Riskcenter
The research group on Risk in insurance and Finance is attached to the Institute of Applied Economics IREA-UB

**UB Riskcenter** Working Paper Series

University of Barcelona

**Research Group on Risk in Insurance and Finance** www.ub.edu/riskcenter

Working paper 2014/07 \\ Number of pages 19

# A joint longitudinal and survival model with health care usage for insured elderly

Xavier Piulachs, Ramon Alemany and Montserrat Guillén

Riskcenter

# A joint longitudinal and survival model with health care usage for insured elderly

Xavier Piulachs*, Ramon Alemany and Montserrat Guillén
Department of Econometrics, Riskcenter-IREA
University of Barcelona

August 11, 2014

Draft version

**Abstract**

We study longevity and usage of medical resources of a sample of individuals aged 65 years or more who are covered by a private insurance policy. A longitudinal analysis is presented, where the yearly cumulative number of medical coverage requests by each subject characterizes insurance intensity of care until death. We confirm that there is a significant correlation between the longitudinal data on usage level and the survival time processes. We obtain dynamic estimations of event probabilities and we exploit the potential of joint models for personalized survival curve adjustment.

## 1 Introduction

The gradual development of medical science and technology leads to a larger number of years lived with disabilities, which in turn increases the demand of medical resources. This is a key challenge for health insurance companies

*The support received from the Spanish Ministry of Science/FEDER ECO2013-48326-C2-1-P is acknowledged. Guillén thanks ICREA Academia.

who have to face additional costs in order to meet the care needs in the event of a large cohort of elderly people. Furthermore, it is well known that private insurance policy holders are generally supposed to have a higher socio-economic level compared to the rest of the population because they can afford private health coverage (Schoen et al., 2010). So, the mortality tables of the general population may be biased for the insureds subgroup and insurance companies estimate specific survival probabilities for their portfolios using standard actuarial methods (Yue and Huang, 2011; Denuit, 2009; Denuit and Frostig, 2008). In practice, however, they disregard longitudinal information on their policy holders that is continuously being collected. Health insurance companies accumulate data on the intensity and the type of use of medical resources which can be extremely valuable to predict personalized survival probabilities and to quantify the risk of medical care demand of their clients above the expected values.

The aim of our study is to show how historical and follow-up records, which are in fact repeated measures of a longitudinal marker that counts the number of times that the policy holder has used the insurance policy coverage, can effectively predict personalized survival probabilities. Our proposed joint modeling approach, which is a powerful methodology that has recently been introduced in statistics for medicine (Rizopoulos and Lesaffre, 2014, see), allows to examine the association between a given medical care usage trend and longevity prospects.

In is well known that medical usage intensity increases substantially at older ages (Blane et al., 2008) and end-of-life care expenditures is significantly larger than throughout life (Dao et al., 2014; Murphy, 2012), but according to Bird et al. (2002) men's and women's health care experiences differ as

they age. While increasing attention has been focused on gender differences in health status, prevalence of illnesses, and access to quality care among older adults, little is known about differences in their health care in the last years of their lives. This is precisely what we study.

The dynamic personalized predictions that we are aiming at are based on both baseline subject's time-to-event covariates, recorded at the start of the study, and subject's longitudinal information measured at fixed time points within an observation window. Therefore, both the longitudinal and the survival information is part of a single statistical model, which allows :(i) to establish the degree of association between the value of the longitudinal variable and the time to event outcome, (ii) to estimate subject-specific survival probabilities based on personalized longitudinal outcomes and (iii) to update personalized survival estimations as additional longitudinal responses are collected. This can provide a comprehensive risk assessment of a health insurance portfolio using all available information.

To the best of our knowledge, no study has evaluated how health care usage and risk of death can be modeled jointly. The main results of our analysis are: 1) we confirm age and gender are main factors influencing changes in survival for a health insurance member, 2) we find evidence of a significant association between serial measurements of cumulative private insurance care usage and longevity, and 3) we obtain dynamic estimations of event probabilities by exploiting the potential of joint models. In summary, our contribution shows that an increase in health care usage intensity is negatively associated with survival, but that its influence varies as usage accumulates and depending on other factors such as sex and age, as well as previous insurance conditions.

