed by Peña and Hollander (2004).

`GeneratedData` *Simulated data set generated under the minimal repair model*

Description

This contains recurrent times under minimal repair model with probability of perfect repair equal to 0.6. Data are as a list (see gcmrec help).

Usage

```
data(GeneratedData)
```

`addCenTime` *Add censored time equal to 0*

Description

Add a new line to the dataframe with a censored time equal to 0 when the end of follow-up coincides to the last occurrence

Usage

```
addCenTime(datin)
```

Arguments

`datin`	Dataframe containing id, time and event variables. Another covariates are allowed

Value

A data frame with an added line (censored time equal to 0) for those subjects where the end of follow-up coincides to the last occurrence

Examples

```
library(survival)
data(bladder2)

# we compute the interocurrence time
bladder2$time<-bladder2$stop-bladder2$start

# If we execute:
#     gcmrec(Survr(id,time,event)~rx+size+number,data=bladder2,s=2060)

# We will obtain the following error message:
#   Error in Survr(id, time, event) : Data doesn,t match...

# This means that we have some patients without right-censoring time. So,
# we understand that the last event coincides with the end of study.
# Consequently,we need to add a line with time 0 and status value equal
# to 0, too. To do so, we can use the function "addCenTime" as follows:

# for example:
#   bladder2[bladder2$id==12,]

#     id rx number size start stop event enum time
#  45 12  1      1    1     0    3     1    1    3
#  46 12  1      1    1     3   16     1    2   13
#  47 12  1      1    1    16   23     1    3    7

# there is no censored time for 12th patient. So, if we execute

bladderOK<-addCenTime(bladder2)

# we get

#      id rx number size start stop event enum time
#  45  12  1      1    1     0    3     1    1    3
#  46  12  1      1    1     3   16     1    2   13
#  47  12  1      1    1    16   23     1    3    7
#  471 12  1      1    1    16   23     0    3    0
```

`gcmrec-internal` *Internal gcmrec functions*

Description

Internal gcmrec functions

Usage

```
formatData(id, time, event, covariates, parameffage, cancer)
formatData.effage(id, time, status, covariates, effageData)
formatData.i(id, time, event, covariates, parameffage, cancer = NULL)
generlmi(perrep)
List.to.Dataframe(data)
```

Details

These are not to be called by the user

---

gcmrec *General Class of Models for recurrent event data*

Description

Fits the parameters for the general semiparametric model for recurrent events proposed by Peña and Hollander (2004). This class of models incorporates an effective age function which encodes the changes that occur after each event occurrence such as the impact of an intervention, it allows for the modeling of the impact of accumulating event occurrences on the unit, it admits a link function in which the effect of possibly time-dependent covariates are incorporated, and it allows the incorporation of unobservable frailty components which induce dependencies among the inter-event times for each unit.

Usage

```
gcmrec(formula, data, effageData = NULL, s, Frailty = FALSE,
    alphaSeed, betaSeed, xiSeed, tol = 10^(-6), maxit = 100,
    rhoFunc = "alpha to k", typeEffage = "perfect",
    maxXi = "Newton-Raphson", se = "Information matrix",
    cancer = NULL)
```

Arguments

`formula`	A formula object. If a formula object is supplied it must have a Survr object as the response on the left of the ' ' operator and a term on the right. Covariates are needed.

`data` A data frame in which to interpret the variables named in the formula. This data frame must contain the variables called "id","time" and "event" for subject identification, time of interocurrence, and censored status (coded 1: event, 0:censored), respectively. Furthermore, we can have some covariates.

