Application of the Cox Proportional Hazards Model and Competing Risks Models to Critical Illness Insurance Data

David Zapletal

Science and Research Centre, Faculty of Economics and Administration, University of Pardubice

Correspondence
David Zapletal, Science and Research Centre, Faculty of Economics and Administration, University of Pardubice, Studentska 95, 53210 Pardubice, Czech Republic
Email: david.zapletal@upce.cz

ABSTRACT
A commercial insurance company in the Czech Republic provided data on critical illness insurance. The survival analysis was used to study the influence of the gender of an insured person, the age at which the person entered into an insurance contract and the region where the insured person lived on the occurrence of an insured event. The main goal of the research was to investigate whether the influence of explanatory variables is estimated differently when two different approaches of analysis are used. The two approaches used were (1) the Cox proportional hazard model that does not assign a specific cause, such as a certain diagnosis, to a critical illness insured event and (2) the competing risks models. Regression models related to these approaches were estimated by R software. The results, which are discussed and compared in the paper, show that insurance companies might benefit from offering policies that consider specific diagnoses as the cause of insured events. They also show that in addition to age, the gender of the client plays a key role in the occurrence of such insured events.

KEYWORDS
Critical illness insurance, survival analysis, Cox proportional hazards model, competing risks, cause-specific hazard model, subdistribution hazard model, Fine and Gray model

INTRODUCTION

Survival data analysis is an area of statistics that estimates the amount of time that is likely to elapse before a certain event occurs. Models for this type of analysis have often been applied to medical data, and many scientific books and research articles have reported on the results. This article focuses on the application of survival analysis models to data on insurance transactions; such studies are much less common than those on medical data.

The most common use of survival analysis in the insurance industry is for contract failures. For instance, the analysis of customer survival time in an insurance company after a policy cancellation was introduced by Guillen et al. [13]. The Cox proportional hazards model was

used in [17] to investigate China's residential mortgage life insurance prepayment risk behaviour. The dataset of Danish households with multiple insurance policies was studied in Ref. [5]. Haugen and Moger investigated corporate customers holding multiple contracts for automobile insurance with the same insurance company. They used the shared gamma frailty model to study the time-lapse of single-car policies [16].

The competing risks approach was used by Mihaud and Dutang [22]. They modelled the duration of a life insurance contract through the subdistribution hazard model developed by Fine and Gray. Applications of competing risks models for insurance data can also be found in the following articles. Dang investigated the insolvency outcomes of U.S. property-casualty insurers from 1998 to 2010 [9]. He fitted the competing risks models with time-dependent covariates for five specific insolvency outcomes. The performance of the Federal Housing Administration mortgage program and privately insured home purchase mortgages relative to uninsured mortgages was investigated by Park [23]. He used, besides the Kaplan–Meier survival estimate, the subdistribution hazard model.

The main goal of the presented paper was to study the influence of explanatory variables on the occurrence of an insured event and to investigate whether estimates of the effects of covariates differ when they are made using different approaches (Cox proportional hazard and competing risks approach). The following explanatory variables were used in the analysis: the gender of the insured person, the age at which the person entered into an insurance contract, and the region where the insured person lived. Interesting results were obtained in particular in the case of gender variable. The results also confirm that the age of the client, of course, plays a significant role in the occurrence of the insured events. However, the competing risks approach has again yielded some interesting results in this case.

The remainder of the article is organized as follows. In the Methodology section, the used approaches are described and discussed. The data are introduced and described in the section Data. Explanatory variables and used models are listed in section Models. Section Results contains and describes the estimated results of the used models. Finally, the results and the conclusions drawn from them are summarized in the section Summary and Conclusions.

METHODOLOGY

When there is a single outcome of interest or the outcome of interest is not distinguished by a specific cause, the Cox proportional hazards model provides a suitable method for accommodating covariate information. The model, with its unspecified baseline hazard function, was proposed by Cox in 1972 [7]; he introduced the notion of partial likelihood and went on to consider it in great detail in 1975 [8]. A detailed review of the model and its extension is provided in Ref [31]. The Cox model of the hazard at time t for the i-th individual is given by the equation

$$h_i(t) = \exp(\boldsymbol{\beta}\mathbf{X})\, h_0(t), \tag{1}$$

where $h_0(t)$ is the baseline hazard function of the unspecified form. There is a direct correspondence between the effect of a covariate on the hazard of the outcome and the effect of a covariate on the incidence of the outcome due to the following relationship:

$$S(t) = S_0(t)^{\exp(\boldsymbol{\beta}\mathbf{X})}, \tag{2}$$

where $S(t)$ is a survival function and $S_0(t)$ is a baseline survival function [1, 2]. Thus, making inferences about the direction of the effect of a covariate on the hazard function permits one to make equivalent inferences about the direction of the effect of that covariate on the incidence (or probability of occurrence) of the outcome. This direct correspondence allows interpretation without an exact specification of whether a risk denotes the hazard of an event (i.e., the rate of occurrence of the event for those still at risk of the event) or the incidence of the event (i.e., the probability of the occurrence of the event) [1, 2].

