Università Commerciale “Luigi Bocconi”
PhD School

PhD program in: Statistics
Cycle: XXXV
Disciplinary Field (code): SECS-S/01

Risk prediction and family history effect in survival models for disease onset, with a specific focus on breast cancer

Advisor: Marco Bonetti

PhD Thesis by
Maria Veronica Vinattieri
ID number: 3111885

Year: 2024

Acknowledgements

I would like to thank my supervisor Marco Bonetti for pushing me to always do my best, for his enthusiasm and meticulous attention to detail.

I would like to thank Kamila Czene and Keith Humphreys for their precious advice, for inviting and hosting me at Karolinska Institutet with the support of grant 2022-00584 with PI Kamila Czene.

I would like also to thank the Population Dynamics and Health unit of the Carlo F. Dondena research center at Bocconi University for its support.

I would like to thank all the colleagues that I have met during the PhD in Bocconi, you have been my Milanese family. Especially, I would like to thank the colleagues from my year, it has been a pleasure to share this journey with you and I have learned so much from you.

I would like to thank all my loved ones, I won't give names but you know exaclty who you are and how much I am grateful for your support, for listening to me and offering advice when things get tough, for making me laugh and always making me feel right at home.

Abstract

We are interested in a specific aspect related to modelling disease onset, i.e. the study of family-specific risk. Generally, quantifying the family-specific risk of developing a disease is crucial to improve the patient's survival, and healthcare data provide an essential tool to estimate such quantity. Specifically for breast cancer, early detection is particularly important to increase the chances of a successful treatment. In this setting, risk estimation allows for the identification of subjects at higher risk of developing breast cancer, so that they can be the target of tailored and more intensive screening and prevention strategies. We take families as our statistical units of interest, and we assume that these units belong to different risk groups of developing the disease. We investigate models for the age at breast cancer onset, with a structure of familiar dependence among survival times through the risk.

In Chapter 1, we compare cure rate models with a continuous Gamma frailty to the conventional Cox model in terms of risk prediction accuracy. In Chapter 2, we focus exclusively on cure rate models using a binary frailty to model the age at onset. In Chapter 3 we move to the study of the heritability of longevity. Similarly to the previous chapters, we assume that families have different "risk" in terms of life duration. We carry out our analysis through simulation studies and apply the continuous frailty model from Chapter 1 to the available data sourced from the Multi-Generational Breast Cancer Swedish registry. We conclude with an exploration of most powerful tests for survival data with right censoring in Chapter 4.

Table of contents

Introduction

We focus on a specific aspect of modelling disease onset, namely the investigation of family-specific risk. Quantifying the family-specific risk of developing a disease may be crucial to improve early detection of the disease and consequently patient survival. Healthcare datasets provide an essential tool to estimate family-specific risk, and we develop and employ specific methods to estimate this risk from the available data. We focus on breast cancer onset, although the problem can be generalized to a wide range of diseases. In the context of breast cancer development, risk estimation plays a crucial role in identifying families, and thus individuals, who are at high risk of developing breast cancer. We assume that individuals who belong to the same family share a common from-birth risk of being diagnosed with breast cancer, and that such family-specific risk is unchanged from birth and never observable. Because of the latent nature of the risk we call it the frailty risk. Once we identify the highest-risk families, that are those with highest values of the estimated frailty risk, we may target female members of those families for tailored and more intensive screening and prevention strategies, which can improve their chances of a successful treatment and prognosis.

One common approach to stratify subjects into different risk groups of developing breast cancer is to consider the risk factors associated to breast cancer development. In our motivating data, one among the strongest risk factors is available: the breast cancer family history, defined as the indicator of having observed at least one family member who has already experienced breast cancer diagnosis. Specifically, the family history indicator $FH(t)$ takes value one when, by age t of the subject, at least one family member has been diagnosed with (invasive) breast cancer, otherwise it takes value 0. This indicator is based on the notion that individuals with a positive ($FH(t)p = 1$) family history are at higher risk. It is common to insert this simple indicator into a model with the aim of splitting families into two risk groups: the low-risk group and the high-risk group of developing breast cancer. This approach motivates the categorization of families into two latent risk groups, allowing the frailty risk to therefore be binary with two levels $0/1$, for low and high-risk groups respectively, as we illustrate in Chapter 2. On the other side the family history, and consequently the binary frailty risk, cannot capture the complex nature of such a phenomenon. This motivates our main contribution of this thesis, that is the use of continuous frailty risk multivariate cure-rate survival models in this setting, as we illustrate in Chapter 1. Specifically, the continuous frailty risk is assumed to be distributed according to a Gamma distribution with parameters $(shape = \theta,\ rate = \theta)$, with θ the frailty parameter. We carry out a comparative analysis

to assess that our proposed model as expected, outperforms the other models under analysis in terms of inference precision, prediction accuracy, and "explanability" of the phenomenon. The multivariate part of our proposed model refers to jointly modelling the family female member times-to-event (where the event of interest is breast cancer onset) in contrast to modelling only one subject per family, as the Univariate frailty Cure-Rate model or the Univariate Cure-Rate with the FH covariate models do, as we develop in Chapter 1. The Cure-Rate component refers to the peculiar structure that the survival function takes, based on the assumption that not all women will experience breast cancer onset no matter for how long they will be followed. The proportion of women that won´t experience breast cancer onset is called the "cured" fraction which here will however consist of "non-cases". The cured fraction enters survival function formula, making it not proper anymore. Through the Multivariate frailty Cure-Rate model our novel contribution consists of involving the Cure-Rate baseline survival function in a Lehmann structure that unites the different family-specific frailty risk survival functions. In Chapter 1 we also discuss some related aspects involving the nature of the model and its identifiability.

The Multivariate frailty Cure-Rate model allows us to estimate the survival distribution of "cases", that are defined as women who will eventually experience breast cancer onset, and to capture a peculiar tail behaviour by estimating the cured fraction. These measures contribute to accurately describe breast cancer development in the population of interest. In contrast, for example, the already known and developed semiparametric Multivariate frailty Cox model would not allow us to estimate cure rates directly. Moreover, we show that, as expected, if the parametric assumptions hold, Multivariate frailty Cure-Rate model achieves a level of prediction accuracy on par to the Multivariate frailty Cox model.

Modelling a continuous frailty risk allows us to capture the complexity of the phenomenon of the breast cancer development and deal with this problem in a real dataset, beyond simulation studies. In this chapter we provide some preliminary results obtained from the analysis of the Swedish Multi-generational dataset. On the other side, modelling a binary frailty risk might seem simpler and more direct, but with the drawback of not capturing the real differences in risk among families. Splitting families into two categories and identifying the higher-risk families remains clinically convenient, and we have included a section in Chapter 1 discussing the transition from infinitely many values of the risk to only two groups by splitting the families based on a threshold.

We elaborate on this in Chapter 2 as the use of the family history FH motivates the use of a binary frailty risk to stratify families. Indeed here, the comparison between the Multivariate frailty Cure-Rate model and the Univariate frailty Cure-Rate to the Univariate FH Cure-Rate model is more meaningful than in the previous chapter thanks to the same binary nature of the latent risk and the risk factor FH used as its replacement. Nevertheless, the Univariate FH Cure-Rate model may be not accurate in explaining breast cancer development, and so may be the Univariate frailty Cure-Rate model. Our results confirm and quantify how the Multivariate frailty Cure-Rate model outperforms the other two models in terms of accuracy in risk prediction.

In Chapter 3 we move away from the cure rate structure, and we explore the study of family-

specific frailty risk of death or, in more optimistic words, the heritability of longevity within families. We thus assume that families belong to different groups of life expectancy according to their common genetics (internal factors) and environmental and behavioural setting (external factors). Similarly to the first two chapters, the focus in on posterior risk prediction through estimation of the posterior risk distribution. The novel contribution here is the use of a classification algorithm that operates to compute the posterior distribution of the risk of mortality. The algorithm is applied to simulation scenarios. Beyond contributing to the explanation of heritability of longevity, these studies address families with the highest risk of death to strategies of clinical prevention through screening for diseases and behavioural changes.

As a side study, in Chapter 4 we explore the related issue of obtaining the most powerful (MP) tests for survival data with and without right censoring. Under the proportional hazards assumption, we aim to decide from an i.i.d. sample of survival times whether the population survival function is governed by a known survival function denoted as $S_0(t)$, rather than by an unknown survival function denoted by $S_1(t) = [S_0(t)]^\beta$, with β not equal to one. This is linked to the other chapters through the same wider aim of identifying the risk membership groups of individuals.

To better understand aims, findings and connections between the chapters of the thesis Tables 1, and 2) which briefly summarize the work done.

Table 1: Connections among the chapters.

	Lehmann family	Survival function	Event of interest	Frailty risk
Chapter 1	✓	Cure-Rate	Breast Cancer onset	Continuous
Chapter 2	✓	Cure-Rate	Breast Cancer onset	Binary
Chapter 3	✓	Traditional	Death	Binary/Continuous
Chapter 4	✓	Traditional	Death	Binary

Table 2: Aims and findings.