Alternatively, it can also be a list containing the elements "n" and "subjects". Number of subjects must be recorded in "n". The element "subject" must have the following elements: subj, k, tau, caltimes, gaptimes, intercepts, slopes, lastperrep, perrepind, effage, effagebegin, and covariate including this information:

subj: Subject number or identificator.

k: Number of recurrences (time 0 must be included).

tau: Administrative time, time of study termination.

caltimes: Calendar times at each recurrence (time 0 must be included).

gaptimes: Gap times at each recurrence (time 0 must be included).

intercepts: Intercept value for the effect after each recurrence.

slopes: Slope value for the effect after each each recurrence.

lastperrep: Element from Brown and Proschan minimal repair model.

perrepind: Element from Brown and Proschan minimal repair model.

effagebegin: Initial value for effective age.

effage: Effective age after each recurrence.

covariate: covariate value at each recurrence.

See either GeneratedData or hydraulic data sets as an example.

`effageData` List containing the information about effective age. The list must have the elements described in the option 2 of data argument. If NULL we generate these elements under perfect repair model or minimal repair one depending on the 'typeEffage' argument (see below).

`s` A selected calendar time.

`Frailty` Logical value. Is model with frailties fitted? If so parameters for General Class of Models with frailty component are estimated.

`alphaSeed` Seed value for α.

`betaSeed` Seed value for β.

`xiSeed` Seed value for ξ.

`tol` Tolerance for maximization procedures.

`maxit` Maximum number of iterations in maximization procedures.

`rhoFunc` A character string specifying the effects attributable to the accumulating event occurrences, $\rho(k;\alpha)$. Possible values are "Identity" for $\rho(k;\alpha) = 1$ or "alpha to k" for $\rho(k;\alpha) = \alpha^k$. The default is "alpha to k". Only the first words are required, e.g "Id","a". Future versions will include other functions such as Markovian model for tumor occurrences, $\rho(k;\alpha) = \alpha - k + 1$ proposed by Gail et al. (1980).

`typeEffage` Effective age function. Possible value are "perfect" or "minimal" for perfect repair model or minimal repair model, respectively. The default is "perfect". Only the first words are required, e.g "p","m"

`maxXi` Maximization method for marginal likelihood with respect to ξ. Possible values are "Newton-Raphson" for Newton-Raphson maximization procedure or "Brent" for Brent's method maximization in one dimension. The default value is "Newton-Raphson". Only the first words are required, e.g. "N","B"

`se` Standard errors of parameters. Possible values are 'Information matrix' or 'Jacknife' for inverse of the partial likelihood information matrix or jacnife estimates, respectively.

`cancer` Effective age for fitting a cancer model proposed by Gonzalez et al (2005). This variable contains the information of the effect of treatments administred after cancer relapses coded as "CR", "PR" or "SD" depending on if complete, partial, or null response (stable disease) is achieved. See lymphoma data set as an example.

Details

Estimation with frailties are implemented using expectation-maximization (EM) algorithm. In this procedure, we need to maximize the marginal likelihood with respect to ξ. This

maximization is a one-dimensional maximization without derivarives. First we bracket the maximizing value, and the we obtain it using Brent's method in one dimension. When we implement this algorithm, we re-parameterize ξ using $\xi^* = log(\xi)$ to alleviate the problem of getting negative estimates for ξ. Iteration is terminated when successive values of $\xi/(1+\xi)$ differ by no more than the "tol" parameter. Maybe estimation under frailty model can be not too fast.

Value

a gcmrec object. Methods defined for gcmrec objects are provided for print and plot.

References

Peña, E. and M. Hollander (2004). Mathematical Reliability: An Expository Perspective, Chapter 6. Models for Recurrent Events in Reliability and Survival Analysis, pp. 105-123. Kluwer Academic Publishers.

M. Gail, T Santner, and C Brown (1980). An analysis of comparative carfinogenesis experiments based on multiple times to tumor. *Biometrics*, 36, 255-266.

JR Gonzalez, E Peña, E Slate (2005). Modelling treatment effect after cancer relapses, with application to recurrences in indolent non-Hodgkin's lymphomas. *Stat Med*, 2005.