The term *competing risks* implies that a person may experience one of a set of different events or outcomes. In time-to-event analyses, the occurrence of one of the possible events often precludes the occurrence of other possible events. For example, in critical illness insurance, the occurrence of the insured event due to cancer precludes the occurrence of the event due to another disease. Two different methods are usually used in the presence of competing risks: the cause-specific hazard method [24] and the subdistribution hazard method [12]. Both methods account for competing risks, but they do so by modelling the effect of covariates on different hazard functions. Consequently, each model has its unique interpretation [2].

The cause-specific hazard approach was proposed by Prentice et al. [24] and discussed in Refs. [10, 18, 25]. The cause-specific hazard function for the i-th individual from j-th cause is defined as

$$h_{ij}(t) = \lim_{\Delta t \to 0} \frac{\text{Prob}(t \leq T < t + \Delta t, D = j | T \geq t)}{\Delta t}, \tag{3}$$

where D is a variable denoting the type of event that occurred. The cause-specific hazard function denotes the instantaneous rate of occurrence of the j-th event in subjects who are currently event-free (i.e., in subjects who have not yet experienced any of the different types of events).

The regression model for the so-called subdistribution hazard function was introduced in 1999 and has been designated the Fine and Gray model by its developers [12]. Generalization of this model was provided by Scheike and Zhang [27], and a suitable package in R software was also described by them [28]. The subdistribution hazard function for the i-th individual from the j-th cause is defined as

$$\lambda_{ij}(t) = \lim_{\Delta t \to 0} \frac{\text{Prob}(t \leq T < t + \Delta t, D = j | T \geq t \cup (T < t \cap J \neq j))}{\Delta t}. \tag{4}$$

The subdistribution hazard function denotes the instantaneous risk of failure from the j-th cause in subjects who have not yet experienced an event of type j. In this case, the risk set includes those who are currently event-free, as well as those who have previously experienced a competing event. This differs from the risk set for the cause-specific hazard function, which includes only those who are currently event-free.

The subdistribution hazard model has also been described as a cumulative incidence function (CIF) regression model. The CIF for the j-th cause is defined as

$$\text{CIF}_j(t) = \text{Prob}(T \leq t, D = j). \tag{5}$$

The function $\text{CIF}_j(t)$ denotes the probability of experiencing the j-th event before time t and before the occurrence of a different type of event. Thus, the subdistribution hazard model allows one to estimate the effect of covariates on the CIF for the event of interest, and it permits one to recover a relationship similar in form to that described in Equation (2). Thus, if a covariate is associated with an increase in the subdistribution hazard function, it will also be

associated with an increase in the incidence of the event. There is a one-to-one relationship with the CIF for the subdistribution hazard but not for the cause-specific hazard [1]. When using the cause-specific hazard model in the presence of competing risks, it is incorrect to infer that a given variable is associated with an increased or decreased incidence of the event of interest because Equation (2) does not hold in the presence of competing risks [23, 27]. This is because one must account for the effect of covariates on the cause-specific hazard function of each of the different types of events when determining their effect on the CIF for the event of interest [19].

There are two key questions that arise when using competing risks regression models: (1) Which covariates affect the rate at which events occur? (2) Which covariates affect the probability of an event occurring over time? [1] According to Lau et al., cause-specific hazard models are better suited for studying the cause (reason) of the occurrence of the event [21]. The reason is that the cause-specific hazard function denotes the instantaneous rate of the primary outcome. Thus, regression coefficients from the cause-specific hazard model can be interpreted as the relative effect of the corresponding covariate on the relative increase in the rate of occurrence of the primary event in subjects who are currently event-free [2]. On the other hand, Koller et al. suggested that the subdistribution hazard method was preferable when the focus was on estimating the actual risk and prognosis [19]. Furthermore, the subdistribution hazard method may be of great interest if one is interested in the overall impact of covariates on the incidence of the outcome of interest, even when predictions of incidence are not of direct interest [2].

The following predictors (covariates) were used in the analysis: the gender of the insured person, the age at which the person entered into an insurance contract, and the region where the insured person lived. In connection with the variable age, it should be noted that it was considered as a factor (containing four age groups – see below) in the model, and it was fixed to the age at the conclusion of the insurance contract. On the other hand, when the dependence of the hazard function on continuous explanatory variables (or more generally on a variable that takes a wide range of values) to be modelled, we should consider whether it is appropriate to include that variable as a linear term in the model [6]. If the age was considered as a continuous linear term in the model (or at baseline hazard function), then the effect of changing age can be ignored in a survival model due to the structure of partial likelihood [30]. However, if there is evidence that the effect of age on risk is nonlinear, the situation is somewhat more complicated. In this case, it is possible to incorporate fractional polynomials into the model (see Ref [26] for details) or to use one of a wide range of mortality models (see, for example, Refs. [3, 4, 11].

The results presented here are in accordance with the recommendations given in [20] for all causes of the outcome (i.e., no specific cause is distinguished) and for both the cause-specific hazard method and the subdistribution hazard method (i.e., competing risks approach). As pointed out in [2], such a practice enables a complete understanding not only of the effects of prognostic factors but also of the absolute risks of the different outcomes in the study sample. Therefore, in addition to the results obtained for all causes by applying the Cox model and published in [32], both types of competing risks models (i.e., the cause-specific hazard model and subdistribution hazard model) were applied to the insurance data. The results of these models are discussed and compared.

