Joint Modeling of Longitudinal and Time-to-Event Data with Applications in Health Insurance

Xavier Piulachs Lozada-Benavente

Joint Modeling of Longitudinal and Time-to-Event Data with Applications in Health Insurance

Author: Xavier Piulachs Lozada-Benavente

Advisor: Montserrat Guillén Estany

Co-Advisor: Ramon Alemany Leira

Department of Econometrics, Statistics and Applied Economy
University of Barcelona

Date: September 26, 2017

JURY MEMBERS

Dr. Jordi Ocaña Rebull
Member of Jury 1, Universitat de Barcelona

Dr. Guadalupe Gómez Melis
Member of Jury 1, Universitat Politècnica de Catalunya

Dr. Cécile Proust-Lima
Member of Jury 1, Université de Bordeaux

Dr. Montserrat Rué Monné
Member of Jury 2, Universitat de Lleida

Dr. Catalina Bolancé Losilla
Member of Jury 2, Universitat de Barcelona

Dr. Kazem Nasserinejad
Member of Jury 2, Erasmus University Medical Center

ACKNOWLEDGMENTS

First and foremost, I want to thank **Dr. Montserrat Guillén** very much for all the big and constant help she provided me during my PhD period, as well as her patience while guiding, advising, and encouraging me. This thesis has been possible thanks to her. In the same way, I'm very grateful to **Dr. Ramon Alemany** for his support and help since the first day I arrive to the Department of Econometrics at the University of Barcelona.

Special thanks to **Dr. Dimitris Rizopoulos** for his valuable help with some methodological aspects, and also for hosting me at the Erasmus University Medical Center. Also, many thanks to **Dr. Eleni-Rosalina Andrinopoulou** for her support during the last period of my doctoral thesis.

Barcelona, July 14, 2017

To Mònica and Gerard Maria

LIST OF PUBLICATIONS

Piulachs, X., Alemany, R. (2015). *Joint Modeling Scheme by Weighting Cumulative Effects over Time. Research on Mortality for Insured elderly.* Current topics on Risk Analysis: ICRA6 and RISK2015 Conference. Fundación MAPFRE. pp.629-636.

Piulachs, X., Alemany, R., Guillén, M., and Serrat, C. (2015). *Joint Modeling of Health Care Usage and Longevity Uncertainty for an Insurance Portfolio.* Scientific Methods for the Treatment of Uncertainty in Social Sciences. Springer International Publishing Switzerland. 377, pp.289-297.

Piulachs, X., Alemany, R., and Guillén, M. (2016). Joint Modelling of Survival and Emergency Medical Care Usage in Spanish Insureds Aged 65+. *PloS ONE*, 11(4):e0153234. DOI: 10.1371/journal.pone.0153234

Piulachs, X., Alemany R, and Guillén M (2017). Emergency Care Usage and Longevity have Opposite Effects on Health Insurance Rates. *Kybernetes*, 46(1), pp.102-113. DOI: 10.1108/K-06-2016-0149.

Piulachs, X., Alemany, R., Guillén, M., and Rizopoulos, D. (2017). Joint Models for Longitudinal Counts and Left-Truncated Time-to-Event Data with Applications to Health Insurance. *SORT: Statistics and Operations Research Transactions*, Accepted.

ABSTRACT

Health insurance companies accumulate a great wealth of historical data, related both to the intensity of health care usage made by policyholders and to the type of medical claims. Furthermore, the occurrence of death is also monitored, as well as a set of personal characteristics (gender, residence, etc.). In particular, the studies carried out in the sphere of private health care include the medical claims made by each individual over time, these being able to be used as an indirect measure of health status. At the same time, the aging process taking place in developed countries leads to an obvious interest in assessing the relationship between emergency care demand and survival rate, paying particular attention to those policyholders aged 65 and over. From a purely economic perspective, elderly policyholders need to be provided cost-effective premiums according to their individual health status, and insurance companies need to plan for the potential costs of dealing with lifetimes above the mean expectations. Theoretically, the amount needed to cover the life insurance costs of policyholders who require a great deal of emergency care should be compensated due to a lower survival rate, but this compensation is ambiguous due to the heterogeneity between subjects. Indeed, aging and mortality rates are influenced by subject-specific socio-economic and biological variables, which may vary considerably not only between individuals, but also dynamically within a single subject.

Consequently, there is a both medical and economic necessity to assess, in an individualized manner, how the medical demand of the elderly will evolve over time, as they will be the principal beneficiaries of additional medical resources due to the high prevalence of chronic diseases in this age range. On the one hand, pricing of health insurance is measured in terms of premiums, so the individual health status of elderly people must be considered in order to allow them to sign actuarially fair contracts. On the other hand, an insurance company providing retirement pensions and health insurance needs to plan for unexpected costs derived from people having lifespans above mean expectations. Under this interdependency scheme, the joint models for longitudinal and time-to-event data proposed in this thesis provide useful tools to properly address the underlying relationship between the emergency medical demand and the hazard for death event.

This thesis makes a contribution to the statistical methodology in the field of joint modeling techniques, which is applied to a large longitudinal dataset, the HI dataset, provided by a Spanish medical insurance company. From this dataset, we collect those subjects aged 65 and above, living in the city of Barcelona (Spain). For each subject, we have the historical emergency claims information within the study window, as well as the time-to-event information. The longitudinal outcome is of discrete nature, usually restricted to a small range of non-negative integer values which are affected by some degree of overdispersion; that is, the observed variance exceeds the mean. Additionally, counts with a large number of zeros become quite common owing to the nature of health insurance data, in which a lack of information about health status exists for some subjects. Another important characteristic concerning the cohort under study relies on the fact that only policyholders who have reached the age of 65 come under study, so all those entering after this age are considered

as delayed entries, and their time-to-event data are subject to left truncation in addition to the potential right censoring (considered as non-informative) from a certain date onwards.

Then, the implemented joint models must account for the special characteristics of our observed longitudinal response, departing from the common Gaussian responses, together with the specific time-to-event data pattern. This involves a great methodological challenge, as well as demands a huge computational effort, considering the large sample size.

In particular, there are three main tasks to be carried out by the joint analysis of the two outcomes considered:

1. We aim to implement a joint model which allows for the handling of longitudinal counts, also considering the potential overdispersion present at a subject-specific level by means of the specification of a model which considers an excess zeros (zero inflation). Moreover, survival times can also be subject to both left truncation and right censoring, being these features non-informative.

2. We want to assess the functional form to instantaneously associate, in a personalized manner, the expected longitudinal response with death risk. In this regard, we investigate the effect of the cumulative longitudinal response on the current death hazard.

3. As a central focus of this thesis, we propose the existence of a time-dependent relationship between the longitudinal process and the time-to-event outcome. This relationship is defined using penalized B-splines in order for any specific shape to be conferred.

All the analyses included in this thesis have been implemented under the Bayesian framework, in the R and JAGS free-software environments. The software codes are available from the author upon request.

RESUMEN

Las compañías aseguradoras médicas atesoran una valiosa cantidad de datos históricos, relativos tanto a la frecuencia de demanda entre sus asegurados, como a las peticiones médicas que éstos realizan. Además, la muerte del cliente también se registra, así como un conjunto de características de tipo personal (género, población de residencia, etc.). En concreto, los estudios realizados dentro del ámbito de la salud privada recogen las peticiones médicas efectuadas por cada individuo a lo largo del tiempo, pudiendo ser utilizadas como medida indirecta de su estado de salud. Paralelamente, el envejecimiento poblacional que tiene lugar en los países desarrollados conduce a un interés obvio en evaluar la relación existente entre la demanda de peticiones médicas de urgencia y la tasa de supervivencia, con especial atención sobre el grupo de asegurados con edad igual o superior a los 65 años. Desde una perspectiva puramente económica, los asegurados de mayor edad necesitan disponer de tarifas adecuadas a su particular estado de salud, mientras que la compañía aseguradora tiene que planificar aquellos costes adicionales generados por los asegurados que alcanzan edades por encima de las expectativas de vida media. En teoría, la cantidad de dinero que tiene que pagar la aseguradora en el caso de clientes que requieren una gran atención médica de urgencia habría de quedar compensada por una menor tasa de supervivencia de éstos, pero esta compensación resulta ambigua debido a la heterogenidad entre individuos. De hecho, el envejecimiento de una determinada persona depende de factores socio-económicos y biológicos que son inherentes a cada individuo, pudiendo variar considerablemente no sólo entre diferentes individuos, sino también dentro de un mismo sujeto a lo largo del tiempo.

En consecuencia, existe una necesidad médica y económica en poder evaluar, de manera personalizada, la evolución temporal de la demanda de servicios médicos dentro de los individuos asegurados de mayor edad, siendo ellos los principales beneficiarios de esta inversión adicional en recursos médicos debido a la alta prevalencia de situaciones de cronicidad en este rango de edades. Por un lado, el coste de los servicios médicos se mide conforme a las tarifas de la aseguradora, de forma que el estado de salud de un individuo debe considerarse a la hora de firmar contratos justos en términos actuariales. Por otro lado, una compañía que proporciona planes de pensiones y servicios médicos ha de tener en cuenta los costes que se producirán debido a un incremento de la esperanza de vida. Bajo este esquema de interdependencia, los *joint models* para datos longitudinales y de supervivencia propuestos en esta tesis constituyen una herramienta útil para estimar la relación subyacente entre la frecuencia de demanda médica de urgencia y el riesgo de mortalidad.

Esta tesis realiza una contribución en la metodología estadística en las técnicas de *joint modeling*, habiéndose aplicado sobre una extensa base de datos longitudinales, HI dataset, proporcionada por una compañía de seguros médicos de ámbito español. De esta base se consideran los individuos con edad igual o superior a los 65 años y residentes en la ciudad de Barcelona (España). Para cada sujeto se tiene la información histórica de peticiones médicas de emergencia durante el periodo de observación, así como la información referente a su tiempo de vida. La respuesta longitudinal es de tipo discreto, estando habitualmente restringida a un pequeño rango de valores enteros no negativos, afectados por un cierto nivel

de sobredispersión; es decir, la varianza observada excede el valor de la media. Adicionalmente, suele ser bastante habitual una gran presencia de registros nulos debido a la propia naturaleza de los datos de conteo en el campo de los seguros médicos, donde a menudo existe una falta de información relativa al estado de salud de ciertas personas. Otra importante característica relativa a la cohorte analizada reside en el hecho de que únicamente aquellos individuos que alcanzan la edad de 65 años son incorporados al estudio, de manera que aquellos que acceden en edades posteriores son considerados como entradas tardías. En consecuencia, sus tiempos de supervivencia quedan truncados por la izquierda, además de poder estar sometidos a una censura por la derecha (de tipo no informativo) a partir de una determinada fecha.

Así, los *joint models* implementados deben de considerar las características especiales de nuestros datos longitudinales, alejados de la habitual respuesta gaussiana, junto con el patrón específico de los tiempos de supervivencia. Ello supone una gran reto metodológico, demandando igualmente un enorme esfuerzo computacional motivado por el uso de una extensa base de datos.

En particular, se pueden distinguir tres grandes tareas metodológicas en el análisis conjunto de las dos respuestas consideradas:

1. Implementar un *joint model* que permita la inclusión de procesos de conteo, considerando la potencial sobredispersión en el conjunto de respuestas observadas para cada individuo mediante un modelo que considere un exceso de ceros (*zero inflation*). Además, los tiempos de supervivencia pueden estar sujetos tanto a un truncamiento por la izquierda como a una censura por la derecha, siendo ambos fenómenos de tipo no informativo.

2. Evaluar una forma funcional adecuada para asociar, de forma personalizada y en un instante de tiempo, la repuesta longitudinal esperada con el riesgo de mortalidad. En este punto, se investiga el efecto que tiene la respuesta longitudinal acumulada en el riesgo de mortalidad actual.

3. Como parte fundamental de esta tesis, se considera la existencia de una asociación dependiente del tiempo entre el proceso longitudinal y la respuesta de supervivencia. Esta relación temporal se define por medio de B-splines con penalizaciones, permitiendo así que a priori pueda adoptar cualquier tipo de forma.

Todos los análisis incluidos en esta tesis han sido implementados mediante el esquema de trabajo bayesiano con los programas estadísticos de libre acceso R y JAGS. Los códigos de *software* están disponibles mediante su petición al autor.

RESUM

Les companyies asseguradores mèdiques atresoren una valuosa quantitat de dades històriques, relatives tant a la freqüència de demanda entre els seus assegurats, com al tipus de peticions mèdiques que aquests realitzen. A més, la mort del client també queda enregistrada, així com un conjunt de característiques de tipus personal (gènere, població de residència, etc.). En concret, els estudis realitzats dins de l'àmbit de la salut privada recullen les peticions mèdiques efectuades per cada individu al llarg del temps, podent ser utilitzades com a mesura indirecta del seu estat de salut. Paral·lelament, l'envelliment poblacional que té lloc als països desenvolupats condueix a un interès obvi en avaluar la relació existent entre la demanda de peticions mèdiques d'urgència i la taxa de supervivència, amb especial atenció sobre el grup d'assegurats amb edat igual o superior als 65 anys. Des d'una perspectiva purament econòmica, els assegurats de major edat necessiten disposar de tarifes adequades al seu particular estat de salut, mentre que la companyia asseguradora ha de planificar aquells costos addicionals generats pels assegurats que assoleixen edats per sobre de les previsions de vida mitjanes. En teoria, la quantitat de diners que ha de pagar l'asseguradora en el cas de clients que requereixen una gran atenció mèdica d'urgència hauria de quedar compensada per una menor taxa de supervivència d'aquests, pero aquesta compensació resulta ambigua a causa de l'heterogeneïtat entre individus. De fet, el procés d'envelliment depèn de factors socio-econòmics i biològics que són inherents a cada individu, podent variar considerablement no només entre diferents individus, sino també dins d'un mateix subjecte al llarg del temps.

En conseqüència, existeix una necesitat mèdica i econòmica en poder avaluar, de manera personalitzada, l'evolució temporal de la demanda de serveis mèdics dins dels individus assegurats de major edat, essent ells els principals beneficiaris d'aquesta inversió addicional en recursos mèdics degut a l'alta prevalència de situacions de cronicitat en aquest rang d'edats. Per un costat, el cost dels serveis mèdics es mesura en base a les tarifes de l'asseguradora, de forma que l'estat de salut d'un individu ha de considerar-se a l'hora de signar contractes justos en termes actuarials. D'altra banda, una companyia que proporciona plans de pensions i serveis mèdics ha de tenir en compte els costos que es produiran degut a un augment en l'esperança de vida. Sota aquest esquema d'interdependència, els *joint models* per a dades longitudinals i de supervivència proposats en aquesta tesi constitueixen una eina útil per estimar la relació subjacent entre la freqüència de demanda mèdica d'urgència i el risc de mortalitat.

Aquesta tesi realitza una contribució a la metodologia estadística en les tècniques de *joint modeling*, les quals s'han aplicat sobre una extensa base de dades longitudinals, HI dataset, proporcionada per una companyia d'assegurances mèdiques d'àmbit espanyol. D'aquesta base es consideren els individus amb edat igual o superior als 65 anys i residents a la ciutat de Barcelona (Espanya). Per a cada subjecte es té la informació històrica de peticions mèdiques d'emergència durant el període d'observació, així com la informació referent al seu temps de vida. La resposta longitudinal és de tipus discret, estant habitualment restringida a un petit rang de números enters no negatius afectats per un cert nivell de sobredispersió; és a dir, la

variància observada excedeix el valor de la mitjana. Addicionalment, acostuma a ser bastant habitual una gran presència de registres nuls degut a la pròpia naturalesa de les dades de compteig en el camp de les assegurances mèdiques, on sovint existeix una falta d'informació relativa a l'estat de salut de certes persones. Altra important característica relativa a la cohort analitzada resideix en el fet de que únicament aquells individus que assoleixen l'edat de 65 anys són incorporats dins l'estudi, de manera que aquells que accedeixen en edats posteriors són considerats com a entrades amb retard. En conseqüència, els seus temps de supervivència resten truncats per l'esquerra, a més de poder estar sometsos a una censura per la dreta (de tipus no informatiu) a partir d'una determinada data.

Així, els *joint models* implementats han de considerar les característiques especials de les nostres dades longitudinals, allunyandes de l'habitual resposta gaussiana, juntament amb el patró específic dels temps de supervivència. Això suposa un gran repte metodològic, exigint igualment un enorme esforç computacional motivat per l'ús d'una extensa base de dades.

En particular, es poden distingir tres grans tasques metodològiques en l'anàlisi conjunta de les dues respostes considerades:

1. Implementar un *joint model* que permeti la inclusió de processos de compteig, considerant la potencial sobredispersió en el conjunt de respostes observades per a cada individu mitjançant un model que consideri un excés de zeros (*zero inflation*). A més, els temps de supervivència poden estar subjectes tant a un truncament per l'esquerra com a una censura per la dreta, essent ambdós fenòmens de tipus no informatiu.

2. Avaluar una forma funcional adequada per associar, de forma personalitzada i en un instant de temps específic, la resposta longitudinal esperada amb el risc de mortalitat. En aquest punt, s'investiga l'efecte que té la resposta longitudinal acumulada sobre el risc de mortalitat actual.

3. Com a part fonamental d'aquesta tesi, es considera l'existència d'una associació depenent del temps entre el procés longitudinal i la resposta de supervivència. Aquesta relació temporal es defineix d'una forma flexible mitjançant la consideració de B-splines amb penalitzacions, permetent així que a priori pugui adoptar qualsevol tipus de forma.

Totes les anàlisis incloses en aquesta tesi han estat implementades mitjançant l'esquema de treball bayesià amb els programes estadístics de lliure accés R i JAGS. Els codis de *software* estan disponibles mitjançant la seva petició a l'autor.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND GOALS

1.1 Population Aging in the European Union

Several European countries have experienced a significant growth in their elderly population over the last 50 years, and for the moment this process does not appear to have reached its peak. The process of population aging among senior citizens is primarily related to medical and technological advances in health and longevity, which have allowed for an increase in life expectancy by means of curing or controlling diseases which in the past had no treatment. In medical terms, these health improvements have enabled a large number of subjects to transition from having major incurable diseases towards a chronic but manageable status, which tends to be long-term. Then, in many cases the extension of the lifespan of the elderly takes place at the expense of a tendency to face a greater number of years with a potential range of health problems, which require being attended to within the health system of each country by means of an adequate allocation of economic resources. Particularly, a wide interest arises regarding those subjects aged over 65 years, since that threshold has been commonly assumed as the statutory retirement age, and consequently, is taken as the reference point from which the population is designated as elderly. In general, the risk of multimorbidity (i.e. suffering from more than one chronic condition at the same time) becomes higher with age (Salisbury et al., 2011; Fabbri et al., 2015). At least 60% of the European Union states (EU-28) population reaching retirement age will have at least two chronic conditions in the coming (WHO, 2011). As pointed out by Koller et al. (2014), multimorbidity is related to a higher risk of care dependency, which inevitably leads to an increase in demand for medical care and long-term care services.

Parallel to the aging process, European countries have registered a steady decline in birth rates since the middle of the 20th century. In particular, the average birth rate registered in 2015 was 1.58 children per woman, which is below the population replacement birth rate for most industrialized countries, recently stated at 2.1 births per woman. Consequently, the progressive growth of the number of elderly people has taken place at the same time as its total size increased in comparison to the total population, affecting not only the segment recently reaching the retirement age, but also people at very old aging stages. In the particular case of the states from EU-28, the rate of elderly population growth has been particularly high since the early 1990s, when the percentage of subjects over 65 rose above 15%. Thus, the number of inhabitants 65 and over increased from 10.4% (424.7 million total) on January 1, 1965 to 18.9% (508.50 million total) on January 1, 2015.

Figure 1.1 shows the evolution over time of the EU-28's elderly rates during the period 1965-2015, and it is compared with the same rate in different countries around the world (OECD, 2016). In this regard, due to its key role in this thesis, we include the particular case of Spain, which presents a similar case to that summarized by the EU-28 (Garin et al., 2014). We can also infer how Chile, stated as one of the fastest-growing Latin American economies, presents half of the percentages exhibited by the EU-28. In Japan, by contrast, people over the age of 65 make up a quarter of the total population. Finally, the United States of America, with the largest economy in the world, has an elderly population halfway between EU-28 and Chile, benefiting from policies favorable to the birth rate since the end of the last century. From the results, we can see that European trends are mimicked in other parts of the world, although different socioeconomic factors in each country lead to different growth rates, even among EU-28 states.

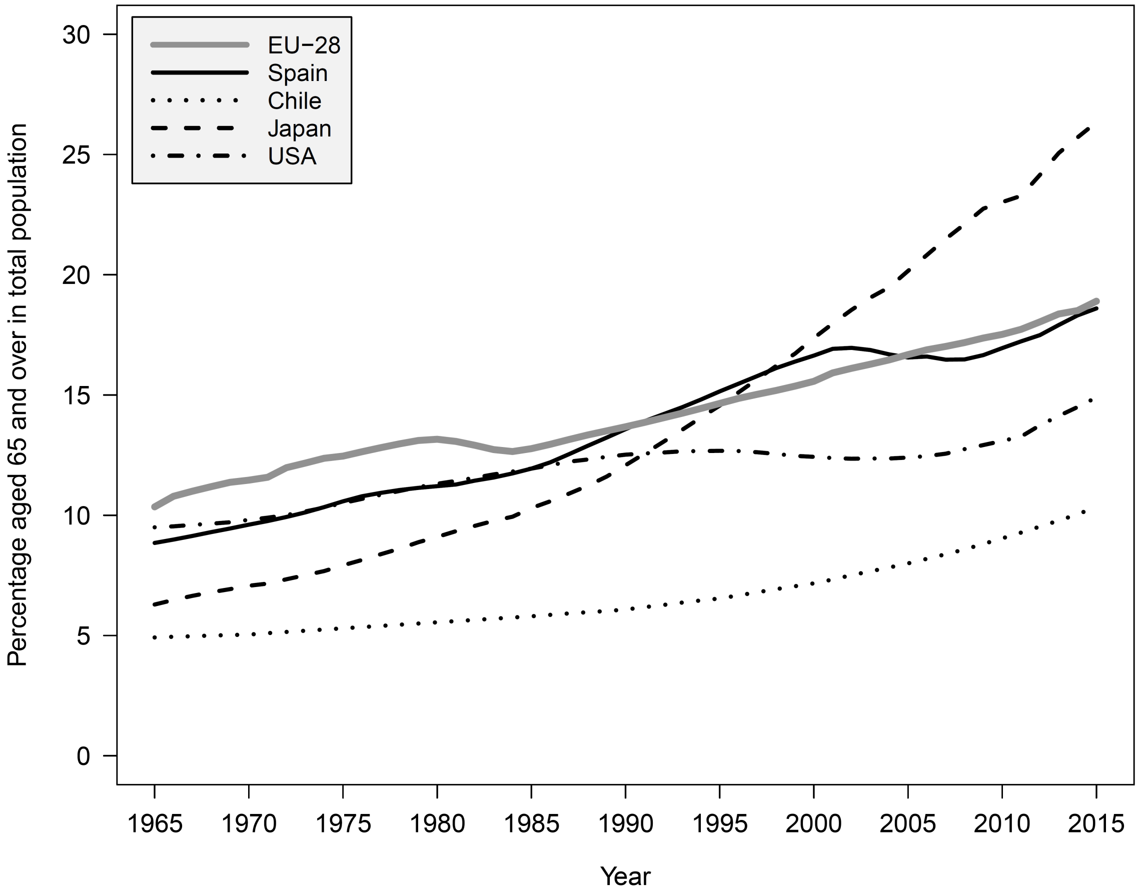


Figure 1.1. Comparison between the evolution of EU-28 elderly rates with some industrialized countries during the period 1965-2015. Source: Data from OECD (2016).

The increase in elderly population, *per se*, would not necessarily entail important changes in the economic structure of a country's health coverage, as long as the generational turnover is maintained. However, in the case of EU-28, the aforementioned increase has been accompanied by a stagnation, if not decline, of the younger population demographics. As an example, the median age of the EU-28 population has risen by 3 years in the last decade, shifting from 39.5 years in 2005 to 42.4 years in 2015 (Eurostat database 2016, `http://ec.europa.eu/eurostat/data/database`).

While it is true that population aging has affected each of the European countries, its evolution over time has not uniformly affected the EU-28 as a whole, reaching higher elderly rates specifically in those countries with the five largest industrial economies: Germany, France, the United Kingdom, Italy and Spain. These five countries each present elderly rates above 16.0% of the total population, as depicted in Figure 1.2.

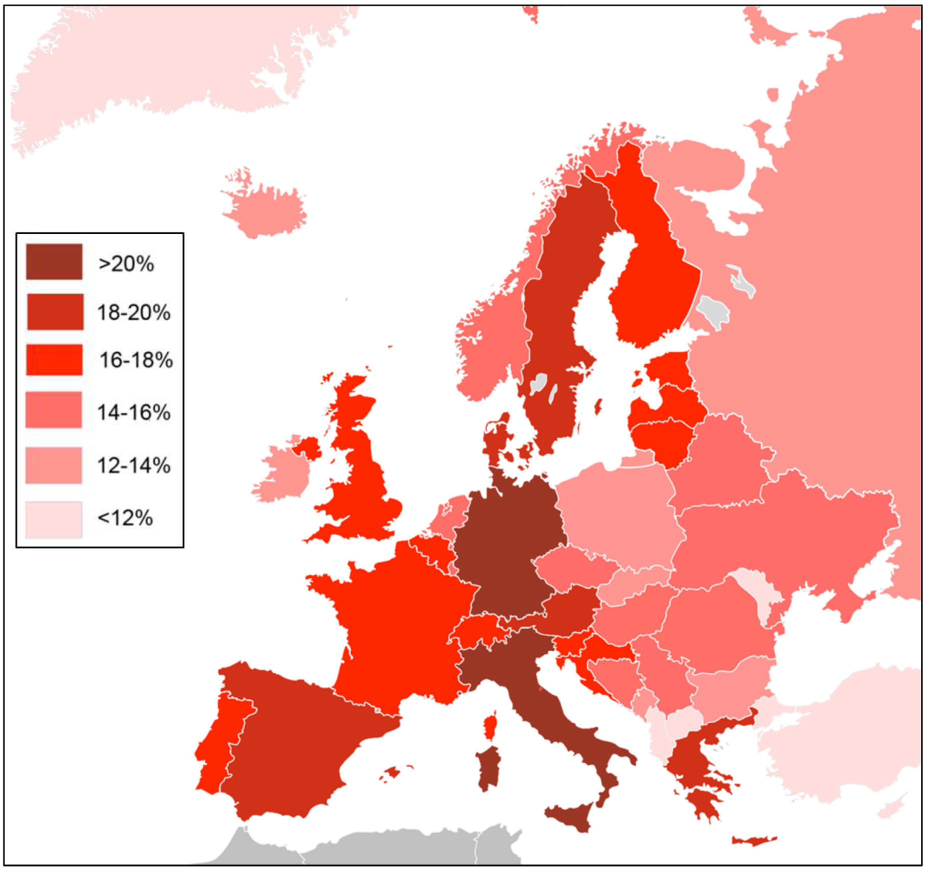


Figure 1.2. Percentage of the total population over 65 years old across the EU-28 member states in 2015. Source: `http://www.wikiwand.com/en/Ageing_of_Europe`

The European aging trend is forecasted to continue rising, although at a slower pace, in the long run. By January 1, 2080, the number of inhabitants is projected to reach 520.1 million, and the median age of the EU-28 population is expected to rise to 46.4 years (Eurostat, 2016). Regarding the percentage of the elderly, it is projected to increase from 18.9% in 2015 to 28.7% in 2080, so Europe as a whole will become an aging society in the widest sense. Aging societies, also named graying societies, can be regarded as a direct consequence of low birth rates combined with higher life expectancy among populations. As a result, the median age of a society increases, and that logically translates into a higher proportion of older people and relatively fewer working-age adults. This demographic revolution poses many challenges in the funding of health care systems, requiring the rethinking of current policies in order to achieve sustainable elderly care in the coming decades.

1.2 Joint Models to Assess Longevity Risk in Health Insurance

In the particular context of health insurance, the demographic shift to a higher life expectancy and lower fertility rates has greatly increased the relative medical demand among the elderly. The increasing usage of private care also extends longevity even further, so that the health insurance system is greatly challenged by those subjects aged over 65. They are indeed in the age range where experiencing critical diseases is most probable, which, at the same time, is usually associated with higher costs. From a purely economic perspective, living longer means additional costs above the expected average, something usually referred to as longevity risk. These two phenomena (longer lifespans and more demand for medical care) have traditionally been studied independently. For example, D'Amico et al. (2009) analyzed a portfolio of policyholders that were covered for disability and found that survival rates could not be separated from impairment conditions. Indeed, standard actuarial methods of health insurance (Yue and Huang, 2011; Pitacco, 2014; Ericson and Starc, 2015) even ignore correlated information about their subjects. Theoretically, savings in emergency care due to a better quality of life should be larger than the increase in the amount needed to cover life insurance costs (Dao et al., 2014), but this compensation is ambiguous due to the heterogeneity between subjects. Indeed, aging and mortality rates are influenced by socio-economic factors, biological variables and health conditions, which may vary considerably not only between subjects, but also dynamically within a single subject. Consequently, there is a need to know, in an individualized manner, how the medical demand of the elderly will evolve over time, as they will be the principal beneficiaries of additional medical resources. On the one hand, pricing of health insurance is measured in terms of premiums (which are related to rising medical expenditures), so the individual health status of elderly people must be considered in order to allow them to sign actuarially fair contracts. On the other hand, an insurance company providing retirement pensions and health insurance needs to plan for unexpected costs derived from people having lifespans above mean expectations.

Health insurance companies accumulate a great wealth of historical longitudinal data on the intensity and type of health care usage through claims made by policyholders. As part of the longitudinal follow-up process, the occurrence of mortality is also monitored, as well as a set of personal characteristics. The combination of this information provides valuable medical information collected at different ages of the same subject, and allows for the assessment of the degree of relationship of these records to their particular health status. From an insurance point of view, it can be extremely valuable for quantifying their clients' medical care demand risks and for predicting individual survival probabilities.

Building on this insight, a basic time-to-event or survival analysis would incorporate longitudinal time-dependent information by means of a time-dependent Cox model (Andersen and

Gill, 1982); however, this assumption is not realistic since it assumes that subject health status remains constant between subsequent measurements, and may lead to biased inferences (Prentice, 1982). In order to ensure that longitudinal information is adequately incorporated into a survival model, a simultaneous approach to both processes is required. Then, joint modeling of longitudinal profiles and time-to-event (also called survival) data stands as the natural field of study for simultaneously analyzing informative dropout (from repeated measurements across time) and event outcomes. Interest in the application of joint models in biostatistics and medical research has dramatically increased in the past two decades, leading to the proliferation of statistical studies that use these models for different types of data. However, the application of joint models in fields other than biostatistics, such as insurance, remains practically unpublished to this day, with few recent exceptions. Among the main reasons for this lack of use in insurance research, we highlight: a) the enormous computational cost that joint modeling techniques require for the large datasets handled by insurance companies, b) the usual departures from the normal distribution in the data, which establish additional difficulties in the already complex data modeling process, and c) in contrast with the biostatistics field, in which a biomarker with strong prognostic capabilities is typically identified, in the case of insurance research, such a clear reference for the main longitudinal outcome does not always exist.

In our particular case, the different modeling difficulties can be solved working under the Bayesian framework, while the demand for medical emergency services is adopted to account for the deterioration of elderly subject's health. We want to assess whether emergency demand and health status have a significant association from the age of 65 onwards, and if so, to evaluate this underlying relationship over time. In general terms, we can assume that a high demand for emergency services will be associated with a poorer health status, and consequently, with a higher mortality risk. The key point relies on the fact that each subject's particular aging process is related to a set of random biological components, so that it is necessary to take into account each subject's longitudinal information (repeatedly measured observations over time) for an accurate assessment of the corresponding mortality risk. The relationship between emergency claims in a subject's medical history and survival responses can be properly addressed in a single model, called a joint model (JM) for longitudinal and time-to-event data, where the association between both outcomes at each time point is traditionally expressed by means of a constant parameter. However, we argue in this thesis that a specific demand for emergency medical services does not necessarily translate into the same increment in mortality risk at any given age within the elderly segment. On the contrary, the mortality risk due to a specific pathology will depend on, among other factors, the age of the person affected. In addition, we must keep in mind that at certain ages the occurrence of chronic illnesses increases; this can lead to peaks in demand for emergency medical services that do not necessarily correlate with a lower survival rate.

We seek a personalized method in order to provide us with knowledge about the risk pattern, thus allowing each subject to negotiate the acquisition of more cost-effective annual premiums according to individual health status. In turn, insurers will have personalized health information about each of their customers, thus being able to accurately allocate the required capital to face the underwriting obligations for each of them.

1.3 Literature Review on Joint Modeling Framework

Although there are several possible approaches to jointly modeling both information sources, the most commonly used one is carried out under the shared-parameter models (SPM) framework. Complete explanations can be found in Wu and Carroll (1988), Little (1995) and Molenberghs and Kenward (2007). The key issue of SPM relies on the conditional independence hypothesis, so that dependence between longitudinal responses and mortality risk is conducted using a common latent structure, described by a set of subject-specific random parameters. Then, given these shared random effects, longitudinal and time-to-event outcomes are independent, as are repeated measurements in the longitudinal process. In such a scheme, the subject-specific longitudinal history can be included as covariate information in a classical proportional hazards (PH) model, so that both processes are linked at each time point by an appropriate association structure.

In the seminal formulation of the JM, the longitudinal hierarchical response is normally distributed, and the time-to-event response is introduced by a proportional hazards model (Tsiatis et al., 1995; Faucett and Thomas, 1996; Wulfsohn and Tsiatis, 1997). The two processes are linked by normally distributed random parameters, thus relating for each subject the current expected outcome of emergency demand at a specific time point to the death hazard or mortality risk (both terms are used interchangeably throughout the different chapters of this thesis). Joint models provide an efficient method not only to assess the relationship between both submodels, but the combined approach also avoids biased estimates from each submodel, as well documented in, for example, Tsiatis and Davidian (2001) and Fieuws et al. (2008). Hence, a JM allows for informative dropout when the longitudinal process is of primary interest, whereas the precision of the estimates of the survival parameters improves when handling the time-to-event process. Finally, these models are an extremely valuable tool in order to assess the relationship between both outcomes.

Since the initial definition of the standard JM, the academic interest in these models has continued to grow. Several case studies followed the seminal articles, most of them aimed at analyzing personalized biological patterns using the relationship between a specific biomarker and the time remaining until the event of interest. Among these earlier contributions, we can find both frequentist and Bayesian strategies to obtain the parameter estimates. One extended approach is the maximization of the joint likelihood. Some key references can be found in Henderson et al. (2000), Yu et al. (2004), Zeng and Cai (2005), and Ding and Wang (2008). A complete account of different functional forms to assess the relationship between longitudinal and survival outcomes is provided by Rizopoulos (2012), in a text book format. Alternatively, other authors opted for a joint model approach under a Bayesian paradigm, as in Xu and Zeger (2001), Wang and Taylor (2001), Brown and Ibrahim (2003) and Ibrahim et al. (2004). An interesting alternative extension of standard joint models is postulated by the so-called latent class joint models (Lin et al., 2002; Proust-Lima et al., 2009), where each subject from a target population is assumed to belong to one and only one latent subpopulation, so that longitudinal and time-to-event outcomes are associated with the corresponding latent class indicator. This joint model approach is particularly useful when focusing on subject-specific predictions, since they allow for very flexible association structures.

With the aim of seeking application to further types of data, the main research lines in the recent past have focused on incorporating the necessary flexibility in the different parts of the JM, i.e., longitudinal response, survival time and the way in which the associative structure between both components is defined. Among the different extensions implemented when tackling the longitudinal outcome, we can point out the inclusion of non-Gaussian longitudinal outcomes (Murawska et al., 2012; Viviani et al., 2012; Rizopoulos, 2016) and the case of multivariate longitudinal outcomes (Song et al., 2002; Brown et al., 2005; Rizopoulos and Ghosh, 2011; Andrinopoulou et al., 2014). More recently, Ivanova et al. (2016) formulated a JM to handle different types of response, i.e., continuous, discrete and ordinal. In the case of time-to-event processes, research has focused on tackling situations such as survival data with competing risks (Elashoff et al., 2008; Williamson et al., 2008; Li et al., 2010; Huang et al., 2011; ProustLima et al., 2016), or cases in which time-to-event data are left-truncated (Piccorelli and Schluchter, 2012; Su and Wang, 2012; Crowther et al., 2016). Finally, a few extensions are worth special attention. These are all those devoted to increasing prognostic capacity of joint models to distinguish between high-risk subjects and those who are more likely to survive, thus making a more reliable personalized prediction for either the longitudinal or the survival outcome. Within this framework, further details on prediction assessment in joint models can be found, among others, in Proust-Lima and Taylor (2009), Rizopoulos (2011), and Sweeting and Thompson (2011).

A key benchmark in the classical joint model formulation is that the association parameter between longitudinal responses and the death hazard is assumed constant across time. However, in our particular problem, we observe that emergency claims peak around 85-90 due to chronic diseases, so that the real impact on survival at this age range is not expected to be as large as at other ages. Consequently, a more realistic approach to describe the underlying relationship between the two responses is achieved by departing from a constant association parameter. Specifically, we propose that the longitudinal response has a time-varying effect on death hazard.

1.4 Motivating Dataset: An Example of the Spanish Situation

We have been provided health care data by a large Spanish insurance company (Piulachs et al., 2016), consisting of a large cohort of subjects aged 65 and over, living in the city of Barcelona. They have a health insurance policy, so they have the right to receive private medical assistance under the conditions of their insurance contract. Regarding the longitudinal information, the individuals included in the study are analyzed over a period of eight years, from January 1, 2006 to February 1, 2014, and we focus on annual records of those claims directly related with subject death hazard, which we refer to as emergency claims per year. These are indeed the largest contributors to the economic costs forecast, and in addition serve to explain the subject's health prognosis in a more definitive way. In this regard, emergency medical claims can be summarized by ambulance services, hospitalizations and non-routine medical visits. So, our longitudinal outcomes are integer values, ranged from 0 to 20, and are affected by some degree of overdispersion, induced by between-subject unobserved heterogeneity and an excess of zeros. At this point, the demand for emergency medical services is a rather infrequent event in elderly people, especially if we consider ambulance services and hospitalization in particular. Moreover, the Spanish public health system offers universal coverage, so many of the policyholders can opt to access public health services, leading to a zero count which does not reflect the real health status of the subject.

The insurance company also provides an additional set of personal information where the age at which each subject enters the study is recorded, as well as their lifespan inside the window study. Hence, we have two types of observations:

- If the subject's death is recorded during the study period, we can know the age of the subject at which the event of interest occurs.
- If the subject's death is not observed, we can know either the age of subject at the study's closing date, established on February 1, 2014 due to administrative reasons, or the age at which the subject has an early dropout, here assumed to be caused by reasons unrelated to the event of interest. In both cases all we know is a certain period of time during which the subject is still alive, but we do not have any further information. These situations correspond to a non-informative right censoring.

In addition, in our particular case the age of 65 is posed as time zero, so all subjects entering the study after that threshold age are considered as delayed entries, and, consequently, their time-to-event data are left-truncated.

1.5 Scope and Specific Goals of the Thesis

The research conducted in this thesis is motivated by the current demographic challenges that must be addressed by European private health insurance companies, where aging has exposed the sector to unexpected financial costs. In particular, we focus on the Spanish situation, where advances in clinical knowledge have enabled the extension of customers' lifespan above mean expectations, so insurers must adjust their premiums in order to reflect the risks expected at very old ages. In our particular research, we analyze real health data containing longitudinal and time-to-event information of elderly policyholders in the city of Barcelona. The main aim of this research is to find the key patterns regarding the mortality risk. To achieve this, we use a simultaneous approach to longitudinal and survival processes, which leads to further extensions of the standard JM.

Specifically, there are three main challenges to face in our analysis:

1. We seek an adequate model to accommodate correlated counts observed in the longitudinal outcome, taking into account the potential overdispersion at subject-specific level, that is, when the within-subject variability is larger than the mean. The two main causes of overdispersion derive from an inherent heterogeneity among measurements and an abundance of zeros. Additionally, time-to-event data should account not only for the usual right censoring, but also for the left truncation caused by the late entry into the study of a large percentage of subjects.

2. We analyze the adequate functional form to relate the subject's claim history within the study window to death risk. Standard joint models assume a constant relationship between the current expected value and the survival rate, but in our case it does not seem reasonable to summarize the health status by only considering the longitudinal information from a single time point. Instead, we can consider the impact of past health status on the current death hazard. Moreover, all past medical information does not have the same importance; the closer measurements are to the current time, the more weighted their consideration should be compared to those that are more distant.

3. As a main issue in this thesis, we want to incorporate a time-varying association parameter between longitudinal and time-to-event outcomes, hence allowing for a more flexible relationship between emergency demand and death hazard. This point becomes essential in the insurance field, since the result of this connection is the one which may prove that expected costs from subjects with a higher emergency demand are compensated with lower survival rates.

The rest of the thesis is organized as described in the following paragraphs:

Chapter 2 presents a description of the motivating dataset that has been used in this thesis, which was provided by a Spanish medical insurance company. It consists of 5470 subjects, aged 65 or above, who signed a health insurance policy. Consequently, they have the right to request for private medical assistance during the study period, 2006-2014. For each subject, the data contain repeated measurements of emergency claims in a year (count rates), as well as left-truncated and right-censored survival information, thereby providing valuable and useful information about the current state of Spain's private health care. The goal within the insurance field is to relate both processes in order to asses the longevity risk in a personalized manner.