R. Brent. *Algorithms for Minimization Without Derivatives*. Prentice-Hall, New York, 1973.

Examples

```
###################################
## Models using different data formats
###################################

#
#    Data input as a data frame
#

#   We use the well-known bladder cancer data set from survival package

library(survival)
data(bladder2)

# we compute the interocurrence time
bladder2$time<-bladder2$stop-bladder2$start
```

```
# If we execute:
#     gcmrec(Survr(id,time,event)~rx+size+number,data=bladder2,s=2060)

# We will obtain the following error message:
#   Error in Survr(id, time, event) : Data doesn,t match...

# This means that we have some patients without right-censoring time. So,
# we understand that the last event coincides with the end of study.
# Consequently,we need to add a line with time 0 and status value equal
# to 0, too. To do so, we can use the function "addCenTime" as follows:

bladderOK<-addCenTime(bladder2)

# Now, we can fit the model using this new data set:

gcmrec(Survr(id,time,event)~rx+size+number,data=bladderOK,s=2060)


#
#    Data as a list. See either GeneratedData or hydraulic data
#                    sets as an example.
#

#
# We can fit the model by transforming our data in a data frame
# using "List.to.Dataframe" function:
#

data(hydraulic)
hydraulicOK<-List.to.Dataframe(hydraulic)
gcmrec(Survr(id,time,event)~covar.1+covar.2,data=hydraulicOK,s=4753)


#
# Our model allows us to incorporate effective age information
#
# To illustrate this example, we will use a simulated data set generated
# under the minimal repair model with probability of perfect repair equal to 0.6
#
# As we have the data in a list, first we need to obtain a data frame containing
# the time, event, and covariates information:
#

data(GeneratedData)
temp<-List.to.Dataframe(GeneratedData)

# then, we can fit the model incorporating the information about the effective
#        age in the effageData argument:

gcmrec(Survr(id,time,event)~covar.1+covar.2, data=temp,
                    effageData=GeneratedData, s=100)
```

```
#########################################################################
##  How to fit minimal or perfect repair models, with and without frailties
#########################################################################

# Model with frailties

mod.Fra<-gcmrec(Survr(id,time,event)~rx+size+number,data=bladderOK,s=2060,Frailty=TRUE)
print(mod.Fra)

# effective age function: perfect repair and minimal repair models
# (models without frailties)

data(readmission)

# perfect
mod.per<-gcmrec(Survr(id,time,event)~as.factor(dukes),data=readmission,
    s=3000,typeEffage="per")
print(mod.per)

# minimal
mod.min<-gcmrec(Survr(id,time,event)~as.factor(dukes),data=readmission,
   s=3000,typeEffage="min")
print(mod.min)

#########################################################################
##  How to fit models with \rho function equal to identity
#########################################################################

data(lymphoma)

gcmrec(Survr(id, time, event) ~ as.factor(distrib),
    data = lymphoma, s = 1000, Frailty = TRUE, rhoFunc = "Ident")

#########################################################################
##  How to fit cancer model
#########################################################################

mod.can<-gcmrec(Survr(id,time,event)~as.factor(distrib), data=lymphoma,
     s=1000, Frailty=TRUE,  cancer=lymphoma$effage)

#  standard errors can be obtained by adding se="Jacknife".
# This procedure can be very time consuming
```

`graph.caltimes` *Plot of recurrent events*

Description

Plots calendar times at succesive recurrences from a data set. Information about effective age and categories of covariates are allowed.

Usage

```
graph.caltimes(data, var = NULL, effageData = NULL, width = 2,
        lines = TRUE, sortevents = TRUE, ...)
```

Arguments

`data`	data frame containing id, time, event variables and some other covariates
`var`	categorical variable
`effageData`	effective age function information
`width`	point width
`lines`	Are horizontal lines printed? The default is TRUE
`sortevents`	Are events sorted? The default is TRUE
`...`	other graphical parameters

Examples

```
# with data in a data frame
library(survival)
data(bladder2)
bladder2$time<-bladder2$stop-bladder2$start

graph.caltimes(bladder2)

# or data in a list

data(hydraulic)
graph.caltimes(hydraulic)

# We can print some covariate as follows:
graph.caltimes(bladder2,bladder2$rx)
```

`hydraulic` *hydraulic load-haul-dump (LHD) subsystems*

Description

Hydraulic load-haul-dump (LHD) subsystems used in moving ore and rock in underground mines in Sweden. The data set provides the calendar times (in hours), excluding repair or down times, of the successive failures of n = 6 such systems during the two-year development phase.