DATA

A commercial insurance company in the Czech Republic provided data on critical illness insurance. This type of insurance is offered as a supplement to life insurance if the policyholder is an adult under the age of 65 years or as a separate insurance contract for persons under 18 years of age. The insurance covers the risk of 31 critical illnesses such as cancer, heart attack, and stroke. If the client is diagnosed with any of these diseases, the client will receive the agreed sum insured. The insurance indemnity is paid only once, and after its payment, the insurance contract expires. The diagnosis of one of the critical diseases resulting in an insured event, therefore, excludes the occurrence of a different critical illness as an insured event. This fact justifies the use of methods suitable for competing risks.

After a client purchased a critical illness policy from the company in our study, there was a 3-month waiting period during which the company would not pay a claim for that client if an insured event occurred. Critical illnesses that arose as a result of extreme sports or alcohol or drug use were excluded from the critical illness insurance policies.

The analysed dataset in our study contained clients who entered into insurance contracts from July 1, 1997, to April 30, 2017. The date of the start and eventual endpoints of each policy were known. If the endpoint was reached on or before April 30, 2017, its cause was known. The endpoint could be the date that an insured event occurred or the date that the policy was terminated by the client or insurance company. For the purposes of this analysis, however, only the occurrence of an insured event was considered as the endpoint. Other causes of termination of the insurance contract or policy in force were deemed to be censored cases.

Therefore, each policy's duration (in days) was calculated to analyse the time to the endpoint (i.e., to the occurrence of an insured event). Due to the above-mentioned 3-month waiting period, policies that had been in force for less than three months were removed from the dataset. In addition to the policy duration, we also had information about the gender and age of the policyholder and the region in which the policyholder lived. This allowed us to investigate the influence of the gender of the insured person, the age at which the person entered into the insurance contract, and the region of residence of the insured person on the insured event's occurrence.

The dataset contained data for 231,046 persons, and the number of insured events in the monitored period was 1,045. A crucial assumption made when using the Cox model is that of proportional hazards. Hazards are said to be proportional if the ratios of hazards are independent of time. The hazard proportionality assumption was tested by the so-called zph test based on Schoenfeld residuals developed by Grambsch and Therneau [17]. The assumption of hazard proportionality was not rejected for the variables of gender and region, but it was rejected for the variable of age. For this reason, the dataset was stratified into two groups: clients under age 18 years and clients 18 years of age and older. The subset of clients under age 18 years had 91,083 clients with 114 insured events. The subset of clients who were 18 years old or older had 139,963 persons with 931 insured events. One can see that there was a large disproportion in the number of insured events, given the total number of clients in these two groups, and this led to a rejection of the proportional hazard assumption.

Therefore, only the data for individuals 18 years of age and older were analysed by the survival models and considered in this paper.

The specific cause of each insured event was also known (i.e., the specific diagnosis of critical illness). There were 21 different causes (of the 31 covered by insurance), but 16 of them did not reach an incidence of 1%. Five diagnoses reached an incidence of 2% or more, and together these diagnoses accounted for over 96% of all occurrences of insured events (see Table 1). The most common diagnosis was cancer at more than 70% of all occurrences. The next most common diagnosis was a heart attack (>11%), stroke (~8%), benign brain tumor (~4%), and coronary artery disease requiring operation (>2%). Therefore, two groups of causes (diagnoses) were created (see Table 1): cancer (labelled Cancer in the table) and all other diagnoses (labelled Other).

MODELS

In accordance with the recommendations in [20], two different approaches to the outcome of interest (i.e., the occurrence of the insured event) were used. The first approach does not assign a particular cause to the insured event, and the second one does assign a specific cause (i.e., diagnosis of a critical illness). The results of the competing risks approaches were obtained via the cause-specific hazard model and subdistribution hazard model and were compared with the calculations of the first approach, the Cox model, published in Ref. [32].

The influence of the gender of the insured person, the age at which the person entered into the insurance contract, and the region where the insured person lived at the time of occurrence of the insured event were investigated. The Cox model of the hazard at time t for the i-th individual is in this case given by the equation

$$h_i(t) = \exp(\beta_1 Gender_i + \beta_2 Age_i + \beta_3 Gr\_Region_i)\, h_0(t), \qquad (7)$$

where $h_0(t)$ is a baseline hazard function of unspecified form. The variables $Gender_i$, Age_i and Gr_Region_i are the categories of explanatory variables (see below) for the i-th individual.

In the competing risks approach, the cause-specific hazard model is the most direct. Two separate Cox models are developed for each cause of occurrence of an insured event, in our case, one for cancer and one for the other group of critical diagnoses (see Table 1). The following two models for the cause-specific hazard function are

$$h_{ij}(t) = \exp\left(\beta_{1j} Gender_i + \beta_{2j} Age_i + \beta_{3j} Gr\_Region_i\right) h_{0j}(t), \qquad (8)$$

where $h_{ij}(t)$ is the hazard of the insured event for the i-th individual from the j-th cause ($j = 1$ means cancer, $j = 2$ means the other causes) and $h_{0\mathrm{j}}(t)$ is the baseline hazard function for the j-th cause.

