On a general structure for hazard-based regression models: an application to population-based cancer research

Francisco J. Rubio[1], Laurent Remontet[2], Nicholas P. Jewell[3], and Aurélien Belot[1]

Abstract

The proportional hazards model represents the most commonly assumed hazard structure when analysing time to event data using regression models. We study a general hazard structure which contains, as particular cases, proportional hazards, accelerated hazards, and accelerated failure time structures, as well as combinations of these. We propose an approach to apply these different hazard structures, based on a flexible parametric distribution (Exponentiated Weibull) for the baseline hazard. This distribution allows us to cover the basic hazard shapes of interest in practice: constant, bathtub, increasing, decreasing, and unimodal. In an extensive simulation study, we evaluate our approach in the context of excess hazard modelling, which is the main quantity of interest in descriptive cancer epidemiology. This study exhibits good inferential properties of the proposed model, as well as good performance when using the Akaike Information Criterion for selecting the hazard structure. An application on lung cancer data illustrates the usefulness of the proposed model.

1 Introduction

The analysis of time-to-event data has been dominated by the use of the semi-parametric Cox model during the last decades. While extensions to remove the assumption of *proportional hazards* (PH) were already discussed in Cox's original paper, a lot of work has been devoted to improve the flexibility of hazard-based regression models using flexible functions for both the baseline hazard and the inclusion of time-dependent parameters, mainly using splines or fractional polynomials[1–5]. However, the general structure of those models remained the same: the hazard function was expressed with a baseline hazard h_0 multiplied by the exponential of a flexible function g of time t and covariates $\mathbf{x}$: $h(t; \mathbf{x}) = h_0(t) \exp(g(t; \mathbf{x}))$. Alternative hazard structures that directly account for time-dependent effects of covariates have been proposed; one of the earliest alternative to the PH model is the *accelerated failure time* (AFT) model, in which the variables have a direct effect on the time to event, in contrast to the PH model where the covariates affect the hazard function[6]. More recently, in a series of papers[7–10], Y. Chen and co-authors proposed and studied semiparametric *accelerated hazards* (AH) models, where the covariates have a time-scaling effect on the hazard function, thus allowing for a time-varying effect.

In cancer epidemiology using population-based registry data, the quantity of interest is usually the excess hazard (of death) instead of the overall hazard[11–14]. The basic idea behind excess hazard models consists of decomposing the hazard function of death associated with an individual, $h(t; \mathbf{x})$, as the sum of an excess hazard, $h_E(t; \mathbf{x})$, and the general population hazard $h_P(t; \mathbf{z})$. The general population hazard is supposed to correctly reflect the other-causes hazard of death in our population of interest, assuming that the contribution in the general population hazard of the disease of interest (*e.g.* a specific cancer type) is small compared to all others. This excess hazard is interpreted as the hazard that could be due, directly or indirectly, to the cancer under study. Formally, this can be written as:

$$h(t; \mathbf{x}) = h_P(\text{age} + t; \text{year} + t; \mathbf{z}) + h_E(t; \mathbf{x}), \tag{1}$$

[1] Cancer Survival Group, Faculty of Epidemiology and Population Health,
Department of Non-Communicable Disease Epidemiology,
London School of Hygiene & Tropical Medicine,
London, WC1E 7HT, United Kingdom.
[2] Hospices Civils de Lyon,
Service de Biostatistique-Bioinformatique,
Lyon, 69003, France.
[3] University of California, Berkeley,
School of Public Health,
Berkeley, CA 94720-7360.

Corresponding author:
Francisco J. Rubio, Cancer Survival Group, Faculty of Epidemiology and Population Health,
Department of Non-Communicable Disease Epidemiology,
London School of Hygiene & Tropical Medicine,
London, WC1E 7HT, United Kingdom.
Email: Francisco.Rubio@lshtm.ac.uk

where $\mathbf{x}$ and $\mathbf{z}$ are vectors of covariates, $\mathbf{z}$ typically corresponding to a subset of covariates of $\mathbf{x}$, "age" is the age at diagnosis and "year" is the year of diagnosis (thus "age + t" and "year + t" are respectively the age and the year at time t after diagnosis). The population hazard $h_{\mathrm{P}}(\cdot; \mathbf{z})$ is typically obtained from national life tables based on the available sociodemographic characteristics ($\mathbf{z}$ in addition to age and year). A survival function derived from the excess hazard $h_{\mathrm{E}}(\cdot; \mathbf{x})$ is called the *net survival*. Net survival represents a useful way of reporting the probability of survival of cancer patients since this allows for a fairer comparison of survival rates between different populations or countries [15–18], and a non-parametric estimator of net survival has been proposed recently [16]. Although the interest in cancer epidemiology is on a slightly different quantity (the excess hazard instead of the overall hazard), the methodological developments of regression models have followed the same path than those described above [13,19–23] in terms of the hazard structure adopted.

Our aim is to provide a valuable supplement in the available toolbox for analysing survival data, and which is applicable for both overall and excess hazard regression models. The proposed approach builds on top of two recent developments: (i) the general parametrisation of hazard functions [7,8], combined with (ii) the use of a flexible parametric distribution for modelling times to event, the exponentiated Weibull distribution [24]. The paper is organised as follows. In Section 2, we describe the general hazard structure, discuss how the accelerated failure time model, the proportional hazards model, and the accelerated hazards model are nested within this general structure, and discuss the parameter interpretations for these models. In this section, we also introduce the exponentiated Weibull distribution, which will be used to model the "baseline" hazard in the general hazard structure (and the corresponding sub-models) based on its flexibility and ease of implementation. We also discuss maximum likelihood estimation and inference associated with these models. In Section 3, we present an extensive simulation study where we illustrate the inferential properties of the proposed models. In Section 4, we present a data example from lung cancer epidemiology. Here, we illustrate the usefulness of the proposed models and compare them against appropriate competitor approaches. We conclude with a general discussion and point to possible extensions in Section 5. Additional simulations and results are provided in the Supplementary Material.

2 Methods

2.1 The different model structures

We considered the following excess hazard structures (see [7],[8] and [9] for extensive discussion). We express the different structures below via the hazard function $h()$ and the cumulative hazard function $H()$, respectively, according to time t and a vector of covariates $\mathbf{x}_j$. We assume that the vector of covariates does not contain an intercept, in order to avoid identifiability issues. The survival function can be obtained from the well-known relationship $S(t) = \exp[-H(t)]$. The vector β denotes the unknown regression parameters.