## 2 Data and methods

The motivating dataset corresponds to the information provided by a Spanish private health insurance mutual company, containing historical data which started being collected on January 1st, 2006 and ended on February 1st, 2014. In particular, our study is limited to 39,399 insurance policy holders (39.8% men and 60.2% women) who had reached the age of 65 before the observation period started.

Table 2 presents the definition of the variables that are used in the analysis. Two variables are central in our study. First, the longitudinal process which counts the number of times that the health insurance company has provided a service to the policy holder. The unit service can be a variety of possible coverage functions such as a doctor visit, a blood or an X-ray test, a prescribed therapy, a hospital stay and any other treatment that is established in the insurance contract. We do not distinguish between different types of services at this stage, but obviously the accumulated number of unit services provided over time to a given patient is strongly correlated with her health condition. Due to the right skewed shape exhibited by the longitudinal outcome, a logarithmic scale is applied (Verbeke and Molenberghs, 2009). Second, we also consider the survival time, where the event of interest is death. Information is censored because the majority of individuals survive beyond the end of the study period. Some other cancel their insurance policy and therefore they quit the study automatically. These dropouts are considered random, as they are generally due to personal reasons such as the decision not to renew the policy, or a change of company.

In one part of our study, variables such as gender, age and the cumulative

Table 1: Variables in the private insurer data set (2006-2014)

| Variable name | Definition |
|---|---|
| *ID* | Subject identifier: $i = 1, 2, \ldots, 30580$ |
| *SEX* | Gender of the subject: 0 = Male, 1 = Female |
| *OBSTIME* | Age (years) in excess of 65 at each time point |
| *CUM0* | Cumulative number of private health service usage units over the four years previous to entering the sample |
| *CUM* | Cumulative number of private health service usage units at each observation time point |
| *TIME* | Final observation time (years), which may correspond to an event(death) or to a right-censored data. |
| *CENS* | Censoring indicator: 0 = Right-censored, 1 = Event |

A private health service usage unit is a visit to a GP or a specialist, a hospital spell, a medical test and so on

number of medical care service units play the role of baseline covariates and become part of a first survival analysis. In another part, a longitudinal analysis is presented, where the cumulative number of medical visits observed in annual periods for each subject characterizes insurance intensity of use until death. Finally, both models are considered jointly, thus establishing an association parameter between the longitudinal and the survival processes. The application of joint modeling techniques allows to determine whether a pronounced increase in the cumulative number of coverage usage units also implies a simultaneous increased risk of death for the subject. The simultaneity is a fundamental part of joint modeling. When the two processes are endogenously determined, they cannot be modeled separately.

The added value of joint models has been empirically illustrated by Fieuws et al. (2008) who noted that predictions of failure in a kidney transplant study based on a joint model using all recorded biomarkers of kidney functioning substantially outperformed the separate analyses per marker. In addition,

in a similar context, Rizopoulos (2011) and Proust-Lima and Taylor (2009b) showed that joint models can also be aimed to dynamically update predictions of survival probabilities and help in discriminating between patients who have a high risk of experiencing the event of interest (e.g. death) in relatively short time interval, from patients whose risk is rather minimal.

Let $y_i(t) = log.CUM_i(t) = \log\{1 + CUM_i(t)\}$ be the response variable of the $i$-th subject, $i = 1, \ldots, n$, observed at time $t$, where $n$ is the total number of observed individuals in the sample. The outcome is linearly related to a set of $p$ explanatory covariates and $q$ random effects. In our application the first response that is modeled is the number of private health usage service units after a logarithmic transformation.