Chapter 3 introduces the standard formulation of joint models for longitudinal and time-to-event data. The application of these models to our longitudinal information is first carried out by applying a logarithmic transformation to the count rates. Moreover, we outline the special features of our time-to-event data, focusing on left truncation and right censoring, and how the inclusion of both issues must be considered in order to avoid biased inferences in estimating survival parameters. The simultaneous approach to longitudinal and event times is carried out under the shared-parameter JM framework, in which traditionally the expected longitudinal outcome and the event hazard (and, consequently, the time-to-event) are instantaneously related using a constant association parameter. In addition, we extend the aforementioned relationship to account for the weighted cumulative effects on the event hazard. Thus, we also analyze a constant relationship between the recency-weighted area under the expected longitudinal profile until a time point and mortality risk at this point. The parameter estimation is performed under the Bayesian framework using Markov chain Monte Carlo methods.

Chapter 4 outlines the necessary concepts for the inclusion of counting outcomes in the longitudinal submodel of the Bayesian JM. Initially, the Poisson mixed-effects model is introduced, this being the most used probability distribution to accommodate panel discrete data in longitudinal studies. Since the real data are usually affected by some degree of overdispersion, the negative binomial mixed-effects model is also introduced. This model represents a further step when dealing with counting data, emerging as the simplest way to account for the overdispersion effects in the observed responses. Regarding the time-to-event submodel, the main outcome is subject to both left truncation and right censoring. Building on these assumptions, longitudinal and survival responses are joined into a single statistical model by a common set of random effects. The resulting JM clearly departs from the standard approach, and its consideration allows for a more accurate modeling of our data.

Chapter 5 presents a Bayesian JM to relate, with a time-varying association parameter, the emergency claims per year with left-truncated lifetimes. Specifically, this is achieved by allowing a flexible shape for the unknown association structure by its expansion into B-splines with discrete penalties, namely P-splines. The counting sequence is undertaken by a hierarchical zero-inflated response, which includes a mixture of a point mass at zero and a classical counting distribution. The zero-inflated models considered are the zero-inflated Poisson and the zero-inflated negative binomial models with random effects. The time-to-event outcome is, as in previous chapters, subject to both left truncation and right censoring. The main results of the JM with a time-varying association are reported, and its benefits with respect to a constant link are commented.

Chapter 6 illustrates how the joint model approach provides dynamic predictions of survival probabilities. In particular, the predictions are obtained from the JM approach which considers the recency-weighted effects of a longitudinal count response. From the subject-specific predictions for a theoretical subject, we provide an illustrative example of how emergency demand and longevity have opposite effects on health insurance rates.

Finally, **Chapter 7** contains a discussion focused on the impact of emergency claims on death hazard, and we also comment the potential areas of future research.

CHAPTER 2
THE HEALTH INSURANCE DATA

2.1 Obtaining the Dataset From External Sources

In this chapter, we introduce our health insurance data, the HI dataset, which has ultimately motivated the different methodologies and statistical models that are studied in this thesis. The information encompassed a large number of contracts from subjects who signed a health care policy with a Spanish insurance company, and therefore had the right to receive private medical coverage during the period in which our study takes place, from January 1, 2006 to February 1, 2014. Due to privacy laws, confidentiality agreements were established between the health insurance company and the University of Barcelona, as a necessary prerequisite for academic uses of these data. All the contracts, each of them directly related to a specific subject, were randomly assigned a unique and anonymous identifier code by the company. The contract information given by the company was originally split into multiple files, so we undertook the difficult process of data linking to store all the information in a single file.

One the one hand, we were provided a claims file, which essentially reported the historical medical claims recorded within each contract during the study window, as well as different types of personal information associated with each subject's contract, such as age, gender, residence and length of time with company. This claims file was arranged in the long-format, so the repeated measurements collected within each contract were stored in multiple lines. On the other hand, a time-to-event file was also given by the company, providing information about insurance dropouts and the reasons behind them, or whether the subject's death occurred before the end of the study. An exhaustive scrubbing process was conducted in both data sources by removing incomplete, improperly formatted, or duplicated records, as well as that information occurring outside the established study period. The algorithms implemented to undertake all these tasks also considered the identification and correction of missing data, redundant information, and the potential existence of contradictory records. After this scrubbing, the final dataset was obtained through the appropriate combination of the information contained in both files.

The whole process for obtaining the HI dataset entailed a huge programming and computational effort, given the high volume of observations. Basically, the main tasks that we had to carry out in order to obtain the final dataset consisted of the following:

1. Importing the claims information.

2. Importing the time-to-event information.

3. Merging different data sources into a single file.

4. Selection and scrubbing of the medical claims to perform our research.

5. Construction of a longitudinal mesh which is used to collect subjects' measurements.

All the above steps and the challenges related to their proper implementation are briefly summarized in the next subsections. The different algorithms for scrubbing and adequately merging different data sources are implemented by means of the `R` software (R Core Team, 2017), and are explained in detail by the software code supplied in the Appendix A.

2.1.1 Importing the Claims File

In this first phase of data management, we consider all the available medical information during the 8.1-year observation period. For this, we take into account both routine and emergency medical demand observed within the contracts, each of which is referred to by a single and anonymous identifier, `id`. The longitudinal information of claims related to the same contract is organized in long-format. Hence, within each contract represented in the file, there were as many rows as claims observed during their monitoring. The type of medical claim is reported by means of three different codes: the general type of medical claim, `cfam`, the medical specialty, `cspe`, and the specific medical service, `cclaim`. In addition, the claim date occurrence, `dclaim`, is included in the resulting file, as well as different variables related to the subject of the contract, such as the birth date, `dborn`, the gender, `sex`, and the geographical information, such as the postal code `cp` or the municipality `town` (where unusual characters and punctuation in place names were removed).

We only select those contracts of subjects living in the city of Barcelona (Spain) to guarantee population homogeneity. This results in a total of 33 311 different contracts, whose information was spread across 2 162 538 observations (rows). The data importation is conducted in the `R` environment using the package `data.table` (Dowle et al., 2017). This package offers important advantages over the traditional `data.frame` format when manipulating big data, requiring a lower computational demand in combination with simple syntax (further details on this package can be found in `http://r-datatable.com`).

2.1.2 Importing the Time-to-Event Information

The provided file regarding the information of lifetimes contains one single observation for each policy contract, using the same identifier coding system (`id`) as in the claims file to denote the contract held by a specific subject. The starting date of the contract with the insurance company is supplied through the variable `dini`, as well as the date on which contract's progress of medical demand stops being observed, `dfinal`. In this regard, the company provides, for each contract, the cause due to which the monitoring ends. This allows us to introduce a dichotomous variable, `status`, indicating in each case if the conclusion of follow-up is due to the occurrence of the event of interest. So, we have `status = 1` in all those observations associated with contracts that have left the study due to the death of the subject, implying the occurrence of an event, whereas we have `status = 0` if a contract's follow-up interval has concluded with the subject still being alive, leading to administrative censoring. In our case, most of the censored data derive from the closing date of the study, but some of them come from loss of follow-up after a known date due to causes not related to the subject's health status (mainly unpaid premiums and disagreements between the subject and company about the premium price).

The main task we face when managing the time-to-event file comes from the correct assignment of initial and final monitoring dates for each contract, as well as the corresponding causes of follow-up cessation. This difficulty arises from the fact that subjects in the study had the possibility of changing their policy terms from a certain date onwards, so each change in the contractual terms was kept in the file as another observation, but with a new identifier code. In these cases, the effective date of change was established as: a) the final follow-up in the previous subject's observation, which remains right-censored due to an "artificial" loss of follow-up (since the subject's contract is actually being followed with another identifier code), and b) the initial follow-up date in the new observation generated by the change in contractual terms. Therefore, there are contracts who have multiple dates associated for both enrollment in the company and for ending their follow-up interval. In all cases we keep the most conservative dates, by considering the oldest enrollment date, and also the oldest following-up date in the cases where the death is not recorded in any of the multiple observations.

Once the possibility of multiple contracts per subject is considered and adequately scrubbed, we get a file containing 145 742 single contracts, each of them uniquely associated with a specific subject of any age. Note that the final number of contracts is a much larger quantity than that obtained in the claims file (a difference of 112 631 contracts), due to the following reasons:

- The file which contains lifetime information does not report about residence, so we are actually dealing with contracts of subjects included living in any part of Spain (subjects living outside the city of Barcelona are later excluded).
- We are considering contracts who belong to subjects of all ages, thus including an important percentage of policyholders at younger ages who have not needed medical coverage yet, and consequently no historical information about them has been recorded in the claims file (in the final dataset, only subjects aged 65 and over will be considered).
- The provided contracts belong to subjects that may use a combination of private and public medical care, since the Spanish system offers universal health care coverage as a constitutionally-guaranteed right. Hence, a subject's contract could not have any record in the claims file, and yet the subject could have received medical care within the public system.

2.1.3 Merging Longitudinal and Survival Data: From Contracts to Subjects

The longitudinal information is merged with the time-to-event information, so from this moment on we are not dealing with contracts any more, but working directly with subject's information. To properly merge both type of data, algorithms are implemented to eliminate demand-related time inconsistencies present in the original information. In particular, we make sure that for each subject, all dates associated with medical claims, denoted by `dclaim`, are indeed placed within the time interval in which the subject is with the insurance company, delimited between the dates `dini` and `dfinal`.

Medical data collection across the study is conducted using a mesh of eight equally-spaced *control points*, respectively allocated at every year end throughout the study period. In this step, at each of these points we account for all types of medical records occurring during the immediately preceding year (later only emergency claims will be used). The establishment of these control points entails the annual discretization of medical demand observations for each subject, originally recorded by the insurance company in a continuous manner. We are then working with longitudinal profiles of subjects for whom repeated measurements of medical services are annually recorded at each of the years covered by their trajectory.

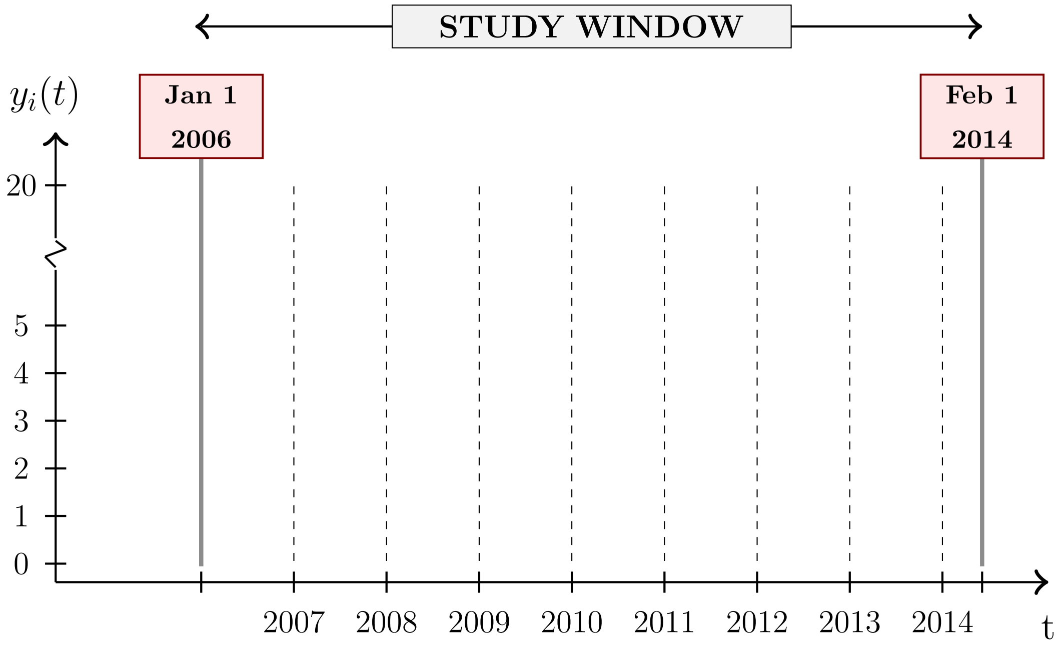


Figure 2.1. Mesh of control points to gather information about annual demand for medical services for each subject.

This response partitioning confers important advantages in dealing with longitudinal claims:

1. It subsequently allows to annually summarize the longitudinal demand, thus working with rates of claims (claims per unit of time) instead of raw claims. This is particularly useful in panel counts, where the time periods during which counts are recorded may vary. As

an example, let us consider two longitudinal profiles, respectively belonging to a couple of subjects who have the same gender and reach 65 years on exactly the same day, before the study begins. Let us additionally suppose that the first subject enrolled with the company on 1990/01/01 (i.e., subject's profile follow-up is observed from the beginning of the study), and the second one enrolled on 2006/07/01 (i.e., the subject is incorporated into the study after it has begun). Without taking into account their particular biological features, the effect of recording the same number of medical claims at the end of year 2006 from their corresponding longitudinal profiles will not have the same medical impact, since only the first profile has been followed-up during the whole year of 2006. Thus, a direct comparison working in raw claims would not be fair since identical information is collected about the second subject in half the time, so the importance of their measurements must be time-weighted accordingly (later, this will be properly addressed).

2. The selected time unit for each measurement is established as a year, thus coinciding with the reference period that insurance companies use to fix their premium pricing.

3. We establish only eight potential *starting points* into the study for each of the subjects, placed just after the corresponding control points. Working with a small number of points to enter into the study represent significant computational advantages given the volume of the data to be handled.

4. The admission date of each subject into the study is automatically reflected in their corresponding longitudinal profile by the control point at which the individual is enters the study, thus providing a longitudinal response. This response is usually assumed to remain constant until the next control point.

5. It proves easier to impose minimum follow-up conditions to allow a specific subject to enter into the study, establishing the constraint that a subject's longitudinal profile must have crossed, at least, one of the prespecified control points.

Once the system to collect all types of medical claims, the admission of each subject into the study is reflected by the date of the first control point in which an elderly 65 subject provides longitudinal information. In particular, each subject's entry date into the study is denoted by `dent`, assigned as the date of the first control point at which an specific individual is 65 or older, and has, at least, half a year of previous following in their medical demand profile. The time frame in which each subject is observed is delimited between the dates `[dent, dfinal]` (logically, the date `dent` is always after or equal to the date `dini`). At the end of this step, we have information about 5496 subjects and 531 580 observations.

2.1.4 Selection and Scrubbing Process of Medical Claims

For each of the previous subjects, we aim to record only those medical claims directly related to emergency care: ambulance services, hospitalizations and non-routine medical visits. Before focusing on selecting emergency claims, however, we have to take into account the subjects who do not have any emergency claim recorded. These subjects should be accounted for in the final dataset, and at this stage a single observation is incorporated into the dataset by assigning a particular medical code and a null value in the quantity of claims. The addition of such an "artificial" row to the claims file is the way to denote that these subjects will provide observations with zeros in the final dataset.

Once these zero values are recorded, we select the three general codes related to the corresponding type of claim we wanted to account for, and then we focus on a scrubbing process within each of these. The process of cleaning up each of the codes involves removing duplicated or inconsistent values which could distort the results of our analyses. In addition, the standardization of the way to account for a claim is another important point, since some of the claims units do not correspond to a real demand (for instance, the ambulance service was given in hours, whereas we are just interested in the request for this service). We keep the previous 5496 subjects, but we have now only 31 170 observations.

2.1.5 The HI Dataset

Within each subject, we add up the emergency claims recorded in the whole calendar year immediately before control point covered by subject's longitudinal profile, and we assign such amount of claims to the corresponding control point. In case of subjects without any record, a zero value is assigned. The specific amount of emergency claims recorded at each of subject's control point measurements is accounted for by the variable `claimyr`. Hence, the main longitudinal outcome in our study is defined as the emergency claims per year. We assume that each measurement associated with a control point will remain constant until the next control point. This *last observation carried forward* approach (LOCF) is the common working scheme for handling both longitudinal and time-to-event information. Further, the longitudinal profile of those subjects who had a single and artificial measurement without claims are expanded across their life span in the study window, thus systematically recording zeros at each of the control points crossed by their longitudinal profiles. In order to avoid anomalous results, of the 5496 people recruited so far, we removed the 19 subjects who have over 20 claims per year in any of their measurements, and also those 7 subjects whose follow-up interval started after having reached the age of 100 years. We finally retrieve the longitudinal and time-to-event information related to a homogeneous population of **5470 insured subjects aged 65 and over**, representing a total of **32 269 measurements** (note that here the number of measurements increases in comparison with the previous step).

However, not all subject-profiles start to be observed at the beginning of a specific calendar year. Consequently, as noted in Section 2.1.3, the time periods during which emergency claims are collected might be different. We therefore need to explicitly consider exposure time effects in order to avoid spurious estimates. This procedure is carried out by relating the amount of emergency claims observed at the end of a calendar year to the corresponding exposure time, thus taking into account the real period-at-risk in which the aforementioned amount is collected. The longitudinal value of the exposure time variable, denoted in the file by `expo`, ranges in our case from `expo = 0.5` (i.e., the subject's medical information is only recorded during half a year) to `expo = 1` (i.e., the medical information is recorded during the whole year). Finally, at the end of each subject's follow-up interval we know the value of dichotomous variable `status`, which informs either of the subject's death or that the individual is still alive.

The longitudinal information of the final dataset is stored in terms of start-stop coding, thus meaning that each subject's longitudinal observation of the response variable is assumed to have been collected inside a time interval (start, stop] during which the value of the variable remain constant. This nomenclature is usually referred to as counting process coding, which assumes the occurrence of events as the realization of a very slow process (Andersen and Gill, 1982; Fleming and Harrington, 1991; Andersen et al., 1993). Additionally, all those subjects whose interval follow-up start at older ages than 65 are assumed to have delayed entries, and consequently their survival times are left-truncated. An appropriate way to handle different starting ages in our study is using the age above 65 years as our particular time scale, so that all our timing references are adapted to this threshold age. Such features yield the final health insurance data source used in this thesis, the HI dataset, where the 10 variables showed in Table 2.1 are included.

Name	Description
`id`	Subject identifier.
`sex`	Gender of the subject: 0 = Man, 1 = Woman
`age_dent`	Age of the subject at study entry (in years).
`obstime`	Observation time (in years above the age of 65) of measurement recording.
`claimyr`	Emergency claims (ambulance services, hospitalizations and non-routine medical visits) per year, denoted by $y_i(t)$ for the i-th subject at the observation time t.
`expo`	Exposure time (value between 0.50 and 1.00 years), denoted by $e_i(t)$ for the i-th subject at the observation time t.
`start`	Lower limit (in years above the age of 65) of the time interval during which the corresponding measurement of emergency claims is recorded.
`stop`	Upper limit (in years above the age of 65) of the time interval during which the corresponding measurement of emergency claims is recorded.
`event`	Event indicator at the end of each of the follow-up intervals (`start`, `stop`] from a specific subject. Its dichotomous value across all intervals related to the same subject will be by 0 if subject is still alive at `stop`, and will vary to 1 when death is recorded.
`status`	Event indicator at the end of whole follow-up of specific subject. Its value is constant across all subjects measurements, and is categorized by: 0 = The subject is still alive at the end of follow-up: Righ-censored data 1 = The subject's death was recorded at the end of follow-up: Event data

Table 2.1. Names and description of the variables provided by the HI dataset.

In Figure 2.2 we present three possible situations of longitudinal profiles contained in the HI dataset, exemplified by subject ①, subject ②, and subject ③. We realize that in all cases the corresponding measurement points match to the control points, equally-spaced over the study period.

Subject ① is a man who enrolled with the company before 2006/01/01, and entered into the study at the second control point, at the age of 65, so his time-to-event data is not left-truncated. The follow-up of this man extends from the date of his first measurement, on 2008/01/01, until the administrative closing date, on 2014/02/01, with him still being alive at the age of 71.1 years. Consequently, his event time is right-censored.

Subject ② is a man who is exactly the same age as subject ①, but he enrolled with the company on 2007/07/01. He also entered into the study at the age of 65, at the second control point, and his follow-up extends from the date of first measurement (where the emergency claims recorded are corrected by an exposure of 0.5), on 2008/01/01, to the date of his dropout from the study, on 2011/10/01. Then, we only know that he was still alive at the age of 68.8, so his time-to-event outcome is right-censored.

Finally, subject ③ is a woman who enrolled with the company before 2006/01/01, and entered into the study at the first control point, at the age of 80. Consequently, her time-to-event data is affected by left truncation (caused because 15 years have elapsed since the time zero of study). Her follow-up extends from the date of first measurement (obtained after having been followed the whole year of 2006), on 2007/01/01, to the date of her death, on 2012/07/01. In this case, the woman's death was recorded when she was 85.5 years old, so her time-to-event response corresponds to an event of interest.

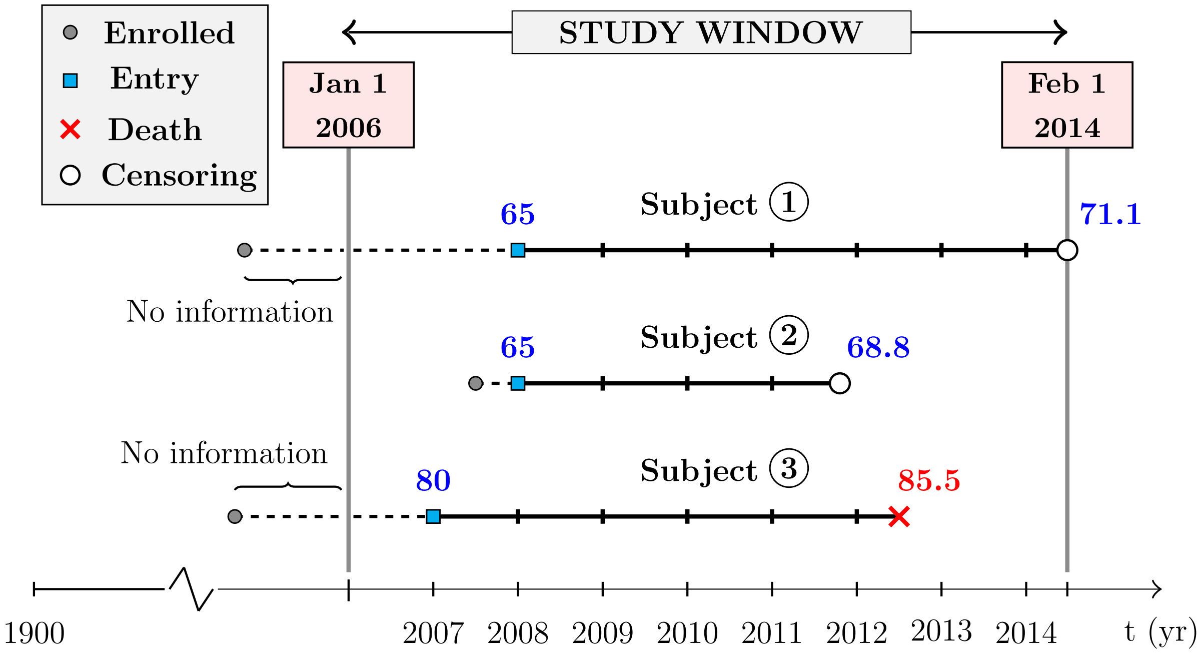


Figure 2.2. Observed time-to-event profiles for three subjects in the HI dataset.

Due to the nature of our dataset (where two different outcomes are reported), each subject's profile represented in Figure 2.2 can be thought of both from a longitudinal and time-to-event perspective, as illustrated by Figure 2.3:

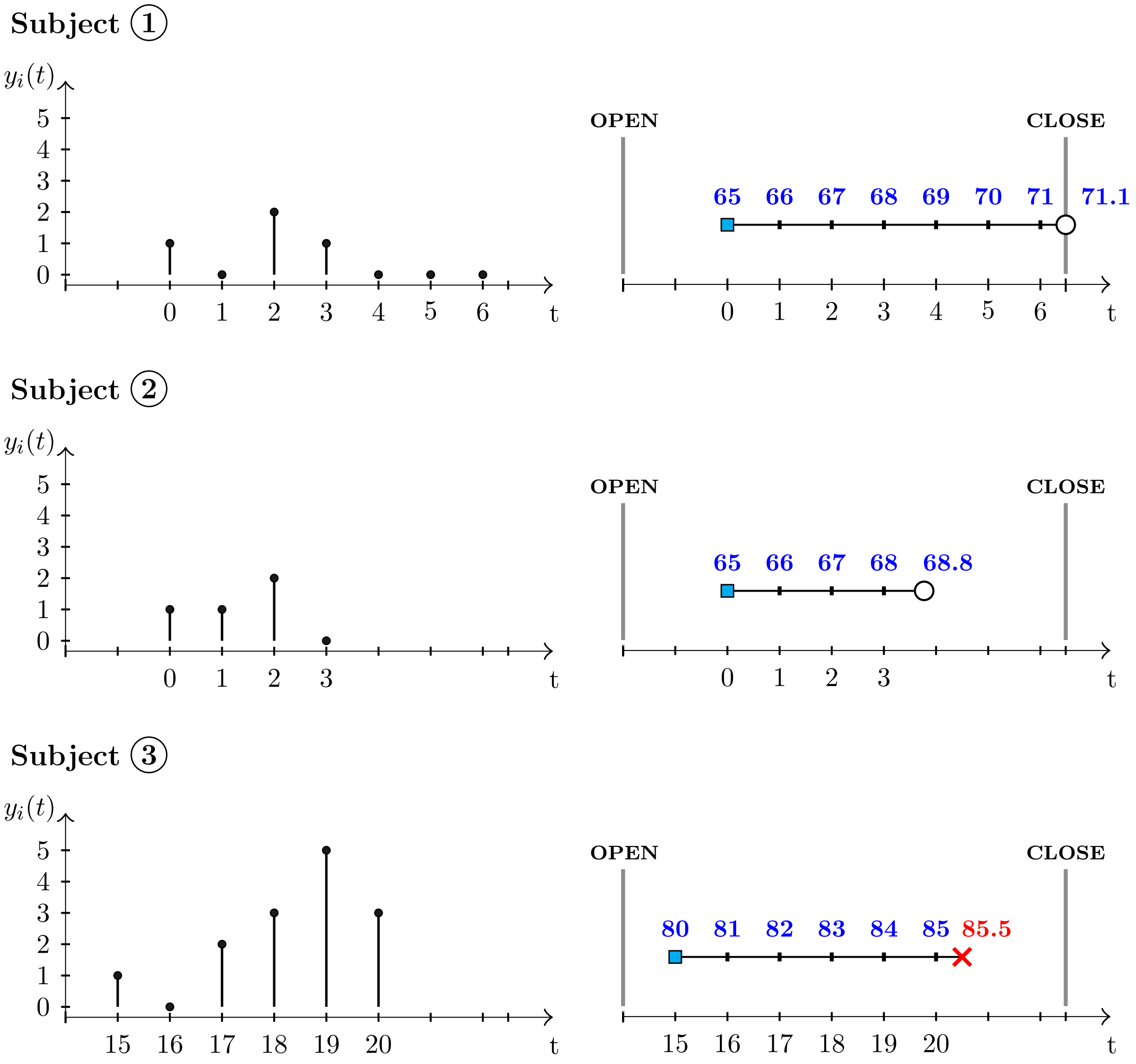


Figure 2.3. Longitudinal and time-to-event information about the three subjects from the HI dataset represented in Figure 2.2. The panels on the left side show the evolution over time of emergency claims per year, whereas the panels on the right side depict the corresponding survival information.

Taking the three profile examples discussed above, Table 2.2 displays the format and the structure with which the information in the HI dataset is stored. In that scheme, the follow-up period for each individual within the study window is divided up into a sequence of shorter time intervals, each of them characterized by an entry time (`start`) and an exit time (`stop`), during which the value of the longitudinal outcome remains constant, as well as the explanatory covariates (in our case only the variable `sex`). Thus the longitudinal and time-to-event information for each subject is represented by a number of right-censored intervals and possibly one interval ending with the death event.

id	sex	age_ent	obstime	claimyr	expo	start	stop	event	status
1	0	65.00	0.00	1	1.00	0.00	1.00	0	0
1	0	65.00	1.00	0	1.00	1.00	2.00	0	0
1	0	65.00	2.00	2	1.00	2.00	3.00	0	0
1	0	65.00	3.00	1	1.00	3.00	4.00	0	0
1	0	65.00	4.00	0	1.00	4.00	5.00	0	0
1	0	65.00	5.00	0	1.00	5.00	6.00	0	0
1	0	65.00	6.00	0	1.00	6.00	6.08	0	0
2	0	65.00	0.00	1	0.50	0.00	1.00	0	0
2	0	65.00	1.00	1	0.50	1.00	2.00	0	0
2	0	65.00	2.00	2	0.50	2.00	3.00	0	0
2	0	65.00	3.00	0	0.50	3.00	3.80	0	0
3	1	80.00	15.00	1	1.00	15.00	16.00	0	1
3	1	80.00	16.00	0	1.00	16.00	17.00	0	1
3	1	80.00	17.00	2	1.00	17.00	18.00	0	1
3	1	80.00	18.00	3	1.00	18.00	19.00	0	1
3	1	80.00	19.00	5	1.00	19.00	20.00	0	1
3	1	80.00	20.00	3	1.00	20.00	20.50	1	1

Table 2.2. Layout information supplied in the HI dataset in accordance with the information provided in Figure 2.3.

The whole process summarized in this section entailed a huge programming and computational effort given the high volume of both the measurements and the different issues to consider in each step. All the scrubbing tools were implemented by means of the R software, and the different steps to be followed for obtaining the final dataset are detailed in the R code supplied in Appendix A.

2.2 Scope of the Data: The HI dataset

The HI dataset becomes an extremely rich and valuable source of medical information to reflect the current situation of private health care providers in Spain, thus helping us to answer the stated research questions. Specifically, these data allow for modeling the given information using a three-pronged approach:

- Longitudinal approach: The HI dataset contains repeated counts of the annual demand for emergency care for each of the considered subjects. We can assume that claim counting for each subject plays the role of a classical biomarker, in the sense that it provides essential information regarding the individual health status during the study period. Given that the final aim is relating the demand evolution to the death hazard, the motivating dataset is constructed in order to carry out a longitudinal study of different time-dependent information associated with each subject.

- Time-to-event approach: The resulting dataset contains the age at which each subject enters the study, coinciding with the time point of their first longitudinal observation, and the age of the subject's death (defined here as the particular event of interest) or right censoring (for those subjects whose death event is not observed). Besides the usual censorship mechanism, our study is unique in that it only focuses on those subjects who have reached the age of 65. This has two direct implications on the survival times ultimately observed: a) all subjects who have not reached 65 years of age within the study window are excluded, and b) all those entering after the age of 65 are assumed to be delayed entries in relation to time zero, set at the age of 65. Both factors, closely related, mean that the presence of left truncation issue should be kept in mind when analyzing our right-censored survival data. The time-to-event modeling is then conducted as the elapsed time from the age of 65 until the death event.

- Joint approach: A simultaneous approach of the longitudinal and time-to-event information can also be considered (including parameters that control their correlation), this being the primary goal over the following chapters. The joint modeling is tackled by assuming a shared latent structure between these two processes, so that longitudinal responses can be taken into account in the survival analysis as a subject-specific covariate information. In such a working scheme, longitudinal and time-to-event data are time linked by an appropriate association structure, which collects the degree of relationship between the two responses. The functional form adopted to collect this association will be the one that most clearly helps to understand the patterns of health care usage by the elderly.

2.3 Longitudinal Information Provided by the HI dataset

2.3.1 Definition and Recording Scheme of the Longitudinal Outcome

The HI dataset derives from the information provided by a Spanish medical insurance company, and is circumscribed within a period of time that, for administrative reasons (and therefore not related to mortality risk), encompasses from the January 1, 2006 until February 1, 2014. The length of time between these two dates is called *study period* or *study window*, and corresponds to the period when an active monitoring of the target population is performed.

The cohort under study consists of a homogeneous population of $n = 5470$ subjects (37.6% men and 62.4% women), aged between 65 and 100 years when entering the study, and living in the city of Barcelona. All these subjects have the right to receive medical assistance during the 8.1-year study period, and we are interested on monitoring, within the i-th subject, $i = 1, \ldots, n$, the degree of the medical service's use in order to obtain personalized patterns of mortality risk. More specific, we focus on time evolution of emergency claim counts, so that follow-up recordings of this discrete response are annually conducted annually using a mesh of control points fixed at the end of each calendar year throughout the study period. Hence, the longitudinal outcome is defined as the rate of emergency claims per year, where ambulance services, hospitalizations and non-routine medical visits are considered in an aggregated manner. This longitudinal response involves a count value, which plays the role of a health status indicator during each subject's follow-up time within the study window. Because subjects do not enter the study at the same age, each subject provides a set of n_i responses, observed at time points $t \in \{t_{ij},\ j = 1, \ldots, n_i\}$, and summarized via the vector $\mathbf{y}_i = \{y_i(t_1), y_i(t_2), \ldots, y_i(t_{n_i})\}$. A total of 32 269 observations are conducted during the study period.

The monitoring of the data can be considered adequate, with an average number of measurements by subject of about six measurements (a mean follow-up in time of 5.10 years, and a median of 6.33 years). Nearly half of the subjects are observed across the whole study period, and no significant differences can be inferred by relating the number of measurements to the average age of study entry (Table 2.3).

Measurements per subject	Average age at entry	Subjects (%)	
1	75.00	383	(7.0)
2	75.54	414	(7.5)
3	76.61	447	(8.2)
4	75.36	411	(7.5)
5	75.18	425	(7.8)
6	74.52	406	(7.4)
7	74.62	360	(6.6)
8	75.50	2624	(48.0)
Overall	75.40	5470	(100.0)

Table 2.3. Distribution of the number of measurements in the HI dataset.

2.3.2 Particular Features of the Longitudinal Outcome

When recording the diagnostic information inside the observation period, the within-subject measurements are non-negative integers which precisely range from 0 to 20 emergency claims per year, so higher values can be related to poorest health status. Overall mean and variance values are 0.84 claims/year and 2.66 (claims/year)^2, respectively, thus suggesting that these measurements are affected by a marked degree of overdispersion. In our context of panel data, it entails that within-subject observed variance is significant large compared to the observed mean, so that a same subject can record few counts during most measurements and many at some specific control points. The causes of overdispersion have been pointed out by several authors; e.g., Hinde and Demétrio (1998), Molenberghs and Verbeke (2005), Winkelmann (2008), Frees (2010), and Hilbe (2011). In practice, we can conclude that panel count data are often overdispersed, mainly due to a) the unobserved heterogeneity, b) the correlation between repeated measurements on a same subject, c) an excess of zeros in the observed data. The aforementioned reasons may arise separately or together.

In our motivating dataset, a large number of zeros are exhibited in the longitudinal outcome, representing 63.1% of the overall measurements, and presenting as the main cause of overdispersion. The tendency to overdispersion continues when stratifying by subject's gender, as shown in Table 2.4.

Sex	Subjects (%)	*Emergency claims per year* summary				
		Mean	SD	Min	Max	% Zeros
Man	2055 (37.6)	0.85	1.66	0	19	63.8
Woman	3415 (62.4)	1.84	1.60	0	20	62.7
Overall	5470 (100.0)	0.84	1.63	0	20	63.1

Table 2.4. Descriptive statistics of emergency claims per year stratified by gender indicator in the HI dataset.

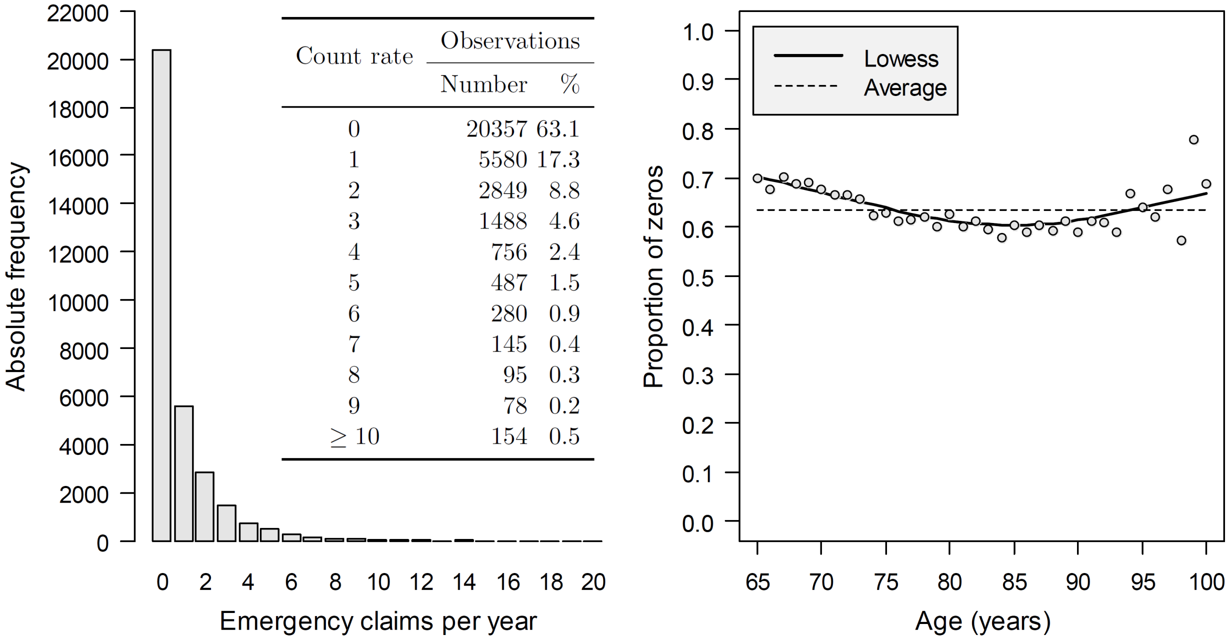
Regarding the observed abundance of zeros, the universal coverage offered by the Spanish health system becomes a key point. It is aimed to ensure that everyone is protected from health issues, regardless of their particular economic situation. Due to this particular feature, in our case zero counts may arise from two different sources:

- If a policyholder has a good health status (here understood as the absence of a critical disease), it means that the subject is only using the insurance coverage for routine medical care, so a zero arises from the considered counting model. It is a classical zero, sometimes referred to as a *sampling zero*, which is obtained randomly due to a subject's health status.

- In the case that a policyholder is being treated in a public medical center, that subject will not have emergency claims recorded inside the health insurance system, and consequently subject's risk can not be observed within the private health care. In that case, it arises a zero which does not come from a counting process because this null value does not appear by chance. These zeros are usually named *structural zeros*, which inflate the response that initially might be expected at this value from a standard count model.

The above circumstance, displayed on the left-hand panel in Figure 2.4, is compounded with by the infrequent nature of the an emergency claim, and is the main source of overdispersion in the longitudinal response. Consequently, this needs to be taken into account when modeling the counting process.

Although it can be assumed that the proportion of zeros is inherent to the health status and use of private services of each individual, the right-hand panel of Figure 2.4 suggests that, in principle, we can not infer any significant trend in the proportion of zeros over time. Indeed, the overimposed lowess curve deviates slightly from the dotted horizontal line representing the average proportion of zeros (0.63). The rate of null values is lowest around 85-90 years old, even though the rates do not excessively depart from the mean rate.



Count rate	Observations	
	Number	%
0	20357	63.1
1	5580	17.3
2	2849	8.8
3	1488	4.6
4	756	2.4
5	487	1.5
6	280	0.9
7	145	0.4
8	95	0.3
9	78	0.2
≥ 10	154	0.5

Figure 2.4. Descriptive plots of the longitudinal outcome of the HI dataset. Left panel: Frequency plot over all measurements and both genders (recall that 5470 individuals were observed over a maximum of 8.1 years). Right panel: Average proportion of zero count rates by age, where a smooth curve has been superimposed.

Although it is commonly accepted that the annual demand for health care increases with age, it has been proven that this pattern is not necessarily observed for any age period over 65, as exposed, for example, by Reinhardt (2003). In this regard, Charpentier (2015) suggests analyzing the evolution of the empirical claims frequency according to policyholder age. In our particular dataset, an initial increase in emergency demand is observed as subject ages when fitting a generalized additive model (GAM) under the Poisson and negative binomial models (Figure 2.5). However, this demand peaks around 85-90 years, and a changing trend is detected at very old ages, so our data show that the use of health insurance services decreases among those of an advanced age. This may reflect the fact that a portion of the elderly population have taken up residence in nursing homes at older ages, and thus, receive personalized care, or it might be a result of a preference for public over private health services for severe treatments (Rodriguez and Stiyanova, 2004). Note that those ages related to maximum demand levels correspond precisely to those ages at which the highest proportion of zeros counts are observed.