(i) Proportional hazards model (PH).

$$
\begin{aligned}
h_{\mathrm{E}}^{\mathrm{PH}}(t; \mathbf{x}_j) &= h_0(t) \exp\left(\mathbf{x}_j^T \beta\right), \\
H_{\mathrm{E}}^{\mathrm{PH}}(t; \mathbf{x}_j) &= H_0(t) \exp\left(\mathbf{x}_j^T \beta\right).
\end{aligned}
\tag{2}
$$

(ii) Accelerated hazards model (AH).

$$
\begin{aligned}
h_{\mathrm{E}}^{\mathrm{AH}}\left(t; \mathbf{x}_j\right) &= h_0\left(t\exp(\mathbf{x}_j^T\beta)\right), \\
H_{\mathrm{E}}^{\mathrm{AH}}\left(t; \mathbf{x}_j\right) &= H_0\left(t\exp(\mathbf{x}_j^T\beta)\right)\exp(-\mathbf{x}_j^T\beta).
\end{aligned} \tag{3}
$$

(iii) Hybrid hazards model (HH).

$$
\begin{aligned}
h_{\mathrm{E}}^{\mathrm{HH}}\left(t; \mathbf{x}_j\right) &= h_0\left(t\exp(\mathbf{x}_{1j}^T\beta_1)\right)\exp(\mathbf{x}_{2j}^T\beta_2), \\
H_{\mathrm{E}}^{\mathrm{HH}}\left(t; \mathbf{x}_j\right) &= H_0\left(t\exp(\mathbf{x}_{1j}^T\beta_1)\right)\exp(-\mathbf{x}_{1j}^T\beta_1 + \mathbf{x}_{2j}^T\beta_2),
\end{aligned} \tag{4}
$$

where $\mathbf{x}_{1j} \subseteq \mathbf{x}_j$, and $\mathbf{x}_{2j} \subseteq \mathbf{x}_j$.

(iv) Accelerated failure time model (AFT).

$$
\begin{aligned}
h_{\mathrm{E}}^{\mathrm{AFT}}\left(t; \mathbf{x}_j\right) &= h_0\left(t\exp(\mathbf{x}_j^T\beta)\right)\exp(\mathbf{x}_j^T\beta), \\
H_{\mathrm{E}}^{\mathrm{AFT}}\left(t; \mathbf{x}_j\right) &= H_0\left(t\exp(\mathbf{x}_j^T\beta)\right).
\end{aligned} \tag{5}
$$

(v) General hazards model (GH).

$$
\begin{aligned}
h_{\mathrm{E}}^{\mathrm{GH}}\left(t; \mathbf{x}_j\right) &= h_0\left(t\exp(\mathbf{x}_j^T\beta_1)\right)\exp(\mathbf{x}_j^T\beta_2), \\
H_{\mathrm{E}}^{\mathrm{GH}}\left(t; \mathbf{x}_j\right) &= H_0\left(t\exp(\mathbf{x}_j^T\beta_1)\right)\exp(-\mathbf{x}_j^T\beta_1 + \mathbf{x}_j^T\beta_2).
\end{aligned} \tag{6}
$$

The assumptions behind each of these models are different in nature. The basic idea is to include effects that affect the level of the hazard (time-fixed effects) and the time scale (time-dependent effects) separately, as follows. In the PH model (2), a unit change in a covariate value have a multiplicative effect on the hazard, thus leading to a level change on the y-axis of the hazard. In the AH model (3), the effect of a unit change in a covariate affects the time scale of the baseline hazard (x-axis), thus assuming that there is a time-dependent effect of each covariate. In other words, $\exp(\beta)$ could be seen as a factor of how much more (or less) time is needed to reach the same hazard level, as compared to baseline, when the covariate is increased by one unit. In addition, the AH model does not assume that the parameters affect the hazard immediately at time $t = 0$, but "gradually", which is not the case with the other hazard structures[10]. While a "gradual" effect can be useful when estimating a treatment effect, it may not be justified for other variables; it highlights the usefulness of the HH structure (4). Indeed, the HH relaxes the assumption of the AH model by allowing some variables to have a proportional hazards effect rather than time-dependent "gradual" effects. In the AFT model (5), the effect is on the survival time directly as it can be, in fact, formulated as a log-linear regression model on survival times[6]. In such models, the estimated regression parameter for a unit change in one covariate accelerates or decelerates the event time, thus leading to a time-dependent effect. Notice that the AFT, PH, and AH models coincide for the case when the baseline hazard is Weibull[8]. The GH model (6) represents a general hazard structure that contains, as particular cases, the PH, AH, HH, and AFT models. More specifically, if $\beta_1 = 0$, then GH = PH; if $\beta_2 = 0$, then GH = AH; and if $\beta_1 = \beta_2$, then GH = AFT.

2.2 The Exponentiated Weibull distribution

We propose modelling the baseline hazard $h_0(t)$ using the Exponentiated Weibull (EW) distribution. The Exponentiated Weibull distribution is simply obtained by exponentiating the Weibull cumulative distribution function to an unspecified positive power[25]. This simple transformation adds a second shape parameter that, interestingly, induces considerable flexibility to the hazard function. The hazard function of the Exponentiated Weibull distribution can capture several basic shapes: constant, increasing, decreasing, bathtub, and unimodal (See Figure 1), making it appealing for survival models[26]. This distribution has been recently used in the context of AFT models[27], while alternative families of flexible parametric AFT models have also been studied in[28,29]. The Exponentiated Weibull probability density and cumulative distribution functions with scale, shape, and power parameters (σ, κ, α) are given, respectively, by:

$$
\begin{aligned}
f_{\mathrm{EW}}(t; \sigma, \kappa, \alpha) &= \alpha \frac{\kappa}{\sigma} \left(\frac{t}{\sigma} \right)^{\kappa-1} \left[1 - \exp\left\{ -\left(\frac{t}{\sigma} \right)^{\kappa} \right\} \right]^{\alpha-1} \exp\left\{ -\left(\frac{t}{\sigma} \right)^{\kappa} \right\}, \\
F_{\mathrm{EW}}(t; \sigma, \kappa, \alpha) &= \left[1 - \exp\left\{ -\left(\frac{t}{\sigma} \right)^{\kappa} \right\} \right]^{\alpha},
\end{aligned}
\tag{7}
$$

where $t, \sigma, \kappa, \alpha > 0$. This distribution reduces to the Weibull distribution for $\alpha = 1$. The corresponding hazard function is obtained, by definition, as $h_{\mathrm{EW}}(t; \sigma, \kappa, \alpha) = f_{\mathrm{EW}}(t; \sigma, \kappa, \alpha) / \left[1 - F_{\mathrm{EW}}(t; \sigma, \kappa, \alpha) \right]$. The EW distribution is identifiable (See Proposition 1 from the Appendix), and general results about the identifiability of the different hazard-based models (2)–(6) are presented in[8]. It was shown that the GH model is identifiable except for the case when the baseline hazard is Weibull[8]. This is not an issue since, as already mentioned, it corresponds to the case when the AFT, PH, and AH models coincide. Moreover, it has also been shown that the AFT and AH models are not identifiable when the baseline hazard is flat[8] (*i.e.* exponential), which corresponds to a case when the shape of the baseline hazard does not change with time, a case of little interest in practice.

2.3 Parameter interpretation

Some clarifications on the interpretation of the parameters estimated from the AH and GH models seem appropriate as these models are not as well known as PH and AFT models. Notice that these interpretations directly translate to the HH model as this model is a special case of the GH model.

We start by interpreting the parameters for the AH model (3), as this will facilitate the interpretation for the GH model. The parameter interpretation depends on the shape of the baseline hazard, which we classify here as monotone (increasing/decreasing) or not (bathtub/unimodal).