	Aims and findings
Chapter 1	The Multivariate continuous frailty Cure-Rate model outperforms the others.
Chapter 2	The Multivariate binary frailty Cure-Rate model outperforms the others.
Chapter 3	An algorithm for risk prediction has been developed.
Chapter 4	Study of MP test in PH non-cure rate survival models.

Chapter 1

An Application of risk prediction models to the Swedish Multi-Generational Breast Cancer Registry data

Joint work with Kamila Czene.[1]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Solna, 171 77, Sweden, email: kamila.czene@ki.se.

We aim to study a specific aspect of the breast cancer onset: family-specific risk. We assume that the family-specific risk of developing breast cancer is latent and unchanged from birth. Our objective is to identify the highest-risk families, i.e. those with the highest values of family-specific risk, and target them towards more intensive screening and prevention strategies to enhance the probability of successful treatment. In contrast with the risk prediction models from the literature which include risk factors associated to breast cancer, we develop a parametric Multivariate frailty Cure-Rate survival model that takes into account the subjects as part of a family. This model incorporates a cure-rate component, allowing us to identify the fraction of the population not susceptible to breast cancer (the "cured" fraction), shedding light on the magnitude of the phenomenon of breast cancer development within the population. As expected, in simulations, our model shows high accuracy in risk prediction when compared to the semiparametric counterpart: the Multivariate frailty Cox model.

We conduct a comparative analysis with other models, including Cox models and models involving a risk factor associated to breast cancer: the indicator defined as a covariate that indicates the presence of family history. Our comparison is illustrated using simulation studies and a real case dataset from the Multi-Generational Swedish Breast Cancer registry. Our assessment criteria include accuracy in risk prediction, as measured by the area under the ROC curve (AUC), and Harrell's Concordance index.

Our findings not only highlight the crucial role of incorporating complete family information in identifying high-risk families, in contrast to the limited utility of a family history indicator, but also demonstrates that the Multivariate frailty Cure-Rate model can elucidate the dynamics of breast cancer development within the population, by directly estimating the fraction of "cured" (i.e. "non-cases") subjects and the distribution of breast cancer cases, thus combining explanation and prediction performance.
keywords: breast cancer, family history, frailty models, risk prediction

1.1 Introduction

Breast cancer risk prediction models have been extensively developed to tailor screening and prevention strategies. Several subject-specific risk factors associated to breast cancer have been involved to explain the variability of the breast cancer development and improve the predictive power of many predictive models. For example among three of the main references in the literature, there is Gail [11] that uses the number of first-degree relatives who experienced breast cancer to predict the long-term probability of developing breast cancer. Rosner and Colditz that [31] use, among the others, several reproductive covariates as the age at menarche, the age at menopause, and the age at childbirth to predict the incidence of breast cancer over a specified time period. And finally, Tyrer and Cuzick [40] have the same aim of predicting the subject-specific risk to develop breast cancer over a specific time period by accounting for risk factors, like reproductive covariates, family history, mammographic breast cancer density (MD) [36], body mass index (BMI), and life-style covariates. What we notice from the the aforementioned models is that they always model subject-specific time-to-event. The limitation here is that they include the hereditary component of breast cancer only through subject-specific covariates. For example, the number of first-degree breast cancer cases is a limited measure of the complete breast cancer history in a family, as it does not allow us to know the specific history of those cancers.

Different from the others, the BOADICEA model [1, 3, 2, 37] tried to directly incorporate the hereditary component of breast cancer, by inferring a subject-specific genetic latent quantity, the polygenic component, that consists in the contribution of several genetics mutations with a singular small effect, on the disease development. They aim to estimate the familial risk of developing breast cancer for predicting cancer risk according to the family history of the disease and other risk factors. The BOADICEA is the most efficient tool for breast cancer risk prediction seen so far. Nevertheless, it is a complex tool that requires access to detailed subject-specific genetic and family history information. We aim to create a tool that is at the same time efficient but also easy to understand and use with accessible data.

Our novel contribution is found in retrieving the familiar relationship between subjects to jointly model subject-specific time-to-events within families. This allows us to know both the specific history of breast cancer cases and the family history.

The available data allows us to jointly model the times-to-breast cancer onset under the conditional independence assumption given the family-specific frailty risk. We aim to ultimately infer the hereditary susceptibility to breast cancer by estimating the posterior distribution of the family-specific frailty risk and predicting the risk. This allows us to target high-risk families that are those with a higher value of the frailty risk. This aims to address the highest-risk families to specific screening and prevention strategies. Also, it is worth noting that we can infer the posterior distribution of the family-specific frailty risk also for women not included in the initial dataset. Therefore, it is crucial for new patients to provide familial breast cancer information to obtain a reliable estimate of the probability of belonging to one of the highest-risk families.

We call our model the Multivariate frailty Cure-Rate model where “multivariate” refers to

jointly modelling the times-to-event within families, in contrast to "univariate" where the subject-specific time-to-event is modelled; "frailty" refers to the frailty risk given its latent nature; and "cure-rate" to the peculiar structure that we give to the survival function as we assume that not all women will experience the breast cancer onset eventually.

A comparative analysis is conducted to evaluate the benefits of using the Multivariate frailty Cure-Rate model. We start from a model with includes a strong risk factor associated to breast cancer development: the first-degree family history indicator. This takes value one if at least one family member has experienced the breast cancer onset, or zero if none have, and it can be measured either at the end of the follow-up period or as time-varying during the follow-up. This model motivates the use of univariate models, such as also the Univariate frailty Cure-Rate model, that is the univariate counterpart of the Multivariate frailty Cure-Rate model. Lastly, also the known and widespread Cox model, in several of its forms, is included in the comparative analysis to assess that the predictive power of the Multivariate frailty Cure-Rate model reaches the same accuracy of the Multivariate frailty Cox model. Furthermore the Multivariate frailty Cure-Rate model, at contrary of the Multivariate frailty Cox model, reaches higher precision in inference and explanation of the phenomenon by estimating the cured fraction and the distribution of breast cancer cases.

Section 1.2 provides an overview of the background of the setting, and the deep description of the specific models employed in the analysis, and in Section 1.3 we focus on the procedure of predicting the posterior risk. The simulation results are presented in Section 1.4, while the findings from the real case dataset, the Multi-Generational Swedish Breast Cancer registry, are discussed in Section 1.5.

1.2 Model specification

We model the age (in years) T of breast cancer onset, where the conditional hazard function is denoted by $\lambda_r(t) = \lambda(t \mid r)$, where $R = r$ indicates the frailty risk.

The Univariate frailty model [8] allows the hazard function to depend on the frailty quantity R to capture the unobserved heterogeneity among subjects. The conditional hazard function at time t is given by

$$\lambda_r(t) = \alpha(r)\lambda_0(t),$$

where $\alpha(r)$ can be any continuous function of the risk, and $\lambda_0(t)$ is the baseline hazard function corresponding to the value of the risk r such that $\alpha(r) = 1$.

In presence of a vector of subject-specific covariates x for a subject, the frailty risk explains the unobserved heterogeneity that the covariates are not able to capture. The conditional (subject-specific) hazard function at time t for the subject, with frailty risk r, results

$$\lambda_r(t) = \alpha(r)\lambda_0(t; x).$$

One can choose several form of the function $\alpha(r)$, as a polynomial form, an exponential form, or other (non-negative) forms. One may implement a hazard function that is linearly dependent

on the frailty, i.e.

$$\lambda_r(t) = r\lambda_0(t; x)$$

[8, 29]. Notice that this readily coincides with the model specification, known also as Lehmann structure, given by

$$\Lambda_r(t) = r\Lambda_0(t; x), \qquad S_r(t) = [S_0(t; x)]^r,$$

where Λ indicates the cumulative hazard function, and S indicates the survival function. Notice that for proving the model identifiability, $\mathbb{E}(R) = 1$ is usually assumed and allows by the frailty risk distribution.

In the Multivariate frailty model, the subject-specific hazard function has the same equation as in 1.2, but the frailty risk r is considered shared among members of the same family. What changes is that one need to compute also the the family-specific joint survival function. This last is obtained under the conditional independence assumption [30]. Consider the case with only two women belonging to the same family. Let $T_1 = t_1$ and $T_2 = t_2$ be the times-to-event of the two women, with survival function S_1 and S_2 from family i, with r_i be their family-specific frailty risk. By the conditional independence assumption one has that the joint survival function factorizes conditional to the frailty risk:

$$S_r(t_1, t_2) \overset{T_1 \perp T_2 | R}{=} S_1(t_1 \mid r) S_2(t_2 \mid r),$$

which, in the case of equality between the survivals, reduces to

$$S_r(t_1, t_2) \overset{T_1 \perp T_2 | R}{=} S_r(t_1) S_r(t_2) = [S_0(t_1)]^r [S_0(t_2)]^r = [S_0(t_1) S_0(t_2)]^r = [S_0(t_1, t_2)]^r.$$

Moreover, one case compute the unconditional joint survival function which is equal to

$$\begin{aligned} S_{12}(t_1, t_2) = \int_0^\infty S_{12}(t_1, t_2 \mid r) g(r) dr &= \int_0^\infty e^{-r(\Lambda_{01}(t_1)\Lambda_{02}(t_2))} g(r) dr \\ &= \mathcal{L}_g(\Lambda_{01}(t_1)\Lambda_{02}(t_2)) \end{aligned}$$

where $\mathcal{L}_g(r)$ is the Laplace transform of the density function $g(r)$ of R evaluated at $\Lambda_{01}(t_1)\Lambda_{02}(t_2)$. To obtain the explicit form of $S_{12}(t_1, t_2)$ one should specify $g(r)$, that is typically a Gamma or Log-normal distribution. This framework can be easily extended to more than two subjects [16].