In addition, let $\boldsymbol{m_i(t)}$ denote the true underlying value of the longitudinal outcome, and $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s \leq t\}$ the complete longitudinal history. The joint modeling approach consists in defining: (i) a model for the marker trajectory, usually a mixed model, (ii) a model for the time-to-event, usually a proportional hazard model, and (iii) linking both models using a shared latent structure (Rizopoulos, 2011).

## 2.1 Longitudinal submodel: Random intercept model

The main goal of linear mixed effects models is to account for the special features of serial evaluations of outcomes over time, thus being able to establish a plausible model in order to describe the particular evolution of each subject included in a longitudinal study. The particular features of these models are that they work with unbalanced datasets (unequal number of follow-up measurements between subjects and varying times between repeated measurements of each subject), and that they can explicitly take into

account that measurements from the same patient may be more correlated than measurements from different patients.

The model is specified as follows:

$$\begin{cases} log.CUM_i(t) = \boldsymbol{m_i(t)} + \varepsilon_i(t) = \boldsymbol{\beta_0 + b_{i0} + \beta_1 t} + \varepsilon_i(t) \\ \boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}} \\ b_{i0} \sim \mathcal{N}(0, \sigma^2_{b_0}) \\ \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

This model is straightforward. It just postulates that, besides individual random effects, a linear time trend governs the rate of increase of the number of accumulated service units provided to insurance policy holders. This seems plausible as we also expect that the older the policy holder the larger the rate at which the number of requested services increases.

## 2.2 Survival submodel: PH Cox Model

The celebrated proportional-hazards Cox model (Cox, 1972) allows to model the conditional hazard rate of survival times given certain baseline covariates. It relies on a fundamental assumption, the proportionality of the hazards, implying that the factors investigated have a constant impact on the risk over time. The model provides the conditional hazard function $h_i(t|\mathbf{w}_i)$ at time $t$ of a subject's profile given by a set of $p$ time-independent explanatory covariates called baseline covariates.

we assume that $T^*$, or $TIME$ in our data set, is a non-negative continuous random variable which represents the exact time until some specified event, which is death in our case. The survival model is specified through the hazard function as follows:

$$h_i(t|\mathbf{w}_i) = h_0(t) R_i(t) \exp\{\gamma log.CUM0_i\},$$

where $h_0(t)$ is an unspecified and non-negative baseline hazard function, representing the hazard function where $\mathbf{w}_i = \mathbf{0}$, $\psi(\mathbf{w}_i)$ is a non-negative function which contains the information about the set of explanatory time-independent covariates that define the $i-$th subject's profile.

This model is defined as a semiparametric because a parametric form is assumed only for the covariate effect, $\psi(\mathbf{w}_i)$. Among the possible parameterizations of function $\psi$, the most widely used is an exponential expression: $\psi(\mathbf{w}_i;\boldsymbol{\gamma}) = \exp(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{w}_i)$, where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)^{\mathrm{T}}$ is the parameter vector.

Survival analysis is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of a specific event of interest, usually designed by $\mathcal{E}$. This time is called *survival time*, time-to event or, simply, event time.

Bearing the above into account, what makes survival data special is that the responses correspond to time-durations and thus they are not measured in the same way as other variables. In practice, this fact has two important consequences, namely that the distribution of survival times is often highly left-skewed, and that the only information available about some subjects is that they have not yet experienced the event $\mathcal{E}$ at the last time point of follow-up, so these are termed censored or incomplete observations. In other words, it is unknown when these remaining subjects will experience the event. Considering these two special features, standard statistical methods can not be applied to survival data.

Although there are various categories of censoring, the present work has only focused on right-censoring mechanism which occurs when the subject has not yet experienced the event of interest at the time when the follow-up period ends. Consequently, all that is known about the true survival time is

that it exceeds the observed survival time, $t$, at the study end. Furthermore, we also assume that the censoring is non-informative. In this regard, it will be assumed that there are only two reasons why right-censoring might occur: The event of interest has not occurred by the end of the follow-up period (study end) or a subject is discontinued of follow-up during the study period due to causes unrelated to the event of interest.