For monotonic baseline hazard, a positive value of β for one unit change in the covariate x means that x has a harmful (beneficial) effect if the baseline hazard is increasing (decreasing) (see panels (a) and (b) in Figures S3- S4). A negative value of β means that x has a beneficial (harmful) effect if the baseline hazard is increasing (decreasing) (see panels (a) and (b) in Figures S1-S2). In other words, a positive value of β accelerates the progression of the hazard, which is beneficial for the patients if the hazard decreases and harmful if it increases. On the contrary, a negative value of β decelerates the progression of the hazard, which is beneficial for the patients if the hazard increases and harmful if it decreases.

If the shape of the baseline hazard is unimodal (or bathtub), a positive value of β accelerates the evolution of the hazard, thus the maximum (minimum) will be reached sooner (see panels (c) and (d)

in Figures S3-S4). A negative value of β decelerates the evolution of the hazard, thus the maximum (minimum) will be reached later (see panels (c) and (d) in Figures S1-S2).

From the AH model, it is worth noticing that the general shape will not change (a unimodal shape will remain unimodal) and that the peak/minimum (if any) reached by the hazards defined by different subgroups will be at the same level. On the other hand, when using the GH model (6) the parameter β_1 is directly multiplying the time t, thus re-scaling the timescale (accelerating or decelerating, *i.e.* they play a role on the x-axis, changing the pace of the hazard's progression), while the parameter β_2 is modifying the level of the hazard (role on the y-axis, changing the magnitude of the hazard, Figure 1). This can be clearly seen in the right panels of Figure 1 for the unimodal shape: in panels (a) and (b) β_1 is negative, thus the peak is reached later for patients' group with $x = 1$ (dot-dashed grey lines) compared to the baseline group (solid grey lines), while the parameter β_2 changes the magnitude of the hazard (thus the level of the peak). In panels (c) and (d), β_1 is positive, thus the peak is reached sooner for patients' group with $x = 1$ compared to the reference group, while the peak level is changed according to β_2. The same interpretation applies for bathtub hazards (black lines in the right panels in Figure 1), and for monotonic hazards (left panels in Figure 1), even though for these later the interplay of both parameters β_1 and β_2 is less obvious to see on the graphs.

The AH and GH models allow a crossover of the hazards (and also of the survival functions). In some cases, this advantage may lead to difficulties in interpreting clearly the parameters[30]. Thus, plotting the hazards according to different covariate patterns is recommended to help clarifying the time-to-event process compared to reporting only the survival functions. Indeed, the survival probability, being a cumulative measure, does not help visualising particular features of the instantaneous process[31].

2.4   The likelihood function

Let $(t_j, \mathbf{x}_j, \delta_j)$, $j = 1, \dots, n$ be a sample of times to event from a population of cancer patients $t_j > 0$, with covariates $\mathbf{x}_j \in \mathbb{R}^p$, and vital status indicators δ_j (1-death, 0-censored). Let also $h_{\mathrm{P}}(\mathrm{age}_j + t_j; \mathrm{year}_j + t_j; \mathbf{z}_j)$ be the corresponding population hazard rates obtained from the national life tables based on the variables $\mathbf{z}_j \in \mathbb{R}^q$, $q \leq p$, where age_j represents the age at diagnosis and year_j the year of diagnosis of patient j. The likelihood function of the full vector of parameters $\boldsymbol{\psi}$ is then given by

$$
\begin{aligned}
\mathcal{L}_0(\boldsymbol{\psi}; \mathrm{Data}) &= \prod_{j=1}^{n} h(t_j; \mathbf{x}_j)^{\delta_j} S(t_j; \mathbf{x}_j), \\
&\propto \prod_{j=1}^{n} \left\{ h_{\mathrm{P}}(\mathrm{age}_j + t_j; \mathrm{year}_j + t_j; \mathbf{z}_j) + h_{\mathrm{E}}(t_j; \mathbf{x}_j) \right\}^{\delta_j} \exp\left\{ -H_{\mathrm{E}}(t_j; \mathbf{x}_j) \right\}.
\end{aligned}
$$

Notice that we have removed the term $\exp\left\{ -H_{\mathrm{P}}(\mathrm{age}_j + t_j; \mathrm{year}_j + t_j; \mathbf{z}_j) + H_{\mathrm{P}}(\mathrm{age}_j; \mathrm{year}_j; \mathbf{z}_j) \right\}$ from the likelihood as it does not depend on the parameters $\boldsymbol{\psi}$.

In order to obtain confidence intervals for the parameters, we reparameterise the baseline EW hazard in terms of $(\tilde{\kappa}, \tilde{\sigma}, \tilde{\alpha}) = (\log\{\kappa\}, \log\{\sigma\}, \log\{\alpha\})$. Appealing to the consistency and asymptotic normality of the maximum likelihood estimators (MLEs) of the EW distribution, and the availability of large samples in the context of cancer epidemiology, we propose the use of asymptotic confidence intervals of the type $\widehat{\boldsymbol{\psi}} \pm Z_{1-\frac{\tau}{2}} \operatorname{diag}\left(J^{-\frac{1}{2}}(\widehat{\boldsymbol{\psi}}) \right)$, where $J(\boldsymbol{\psi}) = -\dfrac{\partial^2}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T} \log \mathcal{L}_0(\boldsymbol{\psi}; \mathrm{Data})$ is the negative of the
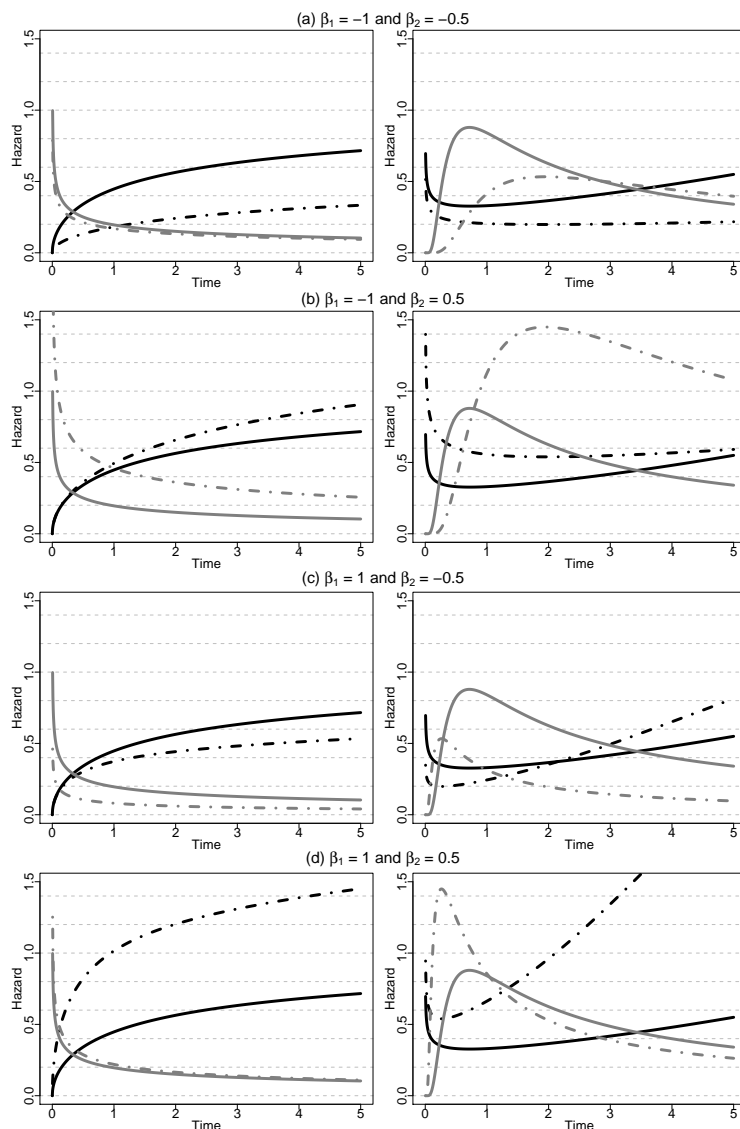
Figure 1. Hazard shapes obtained with different values of the parameters defining the Exponentiated Weibull distribution (σ, κ, α), and different values of the regression coefficients β_1 and β_2 for a binary covariate x. Monotonic hazard (increasing in black and decreasing in grey) are displayed on the left column, and bathtub or unimodal (black or grey, respectively) on the right column, where solid lines represent the baseline hazard (*e.g.* unexposed $x = 0$) and dot-dashed lines represent the hazard for the exposed ones ($x = 1$). The values of (σ, κ, α) are: $(2, 1.2, 1.25)$, $(3.8, 0.5, 1.5)$, $(4.5, 2, 0.4)$, $(0.0002, 0.21, 300)$ for monotonic increasing, monotonic decreasing, bathtub and unimodal shapes, respectively.