Now, we make the notation more complex to explore the formal context where we tackle the problem in-depth.

Let $i = 1, \ldots, n$ identify the family and the main subject of a family, and $\{m, s_1\}$ to identify the two family members "mother" and "first sister", respectively. The generalization of this to families with more than three members is straightforward. The observed survival data is $\underline{X} = (\underline{X}_1, \ldots, \underline{X}_n)^T$, where $\underline{X}_i = (\underline{x}_i, \underline{xs}_{1i}, \underline{x}m_i)^T$. For the generic subject i, $\underline{x}_i = (x_i = \min(t_i, c_i), \delta_i)^T$, $\delta_i = \mathbb{I}(t_i \leq c_i)$ following the usual notation where t_i indicates the survival time, and c_i indicates the administrative (independent) censoring time, both measured from the same origin, that in our case is the birth b_i. The notation for the other family members is obtained by having x, t, c, b,

and δ be followed by m, and s_1. If subject i does not have a sister, then the sister birthday is set as $bs_{1i} = +\infty$ and consequently the time-to-event is $ts_{1i} = +\infty$. The distinction between mother and sister is not strictly needed here as we assume that their time-to-event distributions are equal but it will make the extension to more complex models easier.

We extend the Lehmann structure in terms of the survival functions $S_r(t) = [S_0(t)]^r$ to the Cure-Rate model. We call this model the Lehmann Cure-Rate (LCR) model obtained by applying the Lehmann power transformation to a Cure-Rate (CR) survival function

$$CR: \ S_0(t) = p + (1-p)\widetilde{S}(t)$$

$$LCR: \ S_r(t) = \left[p + (1-p)\widetilde{S}(t)\right]^r, \ \ r > 0,$$

with p the cured fraction, and $\widetilde{S}(t)$ a proper survival function which describes the time-to-event distribution of the cases, which are the subjects who will eventually experience the event, with some parameters $\underline{\gamma}$. Thus, the time-to-event of cases do not admit the value $+\infty$, and their survival function is proper, i.e. $\lim_{t\to+\infty} \widetilde{S}(t) = 0$. Note that for a fixed value r, the survival function $S_r(t)$ also defines a Cure-Rate model. Indeed, $\lim_{t\to+\infty} S_r(t) = p^r$, and $S_r(t)$ can thus be written as

$$S_r(t) = p^r + \left(1 - p^r\right) \widetilde{S}_r(t), \tag{1.1}$$

with proper conditional survival function for the cases equal to

$$\widetilde{S}_r(t) = \frac{\left[p + (1-p)\widetilde{S}(t)\right]^r - p^r}{1 - p^r}, \tag{1.2}$$

and proper conditional density function for the cases equal to

$$\widetilde{f}_r(t) = -\frac{d}{dt}\widetilde{S}_r(t) = \frac{1-p}{1-p^r} r \left[p + (1-p)\widetilde{S}(t)\right]^{(r-1)} \widetilde{f}(t). \tag{1.3}$$

Recall that the cure-rate models as in 1.1 refers to survival random variable T with improper cumulative distribution function. Indeed, we have that $P(T = +\infty) = p > 0$, and also a proper density function does not exist. If $S_0(t) = p + (1-p)\widetilde{S}(t)$ follows a cure-rate structure, we have

$$\lim_{t\to+\infty} S_0(t) = \lim_{t\to+\infty} (p + (1-p)\widetilde{S}(t)) = p > 0. \tag{1.4}$$

Suppose that a proper density function $f(t)$ exists, so that $S_0(t) = \int_t^\infty f(u)\mathrm{d}u$ and $\int_0^\infty f(u)\mathrm{d}u = 1$.

Since $\mathbb{I}(s \geq t)f(s) \leq f(s)\ \forall s \geq 0$ and $\displaystyle\int_0^{+\infty} f(s)\mathrm{d}s = 1,$ one has

$0 \leq \displaystyle\int_0^{+\infty} \mathbb{I}(s \geq t)f(s) \leq \int_0^{+\infty} f(s)\mathrm{d}s = 1$. Then by the DCT [41],

$$\lim_{t\to+\infty} S_0(t) = \lim_{t\to+\infty} \int_t^{+\infty} f(s)\mathrm{d}s = \lim_{t\to+\infty} \int_0^{+\infty} \mathbb{I}(s \geq t)f(s)\mathrm{d}s \overset{DCT}{=} \int_0^{+\infty} \lim_{t\to+\infty} [\mathbb{I}(s \geq t)f(s)]\mathrm{d}s = 0.$$

Hence $\lim_{t\to+\infty} S_0(t) = 0$, which contradicts 1.4. Additional comments on the cure-rate structure and PH assumption can be found in Appendix A.1.

We now specify the models and their likelihood that we will use later for the comparative analysis.

1.2.1 Models

In the literature we can find the use of the family history indicator as a strong risk factor associated to breast cancer. We use the family history as an indicator of the risk of developing breast cancer within a family. The family history indicator FH has value zero until the first case of breast cancer is observed in the family, after which it takes the value one. We thus develop a Univariate FH Cure-Rate model which includes the covariate FH. Notice that the survival function depends on the baseline covariate FH, and it is given by

$$
\begin{aligned}
&S_{FH}(t) = S_0(t)^{(1-fh)} S_0(t)^{fh\beta} = S_0(t)^{fh(\beta-1)+1} = [p + (1-p)\widetilde{S}(t)]^{fh(\beta-1)+1} \\
&\text{or, } S_{FH}(t) = p^{fh(\beta-1)+1} + (1 - p^{fh(\beta-1)+1})\widetilde{S}_{FH}(t) \\
&\text{with } \widetilde{S}_{FH}(t) = \frac{[p + (1-p)\widetilde{S}(t)]^{fh(\beta-1)+1} - p^{fh(\beta-1)+1}}{(1 - p^{fh(\beta-1)+1})}, \\
&\text{and } f_{FH}(t) = (1 - p^{fh(\beta-1)+1})\widetilde{f}_{FH}(t) = (fh(\beta-1)+1)(1-p)\widetilde{f}(t)[p + (1-p)\widetilde{S}(t)]^{fh(\beta-1)}, \\
&\text{with } \widetilde{f}_{FH}(t) = -\frac{\partial \widetilde{S}_{FH}(t)}{\partial t} = \frac{(fh(\beta-1)+1)[p + (1-p)\widetilde{S}(t)]^{fh(\beta-1)}(1-p)\widetilde{f}(t)}{(1 - p^{fh(\beta-1)+1})},
\end{aligned}
$$

for $FH = fh$, and with $\widetilde{S}_{FH}(t)$ the survival function of cases . The parameter β identifies the average risk difference between the group of families with a negative family history ($FH = 0$), and those families with a positive family history ($FH = 1$). Notice that when $FH = 0$ the survival function reduces to the baseline survival function $S_{FH}(t; FH = 0) = [p + (1-p)\widetilde{S}(t)]$, otherwise $S_{FH}(t; FH = 1) = [p + (1-p)\widetilde{S}(t)]^{\beta}$, which allows for a cure-rate model as the previous case.

The Univariate FH Cure-Rate model has a univariate likelihood on the parameter collection $\pi_{FH} = (\underline{\gamma}, p, \beta)^T$, that is given by

$$
\begin{aligned}
L(\pi_{FH}) &= \prod_{i=1}^{n} f_{FH}(x_i)^{\delta_i} S_{FH}(x_i)^{1-\delta_i} \\
&= \prod_{i=1}^{n} \left[(fh_i(\beta-1)+1)(1-p)\widetilde{f}(x_i)\right]^{\delta_i} \left[p + (1-p)\widetilde{S}(x_i)\right]^{fh_i(\beta-1)}.
\end{aligned}
$$

This model motivates the use of a binary risk of developing breast cancer, but we believe that a binary risk is too simplistic. For this reason we move to the development of models with a continuous frailty risk. Moreover, the frailty risk must be positive. We assume the frailty to follow a Gamma($shape = \theta, rate = \theta$) whose density function is given by

$$
g_R(r) = \frac{\theta^{\theta}}{\Gamma(\theta)} r^{\theta-1} \mathrm{e}^{-r\theta},
$$

where we θ the frailty parameter. An extension to other distributions, or with a number of parameters greater than one is straightforward.

We develop two models with the continuous frailty risk: the Multivariate frailty Cure-Rate model and the Univariate frailty Cure-Rate model.