The subdistribution method provides an alternative approach to modelling covariate data with competing risks. It uses the subdistribution hazard for modelling the effects of covariates on a specific cause of a considered event analogously to the Cox proportional hazard model, which in our case led to the model

$$\lambda_{ij}(t) = \exp\left(\beta_{1j} Gender_i + \beta_{2j} Age_i + \beta_{3j} Gr\_Region_i\right) \lambda_{0j}(t), \qquad (9)$$

where $\lambda_{ij}(t)$ is the subdistribution hazard function of the insured event for the i-th individual from the j-th cause ($j = 1$ means cancer, $j = 2$ means the other causes) and $\lambda_{0j}(t)$ is the baseline subdistribution hazard function for the j-th cause. However, increased attention must be paid to interpreting the results of competing risks models, especially to the different constructions of the risk set (see the section on Methodology).

Because of the categorical (factor) explanatory variables, each parameter β_{ij} for $i = 1, 2, 3$ is represented by $q - 1$ estimated parameters, where q means the number of categories of corresponding explanatory variables. The variable $Gender$ included two categories ($q = 2$). The Age variable was made up of four categories ($q = 4$): 18 to 30 years, 31 to 40 years, 41 to 50 years, and over 50 years. As background for the variable Gr_Region, it should be noted that the Czech Republic is divided into 14 territorial administrative regions: Central Bohemia (STC), Hradec Kralove (HKK), Karlovy Vary (KVK), Liberec (LBK), Moravia–Silesia (MSK), Olomouc (OLK), Pardubice (PAK), Plzen (PLK), Prague (PHA), South Bohemia (JHC), South Moravia (JHM), Usti nad Labem (ULK), Vysocina (VYS), and Zlin (ZLK). In our study, we formed categories of the variable Gr_Region based on groups with different rates of risk of an insured event (without distinguishing the specific cause of the insured event; i.e., based on the results of the Cox model) [35]. According to the rate of hazard, we set up three groups of regions ($q = 3$): Group 1 contained the regions of Liberec, Pardubice, Prague, and Zlin, where the hazard rate was the lowest; group 2 included the nine regions of Central Bohemia, Hradec Kralove, Moravia–Silesia, Olomouc, Plzen, South Bohemia, South Moravia, Vysocina, and Usti nad Labem; and group 3 contained only one region, Karlovy Vary, which had the highest hazard rate of occurrence of an insured event. The categories of corresponding explanatory variables with the lowest hazard rate (again based on the approach that did not specify a cause of the critical illness) were determined as reference categories; they were male for the variable $Gender$, 18 to 30 years for the variable Age, and group 1 for the variable Gr_Region.
The R software (package survival [29] and package cmprsk [15]) was used to test the proportional hazard assumption, estimate the CIFs, and estimate all the above-mentioned regression models.

RESULTS AND DISCUSSION

Because the influence of gender, age, and region of residence of the insured person was investigated, the frequency tables of these explanatory variables were calculated ($Gender$ in Table 2, Age in Table 3, and Gr_Region in Table 4). In addition to the absolute frequency of each category of the corresponding variable, the percentage ratios of the individual absolute frequencies were calculated. These ratios are presented for the two diagnostic groups (cancer and other - connected with the competing risks approach) as well as for both groups together (cancer plus other - Cox model approach).
Analysis of the insurance data showed that an insured event did not occur for most clients; they were considered censored cases (see the columns labelled Censored in Tables 2 to 4). Even so, some interesting results can be seen. For example, regarding gender, we can see that if a cause was not specified, the percentage ratios were similar for males (0.64%) and females (0.68%) but that if a cause was specified, the percentage ratios differed greatly (see Table 2). Indeed, the statistical test for the evaluation of differences between proportions rejected the hypothesis of equality of these proportions for both the cancer group and the other diagnosis group (*p*-value < .0001). On the other hand, this hypothesis was not rejected if we did not assign a specific diagnosis (*p*-value = .342). There was a different situation in the case of the age groups, where the ratio increased with increasing age for all three case groups (cancer, other diagnoses, and cancer plus other diagnoses; see Table 3). Regarding the statistical significance of the proportion differences, the statistical test rejected the hypothesis of

equality for all possible pairs of age groups (*p*-value < .0001) except for the age groups 41 to 50 years and over 50 years (*p*-value = .714 for cancer, .365 for other diagnoses, and .421 for cancer plus other diagnoses). The regions are listed in Table 4 according to their increasing ratios (without assigning a specific cause of the insured event in the column for cancer plus other diagnoses), and they form the three groups mentioned above (see the section titled Models). The proportions of these three groups were calculated and compared with each other, showing that all hypotheses of equality were rejected at a level of significance of .05.

Regarding the influence of these variables on the occurrence of the insured event, we can see that the proportion differences among categories were the highest for the age variable, so that, not surprisingly, age had the most significant influence. However, it is essential to note that this simple approach did not take into account the duration of the insurance contract, which was related to the amount of the insurance premium paid by the client until the insured event occurred. That is why survival analysis should be used to account for the duration.