The Cox model is often called a proportional hazards (PH) model because, two individuals $i$ and $i'$ with respective covariate values $\mathbf{w}_i$ and $\mathbf{w}_{i'}$, have a constant hazard rate ratio, so their corresponding hazard rates are proportional to each other and do not depend on time.

## 2.3 Joint model for longitudinal and survival data

To account for the fact that the longitudinal marker is an endogenous time-dependent covariate measured with error (Kalbfleisch and Prentice, 2002) with respect to survival, it is assumed that the risk for an event depends on the true and unobserved value of the endogenous variable at time $t$, denoted by $m_i(t)$. The endogeneity probles is quite intuitive here. There is a latent factor causing a health deterioration, which in turn implies an increase in the risk of death and more intensity of health care service use. So, survival and health care are strongly related to one another, through this latent factor.

The true underlying value of the longitudinal outcome, $m_i(t)$ must be estimated in order to successfully reconstruct the complete longitudinal history $\mathcal{M}_i(t)$. For this purpose, we utilize all available measurements on each subject $\{y_i(t_{ij}),\ \text{for } i = 1, \ldots, n \text{ and } j = 1, \ldots, n_i\}$ and postulate a suitable mixed effects model. We focus on normal data, describing the true subject-specific evolutions by a linear mixed effects model, but we agree that a count

data modeling approach would probably bee more suitable.

In order to quantify the effect of the true outcome $m_i(t)$ on the risk for the event at specific time $t$, we use a relative risk model of the form (Therneau and Grambsch, 2000).

$$h_i(t|\mathcal{M}_i(t), \mathbf{w}_i) = h_0(t) \exp\{\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{w}_i + \alpha m_i(t)\}, t > 0, \tag{1}$$

where $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s \leq t\}$ denotes the history of the true unobserved longitudinal process for subject $i$ up to time point $t$. The parameter $\alpha$ quantifies the degree of association between the longitudinal marker and the risk for the event.

In standard survival analysis, the baseline risk function $h_0(\cdot)$ is typically left completely unspecified (Cox, 1972; Andersen and Gill, 1982). However, within the joint modeling framework (Hsieh et al., 2006) noted that leaving this function completely unspecified leads to an underestimation of the standard errors of the parameter estimates, so it is necessary to explicitly define $h_0(\cdot)$. Although we could have used the hazard function of a standard survival distribution (e.g. Weibull or Gamma), we finally opted for a more flexible solution such a piecewise-constant model:

$$h_0(t) = \sum_{q=1}^{Q} \xi_q I(\nu_{q-1} < t \leq \nu_q), \tag{2}$$

where $0 = \nu_0 < \nu_1 < \ldots < \nu_Q$ denotes a split of the time scale, with $\nu_Q$ being the largest observed time, and $\xi_q$ denotes the value of the hazard in the interval $(\nu_{q-1}, \nu_q]$.

On the basis of the expressed considerations, the true and unobserved outcome at a specific time point $t$ can be modeled by joining the two above

approaches (Rizopoulos, 2012)

$$h_i(t|\mathcal{M}_i(t), \mathbf{w}_i) = h_0(t)R_i(t)\exp\{\gamma log.CUM0_i + \alpha(\boldsymbol{\beta_0} + \boldsymbol{b_{i0}} + \boldsymbol{\beta_1 t})\}. \quad (3)$$

In particular, the hazard at age $t$ for the $i-$th individual, with a true longitudinal profile $\mathcal{M}_i(t)$ up to time $t$, can be expressed as follows:

$$h_i(t|\mathcal{M}_i(t), \mathbf{w}_i) = h_0(t)R_i(t)\exp\left[\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{w}_i + \alpha\{\mathbf{x}_i^{\mathrm{T}}(t)\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}(t)\mathbf{b}_i\}\right]. \quad (4)$$

The models presented in this section can be generalized to higher dimensions (Andrinopoulou et al., 2014). More information on joint modeling fitting can be found in Rizopoulos (2012) and details about the R package implementation are given in Rizopoulos (2010).