We can find the closed form of the multivariate likelihood $L(\underline{y}, p, \theta)$ of the Multivariate frailty Cure-Rate model for $i = 1, \ldots, n$ families of varying size n_i by applying the following steps:

$$
\begin{aligned}
L(\underline{y}, p, \theta) &= \prod_{i=1}^{n} \prod_{j=1}^{n_i} \int_{\mathbb{R}^+} f_r(x_{ij})^{\delta_{ij}} S_r(x_{ij})^{1-\delta_{ij}} g_R(r;\theta)\mathrm{d}r \\
&= \prod_{i=1}^{n} \int_{\mathbb{R}^+} \prod_{j=1}^{n_i} \left[\frac{(1-p)\widetilde{f}(x_{ij})}{p+(1-p)\widetilde{S}(x_{ij})} \cancel{r(p+(1-p)\widetilde{S}(x_{ij}))^{r}} \right]^{\delta_{ij}} [p+(1-p)\widetilde{S}(x_{ij})]^{\cancel{r}(1-\delta_{ij})} g_R(r;\theta)\mathrm{d}r \\
&= \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[\frac{(1-p)\widetilde{f}(x_{ij})}{p+(1-p)\widetilde{S}(x_{ij})} \right]^{\delta_{ij}} \int_{\mathbb{R}^+} \prod_{j=1}^{n_i} r^{\delta_{ij}} S_r(x_{ij}) g_R(r;\theta)\mathrm{d}r \\
&= \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[\frac{(1-p)\widetilde{f}(x_{ij})}{p+(1-p)\widetilde{S}(x_{ij})} \right]^{\delta_{ij}} \int_{\mathbb{R}^+} r^{\sum_{j=1}^{n_i}\delta_{ij}} \prod_{j=1}^{n_i} S_r(x_{ij}) g_R(r;\theta)\mathrm{d}r
\end{aligned}
$$

Thus, given the general distribution $R \sim \text{Gamma}(shape = \alpha, rate = \beta)$, the development of the internal component is given by

$$
\begin{aligned}
&\int_{\mathbb{R}^+} r^{\sum_{j=1}^{n_i}\delta_{ij}} \prod_{j=1}^{n_i} S_r(x_{ij}) g_R(r;\theta)\mathrm{d}r = \int_{\mathbb{R}^+} \prod_{j=1}^{n_i} S_r(x_{ij}) r^{\sum_{j=1}^{n_i}\delta_{ij}} \frac{\beta^{\alpha}}{\Gamma(\alpha)} r^{\alpha-1}\mathrm{e}^{-\beta r}\mathrm{d}r \\
&= \int_{\mathbb{R}^+} \prod_{j=1}^{n_i} S_r(x_{ij}) \frac{\beta^{\alpha+\sum_{j=1}^{n_i}\delta_{ij}}}{\Gamma(\alpha+\sum_{j=1}^{n_i}\delta_{ij})} \frac{\Gamma(\alpha+\sum_{j=1}^{n_i}\delta_{ij})}{\Gamma(\alpha)\beta^{\sum_{j=1}^{n_i}\delta_{ij}}} r^{(\alpha+\sum_{j=1}^{n_i}\delta_{ij})-1}\mathrm{e}^{-\beta r}\mathrm{d}r \\
&= \int_{\mathbb{R}^+} \prod_{j=1}^{n_i} S_r(x_{ij}) \frac{\beta^{\alpha+\sum_{j=1}^{n_i}\delta_{ij}}}{\Gamma(\alpha+\sum_{j=1}^{n_i}\delta_{ij})} \frac{\Gamma(\alpha+\sum_{j=1}^{n_i}\delta_{ij})}{\Gamma(\alpha)\beta^{\sum_{j=1}^{n_i}\delta_{ij}}} r^{(\alpha+\sum_{j=1}^{n_i}\delta_{ij})-1}\mathrm{e}^{-\beta r}\mathrm{d}r \\
&= \prod_{j=1}^{n_i} \frac{\Gamma(\alpha+\sum_{j=1}^{n_i}\delta_{ij})}{\Gamma(\alpha)\beta^{\sum_{j=1}^{n_i}\delta_{ij}}} \int_{\mathbb{R}^+} H(x_{ij};p,\underline{y})^{r} g_{R^*}(r;\alpha,\sum_{j=1}^{n_i}\delta_{ij},\beta)\mathrm{d}r \\
&= \prod_{j=1}^{n_i} \frac{\Gamma(\alpha+\sum_{j=1}^{n_i}\delta_{ij})}{\Gamma(\alpha)\beta^{\sum_{j=1}^{n_i}\delta_{ij}}} \int_{\mathbb{R}^+} \mathrm{e}^{r\log(H(x_{ij};p,\underline{y}))} g_{R^*}(r;\alpha,\sum_{j=1}^{n_i}\delta_{ij},\beta)\mathrm{d}r \\
&= \prod_{j=1}^{n_i} \frac{\Gamma(\alpha+\sum_{j=1}^{n_i}\delta_{ij})}{\Gamma(\alpha)\beta^{\sum_{j=1}^{n_i}\delta_{ij}}} \mathbb{E}_{R^*}[\mathrm{e}^{r\log(H(x_{ij};p,\underline{y}))}] = \prod_{j=1}^{n_i} \frac{\Gamma(\alpha+\sum_{j=1}^{n_i}\delta_{ij})}{\Gamma(\alpha)\beta^{\sum_{j=1}^{n_i}\delta_{ij}}} \text{MGF}(R^*;\log(H(x_{ij};p,\underline{y}))) \\
&= \prod_{j=1}^{n_i} \frac{\Gamma(\alpha+\sum_{j=1}^{n_i}\delta_{ij})}{\Gamma(\alpha)\beta^{\sum_{j=1}^{n_i}\delta_{ij}}} \left(1-\frac{\log(H(x_{ij};p,\underline{y}))}{\beta}\right)^{-(\alpha+\sum_{j=1}^{n_i}\delta_{ij})}
\end{aligned}
$$

where we define the quantity $H(x_{ij};p,\underline{y}) = \prod_{j=1}^{n_i} S(x_{ij}) = \prod_{j=1}^{n_i}\left(p+(1-p)\widetilde{S}(x_{ij})\right)$ for ease of writing. Later we update the parameters of the frailty distribution to obtain the updated frailty risk random variable $R^* \sim \text{Gamma}(shape = \alpha + \sum_{j=1}^{n_i}\delta_{ij}, rate = \beta)$, so that we make use of its moment generating function (MGF) in the point $\log(H(x_{ij};p,\underline{y}))$. Recall that the definition of the MGF implies that

$$MGF(R;y) = \mathbb{E}_R[\mathrm{e}^{ry}],$$

which is exactly what we have. Specifically, for a Gamma distributed random variable $R \sim \text{Gamma}(shape = \alpha, rate = \beta)$ the MGF is given by

$$MGF(R; y) = \left(1 - \frac{y}{\beta}\right)^{-\alpha}.$$

Thus, the multivariate likelihood with $\alpha = \beta = \theta$ is given by

$$L(\underline{y}, p, \theta) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[\frac{(1-p)\widetilde{f}(x_{ij})}{p + (1-p)\widetilde{S}(x_{ij})}\right]^{\delta_{ij}} \frac{\Gamma(\theta + \sum_{j=1}^{n_i} \delta_{ij})}{\Gamma(\theta)\theta^{\sum_{j=1}^{n_i} \delta_{ij}}} \left(1 - \frac{\log(H(x_{ij}; p, \underline{y}))}{\theta}\right)^{-(\theta + \sum_{j=1}^{n_i} \delta_{ij})}.$$

The likelihood, always in the case $\alpha = \beta = \theta$, reduces to the univariate form when there is one subject per family, so that its form is given by

$$L(\underline{y}, p, \theta) = \prod_{i=1}^{n} \left[\frac{(1-p)\widetilde{f}(x_i)}{p + (1-p)\widetilde{S}(x_i)}\right]^{\delta_i} \frac{\Gamma(\theta + \delta_i)}{\Gamma(\theta)\theta^{\delta_i}} \left(1 - \frac{\log(p + (1-p)\widetilde{S}(x_i))}{\theta}\right)^{-(\theta + \delta_i)}.$$

Notice that, at this point, parameter estimation can be achieved through the maximization of the likelihood from the observed independently right-censored data. Once the estimated collection of parameters $\widehat{\underline{\pi}} = (\widehat{p}, \widehat{\underline{y}}, \widehat{\theta})^T$ is obtained, it can be consequently involved in the computation of quantities of interest.

If one is willing to ignore the Cure-Rate structure, one may think of a model that has a proper survival distribution, both conditionally on R and marginally. In other words, we may assume that there exists some finite time $T_{\max}$ (larger than all observed survival and censoring times) such that all conditional survival functions $S(t \mid r) = [S_0(t)]^r$ are proper, i.e. they tend to zero as $t \to +\infty$.