Hessian matrix of the log-likelihood function under the appropriate parameterisation, and $1 - \tau \in (0, 1)$ is the confidence level.

Confidence intervals for the net survival curve at specific time-points are obtained using a simulation-based algorithm[32]. The idea is to simulate from the asymptotic (multivariate normal) distribution of the parameters in order to obtain a Monte Carlo sample of the net survival at specific time-points, which is used to construct the corresponding confidence intervals.

To measure relative goodness-of-fit of a hazard-based model structure among the ones detailed previously, we employ the Akaike Information Criterion (AIC):

$$\text{AIC} = -2 \log \mathcal{L}_0(\widehat{\boldsymbol{\psi}}; \text{Data}) + 2k,$$

where k is the number of parameters (the dimension of $\boldsymbol{\psi}$), and $\widehat{\boldsymbol{\psi}}$ is the MLE of $\boldsymbol{\psi}$.

3 Simulation study

In this section, we present an extensive simulation study where we illustrate the good frequentist properties of the proposed models as well as the ability of the AIC criterion to select models that properly capture the underlying hazard structure. The parameter values are chosen in order to produce scenarios that resemble cancer population studies concerning an aggressive type of cancer (relatively low 5-year survival), such as lung cancer.

3.1 Data generation and simulations designs

We simulated $N = 1000$ data sets of size $n = 1000, 5000, 10000$, assuming the additive hazard decomposition given in (1). The variable "age" was simulated using a mixture of uniform distributions with 0.25 probability on $(30, 65)$, 0.35 probability on $(65, 75)$, and 0.40 probability on $(75, 85)$ years old. The binary variables "sex" and "W" were both simulated from a binomial distribution with probability 0.5 (the variable "W" could be viewed as "treatment" or "comorbidity"). In all the scenarios, we simulated the "other-causes" time to event using the UK life tables based on "age" and "sex", assuming all patients were diagnosed on the same year. The time to event from the excess hazard (cancer event time) was generated using the inverse transform method[33], and assuming effects of the 3 variables "age", "sex" and "W". We assumed either (i) only administrative censoring at $T_C = 5$ years, which induced approximately 30% censoring in all cases, or (ii) an additional independent random censoring (drop-out) using an exponential distribution with rate parameter r. In this latter case, we choose values for r to induce around 50% censoring. We considered 6 generating mechanisms for the excess hazard, all of them with EW baseline hazards: (i) Proportional Hazards (Scenario PH), (ii) Accelerated Failure time (Scenario AFT), (iii) Accelerated Hazard (Scenario AH), (iv) Hybrid Hazard (Scenario HH), (v) General Hazard (Scenario GH), and (vi) a hybrid hazard model where the hazards and the survival curves associated to the variable "W" cross (Scenario CH). The values of the model parameters used for each scenario are summarised in Table S1 from the Appendix. Moreover, tables S2–S7 from the Appendix present the 1-year and 5-year net survival implied by the different combinations of hazard structures and covariate values.

3.2 Analysis of the simulated data

To assess the frequentist properties of the models in all simulation scenarios, we fitted the GH model with EW baseline hazard (GHEW), as compared to the corresponding true generating model. Additionally, to

measure the ability of the AIC to select the true generating model, we also fitted the remaining models in each scenario. In all cases, we used the true parameter values of the generating model as initial points in the optimisation step because our aim was more to study the estimator's properties rather than the properties of the optimisation process. However, as the interest of analysts is also on the overall properties (*i.e.* including the optimisation process properties in the full process of estimation), we present in the Appendix a thorough study where we present three alternative "automatic" choices for the initial points which we will later use in the real data example (see section 5 in the Appendix). We present results about the convergence using these initial points as well as the resulting estimators, which are virtually the same as those obtained with the chosen initial points, as expected, with the expense of a higher computing time.

We performed the analysis using R software. The optimisation step was conducted using the R commands 'nlminb()' and 'optim()', while the Hessian matrix used in the construction of the asymptotic confidence intervals are approximated using the command 'hessian()' from the R package 'numDeriv', which represents an efficient method to calculate the Hessian matrix. The cases where the command 'hessian()' produced 'Inf' or 'NaN' values were excluded as this merely represents a numerical problem associated to numerical differentiation, which mainly affects the most difficult cases (*i.e.* with $n = 1000$ and 50% censoring).

3.3   Measures of performance

We report the mean and median of the MLEs for the corresponding models, as well as the empirical standard deviation, the mean (estimated) standard error, the root-mean-square error, and coverage proportions of asymptotic confidence intervals. In addition, we report the proportion of times the different fitted models are selected using AIC. The excess hazard functions associated with the best models selected using AIC are plotted in Figures S5–S14, for different covariate patterns, and compared to the true generating model.

3.4   Simulation results

The results are presented in Tables S8–S29 in the Appendix. For illustrative purposes, we present in Tables 1–2 and Figure 2 the results for Scenario GH with $n = 5000$. In this scenario, we observe very good performance of our approach, with very small bias and a coverage close to the nominal value of 95% (Table 1 and Figure 2). Regarding performances of model selection using the AIC, the true model was selected in around 1/3 of the simulated datasets in the worst case (sample size of 1000 and 50% of censoring), but this proportion increases to 90% or more in situations with a larger sample size (Table 2). More generally, from this simulation study, we can conclude that the MLEs are close to the true values for moderate samples, even in the case of a high censoring rate (Tables S8, S12, S16, S20, S24, S26) and that the coverage is usually quite close to 95%. As expected, the RMSE decreases as the sample size increases (or with identical sample size but lower censoring percentage). For $n = 5000$, we observe low bias and variance of the MLEs, and that the AIC selects the correct model with a high proportion. For $n = 10000$, the AIC selects the true model in more than 85% of cases for all scenarios and whatever the censoring level. Interestingly, in the GH scenario, as long as the sample size was equal to 5000 or larger, the AIC selected the true model in 90% or more samples, regardless of the censoring level. Moreover, even in cases where the incorrect model is selected using AIC, we see that the baseline hazard is close to

the true generating model, reflecting that the AIC selects the model closest to the true generating model (see figures S5–S14).

