Net Survival Analysis: Parametric Models

Francisco Javier Rubio
@FJavierRubio1

Lecture Aims

- To introduce the relative survival framework.
- To introduce a general hazard structure.
- To present an application in cancer epidemiology.
- To discuss a challenge in the relative survival framework.

Cancer Epidemiology

- We are exposed to many forces of mortality: ageing, illnesses, natural disasters, accidents, crime, COVID19, and etcetera.
- Cancer represents a strong force of mortality that affects large groups of individuals in all countries.
- Cancer patients, the population of interest in cancer epidemiology, are exposed to the additional force of mortality due to cancer.
- The aim is to quantify the survival of cancer patients in order to compare cancer management in different countries. Thus, we need to comparable quantities.

The typical data set

- Sample of **times to event** (possibly right-censored) $(t_1, \ldots, t_n)$ from a group of individuals.
- Vital status (or **censoring** indicators) $(\delta_1, \ldots, \delta_n)$. ($\delta_i = 1$: death, $\delta_i = 0$, right-censored/alive). Censoring may be due to random drop-out, lost to follow-up, or administrative censoring.
- In some cases, we may know some additional characteristics about the individuals, meaning we have access to **covariates** $\mathbf{x}_i = (x_{i1}, \ldots, x_i p)^\top$, (age, sex, deprivation level, comorbidities, tumour stage, $\ldots$).

Survival Analysis Frameworks

- There are three approaches to analyse survival data:
  1. The overall survival framework: all-cause mortality is analysed. Not useful for comparison between populations as they are exposed to different causes of mortality.
  2. The cause-specific framework: where the cause of death is known. Death certificates are highly unreliable in all countries.
  3. The relative survival framework: where we separate the hazard associated to other causes and the hazard associated to cancer.

Relative survival framework

- We would like to quantify the mortality hazard due to the cancer under study, and link this to the patients' characteristics.
- In the relative survival framework, we assume that the expected mortality hazard (other causes than cancer) can be obtained from the general [population life tables].

Excess mortality hazard regression models

- More specifically, we can decompose the individual observed hazard, $h_o(\cdot;\cdot)$, as

$$h_o(t;\mathbf{x}) = h_P(A+t;\mathbf{z}) + h_E(t;\mathbf{x}), \tag{1}$$

  where $A =$ age (at diagnosis), $h_P(A+t;\mathbf{z})$ is the population hazard and it is obtained from the lifetables ($\mathbf{z} \subset \mathbf{x}$, age, sex, deprivation, $\dots$), and $h_E(t;\mathbf{x})$ is the excess hazard.
- Assumptions:
  - The general population hazard is supposed to correctly reflect the other-causes hazard in our population of interest.
  - The excess hazard is interpreted as the hazard due to the cancer under study.

Background Mortality

- In order to calculate $h_P(A + t; \mathbf{z})$, we need to get access to the life tables produced by the country where the patients live in. These are publicly available.
- Depending on the country, these life tables are defined by different characteristics $\mathbf{z}$. Typically: age, sex, and year.
- In the UK, life tables are also available by Deprivation Level (I–V).
- Example: Patient with characteristics $\mathbf{z}_i = (\text{age} = 70, \text{sex} = \text{male}, \text{Deprivation} = \text{III}, \text{Year of Diagnosis} = 2010)$ and $t_i = 1 year$ $\longrightarrow$ 2011 Life table $\longrightarrow$ Extract the corresponding mortality rate.