The popular and widely developed Cox model may be suitable for our case. We include in the comparative analysis a Multivariate frailty Cox model, where the semiparametric estimation of the model is possible, for example using the `emfrail` function available in the software `R` [4, 5], while maintaining the multiplicative frailty structure. This Multivariate frailty Cox model is the exact counterpart of the Multivariate frailty Cure-Rate model, which we would like to prove they have an equal level of prediction accuracy. For sake of completeness of the analysis, we also develop a Univariate FH Cox model, and a Univariate $FH(t)$ Cox model, with $FH(t)$ the time-varying version of the family history indicator. The difference between these two is that the indicator FH is measured at the end of the follow-up, while $FH(t)$ is measured over time and the time of the change-point from zero to one is tracked. Unfortunately it was not possible to include in the analysis the Univariate frailty Cox model because of its non-identifiability due to the combination of the absence of families and an unspecified hazard function. We prove the non-identifiability of the univariate Cox frailty model in the following lines.

Let us start from the multiplicative proportional hazards frailty model that has the observed data likelihood given by

$$L^*(\underline{\pi}; \underline{x}) = \prod_{i=1}^{n} f_X(\underline{x}_i; \underline{\pi}) = \prod_{i=1}^{n} \int_{\mathbb{R}^+} f_X(\underline{x}_i \mid R; \underline{\pi}) g_R(r) \mathrm{d}r = \prod_{i=1}^{n} \int_{\mathbb{R}^+} f_r(x_i)^{\delta_i} S_r(x_i)^{1-\delta_i} g_R(r) \mathrm{d}r,$$

where recall $\underline{\pi} = (p, \underline{\gamma}, \theta)^T$ and $S_r(t) = S(t \mid r)$ and $f_r(t) = f(t \mid r)$ are the (proper) survival function and density function conditional on the cases, respectively.

We note that, however, the Univariate frailty model is not identifiable without some additional structure. We dig deeper into the issue with the following lemma.

Lemma 1. *The Univariate frailty model is not identifiable if one does not specify the form of the baseline survival function $S_0(t)$.*

Proof. Recall the Gamma density function:

$$g_R(r; \alpha, \beta) = \frac{1}{\Gamma(\alpha)} \beta^\alpha r^{\alpha-1} \mathrm{e}^{-r\beta},\ r \geq 0,\ \alpha,\ \beta \in \mathbb{R}^+$$

Letting $\alpha = \beta = \theta > 0$ yields

$$g_R(r; \theta) = \frac{1}{\Gamma(\theta)} \theta^\theta r^{\theta-1} \mathrm{e}^{-r\theta}$$

Recall the moment generator function of the Gamma random variable:

$$MGF_R(y) = \mathbb{E}[\mathrm{e}^{yR}] = \left(1 - \frac{y}{\beta}\right)^{-\alpha},\ y < \beta \text{ or}$$
$$\left(1 - \frac{y}{\theta}\right)^{-\theta} \text{ when } \alpha = \beta = \theta > 0, \text{ for } y < \theta, \text{ and in particular, } \forall\ y \leq 0.$$

The Univariate frailty model has marginal survival function

$$S(t) = \int_{\mathbb{R}^+} S(t \mid r) g_R(r; \theta) \mathrm{d}r = \int_{\mathbb{R}^+} S(t)^r g_R(r; \theta) \mathrm{d}r = \int_{\mathbb{R}^+} \mathrm{e}^{\log(S(t)^r)} g_R(r; \underline{\theta}) \mathrm{d}r$$
$$= \int_{\mathbb{R}^+} \mathrm{e}^{r\log(S(t))} g_R(r; \theta) \mathrm{d}r = \left(1 - \frac{\log(S_0(t))}{\theta}\right)^{-\theta}. \text{ Note that, } \log(S_0(t)) < 0 \forall t \geq 0.$$

Now, for any observed marginal survival function $S(t)$ one has

$$S(t) = \int_{\mathbb{R}^+} S_0(t)^r g_R(r; \theta) \mathrm{d}r = \left(1 - \frac{\log(S_0(t))}{\theta}\right)^{-\theta} \iff S(t)^{-1/\theta} = 1 - \frac{\log(S_0(t))}{\theta}$$
$$\iff S(t)^{-1/\theta} - 1 = -\frac{\log(S_0(t))}{\theta} \iff \theta\left[S(t)^{-1/\theta} - 1\right] = -\log(S_0(t))$$
$$\iff \log(S_0(t)) = \theta\left(1 - S(t)^{-1/\theta}\right) \iff S_0(t) = \exp\left(\theta\left(1 - S(t)^{-1/\theta}\right)\right),$$

for any observed $S(t)$, to each choice of θ there is a corresponding $S_0(t)$. Thus, it exists an infinite set of couple $(\theta, S_0(t))$ with $S_0(t) = \exp\left(\theta\left(1 - S(t)^{-1/\theta}\right)\right)$ that yields to the same $S(t)$. This proves that the Univariate frailty model is not identifiable from the marginal survival function $S(t)$, unless additional constraints are introduced, such as for example on the distribution of $S_0(t)$. □

Thus, the Univariate frailty Cox model is not developed in the end and no results are reported from this model.

Wrapping up, we explore six different models: the Multivariate frailty Cure-Rate model, the Univariate frailty Cure-Rate model, the Univariate FH Cure-Rate model, the Multivariate frailty

Cox model, the Univariate FH Cox model, the Univariate $FH(t)$ Cox model. Our purpose is to observe their difference in inference precision and risk prediction accuracy. We expect the family history indicator to be a weaker indicator compared to the continuous frailty risk, in terms of explaining the phenomenon of breast cancer within a population. The Cox models can not infer some quantities as the cured fraction and the distribution of breast cancer cases, and this is certainly a disadvantage for these models. Also, the univariate models do not use the complete familiar information of the subjects and for this reason we expect them to be worst than the multivariate models in terms of risk prediction accuracy. Summing up all of these thoughts, we expect the Multivariate frailty Cure-Rate model to be the best performing in inference precision to explain the phenomenon breast cancer development, and accuracy in risk prediction.

1.3 Posterior risk prediction

The primary focus of this project is risk prediction. It is interesting to notice that the available data are used for fitting the models, the risk prediction can be applied also to women not included in the initial dataset. The risk prediction procedure is different according to whether the model is semiparametric or parametric. For this reason we make distinction in the following paragraphs.

1.3.1 Semiparametric setting

In the parameter estimation step, by maximizing the partial likelihood we obtain the estimated value of the frailty parameter $\widehat{\theta}$.

With the value of the frailty parameter we can obtain the estimated family-specific posterior frailty risk density. This step is important to quickly predict risk values without estimating parameters when a new family arises. The posterior distribution, for a family of three component, is given by

$$g(r \mid \text{family data}; \widehat{\theta}) = \frac{f(\text{family data} \mid R = r) g(r; \widehat{\theta})}{\int_{\mathbb{R}^+} f(\text{family data} \mid R = r) g(r; \widehat{\theta}) \mathrm{d}r},$$

where

$$f(\text{family data} \mid R = r) \overset{\perp | R}{=} f(\underline{x} \mid r) f(\underline{xm} \mid r) f(\underline{xs_1} \mid r),$$

and

$$f(\underline{x} \mid R = r) = f_r(x)^{\delta} S_r(x)^{1-\delta} = \left(\frac{f_r(x)}{S_r(x)} \right)^{\delta} S_r(x) = (r\lambda_0(x))^{\delta} S_0(x)^r.$$

A similar procedure is applied for the other family members, and also again, this is easily extendable to a higher number of female family members. At this point, an estimator for the baseline survival function $\widehat{S}_0(t)$ and for the baseline hazard function $\widehat{\lambda}_0(t)$ is needed such that one can estimate the subject-specific density function

$$\widehat{f}(\underline{x} \mid R = r) = (r\widehat{\lambda}_0(x))^{\delta} \widehat{S}_0(x)^r.$$

The baseline cumulative hazard function, and consequently the survival function $\widehat{S}_0(t) = \mathrm{e}^{-\widehat{\Lambda}_0(t)}$, can be estimated by the Breslow estimator $\widehat{\Lambda}_0(t)$.

In the following lines, we present the classical estimation of the Breslow estimator with covariates. We consider the expected number of events occurring at time $[\tau_j, \tau_{j+1}]$, given the lth subject "at risk" at time τ_j^-, who belongs the risk group defined as $R(\tau_j)$. Thus, we have

$$\sum_{l \in R(\tau_j)} (\tau_{j+1} - \tau_j)\lambda(\tau_j \mid \underline{x}_l) = \sum_{l \in R(\tau_j)} (\tau_{j+1} - \tau_j) e^{\underline{\beta}' \underline{x}_l} \lambda_0(\tau_j).$$

When setting the aforementioned quantity equal to the observed number of events d_j until time τ_{j+1}, it yields

$$d_j = \lambda_0(\tau_j)(\tau_{j+1} - \tau_j) \sum_{l \in R(\tau_j)} e^{\underline{\beta}' \underline{x}_l}$$

$$\lambda_0(\tau_j)(\tau_{j+1} - \tau_j) \approx \int_{t_j}^{t_{j+1}} \lambda_0(u)\mathrm{d}u = \frac{d_j}{\sum\limits_{l \in R(\tau_j)} e^{\widehat{\underline{\beta}}' \underline{x}_l}} \iff \widehat{\Lambda}_0(t) = \sum_{\tau_j < t} \left[\frac{d_j}{\sum\limits_{l \in R(\tau_j)} e^{\widehat{\underline{\beta}}' \underline{x}_l}} \right]$$

where $\underline{x}_l$ is the covariate vector for subject l, $\underline{\beta}$ is the parameter collection, and $\widehat{\Lambda}_0(t)$ is the Breslow estimator.