Model	Parameter	MMLE	mMLE	ESD	Mean Std Error	RMSE	Coverage
				30% censoring			
	σ (1.75)	1.747	1.726	0.405	0.409	0.405	0.950
	κ (0.6)	0.601	0.600	0.063	0.064	0.063	0.946
	α (2.5)	2.576	2.523	0.461	0.459	0.467	0.947
GHEW	β_{11} (0.1)	0.100	0.100	0.013	0.013	0.013	0.961
	β_{12} (0.1)	0.112	0.109	0.233	0.227	0.233	0.949
	β_{13} (0.1)	0.099	0.104	0.230	0.228	0.230	0.956
	β_{21} (0.05)	0.050	0.050	0.003	0.003	0.003	0.958
	β_{22} (0.2)	0.201	0.202	0.041	0.042	0.041	0.951
	β_{23} (0.25)	0.251	0.251	0.043	0.042	0.043	0.943
				50% censoring			
	σ (1.75)	1.741	1.736	0.457	0.465	0.457	0.957
	κ (0.6)	0.600	0.600	0.074	0.076	0.074	0.948
	α (2.5)	2.601	2.508	0.556	0.539	0.565	0.963
GHEW	β_{11} (0.1)	0.102	0.101	0.016	0.016	0.016	0.959
	β_{12} (0.1)	0.090	0.095	0.267	0.266	0.267	0.957
	β_{13} (0.1)	0.092	0.093	0.269	0.268	0.269	0.948
	β_{21} (0.05)	0.050	0.050	0.004	0.004	0.004	0.945
	β_{22} (0.2)	0.201	0.200	0.045	0.047	0.045	0.956
	β_{23} (0.25)	0.252	0.251	0.049	0.048	0.049	0.949

Table 1. Simulation results for the scenario GH with $(\sigma, \kappa, \alpha) = (1.75, 0.6, 2.5)$, $\beta_1 = (0.1, 0.1, 0.1)$, $\beta_2 = (0.05, 0.2, 0.25)$, and $n = 5000$. Mean of the MLEs (MMLE), median of the MLEs (mMLE), empirical standard deviation (ESD), mean (estimated) standard error, root-mean-square error (RMSE), and coverage proportions (Coverage).

4 Real data example

To illustrate the new proposed models, we analysed a dataset obtained from population-based national cancer registry of lung cancer patients diagnosed in 2012 in the United-Kingdom. We linked these data to administrative data (Hospital Episode Statistics -HES- and Lung Cancer Audit data -LUCADA-) and applied specific algorithms to derive information on stage at diagnosis and presence of comorbidities at the time of diagnosis [34,35]. To retrieve information on comorbidity, we used a 6-year period up to 6 months before diagnosis where we checked for the presence of any of 18 comorbidities that are used to define the Charlson Comorbidity Index (CCI), in addition to obesity [35]. The information was then dichotomised into 2 categories: "no comorbidity" *vs.* "at least 1 comorbidity (comorbidity indicator = 1)" in our illustrative example. We measured deprivation using the Income Domain from the 2010 England Indices of Multiple Deprivation, defined at the Lower Super Output Area level (mean population 1500). The Income Domain measures the proportion of the population in an area experiencing deprivation related to low income, and ranges from 0 to 77% (https://www.gov.uk/government/statistics/

Model	30% censoring	50% censoring
$n = 1000$		
PHEW	1.6	2.7
AHEW	0	0
AFTEW	56.5	64.5
GHEW	41.9	32.8
$n = 5000$		
PHEW	0	0
AHEW	0	0
AFTEW	3.2	10
GHEW	96.8	90
$n = 10000$		
PHEW	0	0
AHEW	0	0
AFTEW	0	0.5
GHEW	100	99.95

Table 2. Percentage of models selected with AIC in the scenario GH with $(\sigma, \kappa, \alpha) = (1.75, 0.6, 2.5)$, $\beta_1 = (0.1, 0.1, 0.1)$, and $\beta_2 = (0.05, 0.2, 0.25)$.

english-indices-of-deprivation-2010). Follow-up was assessed on the 31st of December 2015, at which time patients alived were censored (so the maximum follow-up was 4 years). We restricted our analysis to women with no missing data, and applied the PH model (2), the AH model (3), the AFT model (5) and the GH model (6), with an Exponentiated Weibull distribution for the baseline hazard. The variables included in the models are 'agediagc' (centred age at diagnosis), 'Istage2', 'Istage3', and 'Istage4' (stage at diagnosis), 'INCOME_SCORE_2015c' (centred Income Domain), 'comorbidity' (presence of comorbidity). The time dependent effects are indicated in Table 3 with the subindex 't'.

We observed $n = 14557$ patients with complete cases among which $n_o = 12138$ died before the 31st of December 2015. The median follow-up among patients censored was 3.46 years. The 25%, 50% and 75% quantiles of the patients' age at diagnosis was 64.9, 72.6, 80.2 while the mean was 72.0. Among the patients, 2434 were Stage I, 1131 were Stage II, 3421 were Stage III, and 7751 were Stage IV. Finally, 4318 patients were classified with comorbidity indicator 1. Results of this analysis are presented in Table 3. From this table, we observe that the GHEW is clearly favoured by the AIC (followed by the AFTEW model), thus suggesting the need for including time-dependent as well as proportional effects. The signs of all the estimates are positive in this model. This implies that an increase of one unit in the value of any of the covariates leads to an acceleration of the time to reach the maximum of the hazard as well as an increase of the maximum of the hazard. The magnitude of such acceleration is given by the value of the corresponding estimated parameter, and we noticed that these two effects are different for each variable thus explaining the better fit of the GH model compared to the AFT model. We have compared the GHEW model against alternative models with fewer covariates but they were not favoured by the AIC. For comparison, we have also used the 'mexhaz' R package [23,36] to produce models with time-dependent effects (using the command nph) for some or all the variables and we have found that none of these models provided a better fit than the GHEW in terms of AIC (with 'mexhaz',
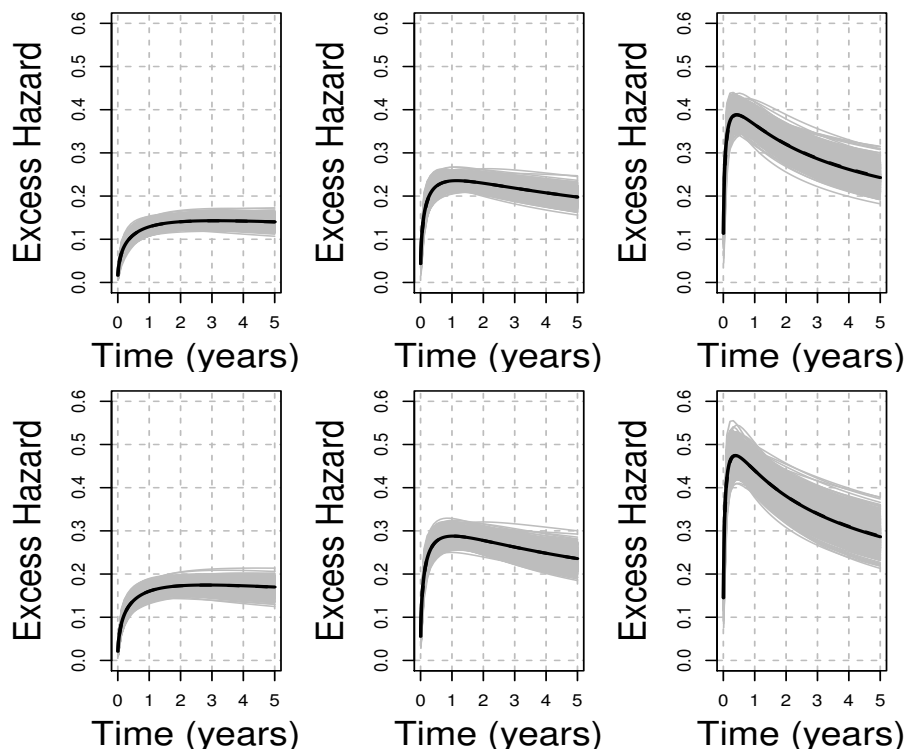
Figure 2. Mean of the best fitted hazards in terms of AIC (dashed lines), compared to the true generating hazard (continuous lines), and 1000 sample-specific fitted hazards (grey lines) in the scenario GH for $n = 5000$ and 30% censoring. Panels from left to right correspond to covariate values (age, sex, comorbidity)=$(60, 0, 0), (70, 0, 0), (80, 0, 0), (60, 1, 0), (70, 1, 0), (80, 1, 0)$, respectively. The dashed and continuous lines are virtually the same.