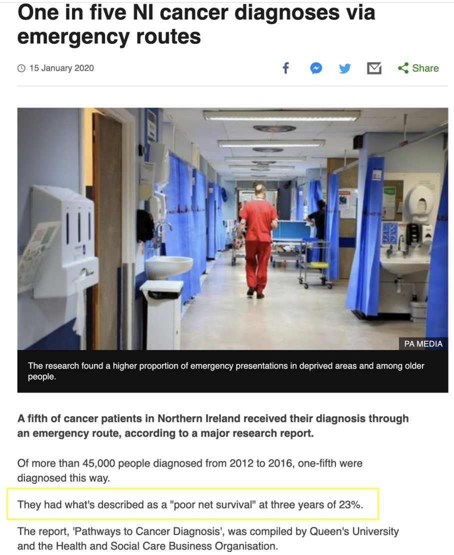
Net survival

- The main quantity of interest is the **net survival**, which is the survival function associated to the excess hazard. Thus,

$$S_N(t;\mathbf{x}) = \exp\left\{-\int_0^t h_E(s;\mathbf{x})ds\right\}.$$

- This quantity only depends on the EHM and it is a useful quantity for comparing the performance of cancer management between countries/period because it is not affected by differences in population mortality hazards.
- Policy-making for cancer management is often based on the population net survival

$$S_N(t) = \frac{1}{n}\sum_{i=1}^{n} S_N(t;\mathbf{x}_i).$$



One in five NI cancer diagnoses via emergency routes

15 January 2020

The research found a higher proportion of emergency presentations in deprived areas and among older people.

A fifth of cancer patients in Northern Ireland received their diagnosis through an emergency route, according to a major research report.

Of more than 45,000 people diagnosed from 2012 to 2016, one-fifth were diagnosed this way.

They had what's described as a "poor net survival" at three years of 23%.

The report, 'Pathways to Cancer Diagnosis', was compiled by Queen's University and the Health and Social Care Business Organisation.

Which parametric excess hazard model?

- Typically modelled as

$$h_E(t;\mathbf{x}) = h_0(t)\exp\left(\mathbf{x}^\top\beta\right).$$

- Fully Parametric models are of interest since they allow for prediction beyond the follow-up period, and they are typically more parsimonious models.
- The main criticisms are: (i) the PH assumption does not take into account time-varying effects, and (ii) the distributions used to model the baseline hazard can only cover specific shapes.

The proposed model

- Aims: (i) Consider more general alternatives to the PH structure, (ii) A parametric model which can cover the basic shapes of interest. [Rubio et al., 2019b]
- (i) In order to avoid the proportional hazards model, we consider the general hazard (GH) structure [Chen and Jewell, 2001]:

$$
\begin{aligned}
h_{\text{E}}^{\text{G}}(t; \mathbf{x}_i) &= h_0\left(t \exp(\mathbf{x}_i^\top \beta_1)\right) \exp(\mathbf{x}_i^\top \beta_2), && (2)\\
H_{\text{E}}^{\text{G}}(t; \mathbf{x}_i) &\overset{\textit{Homework}}{=} H_0\left(t \exp(\mathbf{x}_i^\top \beta_1)\right) \exp(\mathbf{x}_i^\top \beta_2 - \mathbf{x}_i^\top \beta_1). && (3)
\end{aligned}
$$

(i) If $\boldsymbol{\beta}_1 = 0$, then GH $=$ PH.

$$h_{\mathrm{E}}^{\mathrm{PH}}(t;\mathbf{x}_i) = h_0(t)\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right). \tag{4}$$

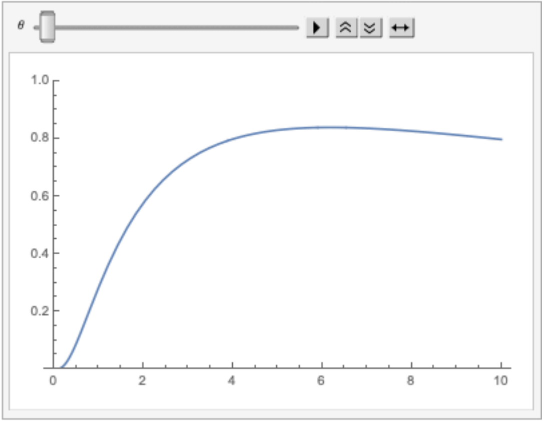
(ii) if $\boldsymbol{\beta}_2 = 0$, then GH $=$ AH (Accelerated hazards).