Usage

```
data(hydraulic)
```

Source

Kumar, D. and B. Klefsjo (1992). Reliability analysis of hydraulic systems of lhd machines using the power law process model. *Reliability Engineering and System Safety* 35, 217- 224.

`lymphoma` *Indolent non-Hodgkin's lymphomas*

Description

This contains cancer relapses times after first treatment in patients diagnosed with low grade lymphoma

Usage

```
data(lymphoma)
```

Format

This data frame contains the following columns:

id identificator of each subject. Repeated for each recurrence

time interocurrence or censoring time

event censoring status. All event are 1 for each subject excepting last one that it is 0

enum which lymphoma

delay delay between first symptom and date of first treatment as a continuous variable

age age at diagnosis

sex gender: 1:Males 2:Females

distrib lesions involved at diagnosis (0=Single, 1=Localized, 2=More than one nodal site, 3=Generalized

effage response achieved after treatment upon relapses, coded as CR: Complete remission, PR: Partial remission or SD: stable disease or null response.

Source

JR Gonzalez, E Peña, E Slate (2005). Modelling treatment effect after cancer relapses, with application to recurrences in indolent non-Hodgkin's lymphomas. *Stat Med*, 2005.

O. Servitje, F. Gallardo, T. Estrach, et al. (2002). Primary cutaneous marginal zone B-cell lymphoma: a clinical, histopathological, immunophenotypic and molecular genetic study of 22 cases. *Br J Dermatol*, 147:1147-1158.

`plot.gcmrec` *Plot Method for an object of class 'gcmrec'.*

Description

Plots estimated baseline survival and hazard functions from an object of class 'gcmrec'.

Usage

```
plot.gcmrec(x, type.plot = "surv", ...)
```

Arguments

`x`	Object of class `gcmrec` (output from calling `gcmrec` function).
`type.plot`	a character string specifying the type of curve. Possible value are ”hazard”, or ”survival”. The default is ”hazard”. Only the first words are required, e.g ”haz”, ”su”
`...`	Other graphical parameters

Value

Print a plot of class `gcmrec`

See Also

`print.gcmrec`

Examples

```
data(lymphoma)

mod<-gcmrec(Survr(id,time,event)~as.factor(distrib),data=lymphoma,s=1000)

# baseline survivor function

plot(mod)

# baseline hazard function

plot(mod,type="haz")
```

`print.gcmrec` *Print a Short Summary of parameter estimates of a general class of models for reccurrent event data*

Description

Prints a short summary of ’gcmrec’ object

Usage

```
print.gcmrec(x, digits = max(options()$digits - 4, 3), ...)
```

Arguments

x	the result of a call to the gcmrec function
digits	number of digits to print
...	other unusued arguments

Value

x, with the invisible flag set

See Also

summary.gcmrec, gcmrec

Examples

```
data(lymphoma)
mod<-gcmrec(Survr(id,time,event)~as.factor(distrib),data=lymphoma,s=1000)
print(mod)
```

---

readmission *Rehospitalization colorectal cancer*

Description

This contains rehospitalization times after surgery in patients diagnosed with colorectal cancer

Usage

```
data(readmission)
```

Format

This data frame contains the following columns:

id identificator of each subject. Repeated for each recurrence

enum which readmission

t.start start of interval (0 or previous recurrence time)

t.stop recurrence or censoring time

time interocurrence or censoring time

event censoring status. All event are 1 for each subject excepting last one that it is 0

chemo Did patient receive chemotherapy? 1: No; 2:Yes

sex gender: 1:Males 2:Females

dukes Dukes' tumoral stage: 1:A-B; 2:C 3:D

charlson Comorbidity Charlson's index. Time-dependent covariate. 0: Index 0; 1: Index 1-2; 3: Index >=3

Source

González, JR., Fernandez, E., Moreno, V. et al. Gender differences in hospital readmission among colorectal cancer patients. J Epidem Community Health, 2005.