The estimated value of the survival function, through Breslow estimator, is obtained through the available function `predictCox` in the package `riskRegression` [25] of the software `R`, which takes as input a Cox model, while the estimated hazard function can be obtained in the function `coxphHaz` of package `biostat3` [38] of the software `R`.

We need also to estimate the other component of formula 1.3.1: the posterior distribution of the frailty risk $g(r; \widehat{\theta})$. The computation of $g(r; \widehat{\theta})$ is solved by relying on the conjugacy property of the Gamma prior for this model. Specifically, the posterior frailty risk distribution for a particular family is easily shown to be distributed as a Gamma distribution with shape parameter equal to $\theta + d_i$, with d_i the number of observed events in family i, and rate parameter equal to $\theta + \sum_{j=1}^{n_i} \Lambda_0(x_j)$, where $\Lambda_0(\underline{x}_i) = \sum_{j=1}^{n_i} \Lambda_0(x_j)$, with $\underline{x}_i = (x_1, \ldots, x_{n_i})^T$ for family size n_i, is the cumulative hazard function at time 0 evaluated at the observed times of the family's members. For a family of three

members the posterior frailty risk distribution is given by

$$g_R(r \mid \underline{X}) = (r\widehat{\lambda}_0(x_i))^{\delta_i}\widehat{S}_0(x_i)^r(r\widehat{\lambda}_0(xm_i))^{\delta m_i}\widehat{S}_0(xm_i)^r(r\widehat{\lambda}_0(xs_{1i}))^{\delta s_{1i}}\widehat{S}_0(xs_{1i})^r g_R(r;\widehat{\theta})$$

$$g_R(r \mid \underline{X}) = r^{\delta_i+\delta m_i+\delta s_{1i}}\left(\widehat{S}_0(x_i)\widehat{S}_0(xm_i)\widehat{S}_0(xs_{1i})\right)^r g_R(r;\widehat{\theta})$$

$$g_R(r \mid \underline{X}) \propto r^{\delta_i+\delta m_i+\delta s_{1i}}\mathrm{e}^{r\left(-\widehat{\Lambda}_0(x_i)-\widehat{\Lambda}_0(xm_i)-\widehat{\Lambda}_0(xs_{1i})\right)}g_R(r;\widehat{\theta}) \iff g_R(r \mid \underline{X}) \propto r^{d_i+\theta-1}\mathrm{e}^{-r\left(\theta+\sum_{j=1}^{n_i}\widehat{\Lambda}_0(x_j)\right)}$$

$$g_R(r \mid \underline{X}) = \frac{r^{d_i+\theta-1}\mathrm{e}^{-r\left(\theta+\sum_{j=1}^{n_i}\widehat{\Lambda}_0(x_j)\right)}}{\int_{\mathbb{R}^+} r^{d_i+\theta-1}\mathrm{e}^{-r\left(\theta+\sum_{j=1}^{n_i}\widehat{\Lambda}_0(x_j)\right)}\mathrm{d}r}$$

$$g_R(r \mid \underline{X}) = \frac{r^{h_i(\theta)-1}\mathrm{e}^{-rl_i(\theta)}}{\left(\frac{1}{l_i(\theta)}\right)^{h_i(\theta)}\Gamma(h_i(\theta))\int_{\mathbb{R}^+}\frac{1}{\left(\frac{1}{l_i(\theta)}\right)^{h_i(\theta)}\Gamma(h_i(\theta))}r^{h_i(\theta)-1}\mathrm{e}^{-rl_i(\theta)}\mathrm{d}r}$$

$$g_R(r \mid \underline{X}) = \frac{r^{h_i(\theta)-1}\mathrm{e}^{-rl_i(\theta)}l_i(\theta)^{h_i(\theta)}}{\Gamma(h_i(\theta))},$$

where $h_i(\theta) = d_i + \theta,\ l_i(\theta) = \theta + \sum_{j=1}^{n_i}\widehat{\Lambda}_0(x_j)$. The final formula is given by

$$g_R(r_i \mid \underline{X}) = \frac{r_i^{d_i+\theta-1}\exp(-r_i(\theta+\widehat{\Lambda}_0(\underline{x}_i)))(\theta+\widehat{\Lambda}_0(\underline{x}_i))^{d_i+\theta}}{\Gamma(d_i+\theta)}.$$

This formula is also computed with a different parametrization of the Gamma distribution [8], for the maximization step of the EM algorithm in order to achieve parameter estimation.

Notice that to obtain the posterior frailty distribution of each family we need to evaluate the cumulative hazard function at the observed times within each family. This means that, if the estimated cumulative hazard value is not available for that observed time, we use linear interpolation on the available data points [22].

Typically, the raw posterior frailty risk distribution has a poor meaning to the end-user. We need to communicate to each family a meaningful summary of it as, for example, the mode, median or mean, whose equations are given by

$$\text{mode} = \arg\max_r g_R(r \mid \underline{X}),$$

$$\text{median} = r_m : \int_0^{r_m} g_R(r \mid \underline{X})\mathrm{d}r = 0.5,$$

$$\text{mean} = \int_0^{+\infty} r g_R(r \mid \underline{X})\mathrm{d}r.$$

The mode of a Gamma distribution is null when $\widehat{\theta}$ has a value lower than one. The mode will therefore not be a good indicator of the posterior frailty density because one may observe several zero values across the families. We therefore exclude it as a valid indicator. On the hand, from the formulas of the Gamma distribution, the mean is immediately obtained as

$$\widehat{r}_i = \frac{d_i + \widehat{\theta}}{\widehat{\theta} + \widehat{\Lambda}_0(\underline{x}_i)}$$

which, following [4], can be rewritten as

$$\widehat{r}_i = w_i \frac{d_i}{\widehat{\Lambda}_0(\underline{x}_i)} + (1 - w_i),$$

where $w_i = \dfrac{\widehat{\Lambda}_0(\underline{x}_i)}{\widehat{\theta} + \widehat{\Lambda}_0(\underline{x}_i)}$ is the weight of the weighted average between the ratio of the observed events d_i on the sum of the cumulative hazard by families $\widehat{\Lambda}_0(\underline{x}_i)$ and one, the prior frailty mean. The median is instead computed numerically quite easily from the estimated family-specific posterior frailty risk distribution.

Because the posterior mean does not really have a scale of reference, we introduce also another meaningful summary of the posterior frailty risk distribution< i.e. the posterior probability of belonging to the highest-risk group by computing the probability that the family-specific risk value is greater than a fixed value from the prior distribution. By setting a percentile value α, the posterior probability of belonging to the highest-risk group is

$$P\left(R \geq F_R^{-1}\left(\alpha; \widehat{\theta}\right) \mid \text{family data}; \widehat{\theta}\right),$$

where $F_R^{-1}\left(\alpha; \widehat{\theta}\right)$ is the estimated prior quantile associated to a probability equal to α on the left side. In the real case study in Section 1.5 we choose an $\alpha = 0.05$, and we compute the corresponding quantile. Once we have the binary classification of families into two risk groups (low-risk and highest-risk group) based on their probability of belonging the the highest-risk group, we can validate it through the receiver operating characteristic (ROC) curve. We can extract the area under the curve (AUC), the positive predictive value (PPV) and the negative predictive value (NPV) from the ROC curve. These measures can assess the accuracy in binary splitting applied by the models. Notice that this validation procedure can be applied only in the case of simulation studies, since the true family-specific risk needs to be known.

1.3.2 Parametric setting

On the contrary to the semiparametric scenario, in the parametric scenario we need to specify a distribution for the baseline survival function. The advantage is that the baseline survival function is directly available by estimating its distribution parameters. This allows one to obtain the posterior mean, mode, and median from the parametric distribution of the posterior risk frailty. Also, the extension to predict the posterior risk frailty for a new family in the dataset in straightforward. The family-specific posterior frailty risk distribution given the whole family data is given

by

$$g_R(r \mid \underline{x}) = \frac{f(\underline{x} \mid r) g_R(r)}{f(\underline{x})} = \frac{f(x_1 \mid r) \dots f(x_{n_i} \mid r) g_R(r)}{\int_{\mathbb{R}^+} f(x_1 \mid r) \dots f(x_{n_i} \mid r) g_R(r) \mathrm{d}r}$$

$$= \frac{\prod_{j=1}^{n_i} \left(f_r(x_j)^{\delta_j} S_r(x_j)^{1-\delta_j} \right) g_R(r)}{\int_{\mathbb{R}^+} \prod_{j=1}^{n_i} \left(f_r(x_j)^{\delta_j} S_r(x_j)^{1-\delta_j} \right) g_R(r) \mathrm{d}r} =$$