the model with the lowest AIC ($= 20270.44$) was the model assuming time-dependent effects for all variables). Figures 3a and 3b show the shapes of the hazard functions associated to the different models for patients aged 70 years at diagnosis, an income score value of 0.15 and with either a stage II or a stage IV cancer, and without comorbidity (panel (a)) or with comorbidity (panel (b)). These figures also show the piecewise excess hazard estimated separately on each of the 4 groups defined by the following values of covariate: age $\in [65, 75]$, Income Score $\in [0.1, 0.2]$, Stage II or Stage IV, and Comorbidity $= 0$ or 1. Those piecewise hazards represent a way to check the quality of the fit of the different models[13]. These figures suggest a good fit of the GH model overall, while the excess hazards obtained from the AH model are always over-estimated after 6 months for the group of stage IV patients. For the group of stage IV patients with comorbidity (panel (b)), the PH model does not capture the high excess hazard just after the diagnosis and the sharp decrease that follows.

Model	PHEW	AHEW	AFTEW	GHEW
scale	0.059 (0.038)	8.482 (0.724)	1.190 (0.175)	1.838 (0.374)
shape	0.188 (0.014)	0.539 (0.046)	0.385 (0.012)	0.442 (0.033)
power	9.175 (1.420)	1.483 (0.129)	4.387 (0.312)	3.593 (0.368)
agediagc$_t$	–	-0.112 (0.006)	0.032 (0.001)	0.041 (0.004)
Istage2$_t$	–	-2.977 (0.282)	0.881 (0.065)	0.691 (0.311)
Istage3$_t$	–	-6.680 (0.337)	1.909 (0.050)	1.707 (0.229)
Istage4$_t$	–	-10.469 (0.416)	3.003 (0.046)	3.413 (0.226)
INCOME_SCORE_2015c$_t$	–	-2.668 (0.416)	0.744 (0.115)	0.822 (0.448)
comorbidity$_t$	–	-1.021 (0.106)	0.289 (0.029)	0.539 (0.114)
agediagc	0.022 (0.001)	–	– ($\star$)	0.034 (0.001)
Istage2	0.721 (0.056)	–	– ($\star$)	0.845 (0.069)
Istage3	1.473 (0.043)	–	– ($\star$)	1.849 (0.053)
Istage4	2.211 (0.041)	–	– ($\star$)	3.073 (0.050)
INCOME_SCORE_2015c	0.527 (0.085)	–	– ($\star$)	0.750 (0.150)
comorbidity	0.192 (0.021)	–	– ($\star$)	0.349 (0.039)
AIC	20523.141	20855.753	20189.124	**20164.911**

Table 3. Maximum likelihood estimates of the parameters (standard errors) for the different excess hazard models, with their corresponding AIC. The time dependent effects are indicated with the subindex 't'. ($\star$) By construction, the effects of covariates are constrained to be the same for the time-dependent and time fixed effects in the AFT model (5). See equations (2)-(6) for more details on the different hazard structures.

Table 4 presents the net survival at $1, 2, 3, 3.9$ years, for the total population (Comorbidity = 0, 1) and for a subgroup (age-group 55-65, Stage I, and Comorbidity = 0, 1,), using the GHEW model and the Pohar-Perme nonparametric estimator[16]. The confidence intervals for the net survival in the GHEW model were obtained using the simulation-based algorithm described in Section 2.4 using $B = 10,000$ samples from the asymptotic distribution of the parameters. The Pohar-Perme estimator and its corresponding confidence intervals were calculated using the 'relsurv' R package. We observe that the results with the parametric and nonparametric approaches are very close, and that the confidence intervals for the parametric model are slightly shorter (as expected), which shows that the proposed parametric model can accurately capture the underlying hazard structure.
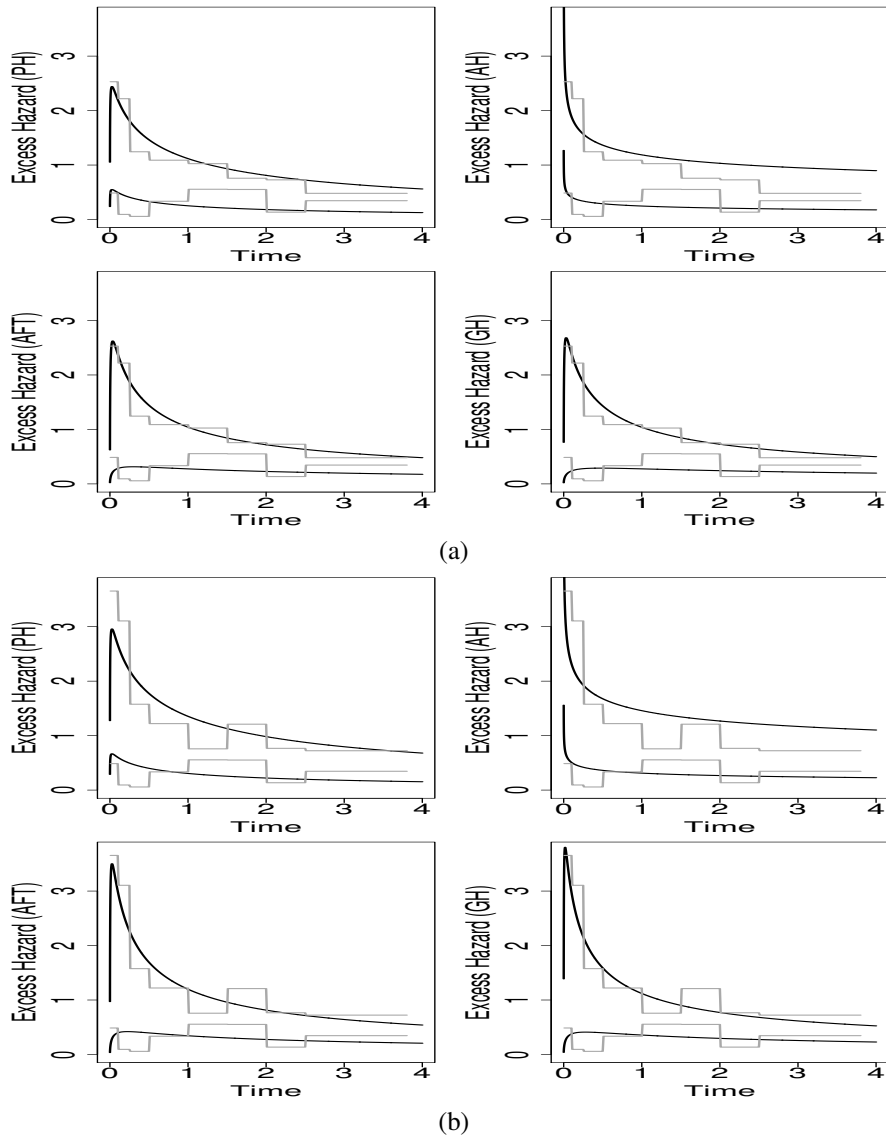
(a)