`summary.gcmrec` *summary of 'gcmrec'*

Description

This function returns hazard rations (HR) and its confidence intervals

Usage

```
summary.gcmrec(object, level = 0.95, len = 6, d = 2, lab="hr", ...)
```

Arguments

`object`	output from a call to gcmrec.
`level`	significance level of confidence interval. Default is 95%.
`len`	the desired number of digits after the decimal point. Default of 6 digits is used.
`d`	the total field width. Default is 6.
`lab`	label of printed results.
`...`	other unusued arguments.

Details

This function calls to intervals.gcmrec

Value

Prints HR and its confidence intervals. Confidence level is allowed (level argument)

See Also

`intervals.gcmrec`

Examples

```
data(lymphoma)
mod<-gcmrec(Survr(id,time,event)~as.factor(distrib),data=lymphoma,s=1000)
summary(mod)

# confidence interval at 99

summary(mod,level=0.99)
```

D.3 The frailtypack Package

This package fits a shared gamma frailty model and Cox proportional hazards model using a Penalized Likelihood on the hazard function. Left truncated, censored data and strata (max=2) are allowed. Clustered and recurrent survival times can be studied (the Andersen-Gill (1982) approach has been implemented for recurrent events). An automatic choice of the smoothing parameter is possible using an approximated cross-validation procedure.

This package also fits the general class of models proposed by Peña and Hollander (2004) using penalized likelihood as it is described in Chapter 4.

`frailtyPenal` *Fit Shared Gamma Frailty model using penalized likelihood estimation*

Description

Fit a shared gamma frailty model using a Penalized Likelihood on the hazard function. Left truncated and censored data and strata (max=2) are allowed. It allows to obtain a non-parametric smooth hazard of survival function. This approach is different from the partial penalized likelihood approach of Therneau et al.

Usage

```
frailtyPenal(formula, data, Frailty = TRUE, recurrentAG=FALSE,
             cross.validation=FALSE, n.knots, kappa1, kappa2, maxit=350)
```

Arguments

`formula`	a formula object, with the response on the left of a ' ' operator, and the terms on the right. The response must be a survival object as returned by the 'Surv' function like in survival package.
`data`	a data.frame in which to interpret the variables named in the 'formula'.
`Frailty`	Logical value. Is model with frailties fitted? If so variance of frailty parameter is estimated. If not, Cox proportional hazards model is estimated using Penalized Likelihood on the hazard function

`recurrentAG` Logical value. Is Andersen-Gill model fitted? If so indicates that recurrent event times with the counting process approach of Andersen and Gill is used. This formulation can be used for dealing with time-dependent covariates. The default is FALSE.

`cross.validation`

Logical value. Is cross validation procedure used for estimating smoothing parameter? If so a search of the smoothing parameter using cross validation is done, with kappa1 as the seed. The cross validation is not implemented for two strata. The cross validation has been implemented for a Cox proportional hazard model, with no covariates. The default is FALSE.

`n.knots` integer giving the number of knots to use. Value required. It corresponds to the (n.knots+2) splines functions for the approximation of the hazard or the survival functions. Number of knots must be between 4 and 20.

`kappa1` positive smoothing parameter. The coefficient kappa of the integral of the squared second derivative of hazard function in the fit (penalized log likelihood). We advise the user to identify several possible tuning parameters, note their defaults and look at the sensitivity of the results to varying them. Value required.