$$h_{\mathrm{E}}^{\mathrm{AH}}(t;\mathbf{x}_i) = h_0\left(t\exp(\mathbf{x}_i^\top\boldsymbol{\beta})\right). \tag{5}$$

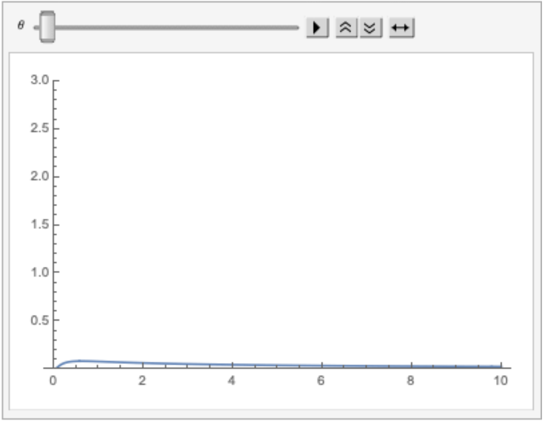
(iii) $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, then GH $=$ AFT (Accelerated failure time).

$$h_{\mathrm{E}}^{\mathrm{AFT}}(t;\mathbf{x}_i) = h_0\left(t\exp(\mathbf{x}_i^\top\boldsymbol{\beta})\right)\exp(\mathbf{x}_i^\top\boldsymbol{\beta}). \tag{6}$$

- The GH structure also includes time-dependent effects through β_1.



- The GH structure also includes hazard-level effects through β_2.



- (ii) for the second aim, we will use the [Exponentiated Weibull distribution] [Mudholkar et al., 1996], which is defined as a simple transformation of the Weibull distribution $F(t \mid \kappa, \sigma)$:

$$G(t \mid \kappa, \sigma, \alpha) = F(t \mid \kappa, \sigma)^{\alpha},$$

- This simple transformation adds a lot of flexibility since the corresponding hazard function can be: unimodal, increasing, decreasing, flat, bathtub. These shapes are often referred to as the "basic shapes" of the hazard function.
- By using the GH structure with EW baseline hazard, we cover both aims. ✓
- There are some alternatives: generalised gamma, [power generalised Weibull], and [generalised Weibull].

Maximum Likelihood Estimation

- The likelihood function of the vector of parameters ψ is $\mathcal{L}_0(\psi; \text{Data})$

$$
\begin{aligned}
&= \prod_{i=1}^{n} h(t_i; \mathbf{x}_i)^{\delta_i} S(t_i; \mathbf{x}_i),\\
&= \prod_{i=1}^{n} h(t_i; \mathbf{x}_i)^{\delta_i} \exp\left\{-H(t_i; \mathbf{x}_i)\right\},\\
&= \prod_{i=1}^{n} \left\{h_{\mathrm{P}}(\mathrm{age}_j + t_i; \mathbf{z}_j) + h_{\mathrm{E}}(t_i; \mathbf{x}_i)\right\}^{\delta_i}\\
&\overset{\textit{Homework}}{\times} \exp\left\{-[H_{\mathrm{P}}(\mathrm{age}_j + t_i; \mathbf{z}_j) - H_{\mathrm{P}}(\mathrm{age}_j; \mathbf{z}_j)]\right\} \exp\left\{-H_{\mathrm{E}}(t_i; \mathbf{x}_i)\right\}\\
&\propto \prod_{i=1}^{n} \left\{h_{\mathrm{P}}(\mathrm{age}_j + t_i; \mathbf{z}_j) + h_{\mathrm{E}}(t_i; \mathbf{x}_i)\right\}^{\delta_i} \exp\left\{-H_{\mathrm{E}}(t_i; \mathbf{x}_i)\right\}.
\end{aligned}
$$

Software and Examples

- The GH model is implemented in the overall survival and relative survival frameworks in the R package [GHSurv].
- A manual on how to simulate times to event from the GH structure can be found at: `https://rpubs.com/FJRubio/GHSim`
- A simulated data example can be found at: `https://rpubs.com/FJRubio/GHGH`
- A manual on how to simulate times to event from a life table can be found at [SimLT].