As mentioned above, the results for the competing risks approaches were compared with the results for the approach that did not assign a specific cause to an insured event (i.e., a diagnosis of a critical illness). Thus, the results of the Cox model (i.e., no particular cause assigned) are once more presented first [32]. Estimations of the coefficients of the Cox proportional hazard model with their statistical significance (*p*-values), hazard ratios, and corresponding confidence intervals are shown in Table 5.

We can see from Table 5 that the time to occurrence of the insured event was not statistically significant (the *p*-value of .678 was greater than the determined level of significance of .05) for male and female insured persons. This means that the risk of critical illness was comparable for males and females (hazard ratio, 1.028; confidence interval contains 1.00). As expected, the situation was different for the explanatory variables Age and Gr_Region. In particular, age plays a significant role in the risk of critical illness. Based on the hazard ratios, we can see that the age category of 31 to 40 years had a risk of occurrence of a critical illness that was more than twice as much as that of the reference age category of 18 to 30 years. For insured persons aged 41 to 50 years, the risk was more than five times greater than the reference category of 18 to 30 years, and for people over 50 years, it was more than ten times higher. In the case of the region of residence, the differences were not so large. The lowest risk of an insured event was in the first group, the region encompassing Liberec, Pardubice, Prague, and Zlin (interestingly, these are not neighbouring regions). The highest risk was in the third group, Karlovy Vary, where the risk of occurrence of critical illness was almost twice as large as in the first group. The other nine regions, which form the second group, had similar risks of an insured event, but this risk was only a little higher (~1.3 times) compared with the first group.

As already mentioned in the section on Methodology, there was a direct correspondence, thanks to Equation (2), between the effect of a covariate on the hazard of the outcome and the effect of a covariate on the incidence of the outcome. This direct correspondence allowed a looser interpretation without an exact specification of whether the risk denoted the hazard of an event or the incidence of the event.

Now, we focus on the results of the competing risks approaches, which can be seen for the cancer group in Table 6 (cause-specific hazard model) and Table 7 (subdistribution hazard model) and for the other diagnoses group in Table 8 (cause-specific hazard model) and Table 9 (subdistribution hazard model). We must keep in mind that the interpretation of the results

of the cause-specific hazard model and the subdistribution hazard model is different (see Methodology). It is worth noting that the estimated regression coefficients of the cause-specific hazard model and subdistribution hazard model were essentially the same for the corresponding diagnoses (compare Table 6 with Table 7 and Table 8 with Table 9). They were caused by a large number of censored cases in the dataset. Once this is taken into account, the differences mentioned above in constructing the risk set were not manifested so much. Therefore, we can say that the considered variables had a similar effect on the cause-specific hazard and on the relative incidence of a given outcome. However, keep in mind that there is a one-to-one relationship with the incidence of an outcome (through the CIF) for the subdistribution hazard but not for the cause-specific hazard. That is why the CIFs are also presented for the cancer group and the other diagnosis group in Figure 1. We can see (from Figure 1) that there was an almost linear increase in incidence in the first 4,400 days (–12 years) after the insurance policy went into effect. The rate of this growth was much higher for the cancer group; however, cancer incidence was almost twice as high as the incidence for the other diagnosis group.

Regarding the explanatory variables of the regression models, we can say that the influence of gender, unlike for the Cox model, was statistically significant and different for cancer and other diagnoses. For cancer, the risk of occurrence was 1.5 times higher for females than for males. For the other diagnoses, the risk of the event was significantly lower for females than for males (0.53 times).

Results for the age variable corresponded to those obtained by testing the proportion differences for both the competing risks approach and the approach that did not assign a cause to a critical illness. For both groups of diagnoses considered, the risk increased with the increasing age of the policyholder. However, the rate of growth was different. For cancer, the increase in risk with increasing age was not as rapid as for the other diagnoses. For example, the risk of occurrence of cancer was approximately 8.5 times higher for clients over 50 years of age than for clients from 18 to 30 years of age, but it was more than 16 times higher from one group to the other for other diagnoses. A closer comparison of the results of the regression models and the results of the proportion differences shows one interesting detail concerning the comparison of age groups of 41 to 50 years and over 50 years. As mentioned above, the proportion differences of these two age groups were not statistically significant in all three cases (cancer, other diagnoses, and cancer plus other diagnoses), which could lead to the merging of these two age groups. All regression models, however, suggest different results based on non-overlapping confidence intervals of hazard ratios. Therefore, all regression models were re-estimated for the reference age group of 41 to 50 years to support this proposition (the resulting tables are not presented here). The statistically significant difference of the hazard ratios for the age group of 41 to 50 years and the age group of over 50 years was confirmed.

Regarding the regions of residence, the differences were again not so significant. Unlike the results of the approach that did not assign a specific cause, there was no statistically significant difference in the risk of an insured event between the first and second groups of regions (for both cancer and other diagnoses). A significantly higher risk was seen in the Karlovy Vary region (third group), and the risk was slightly larger for the other diagnoses than for cancer.

SUMMARY AND CONCLUSIONS

The data on critical illness insurance provided by a commercial insurance company were analysed. First, the dataset had to be divided into two age groups, one for persons under 18 years of age and one for persons 18 years of age or older. The reason for the stratification was the enormous disproportion in the number of insured events (relative to the total number of clients in the group) in these two groups. This proportion was more than five times greater for clients who were 18 years of age or older. Therefore, having persons younger than 18 years in the portfolio of a critical illness insurance company is very beneficial for the company.