`kappa2` positive smoothing parameter for the second stratum, when data are stratified. See kappa1.

`maxit` maximum number of iterations for the Marquardt algorithm. Default is 350

Details

The estimated parameter are obtained using the robust Marquardt algorithm (Marquardt, 1963) which is a combination between a Newton-Raphson algorithm and a steepest descent algorithm. When frailty parameter is small, numerical problems may arise. To solve this problem, an alternative formula of the penalized log-likelihood is used (see Rondeau, 2003 for further details). Cubic M-splines of order 4 are used for the hazard function, and I-splines (integrated M-splines) are used for the cumulative hazard function.

PARAMETERS

As frailtypack is written in Fortran 77 some parameters had to be hard coded in. The default values of these parameters are

maximum number of observations: 60000

maximum number of groups: 5000

maximum number of subjects: 30000

If these parameters are not large enough (an error message will let you know this), you need to reset them in frailtypack.f and recompile. In particular, the statements defining these parameters are `PARAMETER (ndatemax = 60000)`

`PARAMETER (ngmax = 5000)`

`PARAMETER (nsujetmax = 30000)`

Value

an object of class '"frailtyPenal"'. Methods defined for 'frailtyPenal' objects are provided for print and plot. The following components are included in a 'frailtyPenal'object.

`n`	the number of observations used in the fit.
`groups`	the maximum number of groups used in the fit
`n.events`	the number of events observed in the fit
`logVerComPenal`	the complete marginal penalized log-likelihood
`theta`	variance of frailty parameter
`coef`	the coefficients of the linear predictor, which multiply the columns of the model matrix.
`varH`	the variance matrix of theta and of the coefficients.
`varHIH`	the robust estimation of the variance matrix of theta and of the coefficients.
`x1`	vector of times where both survival and hazard function are estimated. By default seq(0,max(time),length=99), where time is the vector of survival times.
`lam`	matrix of hazard estimates at x1 times and confidence bands.

`surv`	matrix of baseline survival estimates at x1 times and confidence bands.
`x2`	see x1 value for the second stratum
`lam2`	the same value as lam for the second stratum
`surv2`	the same value as surv for the second stratum

References

D. Marquardt (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal of Applied Mathematics*, 431-441.

V. Rondeau, D Commenges, and P. Joly (2003). Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime Data Analysis*, 9, 139-153.

McGilchrist CA and Aisbett CW (1991). Regression with frailty in survival analysis. *Biometrics*, 47, 461-466.

See Also

`print.frailtyPenal`, `summary.frailtyPenal`

Examples

```
data(kidney)
#Shared frailty model
frailtyPenal(Surv(time,status)~sex+age+cluster(id),
             n.knots=12,kappa1=1000,data=kidney)

#model without frailties (e.g., Cox proportional hazards
#                          estimated via penalized likelihood)
frailtyPenal(Surv(time,status)~sex+age+cluster(id),
             n.knots=12,kappa1=1000,data=kidney,Frailty=FALSE)

# truncated data

# first, we create a hypothetical truncated data
kidney$tt0<-rep(0,nrow(kidney))
kidney$tt0[1:3]<-c(2,9,13)

# then, we fit the model
frailtyPenal(Surv(tt0,time,status)~sex+age+cluster(id),
             n.knots=12,kappa1=1000,data=kidney)

#stratified data. Let,s use another dataset
data(readmission)
frailtyPenal(Surv(time,event)~as.factor(dukes)+cluster(id)+strata(sex),
```

```
            n.knots=10,kappa1=10000,kappa2=10000,data=readmission)

#Andersen-Gill counting-process approach with time-dependent covariate
frailtyPenal(Surv(t.start,t.stop,event)~as.factor(sex)+as.factor(dukes)+
            as.factor(charlson)+cluster(id),data=readmission, Frail=TRUE,
            n.knots=6,kappa1=100000,recurrentAG=TRUE)

# with the use of the cross validation approach, to find the smoothing parameter
frailtyPenal(Surv(t.start,t.stop,event)~as.factor(sex)+as.factor(dukes)+
            as.factor(charlson)+cluster(id),data=readmission, Frail=TRUE,
            n.knots=6,kappa1=5000,recurrentAG=TRUE,cross.validation=TRUE)
```

`plot.frailtyPenal` *Plot Method for an object of class 'frailtyPenal'.*

Description

Plots estimated baseline survival and hazard functions from an object of class 'frailtyPenal'. Confidence bands are allowed.