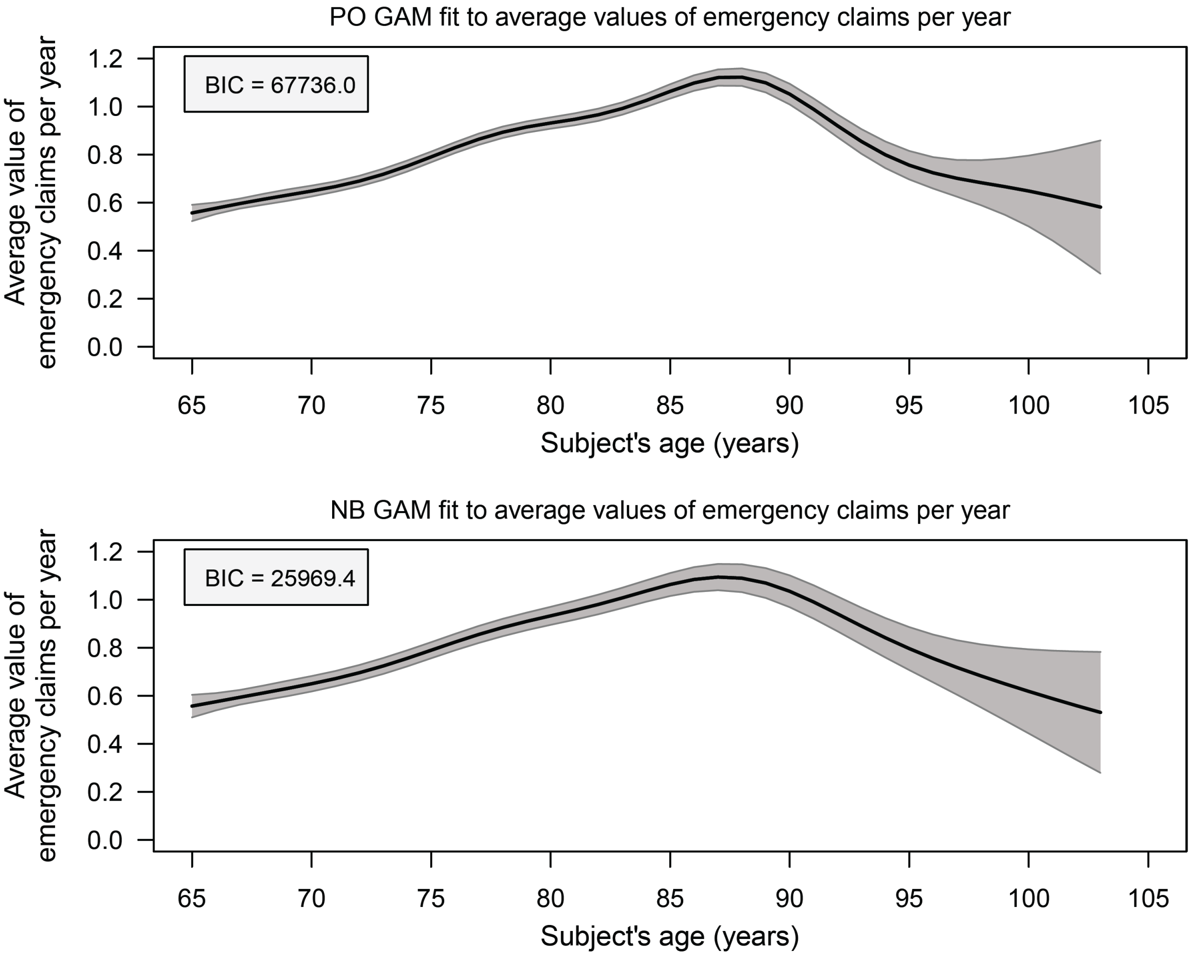


Figure 2.5. Observed annual rates of emergency claims by age in the HI dataset, with Poisson and negative binomial GAM fittings. The 95% confidence regions are presented.

2.3.3 Observed Risk Patterns in the Longitudinal Outcome

Although this thesis focuses on the proper way to account for the subject-specific longitudinal information in estimating the death hazard, it is usually helpful to refer to descriptive results that allow for an initial idea about the type of dependence existing between the two responses. In this regard, when stratifying the descriptive statistics of the emergency claims measurements per year in relation to the death event, the following important results stand out: a) The tendency of the data to overdispersion is maintained as much in subjects during the study as in those whose death event is censored, b) the average number of observed emergency claims in a year is almost double for those subjects who experience the event of interest, and c) the percentage of observed null values is markedly higher for those profiles in which the event of interest is not collected.

Death	Subjects	*Annual claims* summary				
		Mean	SD	Min	Max	% Zeros
No	4961	0.80	1.55	0	20	63.8
Yes	509	1.50	2.45	0	18	52.4
Overall	5470	0.84	1.63	0	20	63.1

Table 2.5. Descriptive statistics of emergency claims per year stratified by event indicator.

In the case of the longitudinal data, we observe a positive correlation within subject's measurements, so a graphical analysis becomes a key element in order to rapidly discern patterns regarding any further explanatory variable (in our case the death event). Figure 2.6 shows the time-plots of various claim profiles, where subject-specific measurements are connected by line segments. The top panel shows the trajectories for a random sample of 100 subjects alive at the end of study, while the bottom panel shows 100 randomly selected profiles of subjects whose death was recorded during the study. Notice that the group of subjects whose death event is recorded presents higher mean values than those presented by the subjects that remain alive.

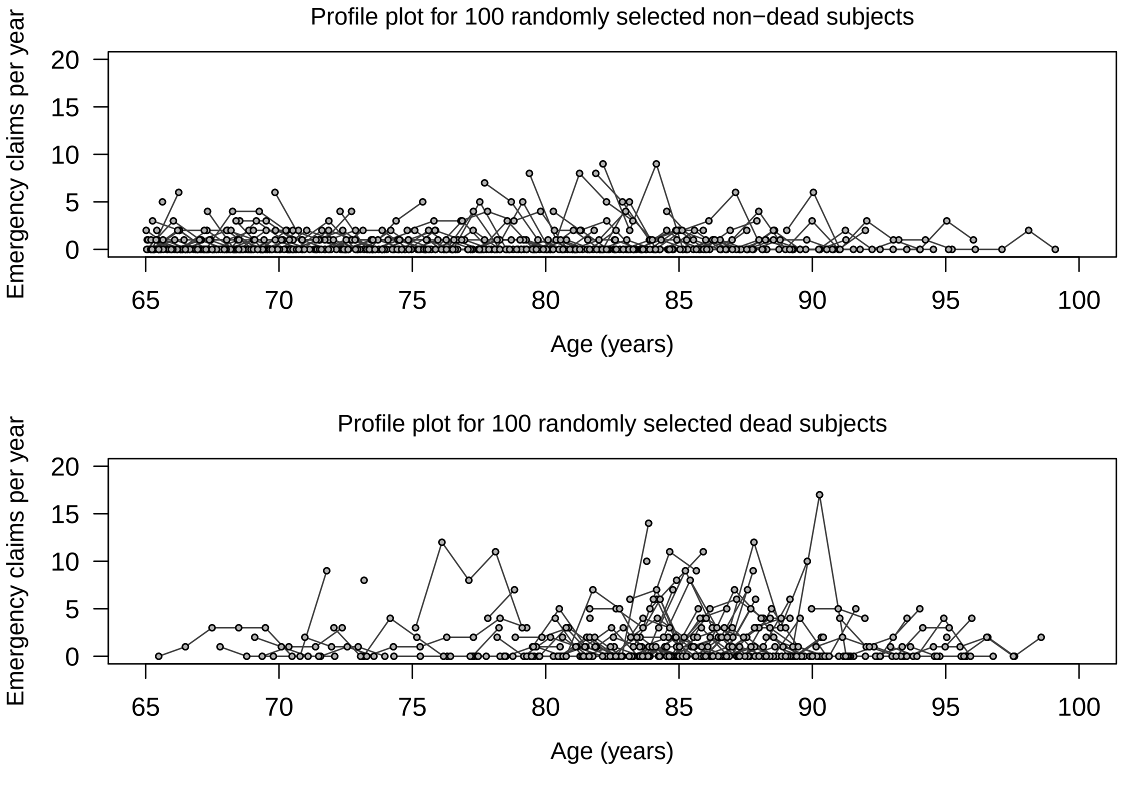


Figure 2.6. Subject-specific claim profiles across time (age in years) for 100 randomly selected subjects still alive during the follow-up period (top panel) and for 100 randomly selected subjects whose death is observed (bottom panel).

2.4 Time-to-Event Information Provided by the HI Dataset

When modeling time-to-event data for the $n = 5470$ subjects who are ultimately included in the study, it is necessary to establish a well defined origin point from which to begin measuring each subject's survival time T_i^*, $i = 1, \ldots, n$, that is, the time that it takes the event of interest to happen for a specific subject. In our particular case, only policyholders living beyond the age of 65 are considered within the study period 2006-2014.

Because our study is carried out in a defined time period of 8.1 years, some subjects die before reaching the age of 65, while others register delayed entries as their follow-up begin after that threshold age. Specifically, a total of 4365 subjects (79.8%) enter the study at ages over 65, thus resulting a mean age at study entry of 75.4 years (i.e., 10.4 years above the minimum threshold required). Therefore, our time-to-event data T_i^* is left-truncated (Uzunoḡullari and Wang, 1992; Klein and Moeschberger, 1997). In order to avoid biased estimates on survival parameters (because those subjects who had died before the age of 65 are not included in the study), a proper consideration of the left truncation issue in the mortality hazard is achieved by setting the time zero at the age of 65 years, and using age (above 65) as our particular time scale (Lamarca et al., 1998; Thiébaut and Bénichou, 2004; Gail et al., 2009). In addition, an important portion of them are still alive at the end of study, while others drop out of the study before having experienced the event, due to causes beyond their health (mainly unpaid premiums and disagreements between the subject and company about the premium price). In each of these cases, we have incomplete follow-up information since we only certainly know that these subjects survival time will greater than their last observed follow-up time. Consequently, the time-to-event data in the HI dataset is subject to left truncation and right censoring. Both survival issues arise here due to causes which are not related to the event of interest, so right censoring and left truncation issues are considered as non-informative.

Once our reference times are adapted to the new time scale due to handle left truncation, the individual profile descriptions for all the subjects cover from their age at first measurement (coinciding with the age at study entry) until the date on which the first of these three possible scenarios happen:

- Scenario A: The subject profile reached the study end, on 2014/02/01, being the subject still alive.
- Scenario B: The subject profile dropped out of the study, due to causes not related with health status, on specific date prior to 2014/02/01.
- Scenario C: The subject profile, which is left-truncated, ended at the date of subject's death, prior to 2014/02/01.

In the first and second scenarios we could only say that up to a determined time point, the subject is still alive, but we have not more information beyond that date (their time-to-event information is incomplete). By contrast, in the third scenario subject experiences the event of interest, and we know the *true* event time.

In the three examples of possible survival profiles shown in Figure 2.2, their respective classification in one of the three proposed scenarios is evident. So, subject ① is framed in scenario A, subject ② is a typical case of the loss to follow-up described in scenario B, and subject ③ is a classic case of scenario C. Because the average life expectancy in Spain is much greater than 65 years (our particular time zero), at the end of follow-up only 509 (9.3% in total) deaths are recorded, while the remaining 4961 (90.7%) subjects lead to right-censored data.

CHAPTER 3

STANDARD JOINT MODEL WITH LEFT-TRUNCATED TIME-TO-EVENT DATA

3.1 Principles of the Standard Joint Model

Many cohort studies focus on assessing the time effect of a longitudinal process on the survival outcome for an specific event of interest. Let $y_i(t)$ denote the observed response for the i-th subject, $i = 1, \ldots, n$, at time t, and let T_i^* be the corresponding event time. A first approach to handling longitudinal covariates in survival analysis consists of including within-subject repeated measurements as time-dependent covariates, which is achieved by an extension of the classical proportional hazards model (Cox, 1972), called the time-dependent Cox model. However, this treatment may lead to inefficient or biased inferences about the underlying data mechanisms, especially in those cases where longitudinal and survival outcomes are strongly associated (e.g. Gould et al., 2015). First, the extended Cox model is thought to accommodate time-dependent exogenous (external) variables in the survival analysis, that is, predictable processes whose value at any time t for the i-th subject does not depend on the occurrence of an event at time s, with $t > s$ (Kalbfleisch and Prentice, 2002). This characteristic is clearly inconsistent with accounting for a subject's health status, since it is a non-predictable process which depends on different unobserved biological features, thus being endogenous (internal) to the mortality risk in our case. Second, the complete covariate path before any t, denoted by $\mathcal{Y}_i(t)$, is not fully observed, but is only intermittently captured by a specific set of time snapshots $t_{ij} \leq T_i^*$, $j = 1, \ldots, n_i$. Finally, an additional drawback of the time-dependent Cox model is that observations of an endogenous process are usually contaminated by a random measurement error. Instead, a typical way of addressing time-dependent survival data with dropouts consists of imputing the last observed value to fill in until the occurrence of a new non-missing value. This imputation method implicitly assumes that the value of the outcome remains constant between two subsequent measurements without taking into account subject's latent effects, which is rather unrealistic in survival analysis when handling endogenous covariates subject to random variability.

Endogenous covariates in the survival analysis are properly addressed in the joint modeling framework, that is, the simultaneous modeling of the longitudinal and time-to-event outcomes. Specifically, the repeated measurements collected from each subject are traditionally related by using a simple linear mixed model (Laird and Ware, 1982; Verbeke and Molenberghs, 2000; Fitzmaurice et al., 2008), whereas the event times are modeled by a classic relative risk model. Then, the key idea relies on assuming that longitudinal and survival processes depend on a common set of latent random effects, so that the expected values $\mu_i(t)$ of the longitudinal process can be included in the relative risk model as a time-dependent information. Building on this insight, the unrealistic step-function assumed in the time-dependent Cox model is now replaced by the true smooth evolution over time of the complete longitudinal history, namely $\mathcal{M}_i(t)$, and the expected longitudinal response $\mu_i(t)$ can be related at each time t with the instantaneous hazard for the event, denoted by $h_i(t)$. The benefits of joint modeling are not restricted to the association between both implicated responses, but the combined approach also allows for the avoidal of biased estimates from each submodel, as has been well documented in, for instance, Tsiatis and Davidian (2001) and Fieuws et al. (2008). When using joint modeling techniques, the point of interest may be focused to address the following concerns: a) adjustment of longitudinal profiles in the presence of informative dropouts, b) determination of time-to-event outcomes by accounting for the subject-specific information, and c) characterization of the association structure between each subject-specific trajectory and the survival time. Figure 3.1 exemplifies the fundamental principle behind joint models on a subject from the HI dataset, where the expected emergency claims per year are related to the subject's death hazard at each time t, given that t exceeds the corresponding left truncation time τ_i. Recall that low values for the annual rate of emergency services are implicitly related to a lower mortality risk.

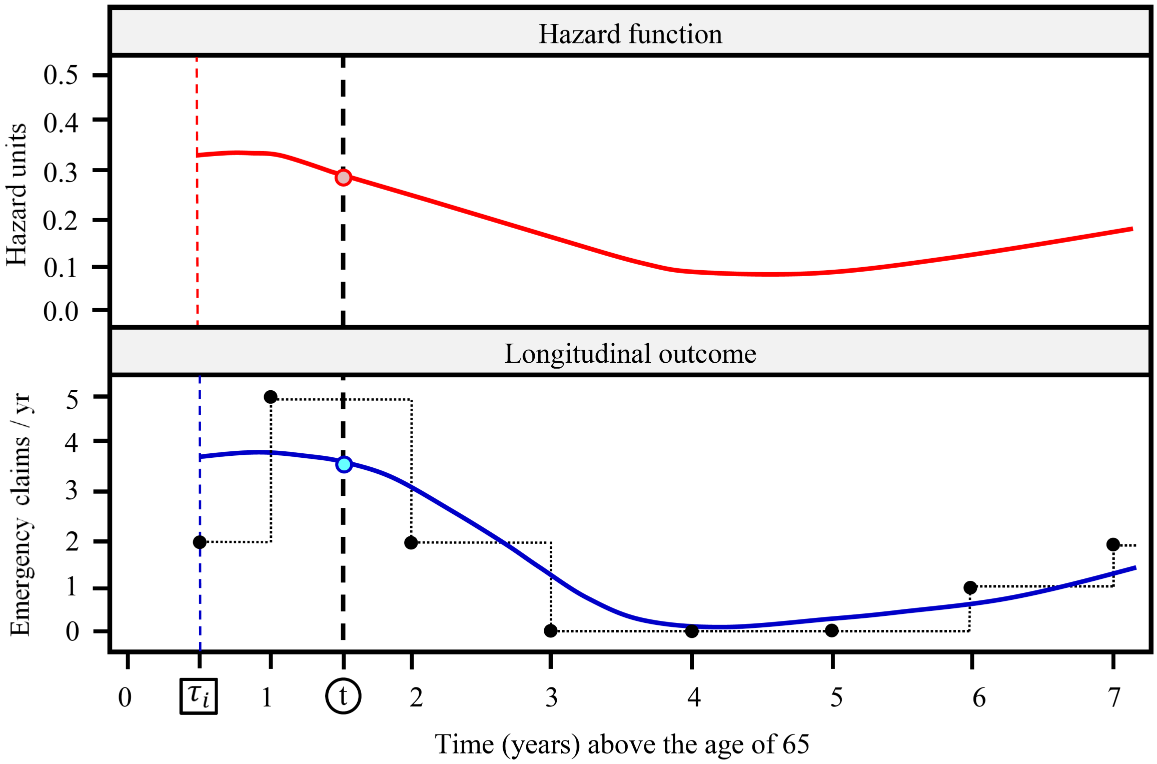


Figure 3.1. Graphical representation of the main idea behind standard joint modeling formulation for a subject from the HI dataset. Top panel: Time evolution of subject's death hazard. Bottom panel: The dotted line depicts the step-function arising from the observed responses in the time-dependent Cox model, while the solid line corresponds to the smooth evolution derived from the expected longitudinal responses.

3.2 Analysis of Longitudinal Continuous Data

3.2.1 Features of Longitudinal Data

Longitudinal studies are based on repeated measurements of subjects taken intermittently over a period of time (called the study window), thereby entailing a (positive) correlation between the subject-specific set of observations. In such study designs, the independence between all outcomes assumed by classic regression methods is no longer valid, and special statistical techniques are required for valid analysis and inferences about parameters.

A quite common characteristic in follow-up studies is the fact that some subjects enter the study at different times or drop out of it prematurely, yielding in both cases unequal number of measurements per subject due to the missing values. Following the missing classification established by Little and Rubin (1987), if the missing data pattern follows a *missing completely at random* (MCAR) or a *missing at random* (MAR) mechanism, the dropout is considered as non-informative or ignorable, and the unbalanced data can be accommodated by the classical longitudinal models. However, in the case of having a *missing not at random* (MNAR) mechanism the dropouts are informative or non-ignorable, and the possible reasons for dropout should be accounted for in order to avoid biased estimates.

This kind of data may arise in medical, economic and social research fields, and its analysis is specially useful in assessing changes over time in the behavior of interest. The longitudinal data structures are indeed clustered within a set of observational units, so they allow for the accommodation of two different variability sources in the response observed over time:

- Between-subject variation: This is the typical source of variation analyzed in cross-sectional studies (where all the measurements relate to one point in time), and reflects the variability of observed responses from one subject to another. In other words, it measures how pronounced the departure of the subject-specific underlying behavior (due to biological, environmental or social factors) is from the population trend.
- Within-subject variation or residual variation: This measures the degree of change in the observed outcomes at the individual level, and collects the inherent variability of each subject that can not be explained by some predictive variable.

In the particular case of claim data analysis, the observed data provide a powerful tool to assess the death hazard according to specific biological and socio-economic latent features of the subject that are difficult to evaluate (Tyree et al., 2006; Shi and Valdez, 2014). Moreover, the incompleteness of some profiles in the HI dataset is caused either by the left truncation of the time-to-event data from those subjects that enter the study after 65, or by the subjects leaving the insurance company due to economic reasons. In both cases, we face an ignorable missing mechanism which can be adequately handled without any imputation or additional assumptions.

3.2.2 Transforming Counts into Continuous Data

Most of the longitudinal studies are designed to accommodate a continuous outcome, which is usually normally distributed around the expected (unobservable) value at each time point within the study window. This is, in fact, the consequence of residuals following a mean-zero normal distribution with a homogeneous variance. However, count processes usually present a highly skewed distribution that clearly departs from a bell-shaped curve, so invalid inferences would be obtained since the data are not in line with the assumptions made in models with a Gaussian response. A first attempt to easily deal with count rates in a continuous manner consists of "normalizing" the observed data by means of a proper transformation function, usually a logarithmic one. Although there are other types of transformations recommended in the literature to normalize the distribution of count data (for example, the square-root transformation or, more generically, the power transformation), the log-transformation is definitively the most widely used in both medical and socio-economic research to make data conform to normality.

One important issue concerning the application of the log-transformation to count rates is dealing with zero values. In this regard, a specific drawback comes from $\log(0) \longrightarrow -\infty$, so a constant value must be added to the observed rates before transformation. Since possible values are restricted to non-negative integers, a one-unit increase is typically used to guarantee a one-to-one transformation mapping, which maintains the same number of zeros and eliminates the possibility of negative values. In this line, the dependent variable for the i-th subject, $i = 1, \ldots, n$, at time t becomes $y_i(t) = \log\{1 + \texttt{claimyr}_i(t)\}$, and we can now assume a linear evolution over time of the expected value of the observation, $\mu_i(t)$.

3.2.3 The Linear Mixed Model

The analysis of longitudinal data is carried out under the linear mixed model (LMM), also referred to as a subject-specific model or hierarchical linear model. The name mixed model comes from the intuitive idea behind this model, in which a combination of fixed and random parameters is considered to properly describe each subject's trend. The fixed-effect coefficients are assumed to be population constants and represent mean trends across all subjects, playing an analogous role as the estimated coefficients in a classic regression. In addition, we also need to take into account the by-subject variability caused by a repeated measurements scheme, this issue being addressed by a set of latent or random effects. These capture the particularities of the same subject and allow for individual predictions.

More specifically, let $\mathbf{y}_i = \{y_i(t_{ij}),\ i = 1, \ldots, n\}$ denote the vector of the observed log-transformed responses for the i-th subject, recorded at a set of time points t_{ij}, $j = 1, \ldots, n_i$, within the study window. Assuming that the expected outcomes are generated by a normal distribution, the LMM is formulated in order to linearly relate each subject-specific observation from the HI dataset to $p+1$ fixed-effect covariates (explanatory covariates) and $q+1$

random effects:

$$\begin{cases} y_i(t) \sim \mathcal{N}\{\mu_i(t),\, \sigma_\varepsilon^2\}, \\ y_i(t) = \log\{e_i(t)\} + \mathbf{x}_i^\top(t)\boldsymbol{\beta} + \mathbf{z}_i^\top(t)\mathbf{b}_i + \varepsilon_i(t) = \mu_i(t) + \varepsilon_i(t), \\ \mathbf{b}_i \sim \mathcal{N}_{q+1}(\mathbf{0},\, \mathbf{D}), \\ \varepsilon_i(t) \sim \mathcal{N}(0,\, \sigma_\varepsilon^2). \end{cases} \tag{3.1}$$

Here, the exposure time effect $e_i(t)$ is logged and included as an offset variable, so the log-response could be divided by the effective length of time. $\mathbf{x}_i^\top(t)$ and $\mathbf{z}_i^\top(t)$ are the row vectors of the fixed and random design matrices, respectively, while $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\top$ and $\mathbf{b}_i = (b_{i0}, b_{i1}, \ldots, b_{iq})^\top$ are the corresponding fixed-effect and random-effect vectors. The coefficients of the random effects can be assumed to follow a standard multivariate normal distribution with an unspecified $(q+1) \times (q+1)$ variance-covariance matrix, $\mathbf{D}$. The $\varepsilon_i(t)$ is the measurement error of an observed value, which denotes the random deviation of the observation $y_i(t)$ from the expected value $\mathbb{E}\{y_i(t) \mid \mathbf{b}_i\} = \mu_i(t)$, and σ_ε^2 represents the within-subject variation (assumed constant across the subjects). Finally, the random effects $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ are assumed to be independent of error terms, $\{\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_n\}$.

3.3 Analysis of Time-to-Event Data

3.3.1 Features of Time-to-Event Data

The terms *time-to-event data* or *survival data* are used interchangeably to refer to those data which measure the time elapsed from a well-defined origin point until the occurrence of a specific event of interest, usually designated by $\mathcal{E}$. The event of interest can be of many kinds, and the elapsed time can be measured in different units (days, months, years, etc.). The statistical techniques developed to deal with these types of data are known as survival analysis.

Although time-to-event is a continuous variable, standard regression procedures can not be used due to such a variable is not measured in the same way as others. First, survival times are restricted to positive values, and their distribution is often highly left-skewed. Second, the probability of survival after a certain point in time may be of more interest than the expected time-to-event itself. Finally, what makes time-to-event data special compared to other types of data is that these times are typically subject to right censoring after a known time, meaning that some patients have not yet experienced the event $\mathcal{E}$ at the end of the follow-up period. These right-censored or incomplete observations are mainly due to subjects who have completed the study free of the event $\mathcal{E}$ (i.e., administrative censoring), or those cases of loss to follow-up without the occurrence of event $\mathcal{E}$ before the study's closing date (i.e., they drop out of the study prematurely).

Besides right censoring, time-to-event data may also be subject to left truncation in observational studies, that is, the time-to-event is incomplete at the left side of the subject's follow-up period. Left truncation arises for those subjects who are not observed from the established time origin of the study (Bull and Spiegelhalter, 1997), but they come under observation at some later known time (called the left truncation time).

Subject information in the HI dataset is collected in calendar time within two prespecified dates, with subjects' death being the particular event of interest. An important percentage of them are still alive at the end of the study, while others drop out of the study before having experienced the event, due to causes not related to their health (mainly unpaid premiums and disagreements between the subject and company about the premium price). In each of these cases, we have incomplete follow-up information since we only know for certain that these subjects' survival times are greater than their last observed follow-up time. Moreover, we focus on subjects aged 65 and older, so those entering into the study after that age, here our time zero, are considered delayed entries. To avoid biased estimates caused by people who die before 65 years old (which in practice leads to an overestimation of the survival rates), we set the age above 65 as our particular time scale, so the survival times are restricted from the age at which the subject enters into the study until the age they leave (having or not experienced the event $\mathcal{E}$). This departure from the classical time scale (which assigns time zero at the study start) allows us to analyze survival evolution parallel to the aging of the subject, becoming left-truncated all the event times of subjects who begin to be followed after the age of 65. Throughout this thesis, our survival analysis methods will be extended to assess both left-truncated and right-censored event times. Both survival issues arise here due to causes not related to the processes which govern the event of interest, so right censoring and left truncation issues are considered non-informative.

3.3.2 Main Functions for Time-to-Event Analysis

Survival and Hazard Functions

From now on, we denote with T_i^* the non-negative continuous random variable which represents the i-th subject's elapsed time from the established time origin until the occurrence of a specific event $\mathcal{E}$. Let us consider the corresponding probability density function, $p_t(t)$, and cumulative distribution function $\mathcal{P}(t)$. In survival analysis, there are three prominent functions used to describe the distribution of the event times:

- Survival function, $\mathcal{S}_i(t)$

  It is denoted by the reverse cumulative distribution function of T_i^*,

$$\mathcal{S}_i(t) = \Pr(T_i^* > t) = \int_t^{\infty} p_t(s)\,\mathrm{d}s = 1 - \mathcal{P}(t), \quad t \geq 0, \tag{3.2}$$

and represents the probability of an individual surviving beyond time t, that is, the probability that the event of interest $\mathcal{E}$ has not yet occurred before t.

- Hazard function or hazard rate, $h_i(t)$

  It is a non-negative function that represents the *instantaneous* rate of occurrence of the event of interest at a given time t, conditional on survival until time t or later (that is, $T_i^* \geq t$). The hazard function can theoretically vary from zero (at the time when the risk is null) up to infinity (at the time when the event $\mathcal{E}$ occurs), and is expressed by:

$$h_i(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T_i^* < t + \Delta t \mid T_i^* > t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T_i^* < t + \Delta t)}{\Delta t \, \Pr(T_i^* > t)}, \quad t \geq 0. \quad (3.3)$$

  Therefore, the hazard rate can be easily related to $\mathcal{S}_i(t)$:

$$h_i(t) = \frac{p_t(t)}{\mathcal{S}_i(t)} = \frac{\mathrm{d}\mathcal{P}(t)/\mathrm{d}t}{\mathcal{S}_i(t)} = -\frac{\mathrm{d}\mathcal{S}_i(t)/\mathrm{d}t}{\mathcal{S}_i(t)} = -\frac{\mathrm{d}}{\mathrm{d}t}\big[\log\{\mathcal{S}_i(t)\}\big]. \quad (3.4)$$

- Cumulative hazard function, $\mathcal{H}_i(t)$

  It measures the total quantity of hazard accumulated until time t, being defined as:

$$\mathcal{H}_i(t) = \int_0^t h_i(s)\,\mathrm{d}s, \quad t \geq 0. \quad (3.5)$$

  In the same way as in the case of the hazard rate, it is also possible to express $\mathcal{H}_i(t)$ in relation to the survival function:

$$\mathcal{H}_i(t) = \int_0^t h_i(s)\,\mathrm{d}s = \int_0^t \frac{p_t(s)}{\mathcal{S}_i(s)}\,\mathrm{d}s = -\int_0^t \frac{1}{\mathcal{S}_i(s)}\left\{\frac{\mathrm{d}}{\mathrm{d}s}\mathcal{S}_i(s)\right\}\mathrm{d}s = -\log\big\{\mathcal{S}_i(t)\big\}. \quad (3.6)$$

All of these functions are interrelated, so that if one is known, then the others are automatically defined. The estimation of these functions constitutes the central axis of the survival analysis carried out over the different chapters.

Survival analysis must be carried out in accordance with the specific characteristics of the time-to-event data contained in the cohort under study. In the particular case of the HI dataset, the survival data are affected by left truncation and right censoring. The left truncation is handled by introducing an independent and non-negative random variable τ_i, which defines the time elapsed between the age of 65 and the age at which a policyholder enters the study, $\tau_i = \text{age.entry}_i - 65$. Furthermore, the random censoring times are denoted by the variable C_i, so for each subject we observe a time $T_i = \min\{T_i^*, C_i\}$, associated with a dichotomous event indicator δ_i which is equal to 1 if the observed time corresponds to a true event (subject's death), and 0 if it corresponds to a censoring.

In the case of left-truncated and right-censored data, a subject will be included in the study if and only if $T_i^* > \tau_i$, giving rise to a left-truncated event time when $\tau_i > 0$. Consequently, the probabilistic distribution of the time-to-death has to be defined according to the proportion of subjects living beyond time point t and is conditional on their being older than the corresponding left truncation time, $\mathcal{S}_i(t \,|\, T_i^* > \tau_i) = \Pr\left(T_i^* > t | T_i^* > \tau_i\right) = \Pr\left(T_i^* > t\right) / \Pr\left(T_i^* > \tau_i\right) = \mathcal{S}_i(t) \,/\, \mathcal{S}_i(\tau_i)$.

Likelihood Function for Right-Censored Data

Let us denote the parametric form of the i-th subject's survival function as $\mathcal{S}_i(t \,|\, \boldsymbol{\theta})$, working with a random n-sample of right-censored time-to-event data, $\{(T_i,\, \delta_i),\ i = 1, \ldots, n\}$. Consequently, the time-to-event distribution comes from the corresponding probability density function, $p_t(t \,|\, \boldsymbol{\theta})$, and cumulative distribution function $\mathcal{P}(t \,|\, \boldsymbol{\theta})$. To make inferences about $\boldsymbol{\theta}$, the overall likelihood function combines the information from censored data and events, so that it is expressed as the product of probabilities, given the observed data:

$$p_t\{T_i,\, \delta_i \,|\, \boldsymbol{\theta}\} = \prod_{i=1}^{n} \{p_t(T_i \,|\, \boldsymbol{\theta})\}^{\delta_i} \, \{\mathcal{S}_i(T_i \,|\, \boldsymbol{\theta})\}^{1-\delta_i}. \tag{3.7}$$

The above function can be also written alternatively by considering the previous relationships between the different survival functions:

$$h_i(t \,|\, \boldsymbol{\theta}) = \frac{p_t(t \,|\, \boldsymbol{\theta})}{\mathcal{S}_i(t \,|\, \boldsymbol{\theta})} \Rightarrow p_t(t \,|\, \boldsymbol{\theta}) = h_i(t \,|\, \boldsymbol{\theta}) \, \mathcal{S}_i(t \,|\, \boldsymbol{\theta}),$$

$$\mathcal{H}_i(t \,|\, \boldsymbol{\theta}) = -\, \log \big\{\mathcal{S}_i(t \,|\, \boldsymbol{\theta})\big\} \Rightarrow \mathcal{S}_i(t \,|\, \boldsymbol{\theta}) = \exp\big\{ -\mathcal{H}_i(t \,|\, \boldsymbol{\theta})\big\} = \exp\bigg\{ - \int_0^t h_i(s \,|\, \boldsymbol{\theta}) \, \mathrm{d}s \bigg\},$$

so that the likelihood function can be completely expressed in terms of hazard function:

$$\begin{aligned} p_t(T_i,\, \delta_i \,|\, \boldsymbol{\theta}) &= \prod_{i=1}^{n} \{h_i(T_i \,|\, \boldsymbol{\theta}) \, \mathcal{S}_i(T_i \,|\, \boldsymbol{\theta})\}^{\delta_i} \, \{\mathcal{S}_i(T_i \,|\, \boldsymbol{\theta})\}^{1-\delta_i} \\ &= \prod_{i=1}^{n} \{h_i(T_i \,|\, \boldsymbol{\theta})\}^{\delta_i} \, \mathcal{S}_i(T_i \,|\, \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \{h_i(T_i \,|\, \boldsymbol{\theta})\}^{\delta_i} \, \exp\big\{ -\mathcal{H}_i(T_i \,|\, \boldsymbol{\theta})\big\} \\ &= \prod_{i=1}^{n} \{h_i(T_i \,|\, \boldsymbol{\theta})\}^{\delta_i} \, \exp\bigg\{ - \int_0^{T_i} h_i(s \,|\, \boldsymbol{\theta}) \, \mathrm{d}s \bigg\}. \end{aligned} \tag{3.8}$$

The overall log-likelihood function is then

$$\log\{p_t(T_i,\, \delta_i \,|\, \boldsymbol{\theta})\} = \sum_{i=1}^{n} \Big\{\delta_i \log h_i(T_i \,|\, \boldsymbol{\theta}) - \int_0^{T_i} h_i(s \,|\, \boldsymbol{\theta})\, \mathrm{d}s\Big\}, \tag{3.9}$$

and the MLE estimation of the parameters $\boldsymbol{\theta}$ can be obtained from $\dfrac{\partial \log\{p_t(T_i,\, \delta_i \,|\, \boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}} = \mathbf{0}$.

Likelihood Function for Right-Censored and Left-Truncated Data

In the presence of left truncation, the likelihood survival function is redefined as:

$$p_t(T_i,\, \delta_i \,|\, T_i^* > \tau_i,\, \boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{\frac{p_t(T_i \,|\, \boldsymbol{\theta})}{\mathcal{S}_i(\tau_i \,|\, \boldsymbol{\theta})}\right\}^{\delta_i} \left\{\frac{\mathcal{S}_i(T_i \,|\, \boldsymbol{\theta})}{\mathcal{S}_i(\tau_i \,|\, \boldsymbol{\theta})}\right\}^{1-\delta_i} = \prod_{i=1}^{n} \frac{\{p_t(T_i \,|\, \boldsymbol{\theta})\}^{\delta_i}\, \{\mathcal{S}_i(T_i \,|\, \boldsymbol{\theta})\}^{1-\delta_i}}{\mathcal{S}_i(\tau_i \,|\, \boldsymbol{\theta})}. \tag{3.10}$$

In the same way as in the case of data only subject to censoring, the log-likelihood can be defined from the hazard function:

$$\begin{aligned}
p_t(T_i,\, \delta_i \,|\, \tau_i > T_i^*,\, \boldsymbol{\theta}) &= \prod_{i=1}^{n} \frac{\{h_i(T_i \,|\, \boldsymbol{\theta})\, \mathcal{S}_i(T_i \,|\, \boldsymbol{\theta})\}^{\delta_i}\, \{\mathcal{S}_i(T_i \,|\, \boldsymbol{\theta})\}^{1-\delta_i}}{\mathcal{S}_i(\tau_i \,|\, \boldsymbol{\theta})} \\
&= \prod_{i=1}^{n} \{h_i(T_i \,|\, \boldsymbol{\theta})\}^{\delta_i}\, \frac{\mathcal{S}_i(T_i \,|\, \boldsymbol{\theta})}{\mathcal{S}_i(\tau_i \,|\, \boldsymbol{\theta})} \\
&= \prod_{i=1}^{n} \{h_i(T_i \,|\, \boldsymbol{\theta})\}^{\delta_i}\, \frac{\exp\big\{-\mathcal{H}_i(T_i \,|\, \boldsymbol{\theta})\big\}}{\exp\big\{-\mathcal{H}_i(\tau_i \,|\, \boldsymbol{\theta})\big\}} \\
&= \prod_{i=1}^{n} \{h_i(T_i \,|\, \boldsymbol{\theta})\}^{\delta_i}\, \exp\Big\{-\int_0^{T_i} h_i(s \,|\, \boldsymbol{\theta})\, \mathrm{d}s + \int_0^{\tau_i} h_i(s \,|\, \boldsymbol{\theta})\, \mathrm{d}s\Big\} \\
&= \prod_{i=1}^{n} \{h_i(T_i \,|\, \boldsymbol{\theta})\}^{\delta_i}\, \exp\Big\{-\int_{\tau_i}^{T_i} h_i(s \,|\, \boldsymbol{\theta})\, \mathrm{d}s\Big\}.
\end{aligned} \tag{3.11}$$

Then, the expression of the overall log-likelihood function when handling left-truncated and right-censored is denoted by

$$\log\{p_t(\tau_i,\, T_i,\, \delta_i \,|\, \boldsymbol{\theta})\} = \sum_{i=1}^{n} \Big\{\delta_i \log h_i(T_i \,|\, \boldsymbol{\theta}) - \int_{\tau_i}^{T_i} h_i(s \,|\, \boldsymbol{\theta})\, \mathrm{d}s\Big\}, \tag{3.12}$$

and the MLE time-to-event parameters are estimated by solving $\dfrac{\partial \log\{p_t(\tau_i,\, T_i,\, \delta_i \,|\, \boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}} = \mathbf{0}$.

3.3.3 Preliminary Survival Results from Non-Parametric Analysis

Let us assume that the cumulative distribution function $F(t)$ is completely unknown, and we aim for a non-parametric estimate of the survival function based on the observed data, which may be subject to (non-informative) right censoring and left truncation. In such a case, the Kaplan-Meier (KM) estimator (Kaplan and Meier, 1958) is obtained by taking age above 65 as an alternative time scale, and then summarizing each subject's survival information by the observed triplet $\{\tau_i,\, T_i,\, \delta_i\}$, $i = 1, \ldots, n$.

The KM estimator is the simplest way of estimating the survival distribution function over time for an n-sample homogeneous population. It is usually referred to as the product-limit estimator, and its construction is based on the cumulative product of the estimated probability of not incurring an event (it can be viewed as the limiting case of the classic actuarial estimator). In our particular dataset, left-truncated and right-censored times are handled. We denote: $\tau = \min\{\tau_1, \ldots, \tau_n\}$; $T^*_{(k)}$, $k = 1, \ldots, D$, as the order statistics of the D unique true event times in the sample so that $T^*_{(1)} < T^*_{(2)} < \ldots < T^*_{(D)}$; $d_{(k)}$ as the number of subjects who die at time $T^*_{(k)}$; and $r_{(k)}$ as the number of subjects at risk of dying (i.e. alive and not censored) just prior to time $T^*_{(k)}$ (these latter subjects are those who enter into the study before time $T^*_{(k)}$). Then, taking the age above 65 as our time scale, for all $t < T^*_{(D)}$, the KM estimator for the survival distribution function takes the form

$$\widehat{S}_{KM}(t\,|\,T^* > \tau) = \begin{cases} 1 & \text{if } t < \tau, \\ \displaystyle\prod_{k:\ \tau < T^*_{(k)} \leq t} \left(1 - \frac{d_{(k)}}{r_{(k)}}\right) & \text{if } t \geq T^*_{(1)}. \end{cases} \tag{3.13}$$

As a result, it provides a non-parametric estimate of the survival curve by defining a step function with jumps at the D ($D < n$) different times where true events are observed in the sample. The D-th true event time might contain: a) one event, b) more than one event, or c) a combination of events and censored observations (in this case, it is assumed that the events precede the censored data). Hence, the KM estimator is very useful to estimate the probability of an event at each time point, and additionally it can be shown that (under certain conditions) the estimate is consistent and asymptomatically normal. See, for example, Breslow and Crowley (1974), Gill (1983), Stute and Wang (1993) and Cai (1998). However, one of the main limitations of the KM estimator is rooted in the fact that it does not allow for a direct inclusion of covariate information in survival estimates, making the use of stratified curves necessary. Thus, other parametric or semiparametric methods are preferable.

One of the characteristics of left-truncated data is that the number of subjects at risk does not follow a decreasing trend within the study window, but may dynamically fluctuate due to a compensation between the number of events and the progressive incorporation of new subjects into the study after the administrative start of the study period.

Figure 3.2 represents the evolution over time of the total number of subjects at risk, stratified by gender, and overall. In all cases, the evolution of the number of subjects at risk in relation to age follows a similar pattern, and reflects a greater percentage of women in the study. The maximum number of men at risk occurs at the age of 76.65, and for women at the age of 79.08 years, with 503 and 863 individuals exposed, respectively. In the total of subjects, it is at 79.22 years of age that a maximum value of 1368 subjects at risk is reached.