Two different survival analysis approaches were used for groups of adults (i.e., clients 18 years of age or older), and the results obtained were compared. If the specific cause of an insured event was not taken into account, the Cox model was estimated; otherwise, the competing risks approach was used. In that case, two models were evaluated: the cause-specific hazard model and the subdistribution hazard model. In general, these two models have different interpretations and can lead to different regression coefficients. In our case, however, this situation did not occur, mainly because of the large number of censored cases. Two groups of diagnoses were differentiated in case of competing risks approach, cancer, and other diagnoses. These groups were formed because more than 70% of the diagnoses were cancer. Interesting results were obtained, especially regarding gender. If the cause of the insured event was not taken into account, there were no differences between males and females in the risk of occurrence of the insured event. If the competing risks approach was used, however, totally different results were obtained. For cancer, the risk was significantly greater for females, but the risk was significantly lower for females in case of other diagnoses.

Regarding age, the risk increased with the increasing age of the client, regardless of the approach used. However, the dynamics of this growth were different. If the competing risks approach was used, the increase in the risk for cancer was not as rapid as for other diagnoses. This was mainly due to the low incidence of diagnoses of other causes in the reference age group of 18 to 30 years. Justification of the use of survival analysis was confirmed, among other things, by the demonstration of a statistically significant difference between the age group of 41 to 50 years and the age group of over 50 years. This difference has not been demonstrated by using a simpler approach.

In the case of regions of residence, the risk differences between the used approaches (including the differentiation of causes within the competing risks approach) were not as great as in the case of age. Regardless of the method, only the region of Karlovy Vary stood out in a negative sense.

Possible recommendations for the insurance company can be formulated as follows. Continue to offer this type of insurance to children and young people. Begin to deal also with products that consider either cancer only or all other diagnoses except cancer as the cause of an insured event. The latter policies could be offered with a significant discount on insurance premiums. For products that focus on a specific diagnosis, it would be appropriate to distinguish between males and females. For the country size of the Czech Republic, extending the model to include the client's regional affiliation is not necessary because the differences between the regions are not so significant. Using a sophisticated statistical method can reveal some interesting details about insurance statistics.

ORCID

David Zapletal https://orcid.org/0000-0002-4795-179X

REFERENCES

1. P. C. Austin and J. P. Fine, *Practical recommendations for reporting Fine-Gray model analyses for competing risk data.* Stat. in Med. 36 (2017), 4391-4400.
2. P. C. Austin, D. S. Lee and J. P. Fine, *Introduction to the analyses of survival data in the presence of competing risks.* Circulation 133 (2016), 601-609.
3. M. Bebbington, C.D. Lai and R. Zitikis, *Modeling human mortality using mixtures of bathtub shaped failure distributions*, J. Theor. Biol. 245(3) (2007), 528-538.
4. M. Bebbington et al., *Beyond the Gompertz law: exploring the late-life mortality deceleration phenomenon*, Scan. Act. J. 3 (2014), 189-207.
5. P. L. Brockett et al., *Survival analysis of a houshold portfolio of insurance policies: How much time do you have to stop total customer defection?*, J. Risk Insur. 75 (2008), 713-737.
6. D. Collet, *Modelling survival data in medical research,* 3rd ed., CRC Press, Boca Raton, 2014.
7. D. R. Cox, *Regression models and life Tables (with discussion)*, J. Royal Stat. Soc. B 34 (1972), 187-220.
8. D. R. Cox, *Partial likelihood*. Biom. 62 (1975), 269-276.
9. H. Dang, *A Competing Risks Dynamic Hazard Approach to Investigate the Insolvency Outcomes of Property-Casualty Insurers*, Gen. Pap. Risk Insur. Iss. Prac. 39(1) (2014), 42-76.
10. L. C. de Wreede, M. Fiocco and H. Putter, *mstate: an package for the analysis of competing risks and multi-state models*, J. Stat. Softw. 38 (7) (2011), 1-30.
11. J. Dolejs, and P. Maresova, *Onset of mortality increase with age and age trajectories of mortality from all diseases in the four Nordic countries*. Clin. Int. in Aging 12 (2017), 161-173.
12. J. P. Fine and R. J. Gray, *A proportional hazards model for subdistribution of a competing risk*, J. Am. Stat. Assoc. 94 (1999), 496-509.
13. M. Guillen et al., *The analysis of customer survival time in the insurance company after a policy cancellation*, Insur. Math. & Econ. 33 (2003), 434-434.
14. P. M. Grambsch and T. M. Thernau, *Proportional hazards tests and diagnostics based on weighted residuals*, Biom. 81 (1994), 515-526.
15. R. J. Gray, *Subdistribution analysis of competing risks. R package version 2.2-7*, 2013, available at http://www.r-project.org.
16. M. Haugen and T. A. Morger, *Frailty modelling of time-to-lapse of single policies for customers holding multiple car contracts*, Scand. Act. J. 6 (2016), 489-501.
17. K. H. Ho and H. Y. Su, *Structural prepayment risk behavior of the underlying mortgages for residential mortgage life insurance in a developing market*. J. Hous. Econ. 15 (2006), 257-278.
18. J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data*, 2nd ed., Wiley, New York, 2002.