Real Data Application

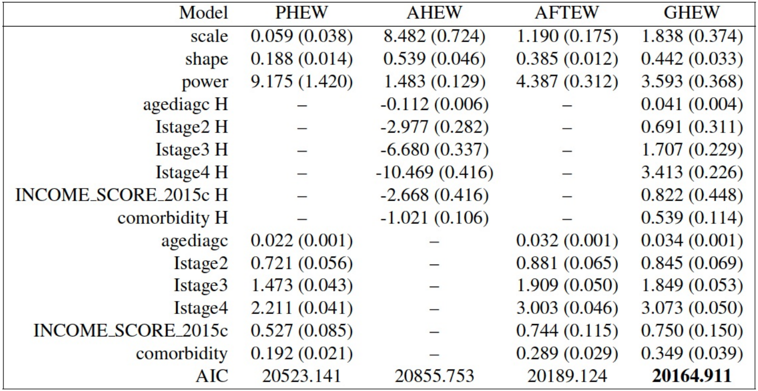
- We analysed a dataset obtained from population-based national cancer registry of lung cancer patients diagnosed in 2012 in the United-Kingdom.
- We restricted our analysis to women with no missing data.
- We observed $n = 14557$ patients with complete cases among which $n_o = 12138$ died before the 31st of December 2015.
- The median follow-up among patients censored was 3.46 years.

Covariates

- **Age at diagnosis.** 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.
- **Tumour stage (I–IV).** 2434 were Stage I, 1131 were Stage II, 3421 were Stage III, and 7751 were Stage IV.
- **Income Score** (Income Domain from the 2010 England Indices of Multiple Deprivation). Continuous (0,1).
- The presence of cardiovascular diseases (**Comorbidity**). 4318 patients were classified with comorbidity indicator 1.

Model

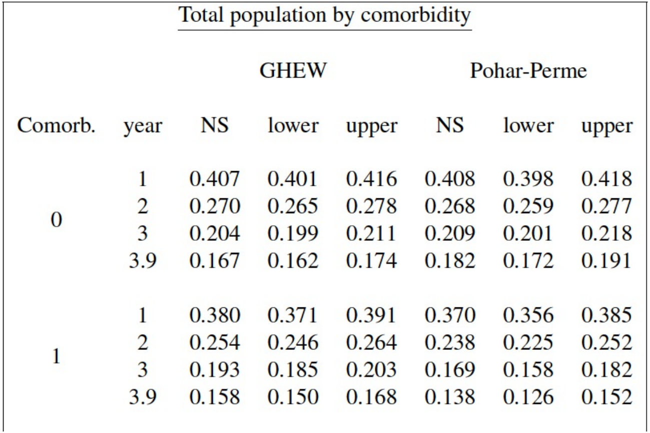
- We use the EW baseline hazard (3 parameters).
- We fit 4 models for the excess hazard: PH, AFT, AH, GH.
- Optimisation of the likelihood function is done using the R command `optim`.
- We will select a model using AIC.



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 H	–	-0.112 (0.006)	–	0.041 (0.004)
Istage2 H	–	-2.977 (0.282)	–	0.691 (0.311)
Istage3 H	–	-6.680 (0.337)	–	1.707 (0.229)
Istage4 H	–	-10.469 (0.416)	–	3.413 (0.226)
INCOME_SCORE_2015c H	–	-2.668 (0.416)	–	0.822 (0.448)
comorbidity H	–	-1.021 (0.106)	–	0.539 (0.114)
agediagc	0.022 (0.001)	–	0.032 (0.001)	0.034 (0.001)
Istage2	0.721 (0.056)	–	0.881 (0.065)	0.845 (0.069)
Istage3	1.473 (0.043)	–	1.909 (0.050)	1.849 (0.053)
Istage4	2.211 (0.041)	–	3.003 (0.046)	3.073 (0.050)
INCOME_SCORE_2015c	0.527 (0.085)	–	0.744 (0.115)	0.750 (0.150)
comorbidity	0.192 (0.021)	–	0.289 (0.029)	0.349 (0.039)
AIC	20523.141	20855.753	20189.124	**20164.911**