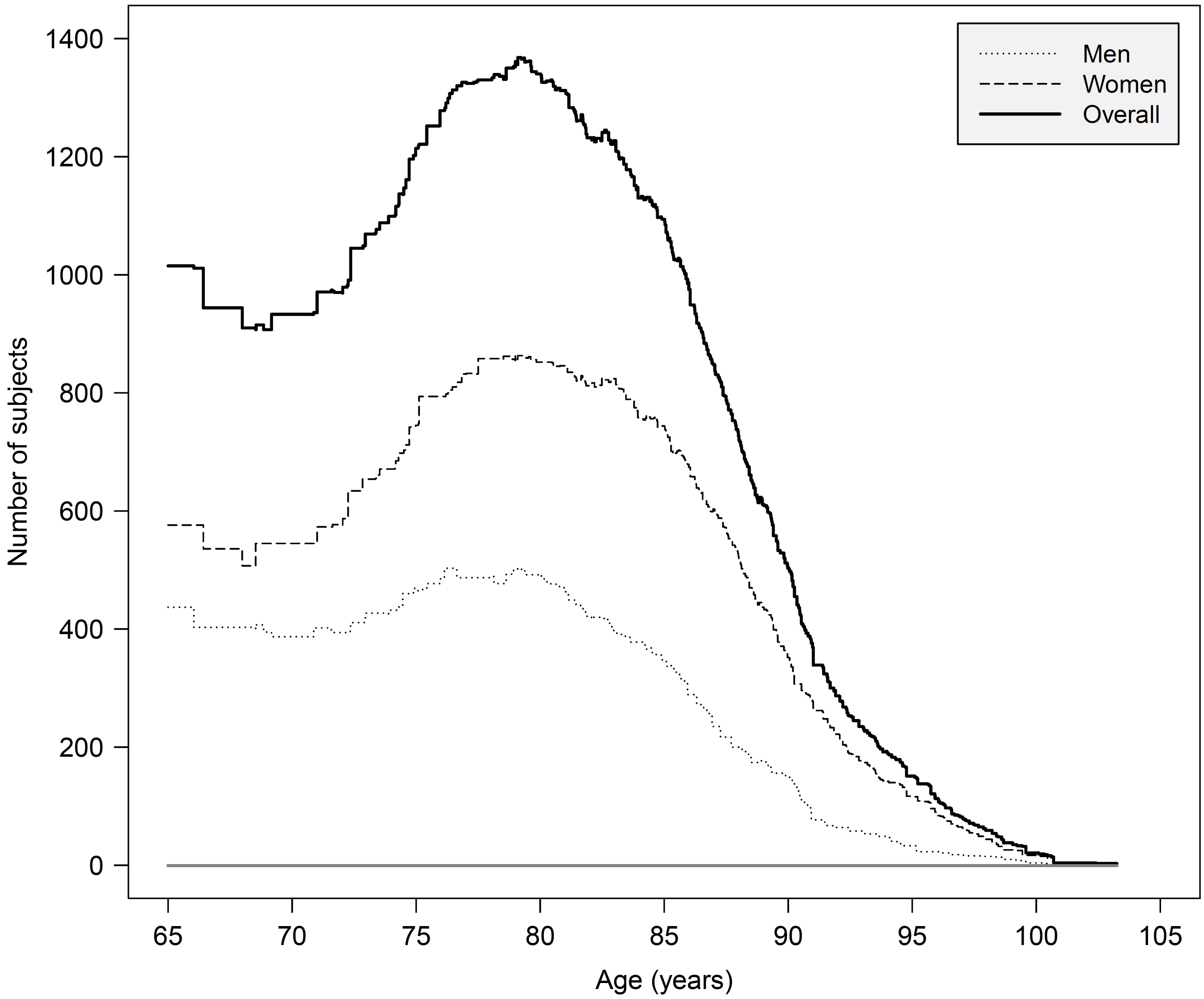


Figure 3.2. Time evolution of the number of subjects at risk in the HI dataset, overall and by gender.

During the study period, the death event is recorded for a total of 509 (9.3%) individuals, which means that 4961 policyholders survive or are no longer in the sample by drop out at the end of the study, i.e. 90.7% right censoring. Of these censored subjects, 3429 (69.1%) are alive at the administrative close of the study window, on February 1, 2014. The remaining 1532 (30.9%) right-censored survival times are attributable to insurance cancellations caused by different reasons not related to the event of death (e.g., dissatisfaction with the service, change of insurance company or an unwillingness to pay), which in practice means that the subject is no longer covered by the insurance policy. Figure 3.3 displays a non-parametric survival curve estimate for the overall sample (on the left) and one stratified by gender (on the right). Although higher survival estimates are registered for women, a comparison between the KM estimates for each gender was performed by means of the corresponding log-rank test (also called Mantel-Cox test), and the results do not suggest a significant difference of survival estimates ($p = 0.242$).

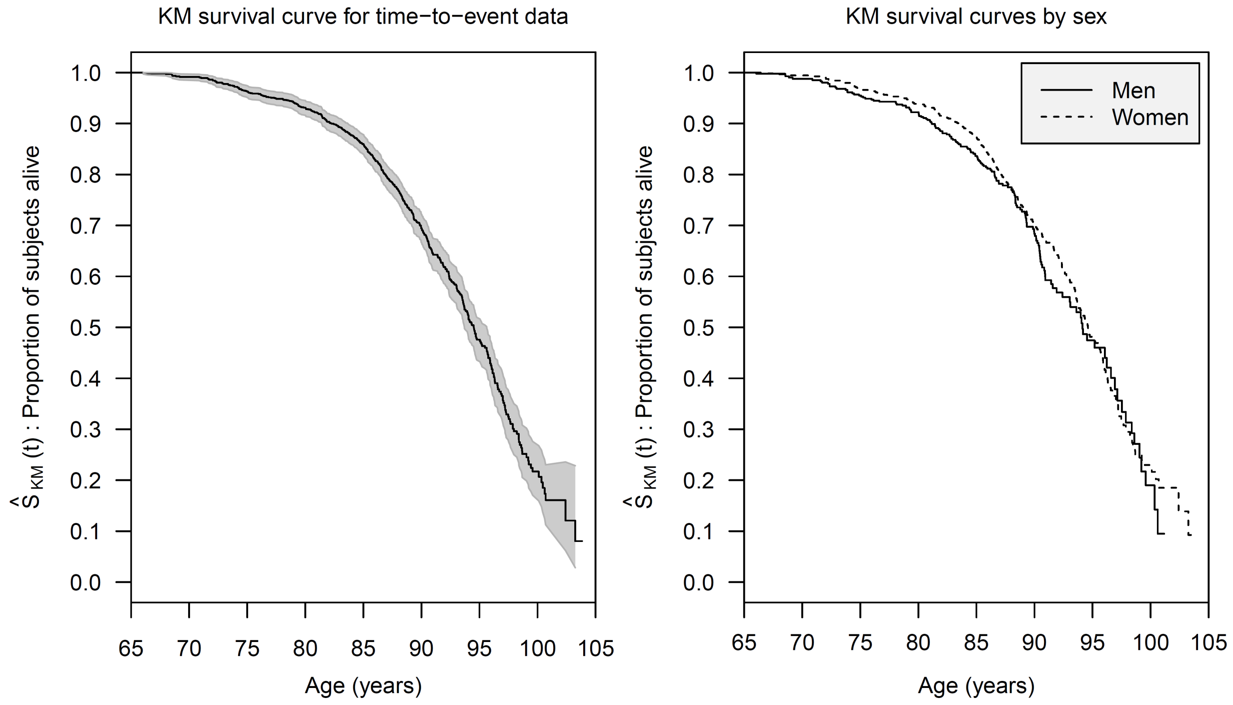


Figure 3.3. Plot of the Kaplan-Meier estimate of the survival function of time-to-death (with 95% confidence intervals) for our overall subjects from the HI dataset (left panel), and stratified by gender (right panel).

3.3.4 The Cox Proportional Hazards Model for Censored Data

Formulation of the Cox Model

The most popular regression model to handle predictor covariates when modeling time-to-event data is the semiparametric PH model, that is, the traditional Cox model. It provides the conditional hazard function at time t of the i-th subject's profile given by p time-independent explanatory covariates or baseline covariates, $\mathbf{w}_i = (w_{i1}, \dots, w_{ip})^\top$, and is given by

$$h_i(t \,|\, \mathbf{w}_i) = h_0(t) \exp\{\boldsymbol{\gamma}^\top \mathbf{w}_i\}, \quad t \geq 0. \tag{3.14}$$

Here, $h_0(t)$ is an unspecified and non-negative baseline hazard function, (that is, the hazard function for a theoretical subject that has $\mathbf{w}_i = \mathbf{0}$), whereas $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)^\top$ denote the vector of regression coefficients. The model then combines a non-parametric functional form in the baseline hazard (usually a step-wise function) with a parametric function in the covariate part, therefore it has a semiparametric character. It is the flexibility conferred by the baseline hazard (it can take any shape as a function of t) that permits us to make inferences about $\boldsymbol{\gamma}$ without knowing the distribution of survival times, and that makes the use of the Cox model so widespread in survival analysis. Moreover, the exponential part ensures that the estimated hazards are non-negative.

The fundamental idea behind this model relies on the proportional hazards assumption, according to which the baseline covariates act in a multiplicative manner on the hazard function and remain constant throughout the entire study period (their effect on hazard does not change). As a result, the relative risk of two subjects, i and i', with respective covariate values $\mathbf{w}_i$ and $\mathbf{w}_{i'}$, is independent of time:

$$HR_{ii'} = \frac{h_i(t\,|\,\mathbf{w}_i)}{h_{i'}(t\,|\,\mathbf{w}_{i'})} = \frac{h_0(t)\exp\{\sum_{k=1}^{p}\gamma_k w_{ik}\}}{h_0(t)\exp\{\sum_{k=1}^{p}\gamma_k w_{i'k}\}} = \exp\left\{\sum_{k=1}^{p}\gamma_k(w_{ik} - w_{i'k})\right\}. \tag{3.15}$$

Then, in the absence of any information for data distribution, the Cox model shows a robust behavior since it allows for to the estimation of the coefficients γ_k in the exponential part of the model, and is usually preferred over parametric models when survival time information is available and there is censoring.

Frequentist Estimation of PH Cox Model Parameters

Inferences about parameters of the Cox model can be based on the equation (3.8), for right-censored data, or on the equation (3.11), in the additional presence of left truncation. However, in his seminal papers, Cox suggested the use of the partial likelihood function (Cox, 1972; Cox and Hinkley, 1974; Cox, 1975), which only depends on $\boldsymbol{\gamma}$, and is defined by generalizing the ideas of conditional and marginal likelihood. Let us consider the set of individuals who are "at risk" for experiencing the event at time t, $\mathcal{R}(t) = \{i : T_i = \min\{T_i^*, C_i\} \geq t\}$. The partial likelihood function is given by:

$$\mathcal{L}_p(\boldsymbol{\gamma}\,|\,T_i,\,\delta_i) \propto \left\{\prod_{i=1}^{n}\frac{\exp\left(\boldsymbol{\gamma}^\top\mathbf{w}_i\right)}{\sum_{j\in\mathcal{R}(T_i)}\exp\left(\boldsymbol{\gamma}^\top\mathbf{w}_j\right)}\right\}^{\delta_i}, \tag{3.16}$$

which is usually expressed in terms of the logarithm of partial likelihood function:

$$\log\mathcal{L}_p(\boldsymbol{\gamma}\,|\,T_i,\,\delta_i) \propto \sum_{i=1}^{n}\delta_i\left[\boldsymbol{\gamma}^\top\mathbf{w}_i - \log\left\{\sum_{j\in\mathcal{R}(T_i)}\exp\left(\boldsymbol{\gamma}^\top\mathbf{w}_j\right)\right\}\right]. \tag{3.17}$$

In particular, the maximum partial likelihood estimators of $\boldsymbol{\gamma}$ are found by solving the equation:

$$\frac{\partial \log \mathcal{L}_p}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{n} \delta_i \left\{ \mathbf{w}_i - \frac{\sum_{j \in \mathcal{R}(T_i)} \mathbf{w}_j \exp(\boldsymbol{\gamma}^\top \mathbf{w}_j)}{\sum_{j \in \mathcal{R}(T_i)} \exp(\boldsymbol{\gamma}^\top \mathbf{w}_j)} \right\} = \mathbf{0}. \tag{3.18}$$

It can be shown that the estimate $\hat{\boldsymbol{\gamma}}$ is unbiased, efficient, and asymptotically normally distributed, so if $n \to \infty$ then $\hat{\boldsymbol{\gamma}} \sim \mathcal{N}\big(\boldsymbol{\gamma},\, \left[\mathbb{E}\left\{\mathcal{I}(\boldsymbol{\gamma})\right\}\right]^{-1}\big)$, where $\mathcal{I}(\boldsymbol{\gamma})$ is the Fisher information matrix.

3.3.5 The Extended Cox Model with Time-Dependent Covariates

Formulation of the Extended Cox Model

So far, we have been considering the traditional Cox model for the handling of right-censored survival times, where the explanatory covariates remain constant over time. However, some cohort studies may present a heterogeneous age distribution at study entry, so the classical PH model must be extended to accommodate delayed entries. The local nature of such a model allows for easy reformulation to account for left truncation by means of the inclusion of time-dependent covariates, that is, a subject-specific vector of predictor variables, $\mathbf{y}_i$, whose values may change over the course of the study period. In this alternative formulation of the classical PH model, age is used as the time-scale (instead of the usual time-on-study) and the subsequent within-subject measurements are stored in terms of start-stop coding, as detailed in Subsection 2.1.5.

The basic idea behind the time-dependent covariates lies in the consideration of a counting process model, where a subject contributes to the risk set for an event as long as an individual is under observation at the time the event occurs, and shares the same baseline hazards function (Andersen and Gill, 1982; Therneau and Grambsch, 2000). The time-dependent relative risk model is formulated as:

$$h_i\{t \,|\, \mathcal{Y}_i(t),\, \mathbf{w}_i\} = h_0(t) \exp\{\boldsymbol{\gamma}^\top \mathbf{w}_i + \alpha\, y_i(t)\}, \tag{3.19}$$

where $\mathcal{Y}_i(t) = \{y_i(s), \tau_i \leq s \leq t\}$, denotes the intermittently observed (and therefore incomplete) covariate history for the i-th subject up to time t, whereas the parameter α is the regression coefficient that accounts for the effect on the hazard of the time-dependent covariate evaluated at t, $y_i(t)$.

In this approach, the hazard ratio at time t between two subjects i and i', with baseline covariates $\mathbf{w}_i$ and $\mathbf{w}_{i'}$, and time-dependent covariates $y_i(t)$ and $y_{i'}(t)$, is no longer constant in time, but depends on the current value at time t of the corresponding covariate values for both subjects:

$$\begin{aligned} HR_{ii'}(t) &= \frac{h_i(t\,|\,\mathcal{Y}_i(t),\, \mathbf{w}_i)}{h_{i'}(t\,|\,\mathcal{Y}_{i'}(t),\, \mathbf{w}_{i'})} = \frac{h_0(t)\exp\{\sum_{k=1}^{p}\gamma_k w_{ik} + \alpha\, y_i(t)\}}{h_0(t)\exp\{\sum_{k=1}^{p}\gamma_k w_{i'k} + \alpha\, y_{i'}(t)\}} \\ &= \exp\left[\sum_{k=1}^{p}\gamma_k(w_{ik} - w_{i'k}) + \alpha\,\{y_i(t) - y_{i'}(t)\}\right]. \end{aligned} \tag{3.20}$$

Frequentist Estimation of Parameters from the Extended Cox Model

The estimation of $\boldsymbol{\gamma}$ and α is again based on the partial log-likelihood function. Let us assume an "at-risk" function $R_i(t)$ so that:

$$R_i(t) = \begin{cases} 0 & \text{if the subject } i \text{ is not at risk at time } t \\ 1 & \text{if the subject } i \text{ is at risk at time } t \end{cases}$$

Then, the partial log-likelihood for a time-dependent relative risk model is:

$$\begin{aligned} &\mathcal{L}_p(\boldsymbol{\gamma}\,|\,\tau_i,\, T_i,\, \delta_i) = \\ &= \sum_{i=1}^{n}\left(R_i(t)\exp\{\boldsymbol{\gamma}^\top\mathbf{w}_i + \alpha\, y_i(t)\} - \log\left[\sum_{j\in\mathcal{R}(T_i)} R_j(t)\exp\{\boldsymbol{\gamma}^\top\mathbf{w}_j + \alpha\, y_j(t)\}\right]\right). \end{aligned} \tag{3.21}$$

The main drawback of the PH model is the fact that this extension can only accommodate exogenous or external time-dependent variables, that is, variables whose whole trajectory is known at every time point without any measurement error. On the contrary, the maximization of the partial log-likelihood function leads to regression coefficient estimates which are asymptotically biased (Prentice, 1982). However, in many longitudinal studies (especially those involving medical research), we do not know all the time-dependent survival covariates history, but only their values at specific time points where the repeated measurements have taken place. The common procedure to circumvent this issue consists of adopting an approach in which the observed values remain constant between two subsequent time measurements and latent terms are not considered, which very often constitutes an inadequate approach in case we are dealing with time-dependent endogenous or internal covariates. Thus, the accuracy of measurements taken on endogenous processes are usually contaminated by a random measurement error, which is inherent in the observed process because the value is affected by unpredictable subject-specific features.

In the case of the HI dataset, the time-dependent variable is the annual rate of emergency claims, here used as an indirect measurement of the individual's health status, which is of endogenous nature. Consequently, we argue that the relationship established between the frequency of use of emergency medical services recorded by subjects and their particular health status will be affected by a random noise due to the inability to instantaneously assess the health of an individual. For example, we must keep in mind that each person has a subjective perception of well-being. This means that one subject experiencing pain may choose to visit the emergency room, whereas another subject suffering from an identical disease may not consider such a visit necessary. Moreover, a superimposition of services with the public health system exists, as commented in previous chapters, which can mean that the insurance company's data may not completely report the level of severity of a subject affected by pain. Consequently, instead of relating the mortality risk to the observed measurements for each subject, $\mathcal{Y}_i(t) = \{y_i(s), \quad \tau_i \leq s < t\}$, we take into account the smooth evolution over time of the expected (underlying) error-free measurement process, $\mathcal{M}_i(t) = \{\mu_i(s), \quad \tau_i \leq s < t\}$, where the complete path of the longitudinal response until time t is now known.

3.4 Specification of the Standard Joint Model Approach

Let us consider that the subject's risk for an event depends on the expected value of the longitudinal variable at time t, denoted by $\mu_i(t)$. Moreover, let us assume a time scale of time (years) over the age of 65, and let T_i^* be the true survival time for the i-th subject. We also define an independent random variable $\tau_i \geq 0$ as the time at which a policyholder enters the study after the age of 65, giving rise to left truncation when $\tau_i > 0$. Only subjects reaching the threshold age can be sampled from the target population, i.e. $T_i^* > \tau_i$, otherwise they can not be observed. In addition, once the subject enters the study, the true survival times are subject to the usual right censorship mechanism, denoted by a potential censoring time C_i. This means we can only know the observed survival time for the i-th recruited individual, $T_i = \min\{T_i^*, C_i\}$, as well as a dichotomous event indicator $\delta_i = \mathbb{I}(T_i^* \leq C_i)$. We assume the conditional independence of the random variables $\{\tau_i, T_i^*, C_i\}$ given the covariate information, as well as that τ_i and C_i are independent of the event of interest. The probabilistic distribution of the event times is conditional on these being older than their left truncation times, $\mathcal{S}_i\{t \,|\, T_i^* > \tau_i, \mathcal{M}_i(t), \mathbf{w}_i\} = \Pr\{T_i^* > t \,|\, T_i^* > \tau_i, \mathcal{M}_i(t), \mathbf{w}_i\}$.

Building on the longitudinal analysis considered in Subsection 3.2.3, repeated count sequences and the mortality risk processes can be connected by assuming the conditional independence hypothesis (SPM framework). The JM for the i-th subject, $i = 1, \ldots, n$, is summarized at each time t within the study by two submodels, one for the longitudinal process and the other for the hazard process:

$$\begin{cases} y_i(t) = \log\{e_i(t)\} + \mathbf{x}_i^\top(t)\boldsymbol{\beta} + \mathbf{z}_i^\top(t)\mathbf{b}_i + \varepsilon_i(t) = \mu_i(t) + \varepsilon_i(t), \\ h_i\{t \,|\, T_i^* > \tau_i, \mathcal{M}_i(t), \mathbf{w}_i\} = h_0(t) \exp\left[\boldsymbol{\gamma}^\top \mathbf{w}_i + \alpha\, \mathcal{F}\{\mu_i(t)\}\right], \\ \mathbf{b}_i \sim \mathcal{N}_{q+1}(\mathbf{0}, \mathbf{D}), \quad \varepsilon_i(t) \sim \mathcal{N}(0, \sigma_\varepsilon^2), \end{cases} \tag{3.22}$$

where $h_0(t)$ denotes the baseline risk function, $\mathbf{w}_i$ the subject baseline survival covariates, $\boldsymbol{\gamma}$ the vector of the corresponding regression parameters, and $\mathcal{F}(\cdot)$ is a functional form which specifies a proper way for the expected longitudinal information, provided by $\mathbb{E}\{y_i(t) \,|\, \mathbf{b}_i\} = \mu_i(t)$, to be accounted for in survival. Because $\mu_i(t) > 0$ in a counting process, $\mathcal{F}(\cdot)$ is positively defined and increases with t. The parameter α quantifies the degree of the (constant) association between the particular longitudinal evolution until time t, and the corresponding death hazard. Specifically, the quantity $\exp(\alpha)$ returns the hazard ratio for a one-unit increase in the value $\mathcal{F}\{\mu_i(t)\}$ at time snapshot t.

Although the function $h_0(t)$ traditionally remains unspecified in the PH literature (Section 3.3.4), this condition is not kept when using joint modeling techniques. The accommodation of random effects makes impossible that the parameter estimates continue to be based on the partial likelihood function, since this would result in an underestimation of the standard errors of the estimates (Hsieh et al., 2006). It is therefore necessary to adopt a parametric specification for the time-dependent baseline risk function. In the absence of any information regarding the distribution of the event times, a B-spline approximation provides enough flexibility to capture the shape that the unknown function $h_0(t)$ may adopt (Rizopoulos, 2012). For this approximation, an equally-spaced vector $\boldsymbol{\lambda}_{h_0}$ of Q_{h_0} knots is placed on $[t_{\min}, t_{\max}]$, and then $\log\{h_0(t)\}$ is approximated through a linear combination of $R_{h_0} = (Q_{h_0} - 1) + d_{h_0}$ B-spline basis functions of degree d_{h_0}:

$$\log\{h_0(t)\} = \sum_{r=1}^{R_{h_0}} \gamma_{h_0,r} \, B_{d_{h_0},r}\left(t, \boldsymbol{\lambda}_{h_0}\right), \tag{3.23}$$

where $\{B_{d_{h_0},r}(t, \boldsymbol{\lambda}_{h_0}),\ r = 1, \ldots, R_{h_0}\}$ denotes the set of B-spline functions, and $\boldsymbol{\gamma}_{h_0} = (\gamma_{h_0,1}, \ldots, \gamma_{h_0,R_{h_0}})^\top$ is the R_{h_0}-dimensional vector of the unknown B-spline coefficients.

The most important advantage of the B-spline implementation is that it allows for local control of the curve shape, since each knot influences the whole curve. However, the crucial point resides in the specification of the number and location of knots in order to avoid overfitting (thus reducing the computational cost). When approximating the baseline hazard function, the knots are usually allocated at the corresponding percentiles of the observed death times, but this is not always an optimum computational strategy since there could be specific time regions where $h_0(t)$ needs to be more defined, so the knots in this region should be placed closer together. In addition, we do not know the total number of knots to be employed, and therefore we can not determine with certainty the type of percentiles to be used (there could be terciles, quartiles, quintiles, etc.). A more formal approach suggests starting with B-spline basis functions of time, and then control the smoothness using a roughness penalty term (which is always positive) based on k-th order differences of adjacent B-spline coefficients. These are the penalized B-splines, also called P-splines (Eilers and Marx, 1996), and they allow for an automatic smoothing selection, as well as provide efficient computational methods comparing to the classical B-splines. In this thesis, the function $h_0(t)$ is approximated by means of quadratic P-splines (section 3.8 provides more details), with equally-spaced internal knots ranging from zero to the highest observed time.

A standard approach to relating longitudinal counts to survival is undertaken by associating the expected longitudinal value to the time-to-death using the identity function, $\mathcal{F}\{\mu_i(t)\} = Id\{\mu_i(t)\} = \mu_i(t)$. However, instead of taking just a single time point, in some cases it may be more relevant to consider the whole path of the expected longitudinal outcome (Figure 3.4). In particular, an extension of the aforementioned approach is to include the entire background previous to measurement at time t (Abrahamowicz et al., 2011). Furthermore, we assume that historical effects fade over time, so the more distant history is less relevant than the more recent. Thus, $\mathcal{F}(\cdot)$ transformation can be defined to account for the cumulative and recency-weighted area under the whole longitudinal profile up to time $t > \tau_i$:

$$\mathcal{F}\left\{\mu_i(t)\right\} = \int_{\tau_i}^{t} \varpi(t-s)\mu_i(s)\,\mathrm{d}s, \qquad \tau_i \leq s \leq t, \tag{3.24}$$

where $\varpi\left(\cdot\right)$ is the selected average weighting function.

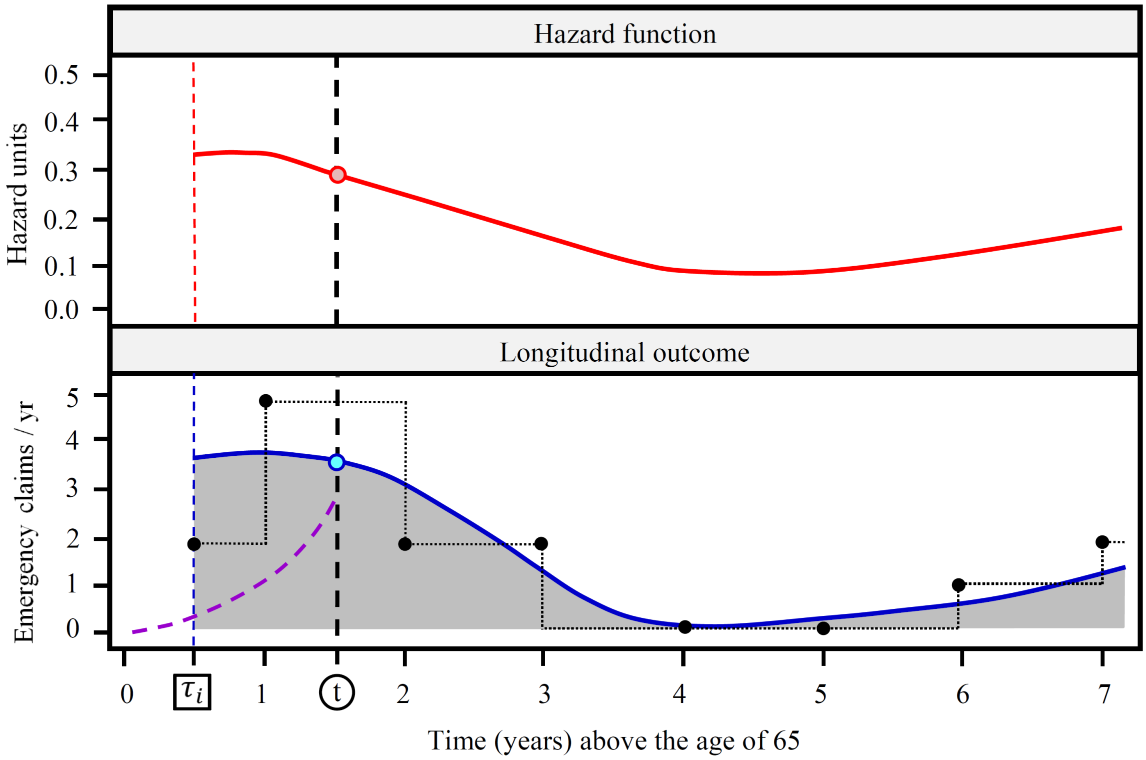


Figure 3.4. Graphical representation of the main idea behind the cumulative and recency-weighted joint modeling formulation for a subject from the HI dataset. Here, we relate the recency-weighted area under the expected longitudinal response (bottom panel) with the instantaneous risk for an event (top panel).

Given the importance of the latest information in explaining current health status, we introduced an exponential function with rate parameter $\nu > 0$ in order to assign different weights for each of the past observed longitudinal values,

$$\varpi(t-s) = \nu \exp\{-\nu(t-s)\}, \quad \tau_i \leq s \leq t,$$

and our functional form to include subject-specific longitudinal information in survival becomes:

$$\mathcal{F}\{\mu_i(t)\} = \int_{\tau_i}^{t} \nu \exp\{-\nu(t-s)\}\big[\log\{e_i(t)\} + \mathbf{x}_i^\top(s)\boldsymbol{\beta} + \mathbf{z}_i^\top(s)\mathbf{b}_i\big]\mathrm{d}s, \quad \tau_i \leq s \leq t.$$

3.5 Bayesian Estimation of the Standard Joint Model

We propose a Bayesian estimation to obtain estimates of the joint model parameters. Let $\mathcal{D}_n = \{(\mathbf{y}_i, \tau_i, T_i, \delta_i),\ i = 1, \ldots, n\}$ denote the information from our original dataset with n subjects. Further, let us denote the unknown parameter vector as $\boldsymbol{\theta} = (\boldsymbol{\theta}_y^\top, \boldsymbol{\theta}_t^\top, \boldsymbol{\theta}_b^\top)^\top$, where $\boldsymbol{\theta}_y$ collects the parameters regarding longitudinal approach, $\boldsymbol{\theta}_t$ collects those parameters related to the time-to-event response, and $\boldsymbol{\theta}_b$ collects the components of the random-effects covariance matrix. As commented in Section 3.4, the formulation of joint models under the SPM assumption states that a common random effects structure accounts for both the association between the longitudinal and event outcomes, and the correlation between the repeated measurements in the longitudinal process. Then, given the shared random effects of the i-th subject, both the longitudinal and time-to-event processes are independent, as well as the subject's n_i longitudinal responses:

$$\begin{aligned} p(\mathcal{D}_n \,|\, \mathbf{b}_i, \boldsymbol{\theta}) &= \prod_{i=1}^{n} p_y(\mathbf{y}_i \,|\, \mathbf{b}_i, \boldsymbol{\theta})\, p_t\,\{T_i, \delta_i \,|\, T_i^* > \tau_i, \mathcal{M}(t), \boldsymbol{\theta}\}, \\ p_y(\mathbf{y}_i \,|\, \mathbf{b}_i, \boldsymbol{\theta}) &= \prod_{j=1}^{n_i} p_y\,\{y_i(t_{ij}) \,|\, \mathbf{b}_i, \boldsymbol{\theta}\}, \end{aligned} \tag{3.25}$$

where $p_y(\cdot)$ and $p_t(\cdot)$ are the conditional likelihood functions for the longitudinal and survival processes, respectively.

In the standard JM, the longitudinal responses are assumed to be normally distributed conditional on random effects, so the joint density for the longitudinal process for the i-th individual can be written as:

$$\begin{aligned} p_y(\mathbf{y}_i \,|\, \mathbf{b}_i, \boldsymbol{\theta}) &= \prod_{j=1}^{n_i} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{\{y_i(t_{ij}) - \mu_i(t_{ij})\}^2}{2\sigma^2}\right] \\ &= \frac{1}{(2\pi\sigma^2)^{n_i/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}\{y_i(t_{ij}) - \mu_i(t_{ij})\}^2\right], \end{aligned} \tag{3.26}$$

whereas the i-th conditional density function for the survival part must account for left truncation, as detailed in Subsection 3.3.2:

$$\begin{aligned}
p_t\left\{T_i,\,\delta_i\,|\,T_i^* > \tau_i, \mathcal{M}_i(t), \boldsymbol{\theta}\right\} &= \frac{p_t\left\{T_i, \delta_i\,|\,\mathcal{M}_i(t), \boldsymbol{\theta}\right\}}{\Pr\left\{T_i^* > \tau_i\,|\,\mathcal{M}_i(t), \boldsymbol{\theta}\right\}} \\
&= \frac{\big[p_t\left\{T_i\,|\,\mathcal{M}_i(t), \boldsymbol{\theta}\right\}\big]^{\delta_i}\,\big[\mathcal{S}_i\{T_i\,|\,\mathcal{M}_i(t), \boldsymbol{\theta}\}\big]^{1-\delta_i}}{\mathcal{S}_i\{\tau_i\,|\,\mathcal{M}_i(t), \boldsymbol{\theta}\}} \\
&= \frac{\big[h_i\{T_i\,|\,\mathcal{M}_i(t), \boldsymbol{\theta}\}\,\mathcal{S}_i\{T_i\,|\,\mathcal{M}_i(t), \boldsymbol{\theta}\}\big]^{\delta_i}\,\big[\mathcal{S}_i\{T_i\,|\,\mathcal{M}_i(t), \boldsymbol{\theta}\}\big]^{1-\delta_i}}{\mathcal{S}_i\{\tau_i\,|\,\mathcal{M}_i(t), \boldsymbol{\theta}\}} \\
&= \big[h_i\{T_i\,|\,\mathcal{M}_i(t), \boldsymbol{\theta}\}\big]^{\delta_i}\,\frac{\mathcal{S}_i\{T_i\,|\,\mathcal{M}_i(t), \boldsymbol{\theta}\}}{\mathcal{S}_i\{\tau_i\,|\,\mathcal{M}_i(t), \boldsymbol{\theta}\}} \qquad (3.27) \\
&= \big[h_i\{T_i\,|\,\mathcal{M}_i(t), \boldsymbol{\theta}\}\big]^{\delta_i}\,\frac{\exp\big[-\mathcal{H}_i\{T_i\,|\,\mathcal{M}_i(t), \boldsymbol{\theta}\}\big]}{\exp\big[-\mathcal{H}\{\tau_i\,|\,\mathcal{M}_i(t), \boldsymbol{\theta}\}\big]} \\
&= \Big(\exp\big[\log\{h_0(T_i)\} + \boldsymbol{\gamma}^\top \mathbf{w}_i + \alpha\,\mathcal{F}\{\mu_i(T_i)\}\big]\Big)^{\delta_i} \\
&\quad \times \exp\Bigg(-\int_{\tau_i}^{T_i} \exp\big[\log\{h_0(s)\} + \boldsymbol{\gamma}^\top \mathbf{w}_i + \alpha\,\mathcal{F}\left\{\mu_i(s)\right\}\big]\,\mathrm{d}s\Bigg).
\end{aligned}$$

Taking advantage of the expressions (3.23), (3.26), and (3.27), the overall joint likelihood conditioned on random effects can be properly formulated to tackle left truncation through the conditional independence assumption:

$$\begin{aligned}
&p\,(\mathcal{D}_n\,|\,\mathbf{b}_i, \boldsymbol{\theta}) = \\
&= \prod_{i=1}^{n}\prod_{j=1}^{n_i} \frac{1}{(2\pi\sigma^2)^{n_i/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}\{y_i(t_{ij}) - \mu_i(t_{ij})\}^2\right] \\
&\times \left(\exp\left[\sum_{r=1}^{R_{h_0}} \gamma_{h_0,r}\, B_{d_{h_0},r}(T_i, \boldsymbol{\lambda}_{h_0}) + \boldsymbol{\gamma}^\top \mathbf{w}_i + \alpha\,\mathcal{F}\left\{\mu_i(T_i)\right\}\right]\right)^{\delta_i} \qquad (3.28) \\
&\times \exp\left(-\int_{\tau_i}^{T_i} \exp\left[\sum_{r=1}^{R_{h_0}} \gamma_{h_0,r}\, B_{d_{h_0},r}(s, \boldsymbol{\lambda}_{h_0}) + \boldsymbol{\gamma}^\top \mathbf{w}_i + \alpha\,\mathcal{F}\left\{\mu_i(s)\right\}\right] \mathrm{d}s\right).
\end{aligned}$$

The mean estimates of parameters and random effects are then derived by using Markov chain Monte Carlo (MCMC) algorithms, which enable inferences to be made by efficiently drawing a sample from the posterior distribution through the Bayes' rule:

$$\pi\left(\boldsymbol{\theta}, \mathbf{b}_i\,|\,\mathcal{D}_n\right) \,\propto\, p\,(\mathcal{D}_n\,|\,\mathbf{b}_i, \boldsymbol{\theta})\,\pi(\boldsymbol{\theta}, \mathbf{b}_i) = p\,(\mathcal{D}_n\,|\,\mathbf{b}_i, \boldsymbol{\theta})\,p_b(\mathbf{b}_i\,|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta}), \qquad (3.29)$$

where $p_b(\cdot)$ is the prior distribution for the random effects, usually assumed to follow a standard multivariate normal distribution, and $\pi(\boldsymbol{\theta})$ is the prior distribution for the full JM parameter vector $\boldsymbol{\theta}$, such that prior independence is assumed between its components.

3.6 Missing Data Mechanism in the Standard Joint Model

Missing or incomplete data are quite common in longitudinal studies, especially in medical research, and present several challenges in the analysis of data from these studies. The presence of missing data results in a reduction in the available data which we use to determine the Bayesian estimate of the parameters, directly affecting the grade of precision of the parameter inferences. Moreover, the lack of this information can be due to causes directly related to the behavior analyzed, in which case the missing pattern can introduce bias and thereby lead to invalid results. It is this problem that most missing data studies focus on, leading to the missing mechanisms outlined in Subsection 3.2.1.

In the particular case of the HI dataset, let $\mathbf{y}_i^{obs} = \{y_i(t_{ij}) : \tau_i < t_{ij} < T_i^*, \, j = 1, \ldots, n_i\}$ denote all the observed longitudinal responses of the i-th subject before the death event $\mathcal{E}$, and let $\mathbf{y}_i^{mis} = \{y_i(t_{ij}) : t_{ij} \geq T_i^*, \, j = 1, \ldots, n_i'\}$ denote all those measurements that would have been observed from the event time to the study close date if the subject had still been alive. The dropout pattern can be treated as a random variable, and is directly obtained by jointly considering the observed and missing longitudinal values:

$$\begin{aligned} p(T_i^* \,|\, T_i^* > \tau_i, \mathbf{y}_i^{obs}, \mathbf{y}_i^{mis}, \, \boldsymbol{\theta}) &\propto p(T_i^*, \boldsymbol{b}_i \,|\, \mathbf{y}_i^{obs}, \mathbf{y}_i^{mis}, \, \boldsymbol{\theta}) \\ &\propto p(T_i^* \,|\, T_i^* > \tau_i, \boldsymbol{b}_i, \, \mathbf{y}_i^{obs}, \mathbf{y}_i^{mis}, \, \boldsymbol{\theta}) \, p_b(\boldsymbol{b}_i \,|\, \mathbf{y}_i^{obs}, \mathbf{y}_i^{mis}, \, \boldsymbol{\theta}) \quad (3.30) \\ &\propto p(T_i^* \,|\, T_i^* > \tau_i, \boldsymbol{b}_i, \, \boldsymbol{\theta}) \, p_b(\boldsymbol{b}_i \,|\, \mathbf{y}_i^{obs}, \mathbf{y}_i^{mis}, \, \boldsymbol{\theta}), \end{aligned}$$

Thus, the time-to-event mechanism depends on both the observed missing longitudinal responses through the posterior distribution of the random effects, which means that the observed data can not be considered a random sample from the target population. In practice, this issue leads us to conclude that SPM joint models allow for the proper accommodation of the MNAR mechanism.

3.7 Goodness-of-Fit for the Standard Joint Model

The deviance information criterion (Spiegelhalter et al., 2002), denoted by DIC, is used in this thesis as a Bayesian goodness-of-fit measure for the model selection. The DIC balances fit with model complexity, and in the JM context, is defined by

$$\mathrm{DIC} = D(\bar{\boldsymbol{\theta}}, \bar{\mathbf{b}}_i) + 2\, p_D, \qquad (3.31)$$

where $D(\boldsymbol{\theta}, \mathbf{b}_i) = -2 \sum_{i=1}^{n} \log p(\mathcal{D}_n \,|\, \boldsymbol{\theta}, \mathbf{b}_i)$ is the Bayesian deviance, and $p_D = \overline{D(\boldsymbol{\theta}, \mathbf{b}_i)} - D(\bar{\boldsymbol{\theta}}, \bar{\mathbf{b}}_i)$ measures the model complexity. Accordingly, the term $\overline{D(\boldsymbol{\theta}, \mathbf{b}_i)}$ denotes the posterior mean deviance of the JM parameters, whereas the term $D(\bar{\boldsymbol{\theta}}, \bar{\mathbf{b}}_i)$ is the deviance evaluated in the posterior mean of the JM parameters. The score provided by DIC serves as the basis for ranking the fitted models, where lower scores of DIC correspond to better model fits. It is important to point out that DIC value is not constant, since it is calculated from the MCMC output, DIC = DIC($\boldsymbol{\theta}, \mathbf{b}_i$).

As a general rule of thumb, if the respective DIC's of two models differ by more than 5 points, the one with the smaller DIC may be chosen (Lesaffre and Lawson, 2012).

3.8 Application of the Standard Joint Model

In order to assess the most adequate way to take into account the subject's claim history when evaluating the death risk, we first consider a standard JM in which the longitudinal submodel assumes linear trends and only accounts for the current expected value, $\mathcal{F}(\cdot) = Id\,(\cdot)$, so that $\mathcal{F}\{\mu_i(t)\} = \mu_i(t)$. This implies considering only the current health status when evaluating the subject's mortality risk, which could result in a naive approximation due to not considering the most recent medical history. As commented in Section 3.4, in our case it may be much more realistic to take into account all the most recent medical history, giving more importance to the most recent emergency claims. In particular, we also considered an extended version of the standard JM, with a longitudinal submodel that accounted for the cumulative and exponentially-weighted expected value:

$$\mathcal{F}\{\mu_i(t)\} = \int_{\tau_i}^{t} \nu \exp\{-\nu(t-s)\}\mu_i(s)\mathrm{d}s, \quad \tau_i < s \leq t. \tag{3.32}$$

The fixed effects of the longitudinal outcome were set at $\{\beta_0, \beta_1\}$, respectively referring to the intercept term and the observation time (directly linked to the subject's age). Initially, we tested a LMM with a single random intercept b_{i0}, and also a random intercept and slope $\{b_{i0}, b_{i1}\}$, with correlation ρ such that $\mathrm{Cov}(b_{i0}, b_{i1}) = \rho\,\sigma_{b_0}\,\sigma_{b_1}$. The Bayesian LMM with two random effects provides a mean estimate for ρ of 0.824 (95% CI: 0.811, 0.838), also leading to a similar DIC than that obtained in the LMM with only a random intercept. For illustrative purposes, we decided throughout this thesis not to include the latent correlation between measurements along time in the JM fitting, and we only considered a single random effect in the longitudinal submodel.