19. M. T. Koller et al., *Competing risks and the clinical community: irrelevance or ignorance?* Stat. Med. 31 (2012), 1089-1097.
20. A. Latouche et al., *A competing risk analysis should report results on all cause-specific hazards and cumulative incidence functions*, J. Clin. Epidemiol. 66 (2013), 648-653.
21. B. Lau, S. R. Cole and S. J. Gange, *Competing risk regression models for epidemiologic data*, Am. J. Epidemiol. 170 (2009), 244-256.
22. X. Mihaud and Ch. Dutang, *Lapse tables for lapse risk management in insurance: a competing risks approach*, Eur. Act. J. 8 (2018), 97-126.
23. K. A. Park, *FHA loan performance and adverse selection in mortgage insurance*, J. Hous. Econ. 34 (2016), 82-97.
24. R. Prentice et al., *The analysis of failure times in the presence of competing risks*, Biom. 34 (1978), 541-554.
25. H. Putter, M. Fiocco and R. B. Geskus, *Tutorial in biostatistics: Competing risks and multistate models*, Stat. Med. 26 (2007), 2389-2430.
26. P. Royston and D.G. Altman, *Regression using fractional polynomials of contionuous covariates: parsimonious parametric modelling (wit discussion)*, App. Stat. 43 (1994), 429-467.
27. T. H. Scheike and M. J. Zhang, *Flexible competing risks regression modeling and goodness-of-fit*, Lif. Data Anal. 14 (2008), 464-483.
28. T. H. Scheike and M. J. Zhang, *Analyzing Competing Risk Data Using the R timereg Pac*, J. Stat. Softw. 38(2) (2011), 1-15.
29. T. M. Therneau, *A Package for Survival Analysis in R. R package version 3.2-3*, 2020, available at https://CRAN.R-project.org/package=survival.
30. T. M. Therneau, C. Crowson and E. Atkinson, *Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model,* Vignette for R survival package https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf, September, 2020.
31. T. M. Therneau and P. M. Grambsch, *Modelling Survival Data: Extending the Cox Model*, Springer, New York, 2000.
32. D. Zapletal and L. Kopecka, *Application of survival analysis to critical illness insurance data*, in *12th Scientific meeting classification and data analysis (CLADAG): Book of short papers*, Cassino, 2019, 472-475.

Table 1 Frequency table of the diagnoses

Diagnosis	Frequency	Percent	Cause	Frequency	Percent
Cancer	665	71.43	Cancer	665	71.43
Heart-attack	104	11.17	Other	266	28.57
Stroke	73	7.84			
Benign brain tumor	35	3.76			
Coronary artery operations	21	2.26			
Various other	33	3.54			
Total	931	100		931	100

Table 2 Frequency table of gender

Variable	Frequency	Ratios from frequencies (in %)			
Gender		**Censored**	**Cancer**	**Other**	**Cancer + Other**
Female	75408	99.32	0.57	0.11	0.68
Male	64555	99.36	0.36	0.28	0.64

Table 3 Frequency table of age

Variable	Frequency	Ratios from frequencies (in %)			
Age		**Censored**	**Cancer**	**Other**	**Cancer + Other**
18-30	46078	99.78	0.18	0.04	0.22
31-40	45944	99.47	0.37	0.16	0.53
41-50	29535	98.81	0.84	0.35	1.19
over 50	18406	98.73	0.87	0.40	1.27

Table 4 Frequency table of regions

Variable	Region	Frequency	Ratios from frequencies (in %)			
Gr_Region			**Censored**	**Cancer**	**Other**	**Cancer + Other**
1st group	PAK	5579	99.55	0.27	0.18	0.45
	PHA	10580	99.52	0.38	0.10	0.48
	LBK	4552	99.52	0.35	0.13	0.48
	ZLK	9964	99.51	0.39	0.10	0.49
2nd group	JHM	12922	99.42	0.43	0.15	0.58
	STC	10095	99.37	0.47	0.17	0.63
	ULK	7834	99.34	0.40	0.27	0.66
	JHC	8425	99.34	0.46	0.20	0.66
	VYS	7038	99.33	0.54	0.13	0.67
	PLK	8571	99.30	0.43	0.27	0.70
	HKK	7932	99.29	0.54	0.16	0.71
	OLK	11736	99.29	0.53	0.18	0.71
	MSK	27830	99.26	0.54	0.20	0.74
3rd group	KVK	6905	98.78	0.75	0.46	1.22

Table 5 Estimations of the Cox proportional hazard model

Variable	Level of Effect	Parameter Estimate	p-value	Hazard Ratio	95% Lower CI	95% Upper CI
Gender	female	0.027	0.678	1.028	0.903	1.170
Age	31-40	0.822	0.000	2.274	1.806	2.864
Age	41-50	1.680	0.000	5.367	4.307	6.687
Age	over 50	2.305	0.000	10.023	7.933	12.664
Gr_Region	2nd gr.	0.240	0.008	1.271	1.064	1.519
Gr_Region	3rd gr.	0.521	0.000	1.684	1.286	2.207