## 2.4 Predicted survival in joint models

Once the model has been specified, estimated and validated, a powerful feature is to derive survival predictions. Thus, considering the sample $\mathcal{D}_n = \{T_i, \delta_i, \mathbf{y}_i; i = 1, \ldots, n\}$ on which the joint model was fitted, the goal consists of predicting conditional probability of surviving time for a new subject $i$ that provides a set of longitudinal measurements, $\mathcal{Y}_i(t) = \{y_i(s);\ 0 \leq s < t\}$ and a vector of baseline covariates, $\mathbf{w}_i$. The flexibility provided by the joint modeling approach is in line with a growing trend towards personalized medicine (Garre et al., 2008; Proust-Lima and Taylor, 2009a; Rizopoulos, 2011). In particular, the real challenge focuses on estimating these probabilities not only at each one of the time points measurements, but also at a generic time $u > t$ given survival up to $t$, i.e.

$$\pi_i(u|t) = \Pr(T_i^* \geq u \mid T_i^* > t, \mathcal{Y}_i(t), \mathbf{w}_i, \mathcal{D}_n; \boldsymbol{\theta}^*), \quad (5)$$

where $\boldsymbol{\theta}^*$ denotes the true parameter values.

This approach therefore allows to obtain the so called survival dynamic predictions for the $i-$th subject, arising from his survival curve which is updated on the basis of any new longitudinal information that is subsequently collected. Hence, as new information at time $t' > t$ is added to existing longitudinal measurements, one can update the estimated survival curve $\pi_i(u \mid t)$ to $\pi_i(u \mid t')$, and therefore proceed in a time dynamic manner.

The estimation of the subject-specific conditional survival probabilities takes full advantage of the conditional independence used to define the joint model. Using a Bayesian formulation (Proust-Lima and Taylor, 2009a; Rizopoulos, 2011), the problem can be written as:

$$\begin{array}{l} \Pr(T_i^* \geq u \mid T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_n) \\ = \int_{\boldsymbol{\theta}} \Pr(T_i^* \geq u \mid T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}_n) \mathrm{d}\boldsymbol{\theta}. \end{array} \tag{6}$$

The first part of the above integrand is given by

$$\begin{array}{l} \Pr(T_i^* \geq u \mid T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) \\ = \int_{\mathbf{b}_i} \frac{S_i\{u \mid \mathcal{M}_i(u, \mathbf{b}_i, \boldsymbol{\theta}); \boldsymbol{\theta}\}}{S_i\{t \mid \mathcal{M}_i(t, \mathbf{b}_i, \boldsymbol{\theta}); \boldsymbol{\theta}\}} \, p(\mathbf{b}_i \mid T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) \mathrm{d}\mathbf{b}_i, \end{array} \tag{7}$$

where $S_i(\cdot)$ denotes the survival function, and furthermore it has been explicitly noted that the true longitudinal history $\mathcal{M}_i(\cdot)$ is a function of both the random effects and the parameters. For the second part of equation (6), it is assumed that the sample size $n$ is large enough, such that $\{\boldsymbol{\theta}; \mathcal{D}_n\}$ can be well approximated by $\mathcal{N}\{\hat{\boldsymbol{\theta}}, \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\}$.

By combining (6), (7) and $\{\boldsymbol{\theta}; \mathcal{D}_n\} \sim \mathcal{N}\{\hat{\boldsymbol{\theta}}, \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\}$, a Monte Carlo estimate of $\pi_i(u \mid t)$ can be derived using the following simulation scheme:

- Draw a value of $\boldsymbol{\theta}^{(l)}$ from a normal distribution $\mathcal{N}\{\hat{\boldsymbol{\theta}}, \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\}$
- Draw a value of $\mathbf{b}_i^{(l)}$ from the pool.