Finally, we have two proposals with a single random intercept each, one with the current value of the underlying longitudinal response, and the other with the weighed cumulative effect. Both of these assumptions are summarized by the following equations:

$$\begin{cases} y_i(t) = \log\{e_i(t)\} + \beta_0 + b_{i0} + \beta_1 t + \varepsilon_i(t) = \mu_i(t) + \varepsilon_i(t), \\ h_i\{t\,|\,T_i^* > \tau_i,\, \mathcal{M}_i(t),\, \boldsymbol{\theta}_t\} = h_0(t)\exp\left[\gamma_g w_{gi} + \alpha\,\mathcal{F}\{\mu_i(t)\}\right], \\ b_{i0} \sim \mathcal{N}(0,\, \sigma^2_{b_0}),\;\; \varepsilon_i(t) \sim \mathcal{N}(0,\, \sigma^2_{\varepsilon}). \end{cases} \tag{3.33}$$

We use independent univariate vague normal priors for the fixed and random regression coefficients, as well as for the measurement error. In the case of the weighting function, the posterior estimation of the rate parameter is performed by adopting a diffuse uniform prior, $\nu \sim \mathcal{U}(a_\nu, b_\nu)$. Because $\nu > 0$, we set $a_\nu = 0$, while for the second hyper-parameter we set $b_\nu = 20$ to expresses the uncertainty around the value of ν.

Following the suggestions of Rizopoulos (2016), the function $h_0(t)$ in the relative risk model is expanded into a quadratic P-spline basis, with $Q_{h_0} = 8$ equally-distanced internal knots ($R_{h_0} = 9$). Specifically, we use Bayesian P-splines (Lang and Brezger, 2004), so $\boldsymbol{\gamma}_{h_0}$ is assumed to follow the improper prior

$$\pi(\boldsymbol{\gamma}_{h_0} \,|\, \tau_{h_0}) \propto \tau_{h_0}^{\mathrm{rk}(\mathbf{M}_{h_0})/2} \exp\left(-\frac{\tau_{h_0}}{2}\boldsymbol{\gamma}_{h_0}^\top \mathbf{M}_{h_0}\boldsymbol{\gamma}_{h_0}\right), \tag{3.34}$$

which in practice yields the hierarchical multivariate Gaussian prior

$$\boldsymbol{\gamma}_{h_0} \,|\, \tau_{h_0} \sim \mathcal{N}_{R_{h_0}}(\mathbf{0},\, \tau_{h_0}\mathbf{M}_{h_0}). \tag{3.35}$$

Here, τ_{h_0} is the smoothing parameter, which is assumed to be $\tau_{h_0} \sim \mathcal{G}(1, 0.005)$, and the penalty matrix $\mathbf{M}_{h0}$ is obtained by solving the following system of equations expressed in matrix form:

$$\mathbf{M}_{h_0} = \Delta_k^\top \Delta_k + 10^{-6}\,\mathbf{I}, \tag{3.36}$$

where Δ_k is the difference matrix of order k, while the term $10^{-6}\,\mathbf{I}$ introduces a small "ridge penalty" to avoid a linearly dependent system. Moreover, the gender information is included as an exogenous covariate in the relative risk model, $\mathbf{w}_i \equiv w_{gi}$ (man:0, woman:1), for which a diffuse normal prior is specified. Finally, the longitudinal and time-to-event outcomes are related using a constant association parameter, for which a diffuse normal prior is assumed. The two longitudinal approaches within the standard JM are fitted using the Bayesian software `JAGS` (Plummer, 2003), version 4.2.0. This software is called from the `R`-environment by means of the package `jagsUI` (Kellner, 2016). The posterior mean estimates of $(\boldsymbol{\theta}, b_{i0})$ are obtained for each JM, and the calculation of the DIC has been programed from the corresponding Bayesian deviance $D(\boldsymbol{\theta}, b_{i0}) = -2\sum_{i=1}^{n}\log\{p(\mathbf{y}_i \,|\, \boldsymbol{\theta}, b_{i0}\})$. For illustrative purposes, the software code to fit the standard JM which considers the current longitudinal value is provided in the Appendix B of this thesis. This code is, in turn, an adaptation of the code used in the `R` package `JMbayes` (Rizopoulos, 2016). In the case of the JM with weighted cumulative effects of the longitudinal response, the necessary changes to introduce an exponential weighting function has been implemented.

The results (Table 3.1) indicate a strong association between emergency claims and survival, so that each unit increase in the current value of the expected log-transformed emergency claims per year involves a $\exp(\bar{\alpha}_{\text{value}}) = 6.73$-fold increase (95% CI: 5.20, 9.18) in the subject's mortality risk when accounting for the current longitudinal value. In the case of accounting for the weighted cumulative effect, the association is also strongly related to the death hazard, so that a one-unit increase in the exponentially weighted area under the log emergency claims per year trajectory leads to a $\exp(\bar{\alpha}_{\text{cum}}) = 6.75$-fold increase in the mortality risk (95% CI: 5.11, 9.03). Thus, we have a positive relationship between the frequency of use of non-routine medical services and the corresponding death hazard. From a goodness-of-fit perspective, the comparison between the fitted joint models is performed using the DIC.

	JM with current value				JM with weighted cumulative effect			
Parameter	Mean	SE	$q_{2.5\%}$	$q_{97.5\%}$	Mean	SE	$q_{2.5\%}$	$q_{97.5\%}$
Longitudinal								
β_0	0.302	< 0.001	0.283	0.320	0.303	< 0.001	0.282	0.318
β_1	0.008	< 0.001	0.007	0.009	0.008	< 0.001	0.007	0.010
ν	–	–	–	–	9.978	0.051	8.087	11.882
σ_{b0}	0.311	< 0.001	0.302	0.320	0.311	< 0.001	0.303	0.320
σ_{ε}	0.498	< 0.001	0.494	0.503	0.497	< 0.001	0.492	0.504
Survival								
γ_g	-0.067	0.004	-0.246	0.107	-0.070	0.003	-0.247	0.112
Association								
α	1.907	0.006	1.648	2.217	1.910	0.008	1.623	2.201
Goodness-of-fit								
DIC	59108.8				59105.4			

Table 3.1. Posterior summaries for all parameters of the standard JM when considering both the expected log-response and the exponentially-weighted cumulative effect of the expected log-response. Mean, standard error, 95% credible interval and DIC are sampled for each parameter from the corresponding posterior distribution.

Note the great similarity between the association parameters of these results and those obtained under the current-value association structure, thus emphasizing that only the most recent past claims have a real influence on the expected survival. The high estimate for the rate parameter of the exponential weighting function indicates so, in practice, it yields that approximately only the 0.30 years (i.e., four months) prior to t are strongly related to the current death hazard. Considering the information provided by the variables, a great similarity is observed between the values obtained. Moreover, we must highlight that the `sex` baseline covariate shows no evidence of different behavior between men and women when analyzing our target sample.

The DIC score obtained when considering the current expected value is slightly less than that resulting from the consideration of a weighted cumulative parametrization of the longitudinal response. However, this last one does not show enough evidence to conclude that there is an improvement in the fitting, since its DIC estimate is not at least 5 points less than that estimated with the first JM approach.

3.9 Discussion of the Standard Joint Model Results

We demonstrate the benefits of taking a joint approach to modeling longitudinal and time-to-event processes by using a large dataset from a Spanish medical company. This contains historical insurance information for 5470 policyholders aged 65 and more (37.6% men and 62.4% women) across the study window. We have implemented a standard joint model with two functional forms; first, we have considered the effect of the longitudinal response on the survival, and we have also incorporated the weighted cumulative effect of emergency claims information, which places greater emphasis on more recent demand than on past service demand. Both models lead to similar conclusions, and no evidence has been obtained that one is clearly preferable over the other. However, the JM which considers the weighted area under the expected profile provides slightly better results.

In methodological terms, we have applied a log-transformation to the longitudinal response in order to allow residuals to conform to normality. However, this naive approach towards handling panel count rates involves in practice a series of methodological criticisms. Firstly, we have to take into account that our observations indeed come from a data generation process that produces counts, which are restricted to a small range of possible values and have a variance related to the mean. Thus, any solution intended to "mask" this fact can not be the best option to collect its real behavior. In fact, this option usually leads to biased inferences, as well as the homokedasticity assumption is usually not achieved. Furthermore, the fact that we work in logarithmic terms makes the direct interpretation of the data difficult, compelling us to constantly undo the applied transformation.

In favor of the use of the logarithmic transformation in our count data, it could be argued that the introduction of a specific counting model, as in the case of the Poisson distribution, also implies a logarithmic transformation of the observed response $y_i(t) = \texttt{claimyr}_i(t)$. When performing a log-transformation, we are using $\log\{y_i(t)\}$ as our response variable in the LMM, that is, we are actually modeling the expected value $\mathbb{E}\big[\log\{y_i(t)\}\,|\,\mathbf{b}_i\big]$, whereas in the Poisson approach we are modeling $\log\left[\mathbb{E}\left\{y_i(t)\,|\,\mathbf{b}_i\right\}\right]$. Consequently, our results are here affected by Jensen's inequality, which in the case of the transformation $g\left\{y_i(t)\right\} = \log\left\{y_i(t)\right\}$ (a concave function), states:

$$\mathbb{E}\left[\log\{y_i(t)\}\,|\,\mathbf{b}_i\right] \leq \log\left[\mathbb{E}\{y_i(t)\,|\,\mathbf{b}_i\}\right], \quad y_i(t) > 0. \tag{3.37}$$

Thus, the results in both cases will not be exactly the same, as will be shown throughout the following chapter.

CHAPTER 4

JOINT MODEL FOR COUNTS AND LEFT-TRUNCATED TIME-TO-EVENT DATA

4.1 Principles of Joint Models for Counts and Delayed Entries

In the previous chapter we have worked with the standard JM approach using the log-transformation of the observed longitudinal response. However, the variable of interest in the longitudinal part, established in the HI dataset as the rate of emergency claims per year (including ambulance services, hospitalizations and non-routine medical visits), can only take into account non-negative integers. So the longitudinal expected response must account for non-Gaussian data of discrete nature. Previous approaches to this issue have been proposed in the field of joint modeling. For example, Rizopoulos and Ghosh (2011) defined a Bayesian JM to relate multiple longitudinal outcomes (of discrete or continuous nature) to a time-to-event. Murawska et al. (2012) presented a two-stage JM where the longitudinal information was summarized by either a non-linear mixed-effects model or a generalized linear mixed model (GLMM) in the first stage, while in the second, Empirical Bayes estimates of the subject-specific parameters were included as predictors in the proportional hazards model. Viviani et al. (2012) implemented an expectation-maximization algorithm to incorporate non-Gaussian data in the longitudinal response, with particular attention to Poisson and binomial responses. More recently, Ivanova et al. (2016) formulated a JM to handle continuous, discrete, or ordinal responses, where parameters were estimated using a likelihood-based approach.

A common feature of the aforementioned extensions is that they do not account for delayed entries in the survival submodel. However, in the time-to-event approach, we consider the lifetime elapsed from the moment when a subject is aged 65 until the death event, and consequently left truncation must be accounted for in all those subjects that enter the study beyond the age of 65, here established as our time zero. Additionally, in our study, most lifetimes cease to be observed at administrative closure, and some are not observed completely due to dropout. Time-to-event data, therefore, are left-truncated further than the usual censorship, and not all individuals present the same number of longitudinal measurements. We have detailed in Chapters 2 and 3 how a proper consideration of the left truncation issue in the mortality hazard can be achieved by using the subject's age as the time scale.

Our goal throughout this chapter is to relate each personalized emergency claims rating profile to time-to-death by postulating an appropriate JM. We also investigate the role played by information contained in medical records and identify a cumulative and fading effect, so that more recent records have a greater influence than older records on the death hazard. From a statistical perspective, this problem requires an innovative approach to the application of a joint framework, where a pronounced dependency pattern between longitudinal and survival outcomes for the elderly is expected. From a methodological perspective, the statistical analysis poses the challenge of handling correlated counts in the longitudinal response, and of having to extend the standard proportional hazards model by incorporating the delayed entries.

As in the standard JM approach, the relationship between longitudinal counts and survival data can be properly assessed using a shared-parameter JM, where a policyholder's emergency claims and survival outcomes are stochastically correlated by a common latent structure. Thus, conditional on the random effects, the longitudinal and time-to-event outcomes are independent, as are repeated measurements in the longitudinal process. Complete overviews of the joint modeling techniques can be found in Tsiatis and Davidian (2004), Yu et al. (2008), and Rizopoulos (2012). In the context of the application of joint modeling techniques to health insurance studies, previous work on elderly policyholders can be found under the frequentist approach in Piulachs et al. (2016), where the counting process was approximated by a log-transformation of the longitudinal outcome to assume a normal response. See also Mukherji et al. (2016), who implemented a Bayesian joint model to explore the relationship between out-of-pocket medical expenditure and hospitalizations in a longitudinal survey of Americans aged 50 or more.

4.2 Analysis of Longitudinal Count Data

4.2.1 Features of Longitudinal Counts

Within a panel data context, let $\mathbf{y}_i = \{y_i(t_{ij}),\ i = 1, \dots, n\}$ denote the observed counts for the i-th subject, recorded at a fixed set of time points t_{ij}, $j = 1, \dots, n_i$. As in the case of LMM, we want to capture both the between-subject variation and the within-subject variation generated by the repetition of measurements on a single subject. We relate each of the observed responses to a set of $p+1$ fixed effect covariates and $q+1$ random effects, respectively denoted by $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^\top$ and $\mathbf{b}_i = (b_{i0}, b_{i1}, \dots, b_{iq})^\top$.

To properly model the correlation between repeated counts measurements, we will apply the same ideas as in the case of a Gaussian response, with the only difference being that now we are dealing with counts. Given the vector $\mathbf{b}_i$ of random effects for the i-th subject, we assume that the observed measurements on this individual derive from a counting process generated by an exponential family (EF) distribution, $y_i(t) \sim \text{EF}\,\{\psi_i(t),\, \phi\}$, with probability mass function:

$$p_y\{y_i(t)\,|\,\mathbf{b}_i, \boldsymbol{\theta}_y\} = \exp\Big(\phi^{-1}\big[y_i(t)\,\psi_i(t) - b\{\psi_i(t)\}\big] + c\{y_i(t),\, \phi\}\Big). \tag{4.1}$$

Here, $b(\cdot)$ and $c(\cdot)$ are known functions, and $\psi_i(t)$ and ϕ are referred to as the canonical and the scale parameter, respectively. It can be directly shown that $\mathbb{E}\{y_i(t)\,|\,\mathbf{b}_i\} = \mu_i(t) = b'\{\psi_i(t)\}$ and $\mathbb{V}\{y_i(t)\,|\,\mathbf{b}_i\} = \sigma_i^2(t) = \phi\, b''\{\psi_i(t)\}$ (Molenberghs and Verbeke, 2005).

In many studies, the recorded counts rely on different periods of time, so the raw expected mean outcome $\hat{\mu}_i(t)$ has to be related to the associated exposure time $e_i(t)$, which plays the role of an offset variable. With this data pattern, modeling the counting rates is more relevant than working with the raw counts, with the longitudinal submodel focusing on expected mean rates of counts per time unit. That is exactly the case we are dealing with in the case of the HI dataset, as explained in Subsection 2.1.3.

4.2.2 The Generalized Linear Mixed Model

In the case of correlated non-Gaussian outcomes, the expected rate value for the i-th subject at time t is related to a set of fixed and random covariates with the introduction of a continuous and differentiable link function $g(\cdot)$,

$$\begin{cases} g\{\mu_i(t)\} = \log\{e_i(t)\} + \eta_i(t) = \log\{e_i(t)\} + \mathbf{x}_i^\top(t)\boldsymbol{\beta} + \mathbf{z}_i^\top(t)\mathbf{b}_i, \\ \mathbb{E}\{y_i(t)\,|\,\mathbf{b}_i\} = \mu_i(t) = g^{-1}\big[\log\{e_i(t)\} + \eta_i(t)\big], \\ \mathbf{b}_i \sim \mathcal{N}_{q+1}\left(\mathbf{0},\, \mathbf{D}\right). \end{cases} \tag{4.2}$$

Here, $e_i(t)$ is the exposure time effect, $\mathbf{x}_i^\top(t)$ and $\mathbf{z}_i^\top(t)$ denote row vectors of the fixed and random design matrices, respectively, while $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ and $\mathbf{b}_i = (b_{i0}, b_{i1}, \ldots, b_{iq})^\top$ are the corresponding fixed-effects and random-effects vectors. The random-effects parameters enable capturing the expression of individual departures from the overall trend, and in most cases they can be assumed to follow a multivariate normal distribution with an unspecified $(q+1) \times (q+1)$ variance-covariance matrix $\mathbf{D}$.

The Poisson Model with Random Effects

The most common choice for modeling panel counts in (4.2) results from a Poisson mixed model for longitudinal data. The marginal mean responses are usually related to the covariates and random effects information using a logarithmic link, $g(\cdot) = \log(\cdot)$, so $g^{-1}(\cdot) = \exp(\cdot)$. This ensures positive outcomes and provides a straightforward interpretation of the estimated regression parameters:

$$\begin{cases} y_i(t) \sim \text{PO}\{\mu_i(t)\}, \ \mu_i(t) = e_i(t) \exp\{\mathbf{x}_i^\top(t)\boldsymbol{\beta} + \mathbf{z}_i^\top(t)\mathbf{b}_i\}, \\ p_y\{y_i(t) \,|\, \mathbf{b}_i\} = \dfrac{\exp\{-\mu_i(t)\}\mu_i(t)^{y_i(t)}}{y_i(t)!}, \quad \mu_i(t) > 0 \\ \mathbb{E}\,\{y_i(t) \,|\, \mathbf{b}_i\} = \mathbb{V}\{y_i(t) \,|\, \mathbf{b}_i\} = \mu_i(t), \\ \mathbf{b}_i \sim \mathcal{N}_{q+1}(\mathbf{0},\, \mathbf{D}). \end{cases} \tag{4.3}$$

The Poisson mixed model allows for robust parameter estimates, even if the underlying distribution is not true, provided that the expectation is correctly specified (Gourieroux et al., 1984). However, the underlying distribution usually has an observed variance greater than the mean, so the counting response presents overdispersion. See, for example, Hausman et al. (1984), Agresti (2012), Booth et al. (2003), Molenberghs and Verbeke (2005), and Harrison (2014), for a related derivation and analysis of this alteration in the equaldispersion hypothesis. This is a common issue affecting count data, mainly due to missing information, aggregate data, or even an excess of zeros in the longitudinal outcome. In such cases, the derived inference under the Poisson model leads to biased parameter estimates or erroneous conclusions about parameter significance. A detailed discussion of these issues can be found in Zuur et al. (2009) and Hilbe (2011).

The Negative Binomial Model with Random Effects

Although there are different models for dealing with the overdispersion relative to Poisson counts, the standard negative binomial (NB) model appears in the literature as being the most obvious choice. See, for example, Ismail and Jemain (2007), Greene (2008), and Hilbe (2011). The NB mixed model for longitudinal data distribution can be easily derived from the Poisson distribution by placing a multiplicative gamma distributed random noise $\varepsilon_i(t)$ in the conditional mean response. Specifically, such a latent variable is defined in terms of shape and rate parameters by $\varepsilon_i(t) \sim \Gamma(\kappa, \kappa),\ \kappa > 0$, with $\mathbb{E}\{\varepsilon_i(t)\} = 1$ and $\mathbb{V}\{\varepsilon_i(t)\} = 1/\kappa$, so that the longitudinal counts are modeled by $y_i(t) \sim \text{PO}\{\varepsilon_i(t)\, \mu_i(t)\}$. This Poisson-gamma mixture presents a closed-form solution which, in practice, leads to the NB distribution with a dispersion parameter κ. In this distribution, the marginal mean responses can be also related to the fixed and random effects using a logarithmic link:

$$\begin{cases} y_i(t) \sim \text{NB}\{\mu_i(t), \kappa\}, \ \mu_i(t) = e_i(t) \exp\{\mathbf{x}_i^\top(t)\boldsymbol{\beta} + \mathbf{z}_i^\top(t)\mathbf{b}_i\}, \\ p_y\{y_i(t) \,|\, \mathbf{b}_i\} = \dfrac{\Gamma\{\kappa + y_i(t)\}}{\Gamma(\kappa)\, y_i(t)!} \, \dfrac{\mu_i(t)^{y_i(t)}\, \kappa^\kappa}{\{\mu_i(t) + \kappa\}^{\kappa + y_i(t)}}, \quad \mu_i(t), \kappa > 0 \\ \mathbb{E}\{y_i\,(t) \,|\, \mathbf{b}_i\} = \mu_i(t), \quad \mathbb{V}\{y_i(t) \,|\, \mathbf{b}_i\} = \mu_i(t) + \dfrac{\mu_i(t)^2}{\kappa}, \\ \mathbf{b}_i \sim \mathcal{N}_{q+1}(\mathbf{0},\, \mathbf{D}). \end{cases} \tag{4.4}$$

The NB distribution has the general canonical form of the exponential family equations for any fixed κ. Because of the quadratic expression for the variance, it is sometimes referred to as NB2 in the literature. Note that the NB distribution can actually be understood as an extension of the Poisson model when overdispersion is accounted for by parameter κ, since the NB model tends towards the Poisson model as $\kappa \to \infty$. This particularity is well-documented by Lawless (1987) and Hinde and Demétrio (1998); see also Boucher et al. (2008) for a numerical application in the field of insurance studies.

4.3 Specification of the Joint Model for Counts and Delayed Entries

Assuming a time scale of time (years) over the age of 65, let T_i^* be the true survival time for the i-th subject. We also define an independent random variable $\tau_i \geq 0$ as the time at which a policyholder enters the study after the age of 65, giving rise to left truncation for a subject when $\tau_i > 0$. Only subjects reaching the threshold age can be sampled from the target population, i.e. $T_i^* > \tau_i$; otherwise they can not be observed. In addition, once the observed subjects enter the study, their survival times are subject to the usual right censorship mechanism, respectively denoted by a potential censoring time C_i. This means we can only know the observed survival time for the i-th recruited individual, $T_i = \min\{T_i^*,\, C_i\}$, as well as a dichotomous event indicator $\delta_i = \mathbb{I}(T_i{}^* \leq C_i)$. Consequently, the probabilistic distribution of the time-to-death has to be defined according to the proportion of subjects living beyond time point t and conditional on their being older than the corresponding left truncation time, $\mathcal{S}_i(t \,|\, T_i^* > \tau_i, \mathcal{M}_i(t), \mathbf{w}_i) = \Pr\left(T_i^* > t \,|\, T_i^* > \tau_i, \mathcal{M}_i(t), \mathbf{w}_i\right)$.

Building on the analysis of longitudinal count data considered in Subsection 4.2.2, repeated count sequences and time-to-event approaches can be coupled by assuming independence between both processes given the shared random effects (conditional independence). The JM for the i-th subject, $i = 1, \ldots, n$, is postulated by a time-dependent relative risk model where the death hazard at time t takes into account the whole expected longitudinal response until t, $\mathcal{M}_i(t) = \{\mu_i(s),\, \tau_i \leq s \leq t\}$:

$$\begin{cases} \mathbb{E}\left\{y_i(t) \,|\, \mathbf{b}_i\right\} = \mu_i(t) = g^{-1}\big[\log\{e_i(t)\} + \mathbf{x}_i^\top(t)\boldsymbol{\beta} + \mathbf{z}_i^\top(t)\mathbf{b}_i\big], \\ h_i\{t \,|\, T_i^* > \tau_i,\, \mathcal{M}_i(t),\, \mathbf{w}_i\} = h_0(t)\exp\left[\boldsymbol{\gamma}^\top\mathbf{w}_i + \alpha\,\mathcal{F}\left\{\mu_i(t)\right\}\right], \\ \mathbf{b}_i \sim \mathcal{N}_{q+1}(\mathbf{0},\, \mathbf{D}). \end{cases} \tag{4.5}$$

In the above equation, $g(\cdot)$ denotes the linking function to relate the expected longitudinal response to the fixed and random effects, respectively $\boldsymbol{\beta}$ and $\mathbf{b}_i$. In case of assuming that the underlying counting process is generated by a Poisson or negative binomial distribution, $g(\cdot) = \log(\cdot)$. Furthermore, the function $h_0(t)$ denotes the baseline risk function, $\mathbf{w}_i$ is the vector with subject's exogenous covariates, and $\boldsymbol{\gamma}$ is the vector of the corresponding regression parameters. Although $h_0(t)$ traditionally remains unspecified in the proportional hazards literature, this condition becomes more flexible when addressed using joint modeling techniques. In particular, the function $h_0(t)$ is approximated by means of quadratic P-splines, in the same way as detailed in Section 3.4.

The functional form $\mathcal{F}(\cdot)$ specifies a proper manner in which the longitudinal information provided by $\mu_i(t)$ is accounted for in survival. Because $\mu_i(t) > 0$ in a counting process, $\mathcal{F}(\cdot)$ is positively defined and increases its value with t. The constant parameter α quantifies the strength of association between the particular longitudinal evolution until time t, and the corresponding death hazard. Specifically, the quantity $\exp(\alpha)$ returns the hazard ratio for a one-unit increase in the value $\mathcal{F}\left\{\mu_i(t)\right\}$ at time snapshot t.

The basic option consists of relating expected longitudinal counts to survival at each time t, which yields $\mathcal{F}(\cdot)$ as the identity function. However, in our particular case, it may be more informative to include the entire underlying profile of counts previous to measurement at time t, and at the same time assume that historical effects fade over time. Thus, the $\mathcal{F}(\cdot)$ transformation is defined to account for the cumulative and recency-weighted area under the expected longitudinal profile until t. Specifically, in our case the area is and weighted by means of an exponential weighting function (see details in Section 3.4).

In the case of handling counts with a Poisson or negative binomial distribution, the weighted area under the whole subject's longitudinal profile is included in the survival process through the following functional form:

$$\mathcal{F}\{\mu_i(t)\} = \int_{\tau_i}^{t} \nu \exp\{-\nu(t-s)\}\, e_i(t) \exp\{\mathbf{x}_i^\top(s)\, \boldsymbol{\beta} + \mathbf{z}_i^\top(s)\, \mathbf{b}_i\}\, \mathrm{d}s, \quad \tau_i \le s \le t, \tag{4.6}$$

4.4 Estimation of the Joint Model for Counts and Delayed Entries

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_y^\top, \boldsymbol{\theta}_t^\top, \boldsymbol{\theta}_b^\top)^\top$ be the JM full parameter vector defined in Subsection 3.5, and let $\mathcal{D}_n = \{(\mathbf{y}_i, \tau_i, T_i, \delta_i),\ i = 1, \dots, n\}$ be the complete information from our original dataset with n policyholders. Taking advantage of the conditional independence assumption and using (3.23), the overall joint likelihood for $(\boldsymbol{\theta}, \mathbf{b}_i)$ is formulated as in the standard JM:

$$\begin{aligned} p\left(\mathcal{D}_n \,|\, \boldsymbol{\theta}, \mathbf{b}_i\right) = &\prod_{i=1}^{n}\prod_{j=1}^{n_i} \exp\Big(\phi^{-1}\big[y_i(t_{ij})\,\psi_i(t_{ij}) - b\{\psi_i(t_{ij})\}\big] + c\{y_i(t_{ij}),\, \phi\}\Big) \\ &\times \left(\exp\left[\sum_{r=1}^{R_{h_0}} \gamma_{h_0,r}\, B_{d_{h_0},r}(T_i, \boldsymbol{\lambda}_{h_0}) + \boldsymbol{\gamma}^\top \mathbf{w}_i + \alpha\, \mathcal{F}\{\mu_i(T_i)\}\right]\right)^{\delta_i} \\ &\times \exp\left(-\int_{\tau_i}^{T_i} \exp\left[\sum_{r=1}^{R_{h_0}} \gamma_{h_0,r}\, B_{d_{h_0},r}(s, \boldsymbol{\lambda}_{h_0}) + \boldsymbol{\gamma}^\top \mathbf{w}_i + \alpha\, \mathcal{F}\{\mu_i(s)\}\right] \mathrm{d}s\right). \end{aligned} \tag{4.7}$$

The mean estimates of parameters and random effects are then derived by Markov chain Monte Carlo (MCMC) algorithms, which enable inferences to be made by efficiently drawing a sample from the posterior distribution through the Bayes' rule:

$$\pi\left(\boldsymbol{\theta}, \mathbf{b}_i \,|\, \mathcal{D}_n\right) \,\propto\, p(\mathcal{D}_n \,|\, \boldsymbol{\theta}, \mathbf{b}_i)\, p_b\left(\mathbf{b}_i \,|\, \boldsymbol{\theta}_{\boldsymbol{b}}\right)\, \pi(\boldsymbol{\theta}). \tag{4.8}$$

In the case of the longitudinal approach, we should stress that the expected outcome is here directly dealt with in terms of emergency claims per year. Thus, we can also assess the survival at time t by incorporating the information provided by the exponentially-weighted area under the expected emergency profile until t.

4.5 Results of the Joint Model for Counts and Delayed Entries

The fixed effects of the longitudinal outcome are again set at $\{\beta_0, \beta_1\}$, respectively referring to the intercept term and the observation time. Initially, we test a single random intercept b_{0i}, and a random intercept and slope $\{b_{i0}, b_{i1}\}$, with $\mathrm{Cov}(b_{i0}, b_{i1}) = \rho\, \sigma_{b_0}\, \sigma_{b_1}$. These different longitudinal approaches are fitted under the Poisson and NB distributions using Bayesian software JAGS, version 4.2.0 (Plummer, 2003), by means of code written by the author. We obtain the posterior mean estimates of the parameters and random effects of the GLMM, and the goodness-of-fit is evaluated in each case by the corresponding DIC score. The results suggest that the models with a single random-intercept provide better fittings, and that the NB mixed model is preferable to consider the Poisson mixed model. This result is unsurprising since the NB model accounts for response heterogeneity through parameter κ, whose mean estimate exhibits strong evidence for overdispersion for both one and two random effects, 0.997 (95% CI: 0.948, 1.049) and 1.008 (95% CI: 0.958, 1.060), respectively. Regarding the two NB models, in both cases the posterior mean estimate for σ_{b_0} shows a significant role in accounting for baseline heterogeneity, with 0.963 (95% CI: 0.930, 0.998) in the random-intercept model and 0.969 (95% CI: 0.933, 1.004) for the two random effects model. However, the mean estimate for σ_{b_1} in the NB model with two random effects displays 0.014 (95% CI: 0.004, 0.026), thus indicating a highly residual role of the latent correlation between measurements along time. Consequently, there is no evidence for including the effect of random-slope when overdispersion is accounted for, and the NB with only a random intercept provides the lower DIC score. In what follows, the longitudinal approach in our JM framework is carried out by a mixed model with a random-intercept per policyholder.

In the case of handling emergency claims per year with a Poisson distribution and considering gender as the only baseline covariate w_{gi} in the survival process, the JM is summarized by

$$\begin{cases} y_i(t) \sim \mathrm{PO}\{\mu_i(t)\}, \\ \mathbb{E}\left\{y_i(t) \,|\, b_{i0}\right\} = \mu_i(t) = e_i(t) \exp\left(\beta_0 + b_{i0} + \beta_1\, t\right), \\ h_i\{t \,|\, T_i^* > \tau_i,\, \mathcal{M}_i(t),\, w_{gi}\} = h_0(t) \exp\left[\gamma_g\, w_{gi} + \alpha \displaystyle\int_{\tau_i}^{t} \nu \exp\{-\nu(t-s)\}\mu_i(s)\, \mathrm{d}s\right], \\ b_{i0} \sim \mathcal{N}(0,\, \sigma_{b_0}^2), \end{cases} \tag{4.9}$$

whereas the JM which uses a NB longitudinal submodel is postulated as:

$$\begin{cases} y_i(t) \sim \mathrm{NB}\{\mu_i(t),\, \kappa\}, \\ \mathbb{E}\,\{y_i(t)\,|\,b_{i0}\} = \mu_i(t) = e_i(t)\exp\left(\beta_0 + b_{i0} + \beta_1\, t\right), \\ h_i\{t\,|\,T_i^* > \tau_i,\, \mathcal{M}_i(t),\, w_{gi}\} = h_0(t)\exp\left[\gamma_g\, w_{gi} + \alpha \displaystyle\int_{\tau_i}^{t} \nu \exp\{-\nu(t-s)\}\mu_i(s)\,\mathrm{d}s\right], \\ b_{i0} \sim \mathcal{N}(0,\, \sigma_{b_0}^2). \end{cases} \tag{4.10}$$

The longitudinal part of the JAGS code to fit the JM with counts and left-truncated time-to-event data, was written and properly implemented, this code being available on request. Both JM's are applied to the HI dataset using JAGS software from the R package jagsUI (Kellner, 2016). The posterior mean estimates of the parameters and random intercepts of each JM, $(\boldsymbol{\theta}, b_{i0})$, are obtained, as well as the corresponding DIC $(\boldsymbol{\theta}, b_{i0})$. The priors and hyper-priors of the parameters are the same as considered in Chapter 3 for the standard JM, adding only a flat prior for the dispersion parameter, $\kappa \sim \mathcal{U}(0, 5)$.

	JM weighted cumulative PO counts				JM weighted cumulative NB counts			
Parameter	Mean	SE	$q_{2.5\%}$	$q_{97.5\%}$	Mean	SE	$q_{2.5\%}$	$q_{97.5\%}$
Longitudinal								
β_0	-1.085	0.001	-1.140	-1.030	-0.985	0.001	-1.041	-0.920
β_1	0.032	< 0.001	0.028	0.035	0.030	< 0.001	0.026	0.034
ν	9.989	0.026	8.083	11.894	10.006	0.026	8.085	11.907
κ	–	–	–	–	1.004	0.001	0.953	1.057
σ_{b0}	1.088	< 0.001	1.058	1.117	0.995	< 0.001	0.972	1.020
Survival								
γ_g	−0.074	0.002	−0.227	0.098	−0.071	0.002	−0.247	0.112
Association								
α	0.429	0.001	0.382	0.477	0.526	0.002	0.475	0.578
Goodness-of-fit								
DIC	99111.0				95723.3			

Table 4.1. Posterior summaries for all parameters of the JM when considering the exponentially-weighted cumulative effect of the expected emergency claims per year. Mean, standard error, 95% credible interval and DIC are sampled for each parameter from the corresponding posterior distribution.

For both joint models we obtain a high estimate for the rate parameter of the exponential weighting function, thus indicating that only the most recent past claims have a real influence on the expected survival. The positive sign of the association parameter in both cases shows that relatively high cumulative demand for ambulance services, hospitalizations, and non-routine visits, is instantaneously related to a deterioration in the subject's health status and, consequently, to lower probabilities of survival. Thus, in the case of the JM with Poisson response, a one-unit increase in the weighted area under the expected emergency claims per year leads to a $\exp(\bar{\alpha}_{\mathrm{PO}}) = 1.54$ - fold increase (95% CI: 1.46, 1.61) in the subject's mortality risk, whereas this risk increases by a factor of $\exp(\bar{\alpha}_{\mathrm{NB}}) = 1.69$ (95% CI: 1.61, 1.78) in the NB model. Again, we can remark that the `sex` baseline covariate shows no evidence of different risk depending on policyholder's gender for those aged 65 and more.

When comparing the DIC score obtained in each JM, we observe a substantial reduction in its value in the case of the NB model, of 3387.7 points. This is due to the fact that the NB model is able to capture the heterogeneity present in the response, using its dispersion parameter, and it is consequently able to have a better approach to the behavior observed in the longitudinal responses.

4.6 Discussion of the Joint Model for Counts Results

We have fitted a Bayesian JM to account for discrete outcomes and delayed entries. The variable of interest in the longitudinal analysis was the annual rate of emergency claims, where Poisson and NB distributions were considered for the fit. The JM was fitted using the JAGS software, applied for the first time to such a large health insurance dataset. We focused on the influence of the cumulative and recency-weighted effect of the whole longitudinal profile until a specific time. The results show that relatively high cumulative demand for ambulance services, hospitalizations, and non-routine medical visits, is positively related to a deterioration in the subject's health status and, consequently, to lower survival rates. The most interesting conclusion is that the emergency demand with the greatest impact is that in the most recent past, and it is this that the JM is able to capture. Moreover, the results confirm the adequacy of assuming a NB distribution in the longitudinal process as a first step to handle overdispersed count data. However, further extensions in the longitudinal part can be considered to specifically deal with zero inflation, as different versions of zero-inflated models. We will delve deep into this issue in the next chapter, and additionally, we will investigate the benefits of including a time-varying association parameter.

CHAPTER 5

JOINT MODEL FOR ZERO-INFLATED COUNTS WITH TIME-VARYING EFFECTS

5.1 Principles of Time-Varying Joint Models with Excess Zeros

The use of hierarchical regression models for handling correlated count data over time is very common in both biomedical and health care contexts. Of special interest is to properly model frequency demand when the longitudinal response can be expressed as the number of claims related to the process we want to characterize. The recorded outcome is then usually restricted to a small range of non-negative integer values per subject, therefore expressing an overall highly right-skewed distribution. Aside from within-subject dependence inherent to hierarchies, panel counts are often affected by unmeasured factors between subjects, which in practice translates into the fact that the observed variance is much larger than the observed mean; that is, the occurrence of overdispersion in the observed response commented in Chapter 4. Furthermore, counts with a large number of zeros become highly usual due both to the highly infrequent nature of the data itself and to the existing interplay between private and universal health coverage in some countries. Then, an extra residual variation arises in the response, which overlaps with the latent heterogeneity among individuals. An appropriate inclusion of these two stochastic sources may be achieved by artificially increasing the expected number of zeros. A common approach advocates for following straightforward generalizations of Lambert's methodology (Lambert, 1992; Greene, 1994), where a zero-inflated response is induced by a mixture of a point mass at zero with a hierarchically structured counts, so an additional subset of null measurements is then recorded. The merge of both can be explained by the existence of a hidden subgroup of subjects whose behavior pattern is not observed during the corresponding trajectory within the overall study period. Consequently, a non-real null response is observed regarding an important proportion of the longitudinal outcomes. These commonly termed structural zeros are not randomly generated by the considered count distribution, but originate from independent external causes with probability equal to 1. These are well differentiated from the sampling zeros derived from a typical counting process and thus randomly arising.

In addition, a key benchmark in the classical joint model formulation is that the survival analysis is undertaken under time-constant coefficients within the analyzed period. Consequently, the coefficient effects on the hazard for survival remain fixed across time. However, in some cases, departure from a stationary association parameter becomes a more realistic approach to describe the underlying relationship between the two responses. At this point, we must refer to Song and Wang (2008), where the interest is focused on inferences about the hazard relationship, and in which time-varying parameters are allowed. Specifically, this is achieved through either a corrected score estimator (Wang, 2006) or a local conditional score estimator (Song et al., 2002). More recently, Andrinopoulou et al. (2016) have proposed a so-called varying-coefficient joint model in which, unlike the standard parametrization, a time-varying association parameter between longitudinal and survival outcomes is introduced. Specifically, this is achieved by allowing for flexible shape for the unknown association structure via its expansion into penalized B-splines. In our paper, such an approximation is taken as the starting point to improve the accuracy with which our model describes the behavior of the data that make up our illustrating example.

As commented in Subsection 2.3, one of the defining characteristics of the HI dataset is the presence of a large quantity of zeros in the observed longitudinal response, representing 63.1% of the overall measurements. In addition, it is not clear that the effect of the emergency demand on the mortality risk is constant with age, as was indicated in Subsection 1.2. In this regard, it seems reasonable to assume that there are age segments that are more predisposed to accumulate emergency claims (from 85 years old onwards, many people are affected by multimorbidity), for which reason the health impact of one of these emergency claims may not be as high as if it would have happened in another age range. Thus, it seems reasonable to assume a time-varying association parameter.

5.2 Analysis of Longitudinal Count Data with Excess Zeros

In a hierarchical discrete context, let $\mathbf{y}_i = \{y_i(t_{ij}),\ i = 1, \ldots, n\}$ accounts for observed longitudinal outcomes per subject at a fixed set of time points t_{ij}, $j = 1, \ldots, n_i$. In many studies the recorded observations relies on different periods of time, so the raw expected mean outcome $\hat{\mu}_i(t)$ must be related to the associated exposure time, $e_i(t)$, which plays the role of an offset variable. With this data pattern it is more relevant to model ratios of counts instead of raw counts, then focusing on the expected mean rates of counts per time unit, $\mu_i(t) = \hat{\mu}_i(t)/e_i(t)$.

Now, let us assume that an excess zeros is observed in the longitudinal response, so each measurement comes from the overlapping of a binary variable $v_i(t)$ and the observed count rate $y_{Ci}(t)$, so that $y_i(t) = v_i(t)\, y_{Ci}(t)$. The variable $v_i(t)$ captures a subject's propensity to use private services and follows a Bernoulli distribution, $v_i(t) \sim \mathcal{Be}\{\pi_{Bi}(t)\}$. Hence, structural zeros arise with probability $1 - \pi_{Bi}(t)$, whereas for each-subject counting sequence, $\{\mathbf{y}_i\}$, a Poisson or standard NB distributions are usually assumed.