(b)

Figure 3. Shapes of the EW excess hazard for PH, AFT, AH, and GH models with baseline covariate values (solid lines) *vs.* Piecewise excess hazard (grey line): (a) comorbidity = 0 and Stage II (the 2 lowest curves/step functions in each cell), comorbidity = 0 and Stage IV (the 2 highest curves/step functions in each cell); and (b) comorbidity = 1 and Stage II (the 2 lowest curves/step functions in each cell), comorbidity = 1 and Stage IV (the 2 highest curves/step functions in each cell).

Total population by comorbidity							
		GHEW			Pohar-Perme		
Comorb.	year	NS	lower	upper	NS	lower	upper
0	1	0.407	0.401	0.416	0.408	0.398	0.418
	2	0.270	0.265	0.278	0.268	0.259	0.277
	3	0.204	0.199	0.211	0.209	0.201	0.218
	3.9	0.167	0.162	0.174	0.182	0.172	0.191
1	1	0.380	0.371	0.391	0.370	0.356	0.385
	2	0.254	0.246	0.264	0.238	0.225	0.252
	3	0.193	0.185	0.203	0.169	0.158	0.182
	3.9	0.158	0.150	0.168	0.138	0.126	0.152
Age group 55-65 at Stage I by comorbidity							
Comorb.	year	NS	lower	upper	NS	lower	upper
0	1	0.924	0.914	0.935	0.928	0.897	0.960
	2	0.842	0.827	0.860	0.837	0.794	0.883
	3	0.770	0.752	0.791	0.797	0.749	0.847
	3.9	0.712	0.692	0.737	0.762	0.705	0.823
1	1	0.881	0.869	0.896	0.901	0.851	0.953
	2	0.772	0.754	0.793	0.744	0.674	0.821
	3	0.684	0.662	0.710	0.670	0.595	0.755
	3.9	0.618	0.595	0.647	0.627	0.541	0.725

Table 4. Net survival (NS) at $T = 1, 2, 3, 3.9$ years and 95% confidence intervals, estimated using the GHEW model and the non parametric Pohar-Perme estimator.

5   Discussion

We have studied a general parametric hazard structure that can capture the basic shapes of the baseline hazard and that contains, as particular cases, the main models of interest in survival analysis: proportional hazards, accelerated failure time, and accelerated hazards models. PH and AFT models already enjoy popularity in survival analysis. However, the limited application of the AH model is mainly due to the lack of efficient and reliable estimation methods[10]. We have shown that, by assuming a flexible parametric distribution function such as the EW, it is possible to conduct classical likelihood inference using already available optimisation algorithms. The combination of such flexible parametric hazard function with the GH structure represents a powerful tool for modelling survival times. Although we have focused on the context of *excess hazard models*, the proposed flexible parametric GH model is also

applicable in the context of *overall survival*. We have employed the EW distribution for modelling the baseline hazard in the GH model (6), however, there exist other flexible parametric distributions, such as the generalised Gamma [26] and the generalised Weibull [24] distributions, that can also capture the basic shapes of the hazard function. Here, we only employed the EW distribution as this choice allows for a parsimonious implementation, facilitates the interpretation of the parameters, and the corresponding maximum likelihood estimators (MLEs) are consistent and asymptotically normal in the presence of censored observations [37]. The simulation study shows that the proposed model has good frequentist properties and that the selection of an appropriate model structure is feasible for large enough samples. The proposed model can also capture cases where the hazard and the survival curves cross, a case of great interest in practice. Cases when the AH and AFT structures produce crossing hazards have been studied in [38], and these are also illustrated in our simulation study. Despite the flexibility of the EW distribution, there still exist baseline hazard shapes that cannot be captured by this model (such as multimodal hazards). We have conducted additional simulation studies in order to assess the effect of model misspecification on the estimation of net survival. We found that in cases where the EW distribution cannot capture the true shape of the baseline hazard, the net survival functions associated to the fitted models tend to be relatively close to the true model, and that the AIC selects the correct hazard structure with high proportion, provided that the departures from the shapes the EW can capture are moderate. If the departures are severe, this may, unsurprisingly, affect the selection of the correct hazard structure but the selected model tend to resemble the shape of the true generating model (see [39] for a general study of fitting parametric survival models under possible misspecification). These additional studies are available from the authors' websites, as well as the R code for fitting the models detailed in this paper.

Although we have centered our attention on the GH structure (6), we point out the more general representation discussed in [9]:

$$h_{\mathrm{E}}(t; \mathbf{x}_j) = \Lambda\left(h_0(t), \mathbf{x}_j^T \beta\right),$$

where $\Lambda(\cdot)$ is a known non-negative function that defines the relationship between the baseline hazard and the covariates. This representation contains, for instance, the additive hazards model [40] $h_{\mathrm{E}}^{\mathrm{AD}}(t; \mathbf{x}_j) = h_0(t) + \mathbf{x}_j^T \beta$, which we do not consider here as it requires additional conditions to guarantee that $h_{\mathrm{E}}^{\mathrm{AD}} > 0$; as well as structures including non-linear relationships. We also point out that the use of splines for including non-linear effects can be coupled with any of the aforementioned hazard structures. A potentially useful choice are B-splines, which are implemented in R.