- Compute $\pi_i^{(l)}(u \mid t)$ as

$$S_i\{u \mid \mathcal{M}_i(u, \mathbf{b}_i^{(l)}, \boldsymbol{\theta}^{(l)}); \boldsymbol{\theta}^{(l)}\} / S_i\{t \mid \mathcal{M}_i(t, \mathbf{b}_i^{(l)}, \boldsymbol{\theta}^{(l)}); \boldsymbol{\theta}^{(l)}\}. \tag{8}$$

The three steeps are repeated $l = 1, \ldots, L$ times, where $L$ denotes the number of Monte Carlo samples. The realizations $\{\pi_i^{(l)}(u \mid t)\ , l = 1, \ldots, L\}$ can be used to derive point estimates of $\pi_i(u \mid t)$, such as the median and the mean values as follows:

$$\hat{\pi}_i^{(l)}(u \mid t) = \text{median}\{\pi_i^{(l)}(u \mid t)\ , l = 1, \ldots, L\} \tag{9}$$

$$\hat{\pi}_i^{(l)}(u \mid t) = \frac{1}{L} \sum_{l=1}^{L} \pi_i^{(l)}(u \mid t). \tag{10}$$

From these estimates, it is also possible to compute the standard errors using the sample standard deviation over the Monte Carlo samples and the confidence intervals through the sample percentiles.

# 3 Results and predictions

The results for the private mutual insurer data set are presented in Table 3 separately for men and women. The association parameter $\alpha$ is positive and significantly different from zero. This indicates that the larger the number of accumulated service units provided the larger the risk of death. This is consistent with intuition as an aggravated patient has a higher probability of death and as a consequence, he demands health care resources. At the same time, who demand health care services are certainly motivated by a deteriorated health condition and therefore expected survival time decreases. Note that a positive parameter factor in the hazard function means that the risk of death increases, whereas a negative parameter has the contrary effect.

Table 2: Results of the joint model estimation in the private insurer data set (2006-2014)

| | Men | | Women | |
|---|---|---|---|---|
| Parameters | Estimate | 95% CI | Estimate | 95% CI |
| $\beta_0$ | 2.140* | (2.100, 2.180) | 2.158* | (2.123, 2.192) |
| $\beta_1$ | 0.170* | (0.167, 0.172) | 0.157* | (0.155, 0.159) |
| $\sigma$ | 0.332* | (0.329, 0.334) | 0.314* | (0.312, 0.316) |
| $\sigma_{b_0}$ | 1.648* | (1.597, 1.698) | 1.791* | (1.748, 1.834) |
| $\gamma$ | -1.174* | (-1.306,-1.042) | -0.964 | (-1.042, -0.887) |
| $\alpha$ | **1.437*** | (1.275, 1.598) | **1.273*** | (1.179, 1.367) |

* indicates significance at the 5% level. CI stands for confidence interval.

We note that the association between the longitudinal process and the survival outcome is slightly higher for men (1.437) than for women (1.273), but the difference is not statistically significant. All other parameter estimates are similar for mean and women except for $\gamma$, which is not significantly different from zero for women. This means that pre-existing conditions, which are represented by $log.CUM0$ and which refer to the number of accumulated services receive during the four years previous to the study, do not influence the survival of women, while they do influence negatively the hazard rate for men. This result seems to indicate that a larger survival is expected for those who were using medical care more intensively than others. This result also means that if a patient accumulates a large number of services, but was able to survive to the starting date of the study, then he has a smaller hazard rate of death compared to another subject who has not accumulated as many services as him. This can be interpreted as the preventive effect or health care or the curative effect, which prove to be efficiently leading to longer life expectancy.

However, if suddenly a patient requires medical care, the number of accumulated services increases and therefore, since parameters $\beta_0$, $\beta_1$ and $\alpha$ are positive, that would synchronize with the hazard rate, which would increase and lead to a higher risk of death.

A log-unit increase in the cumulate number of visits entails a $exp(1.437) =$ 4.2-fold increase in the men risk and $exp(1.273) = 3.6$-fold increase for women.