Table 6 Estimations of the cause specific hazard model for the cancer

Variable	Level of Effect	Parameter Estimate	p-value	Hazard Ratio	95% Lower CI	95% Upper CI
Gender	female	0.410	0.000	1.507	1.285	1.767
Age	31-40	0.687	0.000	1.987	1.530	2.580
Age	41-50	1.546	0.000	4.691	3.662	6.009
Age	over 50	2.147	0.000	8.558	6.558	11.169
Gr_Region	2nd gr.	0.204	0.053	1.226	0.997	1.507
Gr_Region	3rd gr.	0.376	0.027	1.456	1.044	2.029

Table 7 Estimations of the subdistribution hazard model for the cancer

Variable	Level of Effect	Parameter Estimate	p-value	Hazard Ratio	95% Lower CI	95% Upper CI
Gender	female	0.413	0.000	1.510	1.288	1.770
Age	31-40	0.686	0.000	1.990	1.528	2.580
Age	41-50	1.544	0.000	4.680	3.652	6.000
Age	over 50	2.141	0.000	8.510	6.508	11.130
Gr_Region	2nd gr.	0.203	0.054	1.230	0.996	1.510
Gr_Region	3rd gr.	0.373	0.027	1.450	1.043	2.020

Table 8 Estimations of the cause specific hazard model for the other diagnoses

Variable	Level of Effect	Parameter Estimate	p-value	Hazard Ratio	95% Lower CI	95% Upper CI
Gender	female	-0.920	0.000	0.398	0.308	0.515
Age	31-40	1.264	0.000	3.538	2.134	5.866
Age	41-50	2.123	0.000	8.354	5.116	13.641
Age	over 50	2.801	0.000	16.457	9.893	27.376
Gr_Region	2nd gr.	0.340	0.058	1.404	0.988	1.996
Gr_Region	3rd gr.	0.828	0.001	2.288	1.419	3.687

Table 9 Estimations of the subdistribution hazard model for the other diagnoses

Variable	Level of Effect	Parameter Estimate	p-value	Hazard Ratio	95% Lower CI	95% Upper CI
Gender	female	-0.923	0.000	0.397	0.307	0.514
Age	31-40	1.263	0.000	3.535	2.132	5.862
Age	41-50	2.118	0.000	8.317	5.094	13.578
Age	over 50	2.787	0.000	16.238	9.720	27.124
Gr_Region	2nd gr.	0.338	0.059	1.402	0.987	1.992
Gr_Region	3rd gr.	0.824	0.001	2.280	1.415	3.675

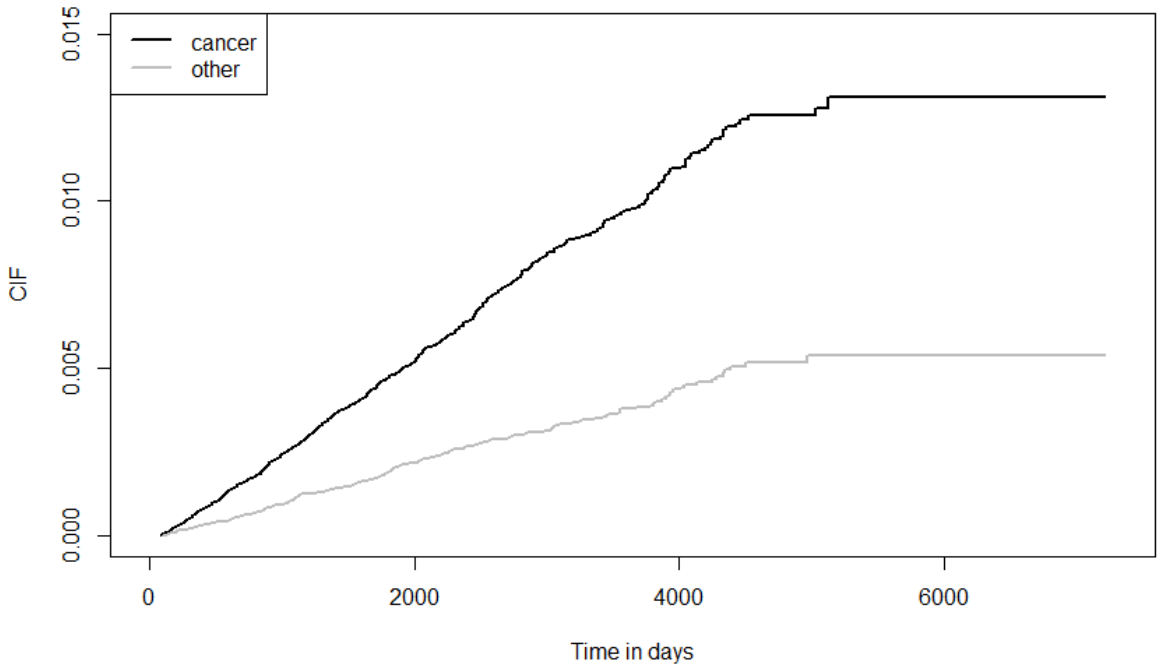

Figure 1 Cumulative incidence functions for the cancer and the other diagnoses