Figure 5.1 displays the zeros generation process, in which the zeros may come from a classic count distribution or from a binary distribution. Then, we say that the expected response at zero value is *inflated*.

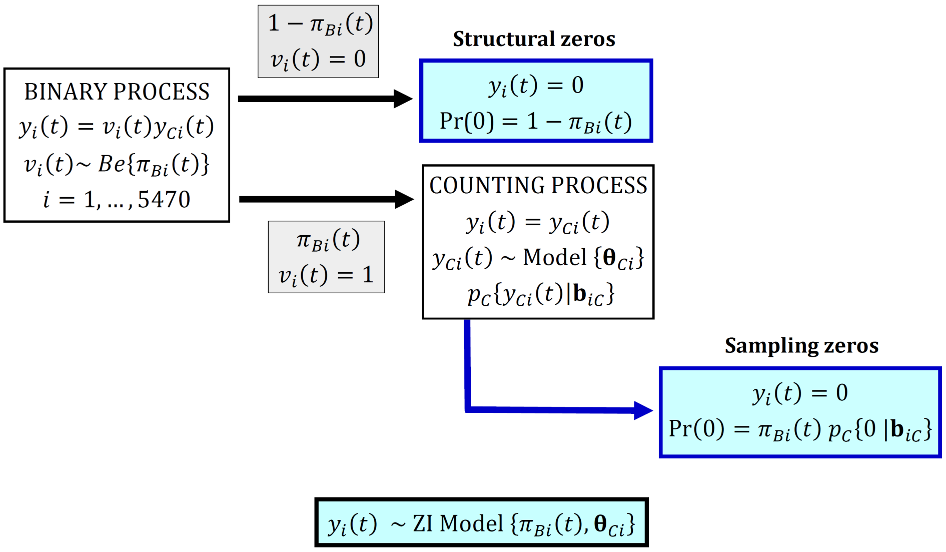

$1-\pi_{Bi}(t)$
$v_i(t)=0$
Structural zeros
BINARY PROCESS
$y_i(t)=v_i(t)y_{Ci}(t)$
$v_i(t)\sim Be\{\pi_{Bi}(t)\}$
$i=1,\ldots,5470$
$y_i(t)=0$
$\Pr(0)=1-\pi_{Bi}(t)$
$\pi_{Bi}(t)$
$v_i(t)=1$
COUNTING PROCESS
$y_i(t)=y_{Ci}(t)$
$y_{Ci}(t)\sim \text{Model}\,\{\boldsymbol{\theta}_{Ci}\}$
$p_C\{y_{Ci}(t)|\mathbf{b}_{iC}\}$
Sampling zeros
$y_i(t)=0$
$\Pr(0)=\pi_{Bi}(t)\,p_C\{0\,|\mathbf{b}_{iC}\}$
$y_i(t)\ \sim \text{ZI Model}\,\{\pi_{Bi}(t),\boldsymbol{\theta}_{Ci}\}$

Figure 5.1. Scheme for zero count generation in a zero-inflated model.

The preponderance of zeros and clustered counts over time are simultaneously modeled, adopting in this thesis a separate random effects extension in each of the two mixture parts. Several authors have focused their research on these hierarchical zero-inflated responses, either by assuming a Poisson counting scheme, e.g. Yau and Lee (2001), Min and Agresti (2005), Liu and Powers (2007), and Neelon et al. (2010), or a NB distribution, e.g. Yau et al. (2003), Fang et al. (2016). Specifically, both the zero-inflated probability and the expected count outcome for the i-th subject at time t are related to a set of both fixed and random explanatory covariates through the corresponding one-to-one monotonic and differentiable link functions, $g_B(\cdot)$ and $g_C(\cdot)$, respectively. The mixture to accommodate many zeros in each subject pattern can be concisely expressed by

$$\begin{cases} \text{Zero-Inflated Model: } p_y\big[y_i(t)\,|\,\{\mathbf{b}_{iB},\mathbf{b}_{iC}\}\big] = \\ \qquad \{1-\pi_{Bi}(t)\}\,\delta_{Bi}(t)+\pi_{Bi}(t)\,f_C\{y_i(t)\,|\,\mathbf{b}_{iC}\}\,\{1-\delta_{Bi}(t)\}, \\ \text{Binary Process: } \pi_{Bi}(t)=g_B^{-1}\big[\mathbf{x}_i^\top(t)\boldsymbol{\beta}_B+\mathbf{z}_i^\top(t)\mathbf{b}_{iB}\big], \\ \text{Counting Process: } \mu_{Ci}(t)=g_C^{-1}\big[\log\{e_i(t)\}+\mathbf{x}_i^\top(t)\boldsymbol{\beta}_C+\mathbf{z}_i^\top(t)\mathbf{b}_{iC}\big], \\ \mathbb{E}\big[y_i(t)\,|\,\{\mathbf{b}_{iB},\mathbf{b}_{iC}\}\big]=\mu_i(t)=\pi_{Bi}(t)\mu_{Ci}(t), \\ \mathbf{b}_{Bi}\sim\mathcal{N}_{q_B+1}(\mathbf{0},\,\mathbf{D}_B),\ \ \mathbf{b}_{iC}\sim\mathcal{N}_{q_C+1}(\mathbf{0},\,\mathbf{D}_C). \end{cases} \tag{5.1}$$

Here, $\pi_{Bi}(t)$ is the subject-specific Bernoulli probability of private care usage at time t, related trough a first monotonic link function $g_B^{-1}(\cdot)$ to a set of of p_B+1 fixed and q_B+1 random effects from binary process, respectively denoted by $\boldsymbol{\beta}_B=(\beta_{B_0},\ldots,\beta_{B_{p_B}})^\top$ and $\mathbf{b}_{iB}=(b_{iB_0},\ldots,b_{iB_{q_B}})^\top$. Further, $\mu_{Ci}(t)$ is the expected value of the underlying count model $f_C\{y_i(t)\,|\,\mathbf{b}_{iC}\}$ (Poisson or NB) to handling the sampling zeros and the non-zero counts, related trough a second monotonic function $g_C^{-1}(\cdot)$ to p_C+1 fixed and q_C+1 random effects, respectively denoted by $\boldsymbol{\beta}_C=(\beta_{C_0},\ldots,\beta_{C_{p_C}})^\top$ and $\mathbf{b}_{iC}=(b_{iC_0},\ldots,b_{iC_{q_C}})^\top$. Additionally,

$\delta_{Bi}(t) = \mathbb{I}\{v_i(t) = 0\}$ acts as an indicator of private-services usage per subject and time-point to undertake the increasing mass at zero value over the default count model. Furthermore, $\mu_i(t)$ is the expected longitudinal value of the zero-inflated model. The distribution of the random effects for both the binary and counting processes is established to follow a multivariate normal distribution, with a zero mean and unspecified variance-covariance matrices, $\mathbf{D}_B$ and $\mathbf{D}_C$, respectively. Notice in the above scheme that both processes are independent, so the associated covariance matrices are not only different, but in general are also obtained by considering different sets of fixed and random covariates (i.e., $p_B \neq p_C$ and $q_B \neq q_C$). Finally, $e_i(t)$ is the exposure time, while $\mathbf{x}_i^\top(t)$ and $\mathbf{z}_i^\top(t)$ respectively denote the row vectors of the fixed and random design matrices, as usual notation in previous chapters.

Building on the above scheme, we can formulate the zero-inflated versions of the Poisson and NB mixed models.

The Zero-Inflated Poisson Model with Random Effects

The Zero-Inflated Poisson (ZIP) mixed model is specially designed to accommodate the overdispersion generated by the presence of excess zeros in our panel count data. In this model, the marginal mean responses for the binary part are usually related to the fixed and random effects using a logit link, $g_B(\cdot) = \text{logit}(\cdot)$, whereas the marginal mean responses for the counting part are related to the fixed and random effects using a logarithmic link, $g_C(\cdot) = \log(\cdot)$. Consequently, we have

$$g_B^{-1}(\cdot) = \text{logistic}(\cdot) = \frac{\exp(\cdot)}{1+\exp(\cdot)},$$
$$g_C^{-1}(\cdot) = \exp(\cdot),$$

and the general form of the ZIP mixed model can be expressed as:

$$\begin{cases} y_i(t) \sim \text{ZIP}\{\pi_{Bi}(t),\, \mu_{Ci}(t)\}, \quad \pi_{Bi}(t),\, \mu_{Ci}(t) > 0, \\ p_y\big[y_i(t)\,|\,\{\mathbf{b}_{iB}, \mathbf{b}_{iC}\}\big] = \\ \qquad \{1-\pi_{Bi}(t)\}\,\delta_{Bi}(t) + \pi_{Bi}(t)\dfrac{\exp\{-\mu_{Ci}(t)\}\mu_{Ci}(t)^{y_i(t)}}{y_i(t)!}\,\{1-\delta_{Bi}(t)\}, \\ \text{Binary Response: } \pi_{Bi}(t) = \dfrac{\exp\{\mathbf{x}_i^\top(t)\boldsymbol{\beta}_B + \mathbf{z}_i^\top(t)\mathbf{b}_{iB}\}}{1+\exp\{\mathbf{x}_i^\top(t)\boldsymbol{\beta}_B + \mathbf{z}_i^\top(t)\mathbf{b}_{iB}\}}, \\ \text{Counting Response: } \mu_{Ci}(t) = e_i(t)\exp\left\{\mathbf{x}_i^\top(t)\boldsymbol{\beta}_C + \mathbf{z}_i^\top(t)\mathbf{b}_{iC}\right\}, \\ \mathbb{E}\big[y_i(t)\,|\,\{\mathbf{b}_{iB}, \mathbf{b}_{iC}\}\big] = \mu_i(t) = \pi_{Bi}(t)\,\mu_{Ci}(t), \\ \mathbf{b}_{iB} \sim \mathcal{N}_{q_B+1}(\mathbf{0},\, \mathbf{D}_B), \quad \mathbf{b}_{iC} \sim \mathcal{N}_{q_C+1}(\mathbf{0},\, \mathbf{D}_C). \end{cases} \tag{5.2}$$

However, in some cases the zero-inflated component from the binary part is not able by itself to allow for data expressing the amount of heterogeneity, this being necessary in order to consider a proper count distribution to explicitly account for this variability source.

The Zero-Inflated Negative Binomial Model with Random Effects

A widely used alternative to consider heterogeneity effects consists of assuming a hierarchical NB distribution for the count part. Despite the fact that the standard NB can be formulated using different parametrization schemes, in most of the cases it establishes a quadratic mean-variance relationship to handle additional sources of variability, and for this reason that distribution is also referred to as NB2 in literature (Cameron and Trivedi, 2005). Combining a hierarchical NB assumption with a random binary process for generating structural zeros, a zero-inflated NB mixed model (ZINB) is derived. This model uses the same link functions as those used in the ZIP mixed model for binary and counting processes:

$$\begin{cases} y_i(t) \sim \mathrm{ZINB}\{\pi_{Bi}(t),\, \mu_{Ci}(t),\, \kappa\}, \quad \pi_{Bi}(t),\, \mu_{Ci}(t),\, \kappa > 0, \\ p_y\big[y_i(t) \,|\, \{\mathbf{b}_{iB}, \mathbf{b}_{iC}\}\big] = \\ \qquad \{1 - \pi_{Bi}(t)\}\, \delta_{Bi}(t) + \pi_{Bi}(t) \dfrac{\Gamma\{\kappa + y_i(t)\}}{\Gamma(\kappa)\, y_i(t)!} \, \dfrac{\mu_{Ci}(t)^{y_i(t)}\, \kappa^{\kappa}}{\{\mu_{Ci}(t) + \kappa\}^{\kappa + y_i(t)}} \, \{1 - \delta_{Bi}(t)\}, \\ \text{Binary Response: } \pi_{Bi}(t) = \dfrac{\exp\{\mathbf{x}_i^\top(t)\boldsymbol{\beta}_B + \mathbf{z}_i^\top(t)\mathbf{b}_{iB}\}}{1 + \exp\{\mathbf{x}_i^\top(t)\boldsymbol{\beta}_B + \mathbf{z}_i^\top(t)\mathbf{b}_{iB}\}}, \\ \text{Counting Response: } \mu_{Ci}(t) = e_i(t) \exp\{\mathbf{x}_i^\top(t)\boldsymbol{\beta}_C + \mathbf{z}_i^\top(t)\mathbf{b}_{iC}\}, \\ \mathbb{E}\big[y_i(t) \,|\, \{\mathbf{b}_{iB}, \mathbf{b}_{iC}\}\big] = \mu_i(t) = \pi_{Bi}(t)\, \mu_{Ci}(t), \\ \mathbf{b}_{iB} \sim \mathcal{N}_{q_B+1}(\mathbf{0},\, \mathbf{D}_B), \quad \mathbf{b}_{iC} \sim \mathcal{N}_{q_C+1}(\mathbf{0},\, \mathbf{D}_C). \end{cases} \tag{5.3}$$

5.3 A Time-Varying Joint Model for Overdispersed Counts with Excess Zeros

Let T_i^* be the true survival time for the i-th individual, defined as a non-negative random variable that collects the time lag from the point an individual reaches t years until subject's death. These subjects can not have been observed since reaching the demanded condition, leading to a lag between the start of follow up and the time origin fixed by the condition of the HI dataset, which leads to random left-truncated survival times, τ_i. As a result, we can only know the observed time to the event of interest for the i-th recruited individual, $T_i = \min\{T_i^*,\, C_i\}$, and we can introduce a dichotomous event indicator $\delta_i = \mathbb{I}(T_i{}^* \leq C_i)$. The probabilistic distribution of the time-to-event is defined by considering the proportion of living subjects beyond time point t and conditional on being older than left truncation time, $\mathcal{S}_i\{t \,|\, T_i^* > \tau_i, \mathcal{M}_i(t)\} = \Pr\{T_i^* > t \,|\, T_i^* > \tau_i, \mathcal{M}_i(t)\}$.

Building on the longitudinal modeling considered in section 5.2, repeated count rates and time-to-event approaches can be joined in one statistical model by assuming independence between both processes given the shared random effects (i.e. conditional independence). From this perspective, we postulate a JM with a time-varying association, (JMTV), in which the association parameter $\alpha(t)$ is assumed time-dependent, so the relationship between the expected longitudinal value, $\mu_{Ci}(t)$, and the hazard outcomes, $h_i\{t \mid T_i^* > \tau_i, \mathcal{M}_i(t), \mathbf{w}_i\}$, is subject to temporal variation. In particular, the time-dependent association parameter is expanded into Bayesian P-splines of degree $d_\alpha = 2$ in the same way as the baseline hazard function:

$$\alpha(t) = \sum_{\tilde{r}=1}^{R_\alpha} \alpha_{\tilde{r}}\, B_{d_\alpha,\tilde{r}}(t, \boldsymbol{\lambda}_\alpha), \tag{5.4}$$

where $\{B_{d_\alpha,\tilde{r}}(t, \boldsymbol{\lambda}_\alpha),\ \tilde{r} = 1, \ldots, R_\alpha\}$ is the same B-spline basis functions that the used for $\log\{h_0(t)\}$, $\boldsymbol{\lambda}_\alpha = \{\lambda_1, \ldots, \lambda_{Q_\alpha}\}$ are the equally-spaced knots on $[t_{\min}, t_{\max}]$ (we take the same 8 knots that we used for $\log\{h_0(t)\}$), and the set of parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{R_\alpha})$ denotes the R_α-dimensional vector ($R_\alpha = 9$) of B-spline coefficients of the association parameter.

Furthermore, the JMTV consists of two submodels: a zero-inflated mixed effects model for the counts, and a semi-parametric relative risk survival model to account for time-dependent covariates. In this thesis we consider two options for the longitudinal submodel: ZIP and ZINB. For the i-th subject, the general expression of our JMTV at each time t is summarized by the set of equations

$$\begin{cases} \pi_{Bi}(t) = \dfrac{\exp\{\mathbf{x}_i^\top(t)\boldsymbol{\beta}_B + \mathbf{z}_i^\top(t)\mathbf{b}_{iB}\}}{1 + \exp\{\mathbf{x}_i^\top(t)\boldsymbol{\beta}_B + \mathbf{z}_i^\top(t)\mathbf{b}_{iB}\}}, \\ \mu_{Ci}(t) = e_i(t)\exp\{\mathbf{x}_i^\top(t)\boldsymbol{\beta}_C + \mathbf{z}_i^\top(t)\mathbf{b}_{iC}\}, \\ \mu_i(t) = \pi_{Bi}(t)\,\mu_{Ci}(t), \\ h_i\{t|T_i^* > \tau_i,\, \mathcal{M}_i(t),\, \mathbf{w}_i\} = h_0(t)\exp\Big[\boldsymbol{\gamma}^\top\mathbf{w}_i + \alpha(t)\,\mathcal{F}\{\mu_i(t)\}\Big] = \\ \qquad h_0(t)\exp\Big[\boldsymbol{\gamma}^\top\mathbf{w}_i + \alpha(t)\displaystyle\int_{\tau_i}^t \nu\exp\{-\nu(t-s)\}\mu_i(s)\,\mathrm{d}s\Big], \\ \mathbf{b}_{iB} \sim \mathcal{N}_{q_B+1}(\mathbf{0},\,\mathbf{D}_B), \quad \mathbf{b}_{iC} \sim \mathcal{N}_{q_C+1}(\mathbf{0},\,\mathbf{D}_C), \end{cases} \tag{5.5}$$

where $\pi_{Bi}(t)$ is the expected value of the Bernoulli model (i.e. the subject-specific probability of using private services), $\mu_{Ci}(t)$ is the expected value of the counting model (i.e. Poisson or NB), and $\mu_i(t)$ is the expected value of the zero-inflated mixed model (ZIP or ZINB). The parameters of the JMTV are estimated under the Bayesian framework, with using non-informative priors whenever possible (see Chapter 3 and Chapter 4). The only novelty here is the presence of the fixed and random effects for the binary part, $\boldsymbol{\beta}_B$ and $\mathbf{b}_{iB}$, respectively, for which univariate and diffuse normal priors are assumed.

5.4 Results for the Time-Varying Joint Model with Excess Zeros

In the same manner we have done in the previous chapters, we have fitted the corresponding JM, but in this case, the parameter association results will be presented in graphical way, by comparing the JM with a constant association parameter and the JMTV (Figure 5.2). For the binary part, we assume a constant subject-specific rate of private care usage, $\pi_{Bi}(t) \equiv \pi_{Bi}$, only considering fixed and random intercept effects. The reason for taking this decision relies on the fact that the longitudinal data analysis performed in Chapter 2 did not provide any evidence to include an slope effect when modeling zero values. For the counting part, we assume fixed effects for intercept and slope terms, and also single random intercept, thus maintaining the same longitudinal structure used for counting distributions in Chapter 4.

In the case of assuming a ZIP longitudinal submodel, the JMTV for the HI dataset is

$$\begin{cases} y_i(t) \sim \text{ZIP}\left\{\pi_{Bi},\, \mu_{Ci}(t)\right\}, \quad \pi_{Bi},\, \mu_{Ci}(t) > 0, \\ \pi_{Bi} = \dfrac{\exp\{\beta_{B_0} + b_{iB_0}\}}{1 + \exp\{\beta_{B_0} + b_{iB_0}\}}, \\ \mu_{Ci} = e_i(t) \exp\left(\beta_{C_0} + b_{iC_0} + \beta_{C_1} t\right), \\ \mu_i(t) = \pi_{Bi}\, \mu_{Ci}(t), \\ h_i\{t \,|\, T_i^* > \tau_i,\, \mathcal{M}_i(t),\, w_{gi}\} = h_0(t) \exp\Big[\gamma_g\, w_{gi} + \alpha(t)\, \mathcal{F}\left\{\mu_i(t)\right\}\Big] = \\ \qquad\qquad h_0(t) \exp\left[\gamma_g\, w_{gi} + \alpha(t) \displaystyle\int_{\tau_i}^{t} \nu \exp\{-\nu(t-s)\} \mu_i(s)\, \mathrm{d}s\right], \\ b_{iB_0} \sim \mathcal{N}(0,\, \sigma^2_{B_0}), \quad b_{iC_0} \sim \mathcal{N}(0,\, \sigma^2_{C_0}). \end{cases} \tag{5.6}$$

and in the case of assuming a ZINB longitudinal submodel, the JMTV expression is:

$$\begin{cases} y_i(t) \sim \text{ZINB}\{\pi_{Bi},\, \mu_{Ci}(t),\, \kappa\}, \quad \pi_{Bi},\, \mu_{Ci}(t),\, \kappa > 0, \\ \pi_{Bi} = \dfrac{\exp\{\beta_{B_0} + b_{iB_0}\}}{1 + \exp\{\beta_{B_0} + b_{iB_0}\}}, \\ \mu_{Ci}(t) = e_i(t) \exp\left(\beta_{C_0} + b_{iC_0} + \beta_{C_1} t\right), \\ \mu_i(t) = \pi_{Bi}\, \mu_{Ci}(t), \\ h_i\{t \,|\, T_i^* > \tau_i,\, \mathcal{M}_i(t),\, w_{gi}\} = h_0(t) \exp\Big[\gamma_g\, w_{gi} + \alpha(t)\, \mathcal{F}\{\mu_i(t)\}\Big] = \\ \qquad\qquad h_0(t) \exp\left[\gamma_g\, w_{gi} + \alpha(t) \displaystyle\int_{\tau_i}^{t} \nu \exp\{-\nu(t-s)\} \mu_i(s)\, \mathrm{d}s\right], \\ b_{iB_0} \sim \mathcal{N}(0,\, \sigma^2_{B_0}), \quad b_{iC_0} \sim \mathcal{N}(0,\, \sigma^2_{C_0}). \end{cases} \tag{5.7}$$

	JMTV with ZIP counts				JMTV with ZINB counts			
Parameter	Mean	SE	$q_{2.5\%}$	$q_{97.5\%}$	Mean	SE	$q_{2.5\%}$	$q_{97.5\%}$
Longitudinal								
β_{B0}	0.406	0.001	0.323	0.481	1.292	0.002	1.112	1.438
β_{C0}	−0.293	0.003	−0.357	−0.225	−0.462	0.001	−0.525	−0.395
β_{C1}	0.030	< 0.001	0.027	0.033	0.031	< 0.001	0.027	0.035
ν	8.237	0.047	7.068	9.421	11.147	0.032	10.287	11.882
κ	–	–	–	–	2.283	0.003	2.046	2.546
σ_{B0}	1.346	0.001	1.255	1.438	1.977	0.002	1.798	2.236
σ_{C0}	0.782	0.001	0.750	0.815	0.775	0.001	0.742	0.810
Survival								
γ_g	−0.083	0.002	−0.267	0.107	−0.081	0.005	−0.249	0.102
Goodness-of-fit								
DIC	115739.4				115253.5			

Table 5.1. Posterior summaries for all parameters of the JMTV when considering the exponentially-weighted cumulative effect of the expected emergency claims per year. Mean, standard error, 95% credible interval and DIC are sampled for each parameter from the corresponding posterior distribution.

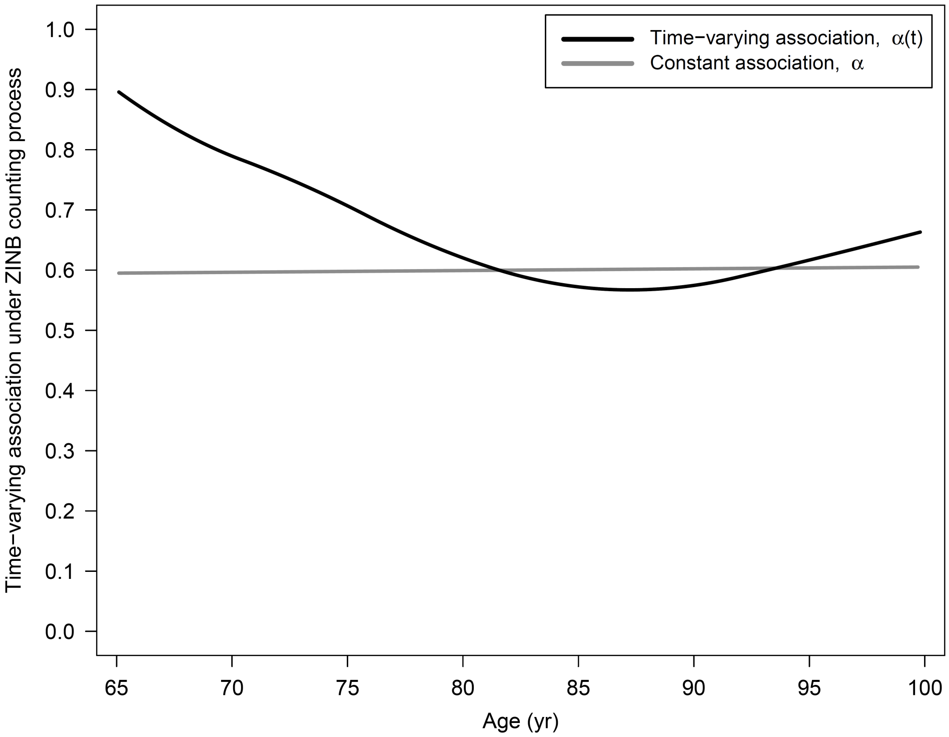


Figure 5.2. Comparison between the constant association parameter of JM with weighted cumulative effects and ZINB longitudinal response, and the time-dependent association parameter of the JMTV with weighted cumulative effects and ZINB longitudinal response.

Firstly, in accordance with the results obtained in previous chapters, we observe that the

JMTV which considers a ZINB longitudinal submodel provides a better fitting than the JMTV which assumes a ZIP longitudinal response. This result is logical, since the ZINB mixed model accounts in a natural way for both inherent overdispersion and excess zeros in the longitudinal observed response. In addition, the gender baseline covariate does not lead to a significantly different mortality risk between men and women, as occurred with the joint models fitted in previous chapters.

However, the main point in this chapter concerns the true behavior of the mortality risk pattern, whose real trend is not constant from the age of 65 onwards (as traditional joint modeling techniques assume). Although the time-dependent association parameter $\alpha(t)$ continues to indicate that the relationship between the expected cumulative emergency demand and death hazard is positive from the age of 65, it also points out that such a risk is much more important for people under 75 years and not so much for ages around 85 years old, precisely the ages range at which the emergency requirement achieves its maximum annual rates (Subsection 3.2.1). In particular, the results show that the mortality risk associated with one-unit increase in the weighted area under the expected longitudinal profile decreases from the age 65 to the age around 85 years old, and then it increases again. Hence, in the last potential ages of life (from the age of 90 onwards) the policyholder's mortality risk is mainly controlled by factors associated with subject's age, being less dependent on emergency services usage. The evolution over time of the association parameter over time is displayed in Table 5.2, and we also provide the value obtained by assuming a constant link between longitudinal and time-to-event outcomes.

Age (years)	Mortality hazard ratio at time t, $\exp\{\alpha(t)\}$, for a one-unit increase in $\mathcal{F}\{\mu_i(t)\}$
65	2.44
70	2.19
75	2.04
80	1.87
85	1.77
90	1.79
95	1.85
100	1.93
Constant	1.84

Table 5.2. Estimation of the mortality hazard ratio from 65 to 100 years for one-unit increase in the weighted area under the expected longitudinal profile.

5.5 Discussion of the Joint Model with Time-Varying Association Parameter

This section has introduced a JM approach which considers a time-varying association parameter between the longitudinal and time-to-event outcomes. In addition, counts affected by an excess of zeros are also accounted for in the longitudinal submodel, as well as left truncation and right censoring in time-to-event submodel. After applying the JMTV to the HI dataset, the results highlight the different effect on subject's death hazard that annual rate of emergency claims has, depending of the subject's age at which this rate claims are recorded (as well as subject-specific features). Although the expected emergency claims per year and time-to-event outcomes are positively related between the ages 65 and 100, we can infer that those emergency services collected under the age of 75 are highly related to death hazard than those observed around 85 years. In fact, as a subject ages from 65 years, it is more common requiring emergency services due by multimorbidity (Section 1.1). Consequently, these type of services will not be so representative of a critical health status as in ages under 75. To conclude, we observe that, from approximately 90 years onwards, the mortality process is mainly an aging issue.

CHAPTER 6

DYNAMIC PREDICTIONS

6.1 Individualized Survival Predictions

One of the key features of the Bayesian joint framework is that personalized and dynamically-updated survival predictions can be obtained by considering each subject-specific longitudinal profile (Proust-Lima and Taylor, 2009; Rizopoulos, 2011). Let us consider a new subject, denoted by $k = i + 1$, not included in the original dataset but sampled from the target population. If emergency claims are recorded until time t, we implicitly know that this new subject is at least alive until t, thus providing a historical set of measurements, $\mathcal{Y}_k(s) = \{y_k(s_{kj}),\ \tau_k \leq s_{kj} \leq t,\ j = 1, \ldots, n_k\}$, as well as a specific value for gender factor w_{gk}. In the case of this information, we can obtain the conditional subject-specific predictions at any future time $u > t$, given survival up to t. This prognosis task can be carried out quite straightforwardly by adopting a Bayesian strategy. Let $\mathbf{\Omega} = (\boldsymbol{\theta}, \mathbf{b}_k)$ denote the mean estimates of full-parameter vector and random effects of the new subject, and let us assume that the inclusion of a new subject does not entail the updating of the previous estimates. Then, the posterior predictive distribution of survival can be written as

$$
\begin{aligned}
\tilde{\pi}_k(u \,|\, s) &= p_t(T_k^* \geq u \,|\, T_k^* \geq s, \mathcal{Y}_k(s), w_{gk}, \mathcal{D}_n) \\
&= \iint_{\mathbf{\Omega}} p_t(T_k^* \geq u \,|\, T_k^* > s, \mathcal{Y}_k(s), w_{gk}, \mathbf{b}_k, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \,|\, \mathcal{D}_n)\, \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\mathbf{b}_k \qquad (6.1) \\
&= \iint_{\mathbf{\Omega}} p_t(T_k^* \geq u \,|\, T_k^* > s, w_{gk}, \mathbf{b}_k, \boldsymbol{\theta})\, p_b(\mathbf{b}_k \,|\, T_k^* > s, \mathcal{Y}_k(s), w_{gk}, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \mathcal{D}_n)\, \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\mathbf{b}_k \\
&= \iint_{\mathbf{\Omega}} \frac{\Pr(T_k^* \geq u \,|\, \mathcal{M}_k(s), w_{gk}, \mathbf{b}_k)}{\Pr\,(T_k^* > s \,|\, \mathcal{M}_k(s), w_{gk}, \mathbf{b}_k)}\, p_b(\mathbf{b}_k \,|\, T_k^* > s, \mathcal{Y}_k(s), w_{gk}, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \mathcal{D}_n)\, \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\mathbf{b}_k,
\end{aligned}
$$

so that a MCMC estimate of $\tilde{\pi}_k(u \mid s)$ has been carried out by combining the previous assumptions.

Using an example, let us consider a male and female policyholder, both aged 70 upon entering the study ($\tau_k = 5$), and both not included in the original dataset. A common history of emergency claims $\mathcal{Y}_s$ is simulated for the next decade, with measurements collected at ages $\{70, \ldots, 80\}$, i.e. $s \in \{5, \ldots, 15\}$. Moreover, a NB counting sequence within the basic JM approach is assumed. We first focus on estimating the survival probability for both subjects at age 90, conditioned on their being alive at s, $\tilde{\pi}_k(u = 25 \,|\, s)$. The results are obtained for the

JM with weighted cumulative parametrization which considers a NB longitudinal submodel, presented in Chapter 4. In particular, we have modified the `survfitJM` function of the `JMbayes` package, and the results show how the Monte Carlo estimates update dynamically as new longitudinal information is considered (Table 6.1). Dynamic updating of this kind emphasizes the need for a well-characterized follow-up for each policyholder when we aim towards personalized decisions and an accurate prediction of the insurance capital needed to cover the corresponding health insurance plan. We conclude that there is an increasing probability of being alive at age 90 when no claims are reported, whereas this probability decreases sharply when a large number of emergency claims are reported. The survival estimates for the female are slightly higher than those for the male policyholder, since the gender coefficient regression indicates that *ceteris paribus* males have a slightly higher mortality hazard than females. Hence, by the age of 80, the survival estimate at the age of 90 for male policyholder is $\tilde{\pi}_{k,m}(u = 25 \,|\, s = 15) = 0.708$, whereas a woman under the same demand process presents an estimate of $\tilde{\pi}_{k,w}(u = 25 \,|\, s = 15) = 0.717$.

Age (yr)	Observed claims, $y_k(s)$	Man's survival at 90 yr			Woman's survival at 90 yr		
		Mean	$q_{2.5\%}$	$q_{97.5\%}$	Mean	$q_{2.5\%}$	$q_{97.5\%}$
70	0	0.775	0.484	0.853	0.781	0.501	0.854
71	0	0.798	0.630	0.857	0.803	0.622	0.858
72	1	0.786	0.597	0.852	0.792	0.590	0.856
73	0	0.801	0.659	0.857	0.806	0.635	0.861
74	2	0.770	0.587	0.848	0.776	0.577	0.854
75	0	0.791	0.645	0.851	0.797	0.633	0.855
76	0	0.803	0.685	0.858	0.808	0.674	0.861
77	8	0.674	0.392	0.816	0.682	0.383	0.823
78	1	0.685	0.430	0.822	0.697	0.421	0.827
79	2	0.683	0.432	0.817	0.692	0.421	0.825
80	0	0.708	0.502	0.824	0.717	0.486	0.832

Table 6.1. Dynamic survival probabilities from the JM considering the NB response with the cumulative and recency-weighted parametrization for expected claims. Mean and 95% CI of being alive at age 90 for a man and a woman with identical claims information collected between the ages of 70 and 80.

If we know for certain that both subjects from the previous example remain alive at age 80, then we can assess their future survival from the information contained in our dataset of policyholders over the age of 80. In this regard, Table 6.2 provides the survival estimates from the age of 80 to the age of 90. Recall that the last row in this table logically provides the same results as those in Table 6.1, since both survival estimates at the age of 90 are performed under the same assumptions.

Age (years)	Man's survival at 90 yr			Woman's survival at 90 yr		
	Mean	$q_{2.5\%}$	$q_{97.5\%}$	Mean	$q_{2.5\%}$	$q_{97.5\%}$
80	1.000	1.000	1.000	1.000	1.000	1.000
82	0.973	0.951	0.983	0.974	0.952	0.984
84	0.934	0.890	0.962	0.938	0.892	0.963
86	0.885	0.807	0.929	0.892	0.797	0.932
88	0.810	0.675	0.885	0.815	0.652	0.887
90	0.708	0.502	0.824	0.717	0.486	0.832

Table 6.2. Prognosis of survival from the JM considering the NB response with the cumulative and recency-weighed parametrization of expected claims. Mean and 95% CI for a man and a woman who remain alive at age 80.

6.2 Assessing the Effect of Emergency Demand and Longevity on Insurance Rates

We provide a brief overview of contributions in the literature that are aimed at studying the statistical analysis of insurance problems. Joint modeling in insurance has been addressed by many authors. Sarabia and Guillén (2008) analyzed models in the general context of risk management. Count data models for insurance are specially relevant in the context of automobile insurance (Boucher et al., 2007). An analysis of a long-term care insurance portfolio was studied by Guillén and Pinquet (2008), and the costs were considered by Guillén et al. (2011). Furthermore, time-varying effects were considered in Guillén et al. (2012). In this section, the relationship between the demand for medical service, the survival, and the price of insurance is addressed.

The results in Table 6.3 present the predicted survival probabilities at age 75 for an example of an insured man and woman, aged 65 at study entry. They both demand a number of emergency claims per year equal to 3, 1, 5 and 15 in the first four years.

Age (yr)	Observed claims, $y_k(s)$	Mean survival at 75 yr for man	Mean survival at 75 yr for woman
65	3	0.964	0.967
66	1	0.968	0.969
67	5	0.958	0.960
68	15	0.942	0.948

Table 6.3. Dynamic survival probabilities from the JM considering the NB response with the cumulative and recency-weighted parametrization for expected claims. Mean probability of being alive at age 75 for a man and a woman with identical claims information collected between the ages of 65 and 68.

If we carefully look at the example of the man, we observe that when at the age of 65, he requests for three emergency claims and he has an estimated probability of survival to 75 years equal to 96.4%; however, when he turns 66 and he has only one claim, then his probability of survival to age 75 increases to 96.8%. On the contrary, if a man reports 15 claims at age 68, his probability to survive to age 75 decreases to 94.2%, as expected by the association between bad health condition and risk of death.

When looking at the example presented in Table 6.3, we observe that when the annual rate of emergency claims increases from 5 to 15, at the age of 67 and 68, then the decrease in the survival probability is $0.958 - 0.942 = 0.016$ for men, and $0.960 - 0.948 = 0.012$ for women. A simple extrapolation implies that an increase in the rate of emergency claims implies a reduction of the probability of survival of roughly 0.167% per claim in men and about 0.125% in women. This result follows from dividing the reduction in survival probability by the initial survival estimate $(0.016/0.958) \cdot 100\% = 1.67\%$, and dividing this percentage by the difference in emergency claims per year, $15 - 5 = 10$. The analysis of the interplay between emergency claims per year and survival reveals that there is an association between them, so that an increase in emergency claims implies a decrease in survival. However, the decrease in survival is low.

Let us consider that the cost of a claim equals C_T. In order to analyze the effect on the price of the premium, we next describe the procedure to follow. Assume that for a given insured the expected number of claims is equal to E and the survival probability is p, then the pure premium price (we do not consider general additional expenses here) would be equal to $C_T\, E\, p$. When the expected claims increase by one then we must consider $E + 1$, but the survival probability of a later age decreases by r, so we must substitute p by $(1 - r)\, p$, and the pure premium is $C_T\,(E + 1)\, p\,(1 - r)$. We obtain that the new premium is

$$C_T\,(E+1)\,p\,(1-r) = C_T\,E\,p - r\,C_T\,E\,p + C_T\,p\,(1-r).$$

So, the premium would be equivalent to the initial one if

$$C_T\,p\,(1-r) = r\,C_T\,E\,p.$$

Then, the survival reduction which exactly compensates the cost of additional claims is:

$$E = \frac{1-r}{r}.$$

This equality would imply perfect diversification of the two processes. For values of r ranging from 0.05% to 20%, we have calculated the corresponding expected number of claims that would imply equilibrium in the premium. The results are presented in Table 6.4.

R(%)	Expected claims
5.0	19
7.5	12
10.0	9
12.5	7
15.0	6
17.5	5
20.0	4

Table 6.4. Reduction in survival probability and the corresponding expected number of claims that implies no change in the premium.

CHAPTER 7
DISCUSSION AND FUTURE RESEARCH

7.1 Contributions to Health Insurance

In this thesis we study health insurance for elderly people together with their survival rates. We defined elderly people as those with a chronological age equal to 65 years or more. In aging societies, the longevity phenomenon implies that elderly people live longer than they did in the previous decades, but with poorer health conditions. In the context of private health care, this shifting demographic trend has given rise to an increasing demand for services, because it is widely known that elderly individuals require medical assistance (they have a larger number of claims) more often than younger ones, thus creating a problem for pricing products correctly. If elderly policyholders were charged proportionally to their own risk, then they would not be able to afford health insurance costs. So, insurers introduce corrections in the premium setting process. In practice, younger members of private health insurance schemes have to subsidize the older ones.

Traditionally, insurance companies have not analyzed the underlying compensation between the demand for health services and the survival rate. This thesis is a unique contribution in this field. It also contributes to the development of statistical methods designed to jointly model non-negative integer outcomes and survival. We want to emphasize that, apart from the methodological results that have been described in detail in the previous chapters, this thesis has a practical contribution to the development and sustainability of health insurance products for elderly people. It has been shown that the demand for health services together with survival rates provides a way for the insurance company to compensate costs. In actuarial terms this is usually called risk compensation. Elderly people who have a poor health status make more claims than the healthy ones, and therefore they cost more to the company than expected. However, due to a deterioration of health conditions, these insured people tend to have bleaker survival prospects than those who are healthy. So, this thesis argues that the higher costs of medical care can be compensated by less costs in the products that are associated with life expectancy. For instance, if an individual has a medical insurance policy and, at the same time, receives a life annuity from the insurance company, the question is whether there is an association between the longevity risk and the medical risk so that both mechanisms can compensate each other. This thesis demonstrates that in the particular dataset that has been analyzed, the association between emergency claims and survival exists, being positive and significantly different from zero. In addition, it is shown that there is no difference between men and women in this respect. Finally, it is also proven that the association intensity varies with age, so that it is much more important for people under 75 years and not so much for ages around 90 years. This result *per se* is a genuine output and can be directly implemented in the real world. However, the thesis is limited in the immediate application because the actuarial methods have not been fully addressed, as they were not the main purpose of this work.

The insurance sector in general stores large portfolios of individual policyholders monitored over long periods of time, thus constituting a benchmark example of the potential applications of joint modeling techniques in a context of large datasets. This feature is also a characteristic of this thesis, where methods have been developed under the premise that the size of the motivating dataset is very large. An increasing number of statistical studies report on the individualized monitoring of time-dependent covariates prior to the occurrence of a particular event. We have shown that the joint analysis of the individualized and expected longitudinal evolution over the lifetime is the proper approach for detecting the strength of association between these two responses.

7.2 Contributions to Joint Modeling

As mentioned before, the thesis makes some contributions to statistical methodology in the field of joint modeling of longitudinal counts and survival outcomes. Looking back to the initial objectives, we summarize the main key findings in the methodology.