Some comments on $\sigma$ and $\sigma_{b_0}$ are needed. Those two parameters can be interpreted as an inherent variability in the random effects of the longitudinal model. Note that $\sigma_{b_0}$ is slightly larger for women than for men, which could also be caused by the fact that the average age is larger for women than for men in this particular sample.

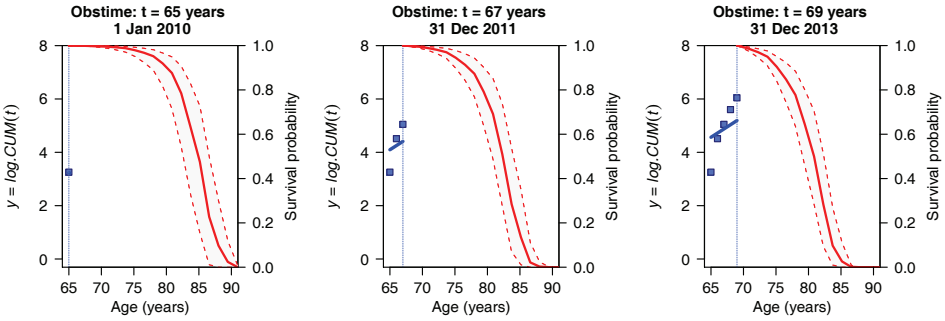
As an illustration, let us consider for instance the case of a woman 65 aged at study start point, for whom her cumulate service received during the four years prior to the study starting time point is known. In Figure 3 we can observe how the model updates the predicted survival probabilities as new longitudinal information is collected. This is a very useful prognostic tool.

# 4 Conclusions

From the analysis of our private insurance longitudinal data sample, we conclude that the observed number of cumulated health care service units provided is strongly as positively associated with the risk of death.

The baseline cumulated number of cumulated health care service units provided to a patient has a protective effect. This is in line with evidence of

Figure 1: Dynamic survival probabilities for a woman aged 65 who is still alive at the end of the study




a preventive affect.

The joint modeling methodology allows to continuously update the predictions of subject-specific survival probabilities, when new information on service usage comes along.

Further work is going to be pursued on the generalization of the statistical model to counting processes and to the implementation of multivariate longitudinal markers, as they seem very natural here. Indeed the number of medical care services needed can be categorized in big groups, those that are routine programmed actions and exceptional treatments, such as surgery or serious procedures.

One of the limitations of our study is the fact that all health services have the same importance in the longitudinal counter. Another practical issues is the fact that insurance customers switch between companies and new policy holders could enter the sample or leave the group motivated by health-related problems. We do not think that this was a problem in this particular data set. Since our policyholder were above 65 years of age, they have been in this mutual company for several year , because it is very infrequent to change

the health insurance provider at this age. we do not expect to have adverse selection in this group of policy holders.

# References


P.K. Andersen and R.D. Gill. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10:1100–1120, 1982.

E.-R. Andrinopoulou, D. Rizopoulos, J.J.M. Takkenberg, and E. Lesaffre. Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in Medicine*, 33(18):3167–3178, 2014.

C.E. Bird, L.R. Shugarman, and J. Lynn. Age and gender differences in health care utilization and spending for medicare beneficiaries in their last years of life. *Journal of Palliative Medicine*, 5(5):705–712, 2002.

D. Blane, G. Netuveli, and S.M. Montgomery. Quality of life, health and physiological status and change at older ages. *Social Science and Medicine*, 66(7):1579–1587, 2008.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34:187–220, 1972.

H. Dao, L. Godbout, and P. Fortin. On the importance of taking end-of-life expenditures into account when projecting health-care spending. *Canadian Public Policy*, 40(1):45–56, 2014.

M. Denuit. Life anuities with stochastic survival probabilities: A review. *Methodology and Computing in Applied Probability*, 11(3 SPEC. ISS.):463–489, 2009.