We have employed the AIC for selecting the hazard structure that better fits the data since this criterion can account for possible model misspecification, as this tool asymptotically selects the model that minimises the Kullback-Leibler divergence between the fitted models and the true generating model, under mild regularity conditions. We point out that other criteria such as the Bayesian Information Criterion (BIC) or cross validation can be employed instead. An interesting feature of the GH model is that, by selecting the active variables, we automatically select the hazard structure that better fits the data. The study of efficient variable selection methods in the general structure (6) would allow the identification of active variables as well as the underlying structure (PH, AH, HH, AFT, or GH) of the hazard function. This points out an interesting research line. The simulation study also illustrates the importance of accounting for sparsity, as the inclusion of spurious variables may bias the estimates of the active variables and the parameters of the baseline hazard. Since, in recent years, more variables such

as comorbidities and types of treatments have become available at the population level, we believe this topic will become relevant in cancer epidemiology.

Acknowledgements

References

1. Durrleman S and Simon R. Flexible regression models with cubic splines. *Statistics in Medicine* 1989; 8: 551–561.
2. Sleeper L and Harrington D. Regression Splines in the Cox Model with Application to Covariate Effects in Liver Disease. *Journal of the American Statistical Association* 1990; 85: 941–949.
3. Hess K. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Statistics in Medicine* 1994; 13: 1045–1062.
4. Abrahamowicz M, Mackenzie T and Esdaile J. Time-Dependent Hazard Ratio: Modeling and Hypothesis Testing with Application in Lupus Nephritis. *Journal of the American Statistical Association* 1996; 91: 1432–1439.
5. Royston P and Sauerbrei W. *Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables*, volume 777. John Wiley & Sons, 2008.
6. Kalbfleisch J and Prentice R. *The Statistical Analysis of Failure Time Data*, volume 360. John Wiley & Sons, 2011.
7. Chen Y and Wang M. Analysis of accelerated hazards models. *Journal of the American Statistical Association* 2000; 95(450): 608–618.
8. Chen Y and Jewell N. On a general class of semiparametric hazards regression models. *Biometrika* 2001; 88(3): 687–702.
9. Chen Y, Jewell N and Yang J. Accelerated hazards model: method, theory and applications. *Handbook of Statistics* 2003; 23: 431–441.
10. Chen Y, Hanson T and Zhang J. Accelerated hazards model based on parametric families generalized with Bernstein polynomials. *Biometrics* 2014; 70(1): 192–201.
11. Hakulinen T and Tenkanen L. Regression analysis of relative survival rates. *Applied Statistics* 1987; : 309–317.
12. Esteve J, Benhamou E, Croasdale M et al. Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine* 1990; 9(5): 529–538.
13. Remontet L, Bossard N, Belot A et al. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in Medicine* 2007; 26(10): 2214–2228.
14. Mariotto A, Noone A, Howlader N et al. Cancer survival: an overview of measures, uses, and interpretation. *Journal of the National Cancer Institute Monographs* 2014; 2014(49): 145–186.

15. Gamel J and Vogel R. Non-parametric comparison of relative versus cause-specific survival in surveillance, epidemiology and end results (SEER) programme breast cancer patients. *Statistical Methods in Medical Research* 2001; 10(5): 339–352.

16. Perme M, Stare J and Estève J. On estimation in relative survival. *Biometrics* 2012; 68(1): 113–120.

17. Danieli C, Remontet L, Bossard N et al. Estimating net survival: the importance of allowing for informative censoring. *Statistics in Medicine* 2012; 31(8): 775–786.

18. Perme M, Estève J and Rachet B. Analysing population-based cancer survival–settling the controversies. *BMC cancer* 2016; 16(1): 933.

19. Bolard P, Quantin C, Abrahamowicz M et al. Assessing time-by-covariate interactions in relative survival models using restrictive cubic spline functions. *Journal of Cancer Epidemiology and Prevention* 2002; 7: 113–122.

20. Giorgi R, Abrahamowicz M, Quantin C et al. A relative survival regression model using B-spline functions to model non-proportional hazards. *Statistics in Medicine* 2003; 22: 2767–2784.

21. Nelson C, Lambert P, Squire I et al. Flexible parametric models for relative survival with application in coronary heart disease. *Statistics in Medicine* 2007; 26(30): 5486–5498.

22. Mahboubi M Aand Abrahamowicz, Giorgi R, Binquet C et al. Flexible modeling of the effects of continuous prognostic factors in relative survival. *Statistics in Medicine* 2011; 30(12): 1351–1365.

23. Charvat H, Remontet L, Bossard N et al. A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Statistics in Medicine* 2016; 35(18): 3066–3084.

24. Mudholkar G, Srivastava D and Kollia G. A generalization of the Weibull distribution with application to the analysis of survival data. *Journal of the American Statistical Association* 1996; 91(436): 1575–1583.

25. Mudholkar G, Srivastava D and Freimer M. The exponentiated Weibull family: a reanalysis of the bus-motor-failure data. *Technometrics* 1995; 37(4): 436–445.

26. Cox C and Matheson M. A comparison of the generalized Gamma and exponentiated Weibull distributions. *Statistics in Medicine* 2014; 33(21): 3772–3780.

27. Khan S. Exponentiated Weibull regression for time-to-event data. *Lifetime Data Analysis* 2017; : 1–27.

28. Rubio F and Genton M. Bayesian linear regression with skew-symmetric error distributions with applications to survival analysis. *Statistics in Medicine* 2016; 35(14): 2441–2454.

29. Rubio F and Hong Y. Survival and lifetime data analysis with a flexible class of distributions. *Journal of Applied Statistics* 2016; 43(10): 1794–1813.

30. Chen YQ and Wang M. Estimating a treatment effect with the accelerated hazards models. *Controlled Clinical Trials* 2000; 21(4): 369–380.

31. Hess KR and Levin V. Getting more out of survival data by using the hazard function. *Clinical Cancer Research* 2014; 20(6): 1404–1409.

32. Mandel M. Simulation-based confidence intervals for functions with complicated derivatives. *The American Statistician* 2013; 67(2): 76–81.

33. Ross S. *Simulation*. Statistical Modeling and Decision Science, Elsevier Science, 2006. ISBN 9780080517223. URL http://books.google.fr/books?id=TMUt5OXVvY0C.

34. Benitez-Majano S, Fowler H, Maringe C et al. Deriving stage at diagnosis from multiple population-based sources: colorectal and lung cancer in England. *British Journal of Cancer* 2016; 115(3): 391.

35. Maringe C, Fowler H, Rachet B et al. Reproducibility, reliability and validity of population-based administrative health data for the assessment of cancer non-related comorbidities. *PloS one* 2017; 12(3): e0172814.

36. Charvat H and Belot A. *mexhaz: Mixed Effect Excess Hazard Models*, 2017. URL https://CRAN.R-project.org/package=mexhaz. R package version 1.4.

37. Qian L. The Fisher information matrix for a three-parameter exponentiated Weibull distribution under type II censoring. *Statistical Methodology* 2012; 9(3): 320–329.

38. Zhang J and Peng Y. Crossing hazard functions in common survival models. *Statistics & Probability Letters* 2009; 79(20): 2124–2130.

39. Hjort N. On inference in parametric survival data models. *International Statistical Review* 1992; : 355–387.

40. Lin D and Ying Z. Semiparametric analysis of the additive risk model. *Biometrika* 1994; 81(1): 61–71.