Usage

```
plot.frailtyPenal(x, type.plot = "hazard", conf.bands=TRUE, ...)
```

Arguments

`x`	Object of class `frailtyPenal` (output from calling `frailtyPenal` function).
`type.plot`	a character string specifying the type of curve. Possible value are "hazard", or "survival". The default is "hazard". Only the first words are required, e.g "haz", "su"
`conf.bands`	logical value. Determines whether confidence bands will be plotted. The default is to do so.
...	Other graphical parameters

Value

Print a plot of class `frailtyPenal`

See Also

print.frailtyPenal

Examples

```
data(readmission)

# Let,s compare shared frailty model with Cox proportional hazards model
mod.sha<-frailtyPenal(Surv(time,event)~as.factor(dukes)+cluster(id),
             n.knots=10,kappa1=10000,data=readmission)
plot(mod.sha,type="surv",conf=FALSE)
mod.cox<-frailtyPenal(Surv(time,event)~as.factor(dukes)+cluster(id),
             n.knots=10,kappa1=10000,data=readmission,Frailty=FALSE)
lines(mod.cox,type="surv",conf=FALSE,col=2)



# Stratified model
mod<-frailtyPenal(Surv(time,event)~as.factor(dukes)+cluster(id)+strata(sex),
             n.knots=10,kappa1=10000,kappa2=10000,data=readmission)
plot(mod)

# no confidence bands
plot(mod,conf.bands=FALSE)
```

---

print.frailtyPenal *Print a Short Summary of parameter estimates of a shared gamma frailty model*

Description

Prints a short summary of 'frailtyPenal' object

Usage

```
print.frailtyPenal(x, digits = max(options()$digits - 4, 3), ...)
```

Arguments

x	the result of a call to the frailtyPenal function
digits	number of digits to print
...	other unusued arguments

Value

x, with the invisible flag set

See Also

`summary.frailtyPenal`, `frailtyPenal`

Examples

```
data(kidney)
mod<-frailtyPenal(Surv(time,status)~sex+age+cluster(id),
     n.knots=8,kappa1=10000,data=kidney)
print(mod)
```

---

readmission *Rehospitalization colorectal cancer*

Description

This contains rehospitalization times after surgery in patients diagnosed with colorectal cancer

Usage

```
data(readmission)
```

Format

This data frame contains the following columns:

id identificator of each subject. Repeated for each recurrence

enum which readmission

t.start start of interval (0 or previous recurrence time)

t.stop recurrence or censoring time

time interocurrence or censoring time

event censoring status. All event are 1 for each subject excepting last one that it is 0

chemo Did patient receive chemotherapy? 1: No; 2:Yes

sex gender: 1:Males 2:Females

dukes Dukes' tumoral stage: 1:A-B; 2:C 3:D

charlson Comorbidity Charlson's index. Time-dependent covariate. 0: Index 0; 1: Index 1-2; 3: Index >=3

Source

González, JR., Fernandez, E., Moreno, V. et al. Gender differences in hospital readmission among colorectal cancer patients. Journal of Epidemiology and Community Health. In press, 2005.