M. Denuit and E. Frostig. First-order mortality basis for life annuities. *Geneva Risk and Insurance Review*, 33(2):75–89, 2008.

S. Fieuws, G. Verbeke, B. Maes, and Y. Vanrenterghem. Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics*, 9(3): 419–431, 2008.

F.G. Garre, A.H. Zwinderman, R.B. Geskus, and Y.W. Sijpkens. A joint latent class changepoint model to improve the prediction of time to graft failure. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171:299–308, 2008.

F. Hsieh, Y-K. Tseng, and J-L. Wang. Joint modeling of survival and longitudinal data: Likelihood approach revisited. *Biometrics*, 62:1037–1043, 2006.

J.D. Kalbfleisch and R.L. Prentice. *The statistical analysis of failure time data, 2nd Edition*, volume 360. John Wiley & Sons, 2002.

M. Murphy. *Proximity to death and health care costs*. Edward Elgar Publishing, 2012.

C. Proust-Lima and J.M.G. Taylor. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of post-treatment PSA: A joint modeling approach. *Biostatistics*, 10:535–549, 2009a.

C. Proust-Lima and J.M.G. Taylor. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of

posttreatment psa: a joint modeling approach. *Biostatistic*, 10(3):535–549, 2009b.

D. Rizopoulos. Jm: An r package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9):1–33, 2010.

D. Rizopoulos. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829, 2011.

D. Rizopoulos. Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule. *Computational Statistics and Data Analysis*, 56(3):491–501, 2012.

D. Rizopoulos and E. Lesaffre. Introduction to the special issue on joint modelling techniques. *Statistical methods in medical research*, 23(1):3–10, 2014.

C. Schoen, R. Osborn, D. Squires, M.M. Doty, R. Pierson, and S. Applebaum. How health insurance design affects access to care and costs, by income, in eleven countries. *Health Affairs*, pages 10–1377, 2010.

T.M. Therneau and P.M. Grambsch. *Modeling survival data: extending the Cox model*. Springer, New York, 2000.

G. Verbeke and G. Molenberghs. *Linear mixed models for longitudinal data*. Springer, 2009.

C.-S.J. Yue and H.-C. Huang. A study of incidence experience for taiwan life insurance. *Geneva Papers on Risk and Insurance: Issues and Practice*, 36 (4):718–733, 2011.

## **UB·Riskcenter** Working Paper Series

### List of Published Working Papers

**[WP 2014/01].** Bolancé, C., Guillén, M. and Pitt, D. (2014) "Non-parametric models for univariate claim severity distributions – an approach using R", UB Riskcenter Working Papers Series 2014-01.

**[WP 2014/02].** Mari del Cristo, L. and Gómez-Puig, M. (2014) "Dollarization and the relationship between EMBI and fundamentals in Latin American countries", UB Riskcenter Working Papers Series 2014-02.

**[WP 2014/03].** Gómez-Puig, M. and Sosvilla-Rivero, S. (2014) "Causality and contagion in EMU sovereign debt markets", UB Riskcenter Working Papers Series 2014-03.

**[WP 2014/04].** Gómez-Puig, M., Sosvilla-Rivero, S. and Ramos-Herrera M.C. "An update on EMU sovereign yield spread drivers in time of crisis: A panel data analysis", UB Riskcenter Working Papers Series 2014-04.

**[WP 2014/05].** Alemany, R., Bolancé, C. and Guillén, M. (2014) "Accounting for severity of risk when pricing insurance products", UB Riskcenter Working Papers Series 2014-05.

**[WP 2014/06].** Guelman, L., Guillén, M. and Pérez-Marín, A.M. (2014) "Optimal personalized treatment rules for marketing interventions: A reviewof methods, a new proposal, and an insurance case study.", UB Riskcenter Working Papers Series 2014-06.

**[WP 2014/07].** Piulachs, X., Alemany, R. and Guillén, M. (2014) "A joint longitudinal and survival model with health care usage for insured elderly", UB Riskcenter Working Papers Series 2014-07.