We have proposed an adequate model to accommodate correlated counts observed in the longitudinal outcome, taking into account the potential overdispersion at subject-specific level, that is, when the within-subject variability is larger than the mean. The two main causes of overdispersion derive from a inherent heterogeneity among measurements and an abundance of zeros. Additionally, time-to-event data should account not only for the usual right censoring, but also for the left truncation caused by the late entry into the study of a moderate percentage of subjects.

We have shown how to analyze the adequate functional form to relate the subject's count history within the study window to death risk. Standard joint models assume a constant relationship between the current expected value and the survival rate, but in our case it does not seem reasonable to summarize the health status by only considering the longitudinal information from a single time point. Instead, we have considered the impact of past health status on the current death hazard. Moreover, all past medical information does not have the same importance; the closer measurements are to the current time, the more weighted their consideration should be compared to those that are more distant.

As a main idea in this thesis, we have incorporated a time-varying association parameter between longitudinal and time-to-event outcomes, hence allowing for a more flexible relationship between emergency demand and death hazard. This point becomes essential in the insurance field, since the result of this connection is the one which may prove that expected costs from subjects with a higher emergency demand are compensated with lower survival rates.

Insights into the underlying trends at this individualized level enable professionals to provide more accurate service demand predictions and to adjust the risk of mortality to capture better the baseline factors of age and gender. To date, joint modeling techniques have usually been used with quite small datasets (in the order of a few hundred individuals) and it is not easy to find applications to larger sets or discussions of the computational obstacles that need to be overcome. Nevertheless, it has been reported the challenging task of implementing a single joint model with a large data sample of policyholders aged 65 and over, and it has demonstrated a statistically significant dependence between a subject's past medical care usage (their use of ambulance and emergency services together with admissions to the hospital) and their current hazard of death. While we have no information about the number of days the subjects spent in hospital or in the emergency room, nor about the condition that required their seeking health care, we are able to provide a personalized survival prediction.

The functional form proposed to include complete historical information in the hazard model considers the cumulative effect of the health-status weighted biomarker outcome over the preceding years. In addition, the impact of health-status exposure on survival response is weighted by an exponential function, so that the most recent claims with respect to the present are assigned greater weights, while the impact of distant outcomes falls sharply with increasing distance from the present. Hence, we have shown that health risks are not greatly affected by long-term cumulative critical claims from the past, but rather by those recorded more recently.

7.3 Further Remarks and Future Research

After looking at the dynamic predictions, we conclude that the reduction in survival should be much greater than that observed in our data to compensate for the increase in the demand for emergency care. This conclusion indicates that risk diversification may not be possible if the insured person only has one medical insurance product, but in case that the insured person has several products linked to longevity risk such as life annuities, then the compensation of risks could be feasible.

Here, it should be noted that we do not consider dependence between the number of events and severity in terms of cost. Further, a limitation of the thesis is that we only consider severe medical care events (namely, ambulance services, hospitalizations, and non-routine medical visits), whereas other medical treatments are not analyzed and, as such, do not form a part of the insurance policy under evaluation.

The conclusion for practitioners is, therefore, that health insurance for the older group remains a matter of pooling the risk with younger customers or of increasing the price of the policy with age. However, we argue that joint modeling of frequency and survival is a method that must be considered when developing dynamic pricing techniques, which are aimed at policy durations shorter than one year, especially for elderly of not-too-advanced ages. In that case, knowledge of the claim history should be combined with an update to the survival prediction.

There are some issues that still remain on the agenda for future research. First, the models proposed here, when implemented in large databases, require an intensive computational effort, so there is a need for optimization of the numerical and estimation procedures in order to have quicker solutions. Second, the inclusion of multiple longitudinal outcomes has not been addressed in the thesis. We think that models could have been addressing the three counts separately (ambulance, hospitalizations, and non-routine medical visits). Third, joint latent class models have not been been analyzed. This could be an area for further improvement. Fourth, financial evaluation had to be discarded from the original objectives due to the lack of information on the costs of the claims. Finally, we believe that other types of models could have also been considered for handling longitudinal count data, such as the Hurdle models or some other generalized alternatives.

Appendix A

R CODE FOR HI DATASET CONFIGURATION

A.1 Importing the Claims File: claims.R

```
import_claim <- function(data, open_info, close_info, keeptown = c("all", "Barcelona")) {

 keeptown <- match.arg(keeptown)

 # Import the claims file via an adequate format and removing those contracts related
 # to subjects who were born before 1900-01-01 (so avoiding impossible ages for being
 # too high), as well as those contracts related to subjects who were born on a date
 # equal to or later than the study closing date, on 2014-02-01.
 # We also keep those contracts whose "town" variable is the one we are interested in.
 # Regard this "town" variable, remove all characters after:
 # 1) Comma: ", E" - ", L" - ", L’" - ", LA" - "," - ", ELS" - ", PARTIDA"
 # 2) Slash (to avoid duplicates in nomenclature)
 # 3) Open parenthesis
 # After that, all the white spaces at the end of a "town" name will also be removed,
 # and the letters are changed from capital to lower case except the first letter of
 # each word.
 # De Morgan’s laws: !(X | Y) is the same as !X & !Y
 # -----------------------------------------------------------------------------------
 claims_ini <- fread(data, na.strings = "", drop = c("edad", "VAR19"))[,
   list(
     id = snip,
     dborn = as.Date(fnacim, format = "%d/%m/%Y"),
     sex = factor(c(1, 0))[match(factor(sexo), c("D", "H"))],
     cp = cpostal,
     town = {
       town0 <- sub("\\s+$", "", gsub("(.*)[/(,].*", "\\1", poblacion))
       factor(gsub("\\b(\\w)", "\\U\\1", tolower(town0), perl = TRUE))
     },
     dclaim = as.Date(facto, format = "%d/%m/%Y"),
     cfam = factor(cfamilia, labels = c("hosp", "prostheses", "non_routine", "hosp_2",
       "visit", "test", "analysis", "treatment", "carnet", "healthcare", "homevisit",
       "homeats", "ambulance", "ats")),
     cspe = factor(cespecialidad),
     cclaim = factor(cacto),
     quantity = cantidad)][!(dborn < as.Date("1900-01-01") | dborn >= close_info)]

 # Multi-replacement of special characters
 # ---------------------------------------
 mgsub <- function(pattern, replacement, x) {
   result <- x
   for (i in 1:length(pattern)) {
     result <- gsub(pattern[i], replacement[i], result)
```

```
  }
result
}

inichar <- c("Ca’N", "D", "D‘", "D’", " d’En ", " De ", "Del ", "Els ", " I ",
             "L’", "La", " Les ", " Na ", " S’", " Sa ", " San ", " Santa ",
             "Sant ", "Santa ", " Ses ")

finchar <- c("Can",  "d’", "d’", "d’", " d’en ", " de ", "del ", "els ", " i ",
             "l’", "la", " les ", " na ", " s’", " sa ", " S. ", " Sta ",
             "St. ", "Sta. ", " ses ")

claims_ini$town <- factor(mgsub(inichar, finchar, as.character(claims_ini$town)))

# Replacement of names from different neighborhoods in the municipality
# "town" = "Palma de Mallorca", as well as names of the three villages
# that belong to "town" = "Gavet de la Conca"
# ---------------------------------------------------------------------
levels(claims_ini$town)[levels(claims_ini$town) %in%
  c("Can Pastilla", "Maravillas las", "S’Aranjassa", "Sa Indioteria", "Son Anglada",
    "Son Espanyolet", "Son Ferriol", "Son Sardina")] <- "Palma de Mallorca"

levels(claims_ini$town)[levels(claims_ini$town) %in%
  c("Aransis", "St. Salvador de Tolo", "St. Serni")] <- "Gavet de la Conca"

# Rename those two levels which have a common string in "town" = "Sta. Eulalia"
# ----------------------------------------------------------------------------
levels(claims_ini$town) <- c(levels(claims_ini$town), "Sta. Eulalia de Aren")
claims_ini[town == "Sta. Eulalia" & cp == "22583", town := "Sta. Eulalia de Aren"]
claims_ini[town %in% c("Cala Llonga", "Es Cana", "Sta. Eulalia"),
  town := "Sta. Eularia des Riu"]

# Multi-replacement of provided geographical names for the correct
# assignment of their municipality and to correct spelling mistakes
# ----------------------------------------------------------------
initown <- c("5100 Jambes", "Aeroport del Prat", "Ampolla l’",
  "Arenal", "Bossa", "Bruselas 1000",
  "Cala Vias", "Castellfollit de la Roc", "Castillo de Bendinat",
  "Chinchilla de Monte Ara", "Es Calo", "Fontpineda",
  "Granyena de les Garrigu", "Lligallo del Ganguill", "Llosses, Les",
  "Madrigal de las Altas T", "Maioris X", "Massalc0reig",
  "Montmany- Figaro", "Nuestra Seora del Pila", "Palmeras Park",
  "Platja de Alcudia", "Poligono Industrial De", "San Fernando",
  "San Lazaro", "San Miguel de Son Carri", "Siesta",
  "Sigena", "St. Carles de la Rapit",
  "St. Fost de Campsentel", "St. Fost de Campsentell",
  "St. Francesc", "St. Francesc de s’Esta",
  "St. Joan", "St. Joan de Les Abad", "St. Joan de Vilatorrad",
  "St. Jordi de ses Salin", "St. Josep de la Atalai",
  "St. Julia del Llor i B", "St. Just",
  "St. Miquel de Campmajo", "St. Pere de Riudebitll",
  "St. Pol", "St. Salvador", "St. Salvador - Coma-Ru",
  "St. Salvador de Guardi", "Sta. Agnes de Malanyan",
  "Sta. Cecilia de Voltre", "Sta. Coloma de Cervell",
  "Sta. Coloma de Gramene", "Sta. Eulalia de Riupri",
  "Sta. Eulalia de Ronan", "Sta. Fe D", "Sta. Gertrudis de Frui",
  "Sta. Margarida de Mont", "Sta. Margarida i els M",
  "Sta. Maria", "Sta. Maria d'Olo",
  "Sta. Maria de Martorel", "Sta. Maria Montmagastr",
  "Sta. Perpetua de Mogod", "Vallcanera", "Vallfogona de Riucorp",
  "Vilars, Els", "Villafruela de Porma")

fintown <- c("Jambes (BE)", "Prat de Llobregat", "Ampolla",
  "S'Arenal", "Platja d'en Bossa", "Bruxelles (BE)",
  "Cala Vinyes", "Castellfollit de la Roca", "Bendinat",
  "Chinchilla de Montearagon", "Es Calo de St. Agusti", "Palleja",
  "Granyena de les Garrigues", "Lligallo del Ganguil", "Llosses",
  "Madrigal de Altas Torres", "Maioris", "Massalcoreig",
  "Montmany-Figaro", "Pilar de la Mola", "Ses Palmeres",
  "Alcudia", "Marratxi", "St. Ferran de ses Roques",
  "St. Llatzer", "St. Miquel de son Carrion", "Torrevieja",
  "Villanueva de Sigena", "St. Carles de la Rapita",
  "St. Fost de Campsentelles", "St. Fost de Campsentelles",
  "St. Francesc de Formentera", "St. Francesc de s'Estany",
  "Alicante", "St. Joan de les Abadesses", "St. Joan de Vilatorrada",
  "St. Jordi de ses Salines", "St. Josep de sa Talaia",
  "St. Julia del Llor i Bonmati", "Pi de St Just",
  "St. Miquel de Campmajor", "St. Pere de Riudebitlles",
  "St. Pol de la Bisbal", "Gelida", "St. Salvador de Coma-Ruga",
  "St. Salvador de Guardiola", "Sta. Agnes de Malanyanes",
  "Sta. Cecilia de Voltrega", "Sta. Coloma de Cervello",
  "Sta. Coloma de Gramanet", "Sta. Eulalia de Riuprimer",
  "Sta. Eulalia de Ronana", "Oluges", "Sta. Gertrudis de Fruitera",
  "Sta. Margarida de Montbui", "Sta. Margarida i els Monjos",
  "Sta. Maria d'Eivissa", "Sta. Maria d'Olot",
  "Sta. Maria de Martorelles", "Sta. Maria Montmagastrell",
  "Sta. Perpetua de Mogoda", "Sils", "Vallfogona de Riucorb",
  "Vilars", "Villafruela del Porma")

for (i in 1:length(initown)) {
  levels(claims_ini$town)[levels(claims_ini$town) == initown[i]] <- fintown[i]
}

# Order alphabetically the levels of the variable "town"
# ----------------------------------------------------
claims_ini$town <- factor(claims_ini$town, levels = sort(levels(claims_ini$town)))

# Order alphabetically the character levels of the variable "cfam"
# --------------------------------------------------------------
claims_ini$cfam <- factor(claims_ini$cfam, levels = sort(levels(claims_ini$cfam)))

# Remove those rows in which variable "quantity" = 0.
# ATTENTION: It has been checked that inside corresponding "id", each of these
# rows is duplicated, so that in one appears "quantity" = 0 and in the other we
# can see "quantity" = 1 (unknown reason). Thus, in this case removing those rows
# does not entail any losses in the number of contracts.
# ---------------------------------------------------------------------------
id_add <- setdiff(unique(claims_ini$id),              # Contracts to recover
  unique(claims_ini[!(quantity == 0)]$id))
claims <- switch(keeptown,
   all = rbind(
     claims_ini[!(quantity == 0)],
     claims_ini[!(duplicated(id)) & (id %in% id_add)]),
   Barcelona = rbind(
     claims_ini[!(quantity == 0)],
     claims_ini[!(duplicated(id)) & (id %in% id_add)])[town == "Barcelona"]
     )[, c("cp", "town") := NULL]

 # Remove contract's rows which contain claims information before the corresponding
 # subject's birthdate (dborn > dclaim), and also the contract's rows in which the
 # claim date "dclaim" is equal to or later than the date of study close. Moreover,
 # drop the levels that do not occur in categoric variables.
 # De Morgan's laws: !(X | Y) is the same as !X & !Y
 # ---------------------------------------------------------------------------------
 catclaims <- which(sapply(claims, is.factor))
 return(subset(claims, !(open_info > dclaim | dclaim >= close_info))[,
   names(claims)[catclaims] := lapply(.SD, factor), .SDcols = catclaims])
}
```

A.2 Importing the Time-to-Event Information: lifetimes.R

```
# NOTE:
# When importing the time-to-event information, we must take into account that there
# are some contracts whose connection between "snip" and "npoliza" is not bijective,
# because the same subject could have been changing the contract conditions during
# subject's stay in the insurance company. Therefore, there are contracts (identified
# by variable "snip" in the original lifetimes file) with multiple dates associated
# for both enrollment in the company (variable "fingreso" in the original file) and
# for ending their follow-up interval (variable "ffinal" in in the original file).
# -----------------------------------------------------------------------------------
import_lifetimes <- function(data, first_control, close_info, lag) {

 # Import the subjects' contracts under the following conditions:
 # 1) All these variable names are replaced: "snip" by "id", "npoliza" by "npolicy",
 #   "fingreso" by "dini", "ffinal" by "dfinal", and "tmotivo" by "endcause".
 # 2) Overwrite the date "close_info" in all contracts where "dfinal" = NA, or where
 #    "dfinal" is later than "close_info". In addition, we change the date format of
 #    the variables "dini" and "dfinal".
 # 3) Overwrite the string "None" in missing values of variable "endcause".
 # Since we are really interested in the minimum initial date (the oldest) of a subject's
 # contract and in its final date within the study period, only one observation for each
 # "id" is considered (even though it has several "npolicy" associated).
 # To obtain this single observation per subject's contract, we distinguish between the
 # following situations:
 # A) If there are more than one "dini", we take the oldest of them.
 # B) Regarding "dfinal" and "endcause" variables, there are three options:
 #    1B) If endcause = "Defuncio del risc" IS NOT present in any of the contract's rows,
 #        then we take the maximum "dfinal" among these rows, with the corresponding
 #        "endcause".
 #    2B) If "endcause" = "Defuncio del risc" is present in ONLY ONE of the contract's
 #        rows, then "dfinal" is that associated with this "endcause".
 #    3B) If "endcause" = "Defuncio del risc" is present in TWO OR MORE of the contract's
 #        rows, then "dfinal" is that associated with the "endcause" = "Defuncio del risc"
 #        with an oldest date.
 # ---------------------------------------------------------------------------------------
 lifetimes_ini <- fread(data, na.strings = "")[order(snip)][, list(
   id = snip,
   npolicy = npoliza,
   dini = as.Date(fingreso, format = "%d/%m/%Y"),
   dfinal = replace(as.Date(ffinal, format = "%d/%m/%Y"), is.na(ffinal) |
     as.Date(ffinal, format = "%d/%m/%Y") > close_info, close_info),
   endcause = as.factor(replace(tmotivo, is.na(tmotivo), "None")))][, {
     if (all(endcause != "Defuncio del risc")) {
        indx <- which.max(dfinal)
        list(dini = min(dini), dfinal = dfinal[indx], endcause = endcause[indx])
     } else {
        if (sum(levels(endcause) == "Defuncio del risc") == 1) {
           indx <- which(endcause == "Defuncio del risc")
           list(dini = min(dini), dfinal = dfinal[indx], endcause = endcause[indx])
       } else {
       indx <- which(id & endcause == "Defuncio del risc" &
         dfinal == min(dfinal[endcause == "Defuncio del risc"]))
       list(dini = min(dini), dfinal = dfinal[indx], endcause = endcause[indx])
       }
     }
 }, by = id]

 # There may be a same "id" which has "endcause" = "Defuncio del risc" in two or more
 # rows, with their corresponding "dfinal" being equally old (and older than the rest
 # of contract's rows). In such cases, duplicities are removed.
 # Moreover, the variable "endcause" is replaced by the dichotomous variable "status"
 # ----------------------------------------------------------------------------------
 lifetimes <- unique(lifetimes_ini[, c("endcause", "status") :=
   list(NULL, as.numeric(endcause == "Defuncio del risc"))], by = "id")

 # Finally, the following rows (contracts) are removed from the time-to-event file:
 # 1) Rows with "dini" equal to "dfinal"
 # 2) Rows with "dini" equal to or later than "lag" days before "close_info"
 # 3) Rows with "dfinal" equal to or earlier than "lag" days after "first_control"
 # --------------------------------------------------------------------------------
 return(lifetimes[!(dini == dfinal | dini >= close_info - lag |
   dfinal <= first_control + lag)])
}
```

A.3 Merging Longitudinal and Survival Data: fusion.R

```
fusion <- function(dtable1, dtable2, age.u, lag, open_info, close_info) {

 # Merge the claims and lifetimes files, and dropping those levels that do not occur.
 # From this moment on we are not dealing with contracts any more, but working directly
 # with subjects' information.
 # ------------------------------------------------------------------------------------
 dt_ini <- merge(dtable1, dtable2, by = "id", allow.cartesian = TRUE)[,
   list(id, dborn, sex, dini, dclaim,
     cfam = factor(cfam, order = TRUE, levels = sort(levels(cfam))),
     cspe = factor(cspe),
     cclaim = factor(cclaim),
     quantity, dfinal, status)][order(id, dclaim)]

 # NATURAL ORDER: dborn ----> dini ----> dclaim ----> dfinal
 # The above chronological order must be always kept for each subject’s row, so that in
 # several cases we will remove those subjects’ information for whom the aforementioned
 # order is not respected. Nonetheless, there are some cases in the provided files in
 # which the logical order between two subsequent dates is altered no longer than "lag"
 # days by administrative issues. These specific cases will be also included in our final
 # dataset after assigning consistent dates.
 # --------------------------------------------------------------------------------------
 # ROWS TO REMOVE                     # ROWS TO KEEP
 # -----------------------------------------------------------------------
 # A) dborn - dini > 30               Logical: dini - dborn >= 0
 #                                    Admitted: 0 <= dborn - dini <= 30
 #                                    (We set "dini" = "dborn")
 # -----------------------------------------------------------------------
 # B) dborn - dclaim > 30             Logical: dclaim - dborn >= 0
 #                                    Admitted: 0 <= dborn - dclaim <= 30
 #                                    (We set "dclaim" = "dborn")
 # -----------------------------------------------------------------------
 # C) dborn - dfinal >= 0             Logical: dfinal - dborn > 0
 # -----------------------------------------------------------------------
 # D) dini - dclaim > 30              Logical: dclaim - dini >= 0
 #                                    Admitted: 0 <= dini - dclaim <= 30
 #                                    (We set "dclaim" = "dini")
 # -----------------------------------------------------------------------
 # E) dini - dfinal >= 0              Logical: dfinal - dini > 0
 # -----------------------------------------------------------------------
 # F) dclaim - dfinal > 30            Logical: dfinal - dclaim >= 0
 #                                    Admitted: 0 <= dclaim - dfinal <= 30
 #                                    (We set "dclaim" = "dfinal - 1")
 # -----------------------------------------------------------------------
 dt_ini_2 <- dt_ini[!(dborn - dini > lag | dborn - dclaim > lag |           # Remove A, B
   dborn - dfinal >= 0 | dini - dclaim > lag |                              # Remove C, D
   dini - dfinal >= 0 | dclaim - dfinal > lag)][(dborn - dini) %in% 0:lag,  # Remove E, F
     dini := as.Date(dborn, origin = "1970-01-01")][,                       # Keep A
     dclaim := as.Date(ifelse((dclaim - dfinal) %in% 0:lag, dfinal - 1,     # Keep F
       ifelse((dini - dclaim) %in% 0:lag, dini,                             # Keep D
       ifelse((dborn - dclaim) %in% 0:lag, dborn, dclaim))),                # Keep B
         origin = "1970-01-01")]

 # WORKING HYPOTHESIS: A subject can enter into the study at any of the eight control
 # points, also called "starting points", placed right at the days January 1 across
 # the time window ("open_info", "close_info"].
 # If someone had been born on February 29, it would not imply any problem since this
 # individual would only be able to enter on the day January 1 from the following year
 # (when the subject will certainly be more than 65 years).
 # The eight starting points are obtained via the components of "sp[i]". It contains
 # all the first days of January located between the years 2007 and 2014, and, for the
 # i-th subject, the first and last components of vector "sp[i]" fit these rules:
 # a) Regard to the first component of the vector: "sp.1[i]"
 #    a1) If "open_info" and the following January 1 are spaced at greater or equal than
 #        "lag" days, then "sp.1[i]" will be located on that January 1.
 #    a2) If "open_info" is located right on a January 1, then "sp.1[i]" will be located
 #        on the January 1 from the following year.
 #    a3) If "open_info" and the following January 1 are spaced at less than "lag" days,
 #        then "sp.1[i]" will be located on the January 1 of the year after the year of
 #        the initially mentioned January 1.
 # b) Regard to the last component of the vector: "sp.n[i]"
 #    b1) If "close_info" and the previous January 1 are spaced at greater or equal than
 #        "lag" days, then "sp.n[i]" will be located on that January 1.
 #    b2) If "close_info" is located right on the January 1, "sp.n[i]" will be located
 #        on the January 1 previous to "close_info".
 #    b3) If "close_info" and the previous January 1 are spaced at less than "lag" days,
 #        then "sp.n[i]" will be located on the January 1 of the year previous to the
 #        year of the initially mentioned January 1.
 # --------------------------------------------------------------------------------------
 sp.start <- as.Date(ifelse(
   as.Date(paste(as.numeric(format(open_info, "%Y")) + 1, 1, 1, sep = "-"))
     - open_info >= lag,
   as.Date(paste(as.numeric(format(open_info, "%Y")) + 1, 1, 1, sep = "-")),
   as.Date(paste(as.numeric(format(open_info, "%Y")) + 2, 1, 1, sep = "-"))),
     origin = "1970-01-01")
 sp.end <- as.Date(ifelse(
   close_info - as.Date(paste(format(close_info, "%Y"), 1, 1, sep = "-")) >= lag,
   as.Date(paste(format(close_info, "%Y"), 1, 1, sep = "-")),
   as.Date(paste(as.numeric(format(close_info, "%Y")) - 1, 1, 1, sep = "-"))),
     origin = "1970-01-01")
 sp <- seq(sp.start, sp.end, by = "year")

 # From the data table "dt_ini_2", only those subjects who reach the age of 65
 # at any of the eight starting points are kept
 # -----------------------------------------------------------------------------
 dt <- dt_ini_2[round((sp[length(sp)] - dborn)/365.25, 2) >= age.u, ]

 # Loop to obtain eight data tables, each of which includes those subjects whose
 # "sp.1[i]" from vector "sp[i]" is respectively associated to one of the eight
 # possible starting points
 # -----------------------------------------------------------------------------
 dt_sp_ini_ <- dt_sp_ <- id_add_sp_ <- list()

 for (i in 1:length(sp)) {
   tp <- c(open_info, sp[- length(sp)])
   date0 <- as.Date("1000-01-01")
   vp <- wp <- as.POSIXlt(c(date0, sp))
   vp$year <- vp$year - age.u          # At least aging 65 years to enter into the study
   wp$mon <- wp$mon - 4                # At least 4 months in insurance before entering

   dt.cut.dborn <- as.numeric(cut(x = as.POSIXlt(dt$dborn),
     breaks = vp, right = TRUE, include.lowest = TRUE))
   dt.cut.dini <- as.numeric(cut(x = as.POSIXlt(dt$dini),
     breaks = wp, right = TRUE, include.lowest = TRUE))
   dt.cut <- pmax(dt.cut.dborn, dt.cut.fini)

   dt_sp_ini_[i] <- split(dt, factor(dt.cut, i))
   dt_sp_ini_[[i]] <- dt_sp_ini_[[i]][!(dfinal <= sp[i])][format(dfinal, "%m") == "01",
     dfinal := as.Date(ifelse(status == 0 | status == 1 & format(dfinal, "%Y") != "2014",
   as.Date(paste(format(dfinal, "%Y"), "02", "01", sep = "-")), dfinal),
       origin = "1970-01-01")]
```

```
 # Getting the longitudinal information regarding the follow-up interval for each subject
 # entering at specific sp[i]: [dent = sp[i], dfinal].
 # To get the subject's first "claimyr" value (that is, the annual "quantity" of medical
 # claims), that corresponds to the date "sp[i]" at which the subject enters into the
 # study, only those "dclaim" which occur from the corresponding "tp[i]" onwards are
 # considered. Therefore, all the longitudinal information occurred on "dclaim" (rows)
 # previous to the date "tp[i]" must be removed.
 # Attention: Removing all subject's claims before "tp[i]" could result in the
 # disappearance of all subject's rows (it would be the case in which the subject only
 # has claims before "tp[i]"). In that case, such a subject must be recovered by one
 # single and "artificial" row such that it automatically assigns "dclaim" = "tp[i]" and
 # "quantity" = 0 (these rows are also given null family codes). This subject will report
 # both longitudinal information (registering "claimyr" = 0 in the subsequent starting
 # points that the subject crosses) and survival information.
 # --------------------------------------------------------------------------------------
 id_add_sp_[[i]] <- setdiff(unique(dt_sp_ini_[[i]]$id),
   unique(dt_sp_ini_[[i]][!(dclaim < tp[i])]$id))      # Subjects to recover

 dt_sp_[[i]] <- rbind(
   dt_sp_ini_[[i]][!(id %in% id_add_sp_[[i]])][!(dclaim < tp[i])],
   dt_sp_ini_[[i]][!(duplicated(id)) & (id %in% id_add_sp_[[i]])][,
     c("dclaim", "cfam", "cspe", "dclaim", "quantity", "expo") :=
   list(tp[i], factor(0), factor(0), factor(0), 0, 1), ])[,
     list(id, dborn, sex,
       age_dent = as.vector(round((sp[i] - dborn) / 365.25, 2)),
           dini, dent = sp[i], dclaim,
       cfam = factor(cfam, levels = c("0", levels(dtable2$cfam))),
       cspe = factor(cspe, levels = c("0", levels(dtable2$cspe))),
       cclaim = factor(cclaim, levels = c("0", levels(dtable2$cclaim))),
       quantity, dfinal,
       age_dfinal = as.vector(round((dfinal - dborn) / 365.25, 3)),
       status)][age_dfinal < 105 & order(id, dclaim)]

 assign(paste0("dt_sp_", 1:length(sp))[i], dt_sp_[[i]])
 }

 # Merging data tables of subjects who enter at different starting points
 # ----------------------------------------------------------------------
 union <- do.call(
   what = rbind,
   args = lapply(X = paste0("dt_sp_", 1:length(sp)),
     FUN = get,
     envir = environment()),
   envir = parent.frame())[, dini := NULL][order(id, dclaim)]

 catunion <- which(sapply(union, is.factor))
 return(union[, names(union)[catunion] := lapply(.SD, factor), .SDcols = catunion])

}
```

A.4 Scrubbing Process and Variable Selection: clean.R

```
clean <- function(dat, keepfam = c("all", "ambulance", "hosp", "non_routine")) {
```

```
cats <- which(sapply(dat, is.factor))
keepfam <- match.arg(keepfam)

# ------------------------------
# Cleaning of the neutral claims
# ------------------------------

# Only those rows whose values of "cfam" belong to one of the levels considered in the
# vector "keepfam" provide non-null emergency claims. Moreover, attention must be paid
# to those "id" in which none of their claims are associated with the mentioned "cfam"
# codes. They are subjects whose variable "quantity" will subsequently record zero, but
# also must be considered, being recorded for analysis using the data table "dt0". At
# the moment, these subjects provide single row with "quantity" = 0 (later, when the
# longitudinal mesh is implemented, they will have as many rows with "quantity" = 0 as
# different starting points crossed by the subject's longitudinal trajectory, i.e.
# ["dent", "dfinal"], within the study period. These rows will be joined the rows with
# the null codes already existing from the previous step.
# --------------------------------------------------------------------------------------
dt0 <- switch(keepfam,
  all = rbind(
    unique(dat[dat$id %in% unique(dat$id[!ave(cfam %in%
      c("0", "ambulance", "hosp", "non_routine"), list(id), FUN = any)])], by = "id")[,
        c("cfam", "cspe", "cclaim", "quantity") := c(lapply(rep(0, 3), FUN = factor), 0)],
    dat[cfam == "0"]),
  ambulance = rbind(
    unique(dat[dat$id %in% unique(dat$id[!ave(cfam %in%
      c("0", "ambulance"), list(id), FUN = any)])], by = "id")[,
        c("cfam", "cspe", "cclaim", "quantity") := c(lapply(rep(0, 3), FUN = factor), 0)],
    dat[cfam == "0"]),
  non_routine = rbind(
    unique(dat[dat$id %in% unique(dat$id[!ave(cfam %in%
      c("0", "non_routine"), list(id), FUN = any)])], by = "id")[,
        c("cfam", "cspe", "cclaim", "quantity") := c(lapply(rep(0, 3), FUN = factor), 0)],
    dat[cfam == "0"]),
  hosp = rbind(
    unique(dat[dat$id %in% unique(dat$id[!ave(cfam %in%
      c("0", "hosp"), list(id), FUN = any)])], by = "id")[,
        c("cfam", "cspe", "cclaim", "quantity") := c(lapply(rep(0, 3), FUN = factor), 0)],
    dat[cfam == "0"])
  )[order(id, dclaim)][, names(dat)[cats] := lapply(.SD, factor), .SDcols = cats]

# -------------------------------------------------------
# Cleaning of the claims associated to ambulance services
# -------------------------------------------------------

if (keepfam %in% c("all", "ambulance")) {

 # All rows are selected whose "cfam" = "ambulance" and whose "cclaim" belongs to one of
 # the following ambulance codes:
 # "880001", "880002", "880003", "880004", "880005", "880006", "880007"
 # The meaning of each one of the previous codes is:
 # Call an ambulance (normal service or ICU): "cclaim" = {"880001", "880002"}
 # Km by normal ambulance or ICU: "cclaim" = {"880003", "880004"}
 # Hours waited for the normal ambulance or ICU: "cclaim" = {"880005", "880006"}
 # Medical fees for the ambulance services: "cclaim" = "880007"

 # Since only absolute frequencies of ambulance claims are of interest, for each single
 # pair ("id", "dclaim") we prior remove those rows whose "cclaim" belongs to the set of
 # codes: {"880003", "880004", "880005", "880006", "880007"}. However, there are some
 # single pairs ("id", "dclaim") that have a set of rows in which the rows with codes
 # "cclaim" to remove are not found together with rows associated to the general codes
 # "880001" or "880002". In such a cases, removing these rows would mean disappearance
 # of any information about ambulance services within the single pair ("id", "dclaim").

 # To begin with, all the single pairs ("id", "dclaim") associated to a single row are
 # identified, and all those codes of "cclaim" not belonging to {"880001", "880002"}
 # are replaced in the following way: Those single rows whose code of variable "cclaim"
 # is {"880003", "880005", "880007"} are replaced by rows with "cclaim" = "880001",
 # while those single rows with "cclaim" = {"880004", "880006"}, are replaced by rows
 # with "cclaim" = "880002".
 # -------------------------------------------------------------------------------------
 sing_amb <- dat[cfam == "ambulance", if (.N == 1) .SD,
   by = .(id, dclaim)][, c("cclaim", "quantity") := list(
     factor(ifelse(cclaim %in% c("880003", "880005", "880007"), "880001",
       ifelse(cclaim %in% c("880004", "880006"), "880002", as.character(cclaim)))), 1)]

 # Later, the single pairs ("id", "dclaim") which have more than one row associated with
 # the ambulance codes are identified. The procedure is then the following:
 # a) In single pairs ("id", "dclaim") in which at least one their rows have the codes
 #    "cclaim" = {"880001", "880002"}, all their rows associated with the codes
 #    "cclaim" = {"880003", "880004", "880005", "880006", "880007"} will be removed.
 # b) In single pairs ("id", "dclaim") in which none of their rows are associated with
 #    "cclaim" = {"880001", "880002"}, all their rows will be replaced with a single
 #    row whose "cclaim" will be either "880001", if the number of rows with "cclaim" =
 #    {"880003", "880005"} is greater than or equal to the number of rows that have
 #    "cclaim" = {"880002", "880004"}, or with "880002" if the opposite is true.
 # -------------------------------------------------------------------------------------
 mult_amb <- rbind(
   dat[cfam == "ambulance", if (.N > 1) .SD,                                   # a)
     by = .(id, dclaim)][ave(cclaim %in% as.character(880001:880002),
       list(id, dclaim), FUN = any)][cclaim %in% as.character(880001:880002)],
   unique(dat[cfam == "ambulance", if (.N > 1) .SD,                            # b)
     by = .(id, dclaim)][!ave(cclaim %in% as.character(880001:880002),
       list(id, dclaim), FUN = any)][, cclaim := factor(
         ifelse(sum(cclaim %in% c("880003", "880005")) >= sum(cclaim %in%
           c("880004", "880006")), "880001", "880002")), by = .(id, dclaim)],
             by = c("id", "dclaim"))
   )[, quantity := 1]

 all_amb <- setcolorder(rbind(sing_amb, mult_amb), names(dat))[,
   names(dat)[cats] := lapply(.SD, factor), .SDcols = cats][order(id, dclaim)]
}

# ---------------------------------------------------------
# Cleaning of the claims associated with non-routine visits
# ---------------------------------------------------------

if (keepfam %in% c("all", "non_routine")) {
  all_nonroutine <- dat[cfam == "non_routine"][, quantity := 1][,
    names(dat)[cats] := lapply(.SD, factor), .SDcols = cats]
}

# -------------------------------------------------------
# Cleaning of the claims associated with hospitalizations
# -------------------------------------------------------

if (keepfam %in% c("all", "hosp")) {

 # STEP 1 with hospitalizations:
 # Firstly, we identify all rows belong to subjects who do not have any rows associated
 # with a "cclaim" code belonging to one of the following group of codes:
 # {"169997", "359997", "790001", "790002", "790007", "790018",
 # "790019", "790021", "790993", "790997", "790998", "790999"}
 # All rows of each these subjects will be replaced with a single row in which "cfam" =
 # "cspe" = "cclaim" = "0", and "quantity" = 0.
 # In the cases in which keepfam = "all", a certain single row will be removed (because
 # is unnecessary) if its ""id" is already present in other "cfam" codes, while it will
 # be kept if the opposite is true.
 # In cases where keepfam = "hosp", all single rows will be kept.
 # ------------------------------------------------------------------------------------
 keephosp <- c("169997", "359997", "790001", "790002", "790007", "790018",
   "790019", "790021", "790993", "790997", "790998", "790999")

 hosp_0 <- unique(dat[cfam == "hosp" &
   id %in% unique(id[!ave(cclaim %in% keephosp, list(id), FUN = any)])][,
     c("cfam", "cspe", "cclaim", "quantity") :=
       c(lapply(rep(0, 3), FUN = factor), 0)], by = "id")

 if (keepfam == "all") {
   sing_0 <- hosp_0[!id %in% unique(rbind(all_amb, all_nonroutine)$id)]
 } else {
   sing_0 <- hosp_0
 }

 # STEP 2 with hospitalizations:
 # Now we focus on those subject's rows that have a minimum of one row whose "cclaim"
 # belongs to the "keephosp" group. Of these rows, we only keep those rows whose code
 # "cclaim" is included in "keephosp".
 # Within each single pair ("id", "dclaim"), we reduce to a single row all those rows
 # with repeated hospitalization codes, in other words, those rows whose "cclaim"
 # levels are associated with the following group of codes:
 # duphosp = c("790001", "790002", "790007", "790018", "790993", "790999")
 # 2a) If the code "790002" is present, The first row with said code is kept.
 # 2b) In the opposite case, the "cclaim" levels of the group of rows are replaced by
 #     the code "790001".
 # 2c) The rows with "cclaim" in "keephosp" but not in "duphosp" are kept.
 # ----------------------------------------------------------------------------------
 hosp_any <- dat[cfam == "hosp" & !(id %in% hosp_0$id) & cclaim %in% keephosp]

 # 2a)
 rhosp_2a <- hosp_any[, if (any(cclaim == "790002")) .SD, by = .(id, dclaim)]
 hosp_2a <- unique(rhosp_2a[cclaim == "790002"], by = c("id", "dclaim"))

 # 2b)
```

```
  # The rows within "rhosp_2a" are removed from the data table "hosp_any",
  # and the rows from single pairs ("id", "dclaim") whose "cclaim" belongs
  # to "790001", "790007", "790018", "790993", "790999" are replaced by
  # a single row with the code cclaim = "790001".
  # ----------------------------------------------------------------------
  rhosp_2b <- hosp_any[!do.call("paste", hosp_any) %in%
    do.call("paste", setcolorder(rhosp_2a, names(dat)))]
  duphosp <- c("790001", "790002", "790007", "790018", "790993", "790999")
  hosp_2b <- unique(rhosp_2b[, if (any(cclaim %in% duphosp[- 2])) .SD,
    by = .(id, dclaim)], by = c("id", "dclaim"))[, cclaim := "790001"]

  # 2c)
  # We finally have single pairs ("id", "dclaim") in which all their rows  have some
  # code from "keephosp", but that in no case the code belongs to the vector "duphosp"
  # ----------------------------------------------------------------------------------
  hosp_2c <- rhosp_2b[, if (!any(cclaim %in% duphosp[- 2])) .SD, by = .(id, dclaim)]

  # STEP 3 with hospitalizations:
  # Join: sing_0, hosp_2a, hosp_2b, hosp_2c
  # --------------------------------------
  all_hosp <- rbind(sing_0,
    setcolorder(rbind(hosp_2a, hosp_2b, hosp_2c), names(dat))[, quantity := 1])[,
      names(dat)[cats] := lapply(.SD, factor), .SDcols = cats][order(id, dclaim)]
 }

 # ------------------------------------
 # RESULT DEPENDING ON THE CHOSEN CLAIMS
 # ------------------------------------
 out <- switch(keepfam,
   all = rbind(dt0, all_amb, all_nonroutine, all_hosp),
   ambulance = rbind(dt0, all_amb),
   non_routine = rbind(dt0, all_nonroutine),
   hosp = rbind(dt0, all_hosp))[, c("cfam", "cspe", "cclaim") := NULL][order(id, dclaim)]
 return(out)
}
```