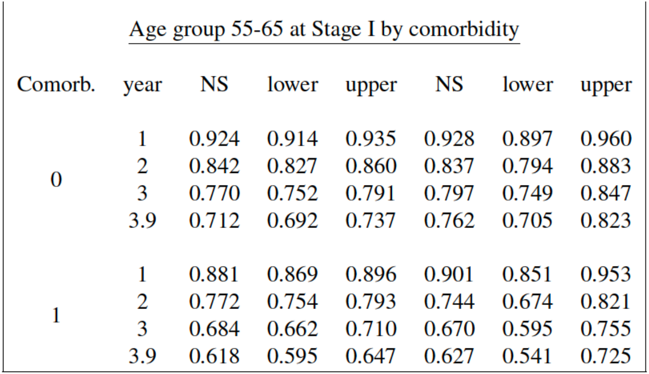
Net survival

- We will now calculate the Net Survival at $t = 1, 2, 3, 3.9$ years after the diagnosis for two groups:
  1. Total population by comorbidity $\{0, 1\}$.
  2. Age group 55-65 at Stage I by comorbidity $\{0, 1\}$.
- We will compare the estimates obtained with the selected parametric model with those obtained with a nonparametric estimator proposed by Pohar-Perme et al. [2012]. This estimator is known as the Pohar-Perme estimator.



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

Discussion

- We have studied the overall and relative survival frameworks. There are underlying assumptions in each of them.
- Survival parametric models are useful and interpretable tools for modelling time to event data. It is important to understand the underlying assumptions of these models.
- We have focused on the parametric framework, however, it is also possible to estimate survival functions using semi- and non-parametric approaches.
- The relative survival framework is popular in policy making, thus it is important to have “good models” (try to reflect about what would make a good model).
- Other challenges: Different types of censoring, informative censoring, competing risks, cure/remission models. Survival analysis is a vast area of active research (see Eletti et al. [2022]).
- In fact ...

A challenge in the relative survival framework

- We have assumed that

  $$h_o(t;\mathbf{x}) = h_P(A+t;\mathbf{z}) + h_E(t;\mathbf{x}),$$

  the hazard associated to other causes and the hazard associated to cancer.
- In some countries, life tables are stratified by age, sex and year of diagnosis.
- Then, in those countries, an extremely poor and an extremely rich patient with the same age, sex and year will be assigned the same $h_P(A+t;\mathbf{z})$.

A challenge in the relative survival framework

- The same happens for a person with diabetes *vs.* a person without diabetes.
- Thus, the population hazard is either overestimated or underestimated.
- A more detailed study of this challenge can be found in Rubio et al. [2019a].

Y.Q. Chen and N.P. Jewell. On a general class of semiparametric hazards regression models. *Biometrika*, 88(3):687–702, 2001.

A. Eletti et al. A unifying framework for flexible excess hazard modelling with applications in cancer epidemiology. *JRSS-C*, 2022.

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

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

F.J. Rubio, B. Rachet, R. Giorgi, C. Maringe, and A. Belot. On models for the estimation of the excess mortality hazard in case of insufficiently stratified life tables. *Biostatistics*, na(na):na–na, 2019a.

F.J. Rubio, L. Remontet, N.P. Jewell, and A. Belot. On a general structure for hazard-based regression models: an application to population-based cancer research. *Statistical Methods in Medical Research*, 28:2404–2417, 2019b.