`summary.frailtyPenal`

summary of 'frailtyPenal'

Description

This function returns hazard rations (HR) and its confidence intervals

Usage

```
summary.frailtyPenal(object, level = 0.95, len = 6, d = 2, lab="hr", ...)
```

Arguments

`object`	output from a call to frailtyPenal.
`level`	significance level of confidence interval. Default is 95%.
`len`	the desired number of digits after the decimal point. Default of 6 digits is used.
`d`	the total field width. Default is 6.
`lab`	label of printed results.
`...`	other unusued arguments.

Details

This function calls to intervals.frailtyPenal

Value

Prints HR and its confidence intervals. Confidence level is allowed (level argument)

See Also

`intervals.frailtyPenal`

Examples

```
data(kidney)
mod<-frailtyPenal(Surv(time,status)~age+sex+cluster(id),
     data=kidney,n.knots=8,kappa1=1000)
summary(mod)

# confidence interval at 99

summary(mod,level=0.99)
```

Appendix E

Additional R functions

```
"Brook.Crowley" <-
function (x,p)
{

# x must be a survival object
# x must contain: time, survival and std.error
# p is the quantile


#
# Asymptotic. Formula (4.3.24 Andersen et. al, 1993)
#
xi<-search.survObject(x,1-p,"time")
se.hat.hat<-search.survObject(x,xi,"se")
num<-(1-p*se.hat.hat)

n.OK<-ceiling(ceiling(length(x$time)/2)/2)
pos.OK<-sum(x$time<=xi)
bn<-x$time[pos.OK]-x$time[pos.OK-n.OK]


den<-(search.survObject(x,xi-bn,"surv")-search.survObject(x,xi+bn,"surv"))/(2*bn)
se.asy<-num/den
ci.asy<-c(xi-1.96*se.asy,xi+1.96*se.asy)


#
# Brookmeyer-Crowley CI (formula page 227 Andersen et. al, 1993)
#

xi.surv<-search.survObject(x,xi,"surv")
#
```

```
# log-log
#
inv.inf<-xi.surv^(exp((1.96*se.hat.hat)/log(xi.surv)))
inv.sup<-xi.surv^(exp((-1.96*se.hat.hat)/log(xi.surv)))
ci.log.log<-c(search.survObject(x,inv.inf,"inv.se"),
              search.survObject(x,inv.sup,"inv.se"))
#
# Identity
#
inv.inf<-xi.surv+1.96*se.hat.hat
inv.sup<-xi.surv-1.96*se.hat.hat
ci.id<-c(search.survObject(x,inv.inf,"inv.se"),
         search.survObject(x,inv.sup,"inv.se"))



temp<-min(3.1415/2,asin(xi.surv^.5)+(0.5*1.96*se.hat.hat*(xi.surv/(1-xi.surv))^.5))
inv.inf<-sin(temp)^2
temp<-max(0,asin(xi.surv^.5)-(0.5*1.96*se.hat.hat*(xi.surv/(1-xi.surv))^.5))
inv.sup<-sin(temp)^2

ci.arcsin<-c(search.survObject(x,inv.inf,"inv.se"),search.survObject(x,inv.sup,"inv.se")


list(percentile=xi,ci95.asymptotic=list(bandwith=bn,ci95=round(ci.asy,2)),
ci95.id=round(c(ci.id),2),ci95.log.log=round(c(ci.log.log),2),
ci95.arcsin=round(c(ci.arcsin),2))

}



"search.survObject" <-
function(x,t,f="surv")
{
if(f=="surv")
 {
  pos<-max(length(x$surv[x$time<=t]),1)
  ans<-x$surv[pos]
 }
if(f=="se")
 {
  pos<-max(length(x$surv[x$time<=t]),1)
  ss<-x$surv[pos]
  se<-x$std[pos]
  ans<-se/ss
```

```
 }
if(f=="time")
 {
  ans<-x$time[x$surv<=t][1]
  if(is.na(ans)) ans<-x$time[length(x$time)]
 }

if(f=="inv.se")
 {
  pos<-length(x$time[x$surv>=t])
  ans<-x$time[pos]
  if(pos==0) ans<-0
 }
return(ans)
}
```