A.5 Obtaining the HI Dataset: mesh.R

```
mesh <- function(data) {

 # a) The following data is joined:
 #    1) Data available after the elimination of possible rows where "dclaim" is after
 #       the date of the last control point, as these do not compute due to no trajectory
 #       reaching the year 2015 (however, we will keep those rows in the described case
 #       if "quantity" = 0, because in that situation they provide information to use in
 #       the longitudinal analysis).
 #    2) A new data table where each "id" is assigned a reference row (that is, provides
 #       a row with "quantity" = 0) for each of the starting points crossed by subject's
 #       profile within the study period.
 # ------------------------------------------------------------------------------------
 all_a <- rbind(
  data[!(dclaim > as.Date("2013-12-31") & quantity > 0)],
  unique(data[, list(dborn, sex, age_dent, dini, dent,
    dclaim = as.Date(c(paste0((year(dent[1]) - 1):(year(dfinal[1]) - 1), "-12-31"))),
    quantity = 0, dfinal, age_dfinal, status), by = id])
  )[order(id, dclaim)]

 # b) After ordering the previous joining, within each single pair ("id", "year") we move
 #    all the values of the variable "quantity" to the corresponding December 31, and we
 #    add all these values for "quantity".
 # --------------------------------------------------------------------------------------
 all_b <- unique(all_a[, c("dclaim", "quantity") := list(dclaim[.N], sum(quantity)),
      by = .(id, as.POSIXlt(dclaim)$year)])

 # c) For each "id", we only keep all those rows whose "dclaim" is a December 31. At
 #    this point, we have all the annual information of emergency claims summarized
 #    just before the control points. In each case, we move this information to the
 #    day immediately after each (a starting point) by means of the variable "start".
 #    Moreover, the variable "quantity" is replaced by the variable "claimyr".
 # ----------------------------------------------------------------------------------
 all_c <- all_b[month(dclaim) == 12 & as.POSIXlt(dclaim)$mday == 31][, list(
   id, dborn, sex, age_dent, dini, start = dclaim + 1 , claimyr = quantity,
   dfinal, status)]

 # d) Almost all the exposure times, "expo" variable, are equal to 1, since most of the
 #    subjects have been observed the whole period between two consecutive "starting
 #    points" (i.e. a whole calendar year). Those subjects whose "expo" value is lesser
 #    than 1 are subjects who have not been observed during the whole calendar year prior
 #    to their entering into the study. Due to this, the value of "claimyr" is corrected
 #    by the corresponding "expo" value, so that yearly claims are weighted depending of
 #    the exposure time.
 #    At this stage, we will have all the longitudinal information situated just at the
 #    beginning of each calendar year within study period.
 # --------------------------------------------------------------------------------------
 all_d <- all_c[, expo := 1][start - dini < 365, expo :=
   ifelse(as.numeric(format(dini, "%m")) >= 7, 0.5, round((start - dini)/365.25, 2))]

 # e) We add the variable "obstime" by setting, for the i-th subject, the time counter
 #    t = 0 at "sp.1[i]" (time at which the subject enters the study). Finally, the
 #    variables "dborn" and "dfinal" disappear.
 #   Attention: Due to the fact that we work with a precision of two decimals in the
 #    timing variables, it is not necessary to remove the last records of those "id"
 #    where practically "obstime" = "stop". They are "id" characterized by some of
 #    these two cases: a) they begin to be observed shortly after a December 31,
 #    b) they stop being observed shortly after a December 31.
 # -----------------------------------------------------------------------------------
 all_e <- all_d[, list(id, dborn, sex, age_dent,
   obstime = as.vector(round((start - dborn)/365.25, 2)) - 65,
   claimyr, expo,
   start = as.vector(round((start - dborn)/365.25, 2)) - 65,
   stop = 0, dfinal, status)][, stop :=
     c(obstime[- 1], as.vector(round((dfinal[.N] - dborn[.N])/365.25, 2)) - 65), by = id][,
       c("dborn", "dfinal") := NULL]

 # f) We obtain a data table in "long format", with the annual rate of emergency claims
 #    for each of the subjects reaching a minimum threshold age of "age_u" years (here 65)
 #    at one of the eight possible starting points within the study, from 2006-01-01 to
 #    2014-02-01.
 #    We also introduce the variable "event", that collects in a longitudinal manner the
 #    information provided by the variable "status" (it may assign a value 1 to the last
 #    subject's record in case the individual has associated "status" = 1)
 # -----------------------------------------------------------------------------------
 all_f <- all_e[, event := 0][!duplicated(id, fromLast = TRUE) & status == 1, event := 1]

 # We only keep those subjects who:
 # a) Do not go over "claimyr" = 20 in any of their measurements (we drop 19 subjects)
 # b) Start to be followed at an age not above 100 years (we drop 7 subjects)
 # --------------------------------------------------------------------------------
 return(all_f[, if (start[1] <= 35) .SD, by = id][!(id %in% all_f[claimyr > 20]$id)])

}
```

Organization of R Source Code

The code of different R scripts detailed in each of the previous sections are subsequently called using a single script, named as `HI_dataset.R`. In the final dataset, we include the longitudinal information provided for each of the subjects who reach a minimum age of 65 within the study window, and therefore meet the criteria to be observed. Additionally, the time-to-event information is recorded for each subject: age at study entry, age at study exit, and the cause due to which the subject is no longer observed.

```
# =============================
# Load the library "data.table"
# =============================

library(data.table)

# ====================================================================
# 1) Import all those medical claims which occur in the study period,
#    from January 1, 2006 to February 1, 2014, and are related to subjects
#    living in the city of Barcelona.
#    SCRIPT: claims.R
# ====================================================================

source("claims.R")
claims <- import_claims(
  data = "ACT20140225.csv",
  open_info = as.Date("2006-01-01"),
  close_info = as.Date("2014-02-01"),
  keeptown = "Barcelona")
# length(unique(claims$id))     # 33311 unique contracts
# nrow(claims)                  # 2162538 measurements

# =============================================================
# 2) Import those subjects whose membership of the health insurance
#    company extends to a date strictly later than January 1, 2006,
#    when the study period starts.
#    SCRIPT: lifetimes.R
# =============================================================
source("lifetimes.R")
lifetimes <- import_lifetimes(
  data = "SOC20140225.csv",
  first_control = as.Date("2007-01-01"),
  close_info = as.Date("2014-02-01"),
  lag = 30)
# length(unique(lifetimes$id))  # 145742 unique contracts (one row by each "id")

# ===============================================================================
# 3) Merging the claims file and the lifetimes information with two conditions:
#    A) Each subject has, at least, 65 years when entering the study in any of the
#       eight starting points. These points are placed on 1st January days from
#       year 2007 to year 2014.
#    B) Each subject's follow-up interval, ["dent", "dfinal"] within the study
#       period entails longitudinal information if the subject's profile crosses,
#       at least, one starting point (regardless of whether the value of emergency
#       claims per year is or not zero).
#    SCRIPT: fusion.R
# ===============================================================================

source("fusion.R")

all_ini <- fusion(
  dtable1 = claims,
  dtable2 = lifetimes,
  age.u = 65,
  lag = 30,
  open_info = as.Date("2006-01-01"),
  close_info = as.Date("2014-02-01"))
# length(unique(all_ini$id))     # 5496 unique subjects
# nrow(all_ini)                  # 531580 measurements

# ======================================================================
# 4) Select those claims associated to the following medical topics:
#    Neutral codes (associated to subjects who register zero counts),
#    ambulance services, hospitalization, and non-routine medical visits.
#    SCRIPT: clean.R
# ======================================================================

source("clean.R")
all <- clean(dat = all_ini, keepfam = "all")
# length(unique(all$id))        # 5496 unique subjects
# nrow(all)                     # 31170 measurements

# ================================================================================
# 5) Annual mesh for the longitudinal follow-up of subjects within the study period
#    [2006-01-01, 2014-02-01]. Each subject enters the study in any of the called
#    starting points, and remains in until the corresponding "dfinal". Thus, the
#    subjects provide both longitudinal and time-to-event information between the
#    corresponding dates associated to their entry into the study and exit of it,
#    that is, in the interval ["dent", "dfinal"]. We finally obtain the HI dataset.
#    SCRIPT: mesh.R
# ================================================================================

source("mesh.R")
```

```
tot <- mesh(all)
# length(unique(tot$id))      # 5470 subjects
# nrow(all)                   # 32269 measurements of emergency claims per year

# ========================================
# HI dataset with longitudinal information
# ========================================

HIdata <- tot
# length(unique(HIdata$id))   # 5470 subjects
# nrow(HIdata)                # 32269 measurements of emergency claims per year

# =========================================
# Time-to-event file with final information
# =========================================

HIdata.id <- HIdata[, .SD[(.N)], by = id][,
  list(id, sex, start = age_dent, stop, status)]
# nrow(HIdata.id)             # 5470 subjects
# sum(HIdata.id[status == 1]) #  509 death events
```

Appendix B

JAGS CODE TO FIT THE STANDARD JM

```
# ##############################################################################
# This code is used to fit the standard joint model which considers a constant
# relationship between the expected current value in the longitudinal response
# (with normal distribution), and left-truncated time-to-event data. The code
# is an adaptation of that contained in the R package JMbayes.
# ##############################################################################

# Simultaneously loading of the R packages which are needed
# ---------------------------------------------------------
lapply(c("data.table", "survival", "splines", "lme4", "jagsUI", "MASS"),
  require, character.only = T)

# The HI Dataset
# --------------
length(unique(HIdata$id))   # 5470 subjects
nrow(HIdata)                # 32269 observations
sum(HIdata$event)           # 509 events

# Definitions regarding the longitudinal process
# ---------------------------------------------
dataL <- HIdata
lmeObject <-                                      # "lme4" package
  lmer(log(1 + claimyr) ~ offset(log(expo)) + obstime + (1 | id), data = dataL)
id <- as.integer(transform(dataL, id = as.numeric(factor(id)))$id)
n <- length(unique(id))
offset <- as.vector(c(1, 1 + cumsum(tapply(id, id, length))))
timeVar <- "obstime"
times <- as.vector(dataL[[timeVar]])

# Definitions regarding the survival process:
# Left-truncated and right-censored event times
# --------------------------------------------
survObject <- coxph(Surv(start, stop, event) ~ sex,
  data = HIdata, x = TRUE, model = TRUE)
W <- survObject$x                                 # Baseline covariates part
SurvInf <- survObject$y
typeSurvInf <- attr(SurvInf, "type")

# Only right-censored event times
if (typeSurvInf == "right") {
    Time <- SurvInf[, "time"]
    Time[Time < 1e-04] <- 1e-04
    nT <- length(Time)
    event <- SurvInf[, "status"]
    LongFormat <- FALSE
    }
```

```
# Left-truncated and right-censored event times
if (typeSurvInf == "counting") {              # TRUE
    idT <- as.vector(unclass(survObject$model$cluster))
    strata <- seq_len(nrow(survObject$model))
    idT <- dataL$id
    idT <- match(idT, unique(idT))
    LongFormat <- length(idT) > length(unique(idT))
    TimeL <- SurvInf[, "start"]
    TimeL <- tapply(TimeL, idT, head, n = 1)
    anyLeftTrunc <- any(TimeL > 1e-07)
    TimeR <- SurvInf[, "stop"]
    TimeR[TimeR < 1e-04] <- 1e-04
    Time <- tapply(TimeR, idT, tail, n = 1)
    nT <- length(Time)
    eventLong <- SurvInf[, "status"]
    event <- tapply(eventLong, idT, tail, n = 1)
    }

# Information regarding the K control points used in the Gauss-Kronrod
# quadrature rule to approximate the integral in the hazard function,
# which does not have a closed-form solution
# --------------------------------------------------------------------
# G-K points
sk <- c(-0.949107912342758524526189684047851, -0.741531185599394439863864773280788,
       -0.405845151377397166906606412076961, 0,
       0.405845151377397166906606412076961, 0.741531185599394439863864773280788,
       0.949107912342758524526189684047851, -0.991455371120812639206854697526329,
       -0.864864423359769072789712788640926, -0.586087254867691130294144838258730,
       -0.207784955007898467600689403773245, 0.207784955007898467600689403773245,
       0.586087254867691130294144838258730, 0.864864423359769072789712788640926,
       0.991455371120812639206854697526329)

# G-K weights
wk <- c(0.063092092629978553290700663189204, 0.140653259715525918745189590510238,
       0.190350578064785409913256402421014, 0.209482141084727828012999174891714,
       0.190350578064785409913256402421014, 0.140653259715525918745189590510238,
       0.063092092629978553290700663189204, 0.022935322010529224963732008058970,
       0.104790010322250183839876322541518, 0.169004726639267902826583426598550,
       0.204432940075298892414161999234649, 0.204432940075298892414161999234649,
       0.169004726639267902826583426598550, 0.104790010322250183839876322541518,
       0.022935322010529224963732008058970)

K <- length(sk)                           # Points in G-K quadrature: K = 15
P <- if (typeSurvInf == "counting" && anyLeftTrunc) {
       (Time - TimeL)/2
    } else Time/2
st <- if (typeSurvInf == "counting" && anyLeftTrunc) {
       outer(P, sk) + c(Time + TimeL)/2
     } else outer(P, sk + 1)
id.GK <- rep(seq_len(nT), each = K)

# Design matrices of fixed effects (X, XT and Xs). We do not consider design matrices
# of random effects since only random intercepts are accounted for in our model
# -----------------------------------------------------------------------------------
formYx <- log(1 + claimyr) ~ offset(log(expo)) + obstime
```

```
mfX <- model.frame(terms(formYx), data = dataL)
TermsX <- attr(mfX, "terms")
X <- model.matrix(formYx, mfX)                    # Design matrix fixed-effects
v.offset <- model.offset(mfX)                     # The offset variable is log(expo)
expo <- exp(v.offset)                             # The exposure time
y.long <- model.response(mfX, "numeric")

dataL.id <- dataL[!duplicated(id)]
dataL.id[[timeVar]] <- pmax(Time, 0)              # Survival time for each subject
mfX.id <- model.frame(TermsX, data = dataL.id)
XT <- model.matrix(formYx, mfX.id)                # Design matrix fixed-effects in Part I
av.expo.T <- unique(dataL[, av := mean(expo),
  by = id], by = "id")$av                         # Average exposure per subject in part I

dataL.id2 <- dataL.id[id.GK, ]
dataL.id2[[timeVar]] <- pmax(c(t(st)), 0)         # K knots from lower start to higher stop
mfX.id2 <- model.frame(TermsX, data = dataL.id2)
Xs <- model.matrix(formYx, mfX.id2)               # Design matrix fixed-effects in Part II
av.expo.s <- replicate(K, av.expo.T)              # Average exposure per subject in part II

# Obtention of a B-spline basis from the computation of truncated power functions
# of degree "p". R Code provided in Eilers and Marx (2010).
# ----------------------------------------------------------------------------------
tpower <- function(x, t, p) (x - t) ^ p * (x > t)

bbase <- function(x, xl, xr, ndx, deg) {
  dx <- (xr - xl) / ndx
  knots <- seq(xl - deg * dx, xr + deg * dx, by = dx) # Extended knots
  P <- outer(x, knots, tpower, deg)
  n <- dim(P)[2]
  D <- diff(diag(n), diff = deg + 1) / (gamma(deg + 1) * dx ^ deg)
  B <- (-1) ^ (deg + 1) * P %*% t(D)
  B }

# Extra design matrices to approximate the log-baseline hazard (W2T and W2s) with
# P-splines of degree dh0 = 2 and placing Qh0 = 8 equally-spaced knots on [tmin, tmax]
# --------------------------------------------------------------------------------------
W2T <-                                                             # Matrix 5470 x 9
  bbase(x = Time, xl = 0, xr = ceiling(max(times)), ndx = 7, deg = 2)
W2s <-                                                             # Matrix 82050 x 9
  bbase(x = c(t(st)), xl = 0, xr = ceiling(max(times)), ndx = 7, deg = 2)

# Fitting a time-dependent Cox model to get initial values for gam.w and alpha
# ------------------------------------------------------------------------------
DF <- data.frame(id = id, Time = Time[id], event = event[id])
Wdat <- as.data.frame(W)
DF <- cbind(DF, Wdat[id, ])
long <- unname(as.data.frame(c(X %*% fixef(lmeObject)) +  b[id, ]))
DF <- cbind(DF, long)
DF$start <- times
splitID <- split(DF[c("start", "Time")], DF$id)
DF$stop <-
  unlist(lapply(splitID, function (d) c(d$start[-1], d$Time[1])))
DF$event <-
  with(DF, ave(event, id, FUN = function (x) c(rep(0, length(x) - 1), x[1])))
```

```
DF <- DF[!names(DF) %in% c("Time", "id")]
tdCox <- coxph(Surv(start, stop, event) ˜ ., data = DF[DF$stop > DF$start, ])
init.gam.w <- unname(coef(tdCox)[1])          # Initial value for the baseline covariate
init.alpha <- unname(coef(tdCox)[2])          # Initial value for the association parameter


# Prior penalty matrices (order k = 2) for coefficients of vector gam.h0
# ----------------------------------------------------------------------
DD <- diag(ncol(W2T))
M.gam.h0 <- crossprod(diff(DD, differences = 2)) + 1e-06 * DD

# ======================================================
# JAGS CODE FOR FITTING A STANDARD JM TO THE HI DATASET
# ======================================================

writeLines("                           # The JAGS code is stored in an external file

model {

 for(i in 1:n) {

   # Longitudinal process
   # --------------------
   for (j in offset[i]:(offset[i + 1] - 1)) {
     y[j] ∼ dnorm(mu[j], tau.e)
     mu[j] <- log(expo[j]) + inprod(beta[1:ncX], X[j, 1:ncX]) + b0[i]
   }

   # Survival process, part I (time T)
   # ---------------------------------
   log.h0T[i] <- inprod(gam.h0[1:ncW2T], W2T[i, 1:ncW2T])      # Baseline log-hazard
   eta.W[i] <- inprod(gam.w[1:ncW], W[i, 1])                   # Baseline covariates
   mu.T[i] <- log(av.expo.T[i]) +
     inprod(beta[1:ncX], XT[i, 1:ncX]) + b0[i]                 # Expected long. outcome
   log.hazard.T[i] <- log.h0T[i] + eta.W[i] + alpha * mu.T[i]  # Log-hazard

   # Survival process, part II (time s): Loop over the K = 15
   # control points of the Gauss-Kronrod quadrature
   # ---------------------------------------------------------
   for (k in 1:K) {
     log.h0s[i, k] <- inprod(gam.h0[1:ncW2s],
       W2s[K * (i - 1) + k, 1:ncW2s])                          # Baseline log-hazard
     # Ws is a null matrix (constant covariates)
     # eta.Ws[i, k] <- 0
     mu.s[i, k] <- log(av.expo.s[i, k]) + inprod(beta[1:ncX],
       Xs[K * (i - 1) + k, 1:ncX]) + b0[i]                     # Expected long. outcome
     SurvLong[i, k] <- wk[k] * exp(log.h0s[i, k] +             # Survival integrand
       alpha * mu.s[i, k])
   }
   approxIntegral[i] <- P[i] * sum(SurvLong[i, ])
   log.Survival[i] <- - exp(eta.W[i]) * approxIntegral[i]

   # Zeros trick to work with the likelihood of the JM without specifying the function
   # ---------------------------------------------------------------------------------
   zeros[i] ∼ dpois(phi[i])
   log.Lik[i] <- (event[i] * log.hazard.T[i]) + log.Survival[i]
   phi[i] <- C - log.Lik[i]

   # Random effects part
   # -------------------
   b0[i] ~ dnorm(0, tau.b0)                                  # Random intercept

 }

 # Priors for the fixed effects of longitudinal submodel
 # -----------------------------------------------------
 beta[1] ~ dnorm(0, 0.001)                                  # Fixed effect for intercept
 beta[2] ~ dnorm(0, 0.001)                                  # Fixed effect for slope

 # Prior for the sd of the perturbation term in longitudinal submodel
 # ------------------------------------------------------------------
  tau.e <- 1 / (sig.e * sig.e)
   sig.e ~ dunif(0, 50)

 # Prior for the sd of the random intercepts
 # -----------------------------------------
 tau.b0 <- 1 / (sig.b0 * sig.b0)
   sig.b0 ~ dunif(0, 50)

 # Prior for the coefficient of baseline survival covariate (gender)
 # -----------------------------------------------------------------
 for (p in 1:ncW) {
      gam.w[p] ~ dnorm(0, 0.001)
 }

 # Priors for the coefficients of the penalized B-splines
 # to approximate the logarithm of baseline risk function
 # ------------------------------------------------------
 gam.h0[1:ncW2T] ~ dmnorm(priorMean.gam.h0[],
                           tau.gam.h0 * M.gam.h0[, ])
 tau.gam.h0 ~ dgamma(a.gam.h0, b.gam.h0)                     # Smoothing parameter

 # Prior for the constant parameter of association
 # -----------------------------------------------
 alpha ~ dnorm(0, 0.001)

}", con = "JM.LOG.txt")

# =============================================================================
# Run two MCMC chains of 25000 iterations each one (burn-in period is included)
# =============================================================================

# Bundled data
# ------------
data.JM.LOG <- list(
  n = n,
  y = log(1 + dataL$claimyr),
  zeros = rep(0, nrow(dataL.id)),
  C = 100000, K = length(wk), P = P, wk = wk, offset = offset,
  X = X, XT = XT, Xs = Xs, ncX = ncol(X),
  W = W, W2T = W2T, W2s = W2s,
  ncW = ncol(W), ncW2T = ncol(W2T), ncW2s = ncol(W2s),
  expo = expo, av.expo.T = av.expo.T, av.expo.s = av.expo.s,
  event = event,
  priorMean.gam.h0 = seq(-8.5, 0, len = ncol(W2s)),
  M.gam.h0 = M.gam.h0, a.gam.h0 = 1, b.gam.h0 = 0.005)

# Initial values for the parameters to estimate
# ---------------------------------------------
inits.JM.LOG <- function() {list(
  beta = as.vector(fixef(lmeObject)),
  sig.b0 = as.data.frame(VarCorr(lmeObject))[1, 5],
  sig.e = as.data.frame(VarCorr(lmeObject))[2, 5],
  gam.w = init.gam.w,
  gam.h0 = numeric(ncol(W2s)),
  alpha = init.alpha)}

# Gibbs sampling with JAGS
# ------------------------
JM.LOG <- jags(
  data = data.JM.LOG,
  inits = inits.JM.LOG,
  parameters.to.save =
    c("beta", "b0", "gam.h0", "gam.w", "sig.b0", "sig.e", "alpha"),
  model.file = "JM.LOG.txt", parallel = TRUE,
  n.thin = 25, n.chains = 2, n.burnin = 5000, n.iter = 30000, n.adapt = 5000)

print(JM.LOG, digits = 3)
```

Bibliography

Abrahamowicz, M., Beauchamp, M.-E., and Sylvestre, M.-P. (2011). Comparison of Alternative Models for Linking Drug Exposure with Adverse Effects. *Statistics in Medicine*, 31(11–12):1014–1030.

Agresti, A. (2012). *Categorical Data Analysis, 2nd Edition*. Wiley, New York (USA).

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

Andersen, P. K. and Gill, R. D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*, 10(4):1100–1120.

Andrinopoulou, E.-R., Rizopoulos, D., Takkenberg, J. J. M., and Lesaffre, E. (2014). Joint Modeling of Two Longitudinal Outcomes and Competing Risk Data. *Statistics in Medicine*, 33(18):3167–3178.

Andrinopoulou, E.-R., Eilers, P. H. C., Takkenberg, J. J. M., and Rizopoulos, D. (2016). Improved Dynamic Predictions from Joint Models of Longitudinal and Survival Data with Time-Varying Effects Using P-splines. Project.

Booth, J. G., Casella, G., Friedl, H., and Hobert, J. P. (2003). Negative Binomial Loglinear Mixed Models. *Statistics in Modelling*, 3(3):179–191.

Boucher, J.-P., Denuit, M., and Guillén, M. (2007). Risk Classification for Claim Counts: A Comparative Analysis of Various Zero-Inflated Mixed Poisson and Hurdle Models. *North American Actuarial Journal*, 11(4):110–131.

Boucher, J.-P., Denuit, M., and Guillén, M. (2008). Models of Insurance Claim Counts with Time Dependence Based on Generalization of Poisson and Negative Binomial Distributions. *Variance*, 2(1):135–162.

Breslow, N. and Crowley, J. (1974). A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship. *The Annals of Statistics*, 2(3):437–453.

Brown, E. and Ibrahim, J. G. (2003). A Bayesian Semiparametric Joint Hierarchical Model for Longitudinal and Survival Data. *Biometrics*, 59(2):221–228.

Brown, E., Ibrahim, J. G., and DeGruttola, V. (2005). A Flexible B-Spline Model for Multiple Longitudinal Biomarkers and Survival. *Biometrics*, 61(1):64–73.

Bull, K. and Spiegelhalter, D. J. (1997). Survival Analysis in Observational Studies. *Statistics* *in Medicine*, 16(9):1041–1074.

Cai, Z. (1998). Asymptotic Properties of Kaplan-Meier Estimator for Censored Dependent Data. *Statistics and Probability Letters*, 37(4):381–389.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press, Cambridge (UK).

Charpentier, A. (2015). *Computational Actuarial Science with R*. Chapman and Hall/CRC The R Series, Boca Raton, Florida (USA).

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society, Series B*, 34(2):187–220.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London (UK).

Cox, D. R. (1975). Partial Likelihood. *Biometrika*, 62(2):269–276.

Crowther, M. J., Andersson, T. M.-L., Lambert, P. C., Abrams, K. R., and Humphreys, K. (2016). Joint Modelling of Longitudinal and Survival Data: Incorporating Delayed Entry and an Assessment of Model Misspecification. *Statistics in Medicine*, 35(7):1193–1209.

D'Amico, G., Guillén, M., and Manca, R. (2009). Full Backward Non-Homogeneous Semi-Markov Processes for Disability Insurance Models: A Catalunya Real Data Application. *Insurance: Mathematics and Economics*, 45(2):173–179.

Dao, H., Godbout, L., and Fortin, P. (2014). On the Importance of Taking End-of-Life Expenditures into Account when Projecting Health-Care Spending. *Canada Public Policy*, 40(1):45–56.

Ding, J. and Wang, J.-L. (2008). Modeling Longitudinal Data with Nonparametric Multiplicative Random Effects Jointly with Survival Data. *Biometrics*, 64(2):546–556.

Dowle, M., Srinivasan, A., Gorecki, J., Short, T., Lianoglou, S., and Antonyan, E. (2017). *data.table: Extension of Data.frame.*

Eilers, P. H. C. and Marx, B. D. (1996). Flexible Smoothing with B-Splines and Penalties. *Statistical Science*, 11(2):89–121.

Elashoff, R. M., Li, G., and Li, N. (2008). A Joint Model for Longitudinal Measurements and Survival Data in the Presence of Multiple Failure Types. *Biometrics*, 64(3):762–771.

Ericson, K. M. and Starc, A. (2015). Measuring Consumer Valuation of Limited Provider Networks. *American Economic Review*, 105(5):115–119.

Eurostat (2016). European Statistical System report 2016. Technical report, European Comission.

Fabbri, E., Zoli, M., Gonzalez-Freire, M., Salive, M. E., Studenski, S. A., and Ferrucci, L. (2015). Aging and Multimorbidity: New Tasks, Priorities, and Frontiers for Integrated Gerontological and Clinical Research. *Journal of the American Medical Directors Association*, 16(8):640–647.

Fang, R., Wagner, B. D., Harris, J. K., and Fillon, S. A. (2016). Zero-Inflated Negative Binomial Mixed Model: An Application to Two Microbial Organisms Important in Oesophagitis. *Epidemiology and Infection*, 144(11):2447–2455.

Faucett, C. L. and Thomas, D. C. (1996). Simultaneously Modelling Censored Survival Data and Repeatedly Measured Covariates: A Gibbs Sampling Approach. *Statistics in Medicine*, 15(15):1663–1685.

Fieuws, S., Verbeke, G., Maes, B., and Vanrenterghem, Y. (2008). Predicting Renal Graft Failure Using Multivariate Longitudinal Profiles. *Biostatistics*, 9(3):419–431.

Fitzmaurice, G. M., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). *Longitudinal Data Analysis*. CRC Biostatistics Series, Chapman and Hall, Boca Raton, Florida (USA).

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York (USA).

Frees, E. W. (2010). *Regression Modeling with Actuarial and Financial Applications*. Cambridge University Press, New York (USA).

Gail, M. H., Graubard, B., Williamson, D. F., and Flegal, K. M. (2009). Comment on Choice of Time Scale and its Effect on Significance of Predictors in Longitudinal Studies. *Statistics in Medicine*, 28(8):1315–1317.

Garin, N., Olaya, B., Perales, J., Moneta, M.-V., Miret, M., Ayuso-Mateos, J.-L., and Haro, J.-M. (2014). Multimorbidity Patterns in a National Representative Sample of the Spanish Adult Population. *PLoS ONE*, 9(1).

Gill, R. D. (1983). Large Sample Behaviour of the Product-Limit Estimator on the Whole Line. *The Annals of Statistics*, 11(1):49–58.

Gould, A. L., Boye, M. E., Crowther, M. J., Ibrahim, J. G., Quartey, G., Micallef, S., and Bois, F. Y. (2015). Joint Modeling of Survival and Longitudinal Non-Survival Data: Current Methods and Issues. Report of the DIA Bayesian Joint Modeling Working Group. *Statistics in Medicine*, 34(14):2181–2195.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). Pseudo Maximum Likelihood Methods: Theory. *Econometrica*, 52(3):681–700.

Greene, W. H. (1994). Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. Technical report, Department of Economics, Stern School of Business, New York University, New York (USA).

Greene, W. H. (2008). Functional Forms for the Negative Binomial Model for Count Data. *Economics Letters*, 99(3):585–590.

Guillén, M. and Pinquet, J. (2008). Long-Term Care: Risk Description of a Spanish Portfolio and Economic Analysis of the Timing of Insurance Purchase. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 33(4):659–672.

Guillén, M., Prieto, F., and Sarabia, J.-M. (2011). Modelling Losses and Locating the Tail with the Pareto Positive Stable Distribution. *Insurance: Mathematics and Economics*, 49(3):454–461.

Guillén, M., Nielsen, J., Scheike, T., and Pérez-Marín, A.-M. (2012). Time-Varying Effects in the Analysis of Customer Loyalty: A Case Study in Insurance. *Expert Systems with Applications*, 39(3):3551–3558.

Harrison, X. (2014). Using Observation-Level Random Effects to Model Overdispersion in Count Data in Ecology and Evolution. *PeerJ*, 2:e616.

Hausman, J. A., Hall, B., and Griliches, Z. (1984). Econometric Models For Count Data with an Application to the Patents-R&D Relationship. *Econometrica*, 52(4):909–938.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint Modelling of Longitudinal Measurements and Event Time Data. *Biostatistics*, 1(4):465–480.

Hilbe, J. M. (2011). *Negative Binomial Regression, 2nd Edition.* Cambridge University Press, Cambridge (UK).

Hinde, J. and Demétrio, C. G. B. (1998). Overdispersion: Models and Estimation. *Computational Statistics and Data Analysis*, 27(2):151–170.

Hsieh, F., Tseng, Y.-K., and Wang, J.-L. (2006). Joint Modeling of Survival and Longitudinal Data: Likelihood Approach Revisited. *Biometrics*, 62(4):1037–1043.

Huang, X., Li, G., and Elashoff, R. M. (2011). A Joint Model of Longitudinal and Competing Risks Survival Data with Heterogeneous Random Effects and Outlying Longitudinal Measurements. *Statistics and Its Interface*, 3(2):185–195.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2004). Bayesian Methods for Joint Modeling of Longitudinal and Survival Data with Applications to Cancer Vaccine Trials. *Statistica Sinica*, 14(3):863–883.

Ismail, I. and Jemain, A. (2007). Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models. In *Casualty Actuarial Society Forum, Winter 2007*, pages 103–158.

Ivanova, A., Molenberghs, G., and Verbeke, G. (2016). Mixed Models Approaches for Joint Modeling of Different Types of Responses. *Journal of Biopharmaceutical Statis-*

tics, 26(4):601–618.

Kalbfleisch, J. D. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data, 2nd Edition*, volume 360. Wiley, New York (USA).

Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation From Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kellner, K. (2016). *jagsUI: A Wrapper Around rjags to Streamline JAGS Analyses.*

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

Koller, D., Schon, G., Schafer, I., Glaeske, G., van den Bussche, H., and Hansen, H. (2014). Multimorbidity and Long-Term Care Dependency: A Five-Year Follow-up. *BMC Geriatrics*, 14(1):70.

Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, 38(4):963–974.

Lamarca, R., Alonso, J., Gómez, G., and Muñoz, A. (1998). Left-Truncated Data with Age as Time Scale: An Alternative for Survival Analysis in the Elderly Population. *Journal of Gerontology: Medical Sciences*, 53A(5):M337–M343.

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, 34(1):1–14.

Lang, S. and Brezger, A. (2004). Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212.

Lawless, J. F. (1987). Negative Binomial and Mixed Poisson Regression. *The Canadian Journal of Statistics*, 15(3):209–225.

Lesaffre, E. and Lawson, A. (2012). *Bayesian Biostatistics.* Wiley, West Sussex (UK).

Li, N., Elashoff, R. M., Li, G., and Saver, J. (2010). Joint Modeling of Longitudinal Ordinal Data and Competing Risks Survival Times and Analysis of the Ninds rt-PA Stroke Trial. *Statistics in Medicine*, 29(5):546–557.

Lin, H., Turnbull, B. W., McCulloch, C. E., and Slate, E. H. (2002). Latent Class Models for Joint Analysis of Longitudinal Biomarker and Event Process: Application to Longitudinal Prostate-Specific Antigen Readings and Prostate Cancer. *Journal of the American Statistical Association*, 97(457):53–65.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* Wiley, New York (USA).

Little, R. J. A. (1995). Modeling the Drop-Out Mechanism in Repeated-Measures Studies.

Journal of the American Statistical Association, 90(431):1112–1121.

Liu, H. and Powers, D. A. (2007). Growth Curve Models for Zero-Inflated Count Data: An Application to Smoking Behavior. *Structural Equation Modeling*, 14(2):247–279.

Min, Y. and Agresti, A. (2005). Random Effect Models for Repeated Measures of Zero-Inflated Count Data. *Statistical Modelling*, 5(1):1–19.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* Springer-Verlag, New York (USA).

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies.* Wiley, New York (USA).

Mukherji, A., Roychoudhury, S., Ghosh, P., and Brown, S. (2016). Estimating Health Demand for an Aging Population: A Flexible and Robust Bayesian Joint Model. *Journal of Applied Econometrics*, 31(6):1140–1158.

Murawska, M., Rizopoulos, D., and Lessaffre, E. (2012). A Two-Stage Joint Model for Non-linear Longitudinal Response and a Time-to-Event with Application in Transplantation Studies. *Journal of Probability and Statistics*, vol. in press, 2012:1–18.

Neelon, B. H., O'Malley, A. J., and Normand, S.-L. T. (2010). A Bayesian Model for Repeated Measures Zero-Inflated Count Data with Application to Outpatient Psychiatric Service Use. *Statistical Modelling*, 10(4):421–439.

OECD (2016). Health Statistics 2016.

Piccorelli, A. V. and Schluchter, M. D. (2012). Jointly Modeling the Relationship between Longitudinal and Survival Data Subject to Left Truncation with Applications to Cystic Fibrosis. *Statistics in Medicine*, 31(29):3931–3945.

Pitacco, E. (2014). *Health Insurance. Basic Actuarial Models.* Springer, Berlin.

Piulachs, X., Alemany, R., and Guillén, M. (2016). Joint Modelling of Survival and Emergency Medical Care Usage in Spanish Insureds Aged 65+. *PLoS ONE*, 11(4):1–11.

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing.* Technische Universität Wien, Vienna (Austria).

Prentice, R. L. (1982). Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model. *Biometrika*, 69(2):331–342.

Proust-Lima, C. and Taylor, J. M. G. (2009). Development and Validation of a Dynamic Prognostic Tool for Prostate Cancer Recurrence using Repeated Measures of Post-Treatment PSA: A joint Modeling Approach. *Biostatistics*, 10(3):535–549.

Proust-Lima, C., Joly, P., Dartigues, J.-F., and Jacqmin-Gadda, H. (2009). Joint Modelling of Multivariate Longitudinal Outcomes and a Time-to-event: A Nonlinear Latent Class Approach. *Computational Statistics and Data Analysis*, 53(4):1142–1154.

ProustLima, C., Dartigues, J.-F., and Jacqmin-Gadda, H. (2016). Joint Modelling of Repeated Multivariate Cognitive Measures and Competing Risks of Dementia and Death: A Latent Process and Latent Class Approach. *Statistics in Medicine*, 35(3):382–398.

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna (Austria).

Reinhardt, U. E. (2003). Does the Aging of the Population Really Drive the Demand for Health Care? *Health Affairs*, 22(6):27–39.

Rizopoulos, D. and Ghosh, P. (2011). A Bayesian Semiparametric Multivariate Joint Model for Multiple Longitudinal Outcomes and a Time-to-Event. *Statistics in Medicine*, 30(12):1366–1380.

Rizopoulos, D. (2011). Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-event Data. *Biometrics*, 67(3):819–829.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data with Applications in R.* CRC Biostatistics Series, Chapman and Hall, Boca Raton, Florida (USA).

Rizopoulos, D. (2016). The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-event Data Using MCMC. *Journal of Statistical Software*, 72(7):1–45.

Rodriguez, M. and Stiyanova, A. (2004). The Effect of Private Insurance Access on the Choice of GP/Specialist and Public/Private Provider in Spain. *Health Economics*, 13(7):689–703.

Salisbury, C., Johnson, C., Purdy, S., Valderas, J. M., and Montgomery, A. A. (2011). Epidemiology and Impact of Multimorbidity in Primary Care: A Retrospective Cohort Study. *British Journal of General Practice*, 61(582):12–21.

Sarabia, J.-M. and Guillén, M. (2008). Joint Modelling of the Total Amount and the Number of Claims by Conditionals. *Insurance: Mathematics and Economics*, 43(3):466–473.

Shi, P. and Valdez, E. A. (2014). Longitudinal Modeling of Insurance Claim Counts Using Jitters. *Scandinavian Actuarial Journal*, 2014(2):159–179.

Song, X., Davidian, M., and Tsiatis, A. A. (2002). A Semiparametric Likelihood Approach to Joint Modeling of Longitudinal and Time-to-Event Data. *Biometrics*, 58(4):742–753.

Song, X. and Wang, C. Y. (2008). Semiparametric Approaches for Joint Modeling of Longitudinal and Survival Data with Time-Varying Coefficients. *Biometrics*, 64(2):557–566.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society, Series B*, 64(4):583–639.

Stute, W. and Wang, J.-L. (1993). The Strong Law under Random Censorship. *The Annals of Statistics*, 21(3):1591–1607.

Su, Y. and Wang, J.-L. (2012). Modeling Left-Truncated and Right-Censored Survival Data with Longitudinal Covariates. *The Annals of Statistics*, 40(3):1465–1488.

Sweeting, M. J. and Thompson, S. G. (2011). Joint Modelling of Longitudinal and Time-to-Event Data with Application to Predicting Abdominal Aortic Aneurysm Growth and Rupture. *Biometrical Journal*, 53(5):750–763.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox model*. Springer-Verlag, New York (USA).

Thiébaut, A. C. and Bénichou, J. (2004). Choice of Time-Scale in Cox's Model Analysis of Epidemiologic Cohort Data: A Simulation Study. *Statistics in Medicine*, 23(24):3803 – 3820.

Tsiatis, A. A., Degruttola, V., and Wulfsohn, M. S. (1995). Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS. *Journal of the American Statistical Association*, 90(429):27–37.

Tsiatis, A. A. and Davidian, M. (2001). A Semiparametric Estimator for the Proportional Hazards Model with Longitudinal Covariates Measured with Error. *Biometrika*, 88(2):447–458.

Tsiatis, A. A. and Davidian, M. (2004). Joint Modeling of Longitudinal and Time-to-Event Data: An Overview. *Statistica Sinica*, 14(3):809–834.

Tyree, P. T., Lind, B. K., and Lafferty, W. E. (2006). Challenges of Using Medical Insurance Claims Data for Utilization Analysis. *American Journal of Medical Quality*, 21(4):269–275.

Uzunoḡullari, Ü. and Wang, J.-L. (1992). A Comparison of Hazard Rate Estimators for Left-Truncated and Right-Censored Data. *Biometrika*, 79(2):297 – 310.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York (USA).

Viviani, S., Alfò, M., and Rizopoulos, D. (2012). Generalized Linear Mixed Joint Model for Longitudinal and Survival Outcomes. *Statistics and Computing*, 24(3):417–427.

Wang, C. (2006). Corrected Score Estimator for Joint Modeling of Longitudinal and Failure Time Data. *Statistica Sinica*, 16:235–253.

Wang, Y. and Taylor, J. M. G. (2001). Jointly Modeling Longitudinal and Event Time Data with Application to Acquired Immunodeficiency Syndrome. *Journal of the American Statistical Association*, 96(455):895–905.

WHO (2011). Global Status Report on Non-Communicable Diseases 2010: Description of the Global Burden of NCDs, their Risk Factors and Determinants. Technical report, Geneva: World Health Organization.

Williamson, P. R., Kolamunnage-Dona, R., Philipson, P., and Marson, A. G. (2008). Joint Modelling of Longitudinal and Competing Risks Data. *Statistics in Medicine*, 27(30):6426–6438.

Winkelmann, R. (2008). *Econometric Analysis of Count Data, 5th Edition.* Springer Science & Business Media, Berlin (Germany).

Wu, M. C. and Carroll, R. J. (1988). Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process. *Biometrics*, 44(1):175–188.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A Joint Model for Survival and Longitudinal Data Measured with Error. *Biometrics*, 53(1):330–339.

Xu, J. and Zeger, S. L. (2001). Joint Analysis of Longitudinal Data Comprising Repeated Measures and Times to Events. *Journal of the Royal Statistical Society, Series C*, 50(3):375–387.

Yau, K. K. W. and Lee, A. H. (2001). Zero-Inflated Poisson Regression with Random Effects to Evaluate an Occupational Injury Prevention Programme. *Statistics in Medicine*, 20(19):2907–2920.

Yau, K. K. W., Wang, K., and Lee, A. H. (2003). Zero-Inflated Negative Binomial Mixed Regression Modeling of Over-Dispersed Count Data with Extra Zeros. *Biometrical Journal*, 45(4):437–452.

Yu, M., Law, N. J., Taylor, J. M. G., and Sandler, H. M. (2004). Joint Longitudinal-Survival-Cure Models and their Application to Prostate Cancer. *Statistica Sinica*, 14(3):835–862.

Yu, M., Taylor, J. M. G., and Sandler, H. M. (2008). Individual Prediction in Prostate Cancer Studies using a Joint Longitudinal Survival-Cure Model. *Journal of the American Statistical Association*, 103(481):178–187.

Yue, C.-S. J. and Huang, H.-C. (2011). A Study of Incidence Experience for Taiwan Life Insurance. *The Geneva Papers on Risk and Insurance Issues and Practice*, 36(4):718–733.

Zeng, D. and Cai, J. (2005). Asymptotic Results for Maximum Likelihood Estimators in Joint Analysis of Repeated Measurements and Survival Time. *The Annals of Statistics*, 33(5):2132–2163.

Zuur, A., Ieno, E., Walker, N., Saveliev, A., and Smith, G. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer-Verlag, New York (USA).