$$= \frac{\prod_{j=1}^{n_i} \cancel{\left((1-p) \widetilde{f}(x_j) \right)^{\delta_j}} \left(p + (1-p) \widetilde{S}(x_j) \right)^{\cancel{(r-1)\delta_j} + r(1-\cancel{\delta_j})} r^{\sum_{j=1}^{n_i} \delta_j} g_R(r)}{\prod_{j=1}^{n_i} \cancel{\left(\frac{(1-p) \widetilde{f}(x_j)}{p + (1-p) \widetilde{S}(x_j)} \right)^{\delta_j}} \frac{\Gamma(\alpha + \sum_{j=1}^{n_i} \delta_j)}{\Gamma(\alpha) \beta^{\sum_{j=1}^{n_i} \delta_j}}} \left(1 - \frac{\log(H(x_j; p, \underline{\gamma}))}{\beta} \right)^{\alpha + \sum_{j=1}^{n_i} \delta_j}$$

$$= \frac{\cancel{\Gamma(\alpha)} \beta^{\sum_{j=1}^{n_i} \delta_j}}{\Gamma(\alpha + \sum_{j=1}^{n_i} \delta_j)} \left(1 - \frac{\log \prod_{j=1}^{n_i} (p + (1-p) \widetilde{S}(x_j))}{\beta} \right)^{\alpha + \sum_{j=1}^{n_i} \delta_j} \cdot$$

$$\cdot \prod_{j=1}^{n_i} \left(p + (1-p) \widetilde{S}(x_j) \right)^r r^{\sum_{j=1}^{n_i} \delta_j} \frac{\beta^\alpha}{\cancel{\Gamma(\alpha)}} r^{\alpha-1} \mathrm{e}^{-\beta r}$$

$$= \left(1 - \frac{\log \prod_{j=1}^{n_i} (p + (1-p) \widetilde{S}(x_j))}{\beta} \right)^{\alpha + \sum_{j=1}^{n_i} \delta_j} \prod_{j=1}^{n_i} \left(p + (1-p) \widetilde{S}(x_j) \right)^r \frac{\beta^{\alpha + \sum_{j=1}^{n_i} \delta_j}}{\Gamma(\alpha + \sum_{j=1}^{n_i} \delta_j)} r^{\alpha + \sum_{j=1}^{n_i} \delta_j - 1} \mathrm{e}^{-\beta r}$$

$$= \left(\frac{\beta - \log \prod_{j=1}^{n_i} (p + (1-p) \widetilde{S}(x_j))}{\cancel{\beta}} \right)^{\alpha + \sum_{j=1}^{n_i} \delta_j} \frac{\cancel{\beta^{\alpha + \sum_{j=1}^{n_i} \delta_j}}}{\Gamma(\alpha + \sum_{j=1}^{n_i} \delta_j)} r^{\alpha + \sum_{j=1}^{n_i} \delta_j - 1} \mathrm{e}^{-(\beta - \sum_{j=1}^{n_i} \log\left(p + (1-p) \widetilde{S}(x_j)\right)) r}$$

which is distributed as a $\text{Gamma}\left(shape = \alpha + \sum_{j=1}^{n_i} \delta_j, rate = \beta - \sum_{j=1}^{n_i} \log\left(p + (1-p) \widetilde{S}(x_j) \right) \right)$.

The density function, in the univariate case, reduces immediately to

$$g_R(r \mid \underline{x}) = \left(1 - \frac{\log(p + (1-p) \widetilde{S}(x))}{\beta} \right)^{\alpha + \delta} \left(p + (1-p) \widetilde{S}(x) \right)^r \frac{\beta^{\alpha+\delta}}{\Gamma(\alpha + \delta)} r^{\alpha + \delta - 1} \mathrm{e}^{-\beta r},$$

which is distributed as a $\text{Gamma}\Big(shape = \alpha + \delta, rate = \beta - \log(p + (1-p) \widetilde{S}(x)) \Big)$.

The posterior mean and mode frailty are extracted from the formula associated to the Gamma distribution as

$$\text{mean} = \frac{\alpha + \sum_{j=1}^{n_i} \delta_j}{\beta - \sum_{j=1}^{n_i} \log\left(p + (1-p) \widetilde{S}(x_j) \right)}, \qquad \text{mode} = \frac{\alpha + \sum_{j=1}^{n_i} \delta_j - 1}{\beta - \sum_{j=1}^{n_i} \log\left(p + (1-p) \widetilde{S}(x_j) \right)}.$$

The univariate case is directly obtained by setting family size $n_i = 1$. Note that when the shape parameter of the distribution is lower than one, the mode is not a reliable summary of the posterior frailty risk. Moreover, the posterior median of a Gamma distributed frailty risk requires numerical methods to be computed as there is not a formula in closed form. For these reasons the posterior mean appears to be the most reliable and easily to compute.

To validate that the mean is the best performing as posterior risk, the mean, median, and mode are compared in terms of various indices. When the analysis in on simulated data, we can compare

the true frailty risk from the data generating step r_i to the predicted frailty risk $\widehat{r}_i$, for n families. For example, we compute the estimated mean squared prediction error (MSPE), whose equation is given by

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{r}_i - r_i)^2.$$

Also, from the linear regression of the true frailty risk r_i on the predicted frailty risk $\widehat{r}_i$, we report the estimated coefficient of determination R^2 as a measure of how much the predicted risk can explain the variability of the true risk. The coefficient of the predicted frailty risk, the linear correlation coefficient of Pearson and the rank correlation coefficient are also extracted and reported.

Once we prove that the mean is the best performing posterior risk summary, also the models under analysis need to be validated in terms of prediction accuracy, depending on whether the analysis is on simulated studies or on the real case.

When the analysis is on simulated studies, we compute the AUC for the binary splitting and the Harrell's concordance index. We choose this index because it can be computed also in the real case application, as it is a measure of model performance for unknown outcomes [27, 15, 32, 39], and it makes the comparison between simulation study and real case much easier. This index estimates the probability that, among a randomly selected pair of subjects with different right-censored survival times and predicted risk values, the one with a lower risk score will outlive the other [15]. This define this pair as "concordant". Notice that this is similar to what is done in the modified version of the Wilcoxon test on survival data with right censoring.

The estimate relies on the ratio of concordant pairs to all pairs, which includes discordant and tied pairs. The formula is given by [33]

$$C = \frac{\sum_{i\neq j}\mathbb{I}(\widehat{r}_i < \widehat{r}_j)\mathbb{I}(x_i > x_j)\delta_j}{\sum_{i\neq j}\mathbb{I}(x_i > x_j)\delta_j},$$

where $\widehat{r}_i$ and $\widehat{r}_j$ denote the predicted risk for two subjects i and j who are not members of the same family. It is of interest to notice that from theoretical result the Harrell's concordance index does not depend on the baseline survival function, but only on a random variable $Y^* \sim \text{Beta}(\theta+1, \theta)$, such that

$$C(\theta) = S_{Y^*}\left(\frac{1}{2}\right).$$

An extended version of this result can be found in Appendix A.2.

1.4 Simulation study

We present results from the comparative analysis on the six models we developed or included from literature. In Table 1.1 we present the six models highlighting their commonalities and differences.

Name of the model	Likelihood	Frailty risk	FH	Survival
Multivariate frailty Cure-Rate	multivariate	✓	✗	cure-rate
Univariate frailty Cure-Rate	univariate	✓	✗	cure-rate
Univariate FH Cure-Rate	univariate	✗	constant	cure-rate
Multivariate frailty Cox	multivariate	✓	✗	traditional
Univariate FH Cox	univariate	✗	constant	traditional
Univariate $FH(t)$ Cox	univariate	✗	time-varying	traditional

Table 1.1: Commonalities and differences among the six models compared.

Notice that the Cure-Rate models we develop, i.e. the Multivariate frailty Cure-Rate, the Univariate frailty Cure-Rate, and the Univariate FH Cure-Rate model are parametric and thus in the analysis we explore several baseline survival distributions to be involved in the parametric models, including Weibull, Gamma, Lognormal, three-parameter Gamma, also known as the Pearson type III distribution (see, e.g., [6]), and three-parameter Lognormal. These two last distributions are included to assess whether an increased flexibility in the distribution could benefit both inference precision and prediction accuracy.

The section is organized starting with the data generating process, above all for the cure-rate structure, and then progressing to the fitting of semiparametric models and then parametric models. We need to make a distinction between them because only in the parametric scenario we analyse different time-to-event distributions. All over the section we compare the model goodness of fit, and their predictive accuracy through the AUC, PPV, NPV and Harrell's Concordance index.

1.4.1 Data generating process from the Lehmann Cure-Rate model

A critical aspect to consider when generating data from a cure-rate survival model [24] is that one needs to employ distinct methods for generating survival times for cases and cured observations.

Recall that the model takes the Lehmann form

$$S(t;r) = [S_0(t)]^r\,, \text{ with } S_0(t) = \left[p_0 + (1-p_0)\widetilde{S}(t)\right],\ i.e.\ \alpha(r) = r \text{ for } r > 0.$$

To generate survival times from this model (with a set of parameter values with varying lenght), one may proceed as follows:

1. generate a family-specific (or subject-specific if the model is univariate) frailty risk term r from the Gamma$(shape = \theta, rate = \theta)$ distribution;

2. generate a subject-specific value u from the Unif$(0,1)$ distribution;

3. if $u \leq p_0^r$

   then set $t = +\infty$ (a cured observation);

else solve $u = \left[p_0 + (1-p_0)\widetilde{S}(t)\right]^r$, which yields immediately

$$t = \widetilde{S}^{-1}\left(\frac{u^{1/r} - p_0}{1-p_0}\right) = \widetilde{F}^{-1}\left(\frac{1-u^{1/r}}{1-p_0}\right),$$

with $\widetilde{F}^{-1}(\cdot)$ the quantile function of the time-to-event distribution. Since we explore several distributions, we can easily find the two-parameter distributions in the basic `R` software, while the package `FAdist`, still from the `R` software, contains functions for the cumulative distribution function (CDF), density function, quantile function, and random sample function from the three-parameter Gamma and three-parameter Lognormal.

Independent (typically, administrative) right censoring of all observations can be applied to obtain the observed data. Censoring is the minimum between the date of death and the end year of the follow-up, set at 2020. For building the follow-up we need to fix a starting point for all the individuals, which is date of birth. We start from generating the mother date of birth from a Uniform between 1905 and 1945. The daughter dates of birth are generated summing up to their mother birth date a value from a Uniform between 25 and 35, which coincides to the age the mother gave birth. If one takes into account the fact that family members from different generations have different birth dates (hence in calendar time), follow-up is clearly longer for members born earlier. Date of death is generated similarly keeping a time coherence between mothers and daughters.

In the multivariate scenario, the data generating follows easily from the conditional independence assumption within each family, assuming for all family members the same frailty risk r. Also, in this scenario is crucial to random generate the size of families included in the sample. To this end, we randomly generate family size as one plus the minimum of a fixed number $(n_F - 1)$ and a value generated from a Poisson random variable with parameter λ_F. To give an insight, one can see n_F as the maximum family size all over the sample, and λ_F the average number of first-degree female relatives a subject can have.

The parameters involved in the models are set to $p = 0.85$, $shape = 8$, $scale = 6$, $threshold = 15$, $\theta = 0.2$ where, recall, p is the cured fraction, $shape$, $scale$, $threshold$ are the three parameters of the baseline survival function on cases. The threshold parameter is only used for a three-parameter distribution. And, θ is the frailty risk parameter. Also, notice that when we employ the Lognormal distribution the parameters are called differently $\mu = 6$ and $\sigma^2 = 8$, replacing respectively $shape$ and $scale$.

On the other hand, if a traditional survival function is involved, the data generation would typically be simpler. For example, if $S_r(t) = [S_0(t)]^r$ with $S_0(t)$ the traditional survival function of the Weibull(k, λ) random variable, it is immediate to check that $T \mid R = r \sim$Weibull$\left(k, \lambda\sqrt[k]{r}\right)$. Usually, a reparametrization dependent on the frailty risk is sufficient to obtain the updated distribution.

1.4.2 Semiparametric setting

Among the six models we compare, three of them are semiparametric: the Multivariate frailty Cox models, the Univariate FH Cox model and the Univariate $FH(t)$ Cox model.

In Table 1.2 we compare the models in terms of recovering the true parameter values set in the data generating process. We generate a sample of 100,000 families, with family size varying between 1 and $n_F = 5$, with a mean of $\lambda_F = 0.8$ relatives. Our experience suggests that a sample size of 100,000 families is sufficient to ensure accurate results with reasonable computational speed. The simulation is repeated for 100 times.

Parameter estimation is performed through an EM algorithm, that can be carried out using functions from the `R` package `frailtyEM` [4, 5]. However, the functions of the `frailtyEM` package are time and memory-consuming for large datasets. In contrast to the approach commonly found in the literature, we speed up this process by utilizing the `cox` function in the `survival` package. This allows us to efficiently obtain all the necessary quantities within a reasonable time and without excessive memory usage.

	θ	$\widehat{\theta}$ (se)	$\widehat{\beta}_{FH}$ (se)
Multivariate frailty Cox			
	0.2	0.199 (0.004)	-
	0.5	0.499 (0.012)	-
	0.8	0.802 (0.029)	-
Univariate FH Cox			
	0.2	-	2.835 (0.081)
	0.5	-	1.455 (0.040)
	0.8	-	1.102 (0.033)
Univariate $FH(t)$ Cox			
	0.2	-	5.846 (0.166)
	0.5	-	2.956 (0.083)
	0.8	-	2.227 (0.066)

Table 1.2: Mean and standard error of the estimated parameters and Harrell's Concordance index (C) in the last column. The first column θ stands for the true value set at the data generating step and varying among (0.2, 0.5, 0.8).

Results from Table 1.2 show that the Multivariate frailty Cox model can accurately estimate the value of θ from the given data, while the Univariate models only estimate the parameter β_{FH} and nothing can be said about parameter recovery. Even though the parameter β_{FH} has little meaning because is not part of the data generating process, it is interesting to notice is always estimated

positive. This means that the family history is a valid and positive indicator for breast cancer development.

Our focus now shifts towards the core of the study, which is frailty prediction. In this regard, the algorithm developed in Section 1.3 is employed to estimate the risk of developing breast cancer with the current dataset and to predict the risk for a new subject. The mean risk and median risk are predicted for each family. In terms of the Multivariate frailty Cox model, these two summary measures are compared in terms of MSPE, R^2 from the linear regression of the true risk on the predicted risk, the coefficient $\widehat{\beta}$ of the predicted risk, and both the Pearson's and the rank correlation coefficient. Clearly, this is not possible for the Univariate models since they compare the estimated family history to the true risk. The comparison is on a sample of 100,000 families and varying family size among $(n_F, \lambda_F) = ((2, 0.8), (5, 0.8), (10, 5), (20, 10))^T$, since differences emerge when family size increases. Average and standard error of the estimated measures are reported. Results in Table 1.3 show that higher is the average family size then higher are the performances of both mean and median risk, meaning that the family size affects in a positive way the risk prediction accuracy, as expected.

Results show that the median performs poorly as a risk predictor, especially when family size is small. However notice that, as the family size decreases, the mean is known to exhibit a shrinkage problem due to the low number of events occurring within families. Figure 1.1 illustrates this phenomenon, in which the predicted mean risk do not reach zero when the true risk is close to zero, and all values are shrunken towards the true (prior) mean of one. Notice that the second plot of Figure 1.1, resulting from the maximum family size equal to twenty, has a lower shrinkage compared to the first plot which has a maximum family size equal to five. This well-known phenomenon is due to the fact that a large family size is required to increase the sum of cumulative hazards on observed time and allow the mean to shrink towards zero [4, 5]. One should keep in mind this limitation of the posterior frailty risk mean when running risk prediction. For large dataset, the shrinkage effect becomes smaller and smaller towards zero and thus we do not encounter this problem on the real case data because of its high dimension.

	(n_F, λ_F)			
	(2, 0.8)	(5, 0.8)	(10, 5)	(20, 10)
Multivariate frailty Cox				
MSPE(mean)	3.80 (0.0020)	3.50 (0.0019)	2.85 (0.0024)	3.63 (0.0032)
MSPE(median)	5.79 (0.0025)	5.64 (0.0034)	5.01 (0.0031)	3.89 (0.0023)
R^2(mean)	0.23 (0.0002)	0.26 (0.0001)	0.52 (0.0001)	0.65 (0.0002)
R^2(median)	0.16 (0.0001)	0.16 (0.0001)	0.26 (0.0002)	0.29 (0.0002)
$\widehat{\beta}_{\text{mean}}$	0.94 (0.0013)	0.96 (0.0008)	0.74 (0.0003)	0.60 (0.0002)
$\widehat{\beta}_{\text{median}}$	11.67 (0.0120)	8.81 (0.0108)	3.08 (0.0015)	1.46 (0.0005)
ρ(mean)	0.49 (0.0002)	0.55 (0.0002)	0.74 (0.0001)	0.82 (0.0001)
ρ(median)	0.42 (0.0002)	0.43 (0.0002)	0.53 (0.0002)	0.55 (0.0002)
Rank ρ(mean)	0.24(0.0001)	0.28 (0.0001)	0.48 (0.0001)	0.58 (0.0001)
Rank ρ(median)	0.11 (0.0001)	0.15 (0.0001)	0.36 (0.0001)	0.49 (0.0001)
Univariate FH Cox				
MSPE	5.91 (0.0022)	5.87 (0.0038)	5.67 (0.0031)	5.23 (0.0027)
R^2	0.03 (0.0001)	0.04 (0.0001)	0.13 (0.0001)	0.16 (<0.0001)
$\widehat{\beta}_{FH}$	3.59 (0.0039)	3.34 (0.0035)	3.03 (0.0014)	2.70 (0.0010)
ρ	0.17 (0.0002)	0.19 (0.0002)	0.36 (0.0001)	0.40 (0.0001)
Rank ρ	0.14 (0.0001)	0.16 (0.0001)	0.35 (0.0001)	0.43 (0.0001)
Univariate $FH(t)$ Cox				
MSPE	2.32 (0.0067)	2.30 (0.0065)	2.08 (0.006)	2.22 (0.0064)
R^2	0.03 (0.0004)	0.04 (0.0003)	0.05 (0.0002)	0.04 (0.0001)
$\widehat{\beta}_{FH}$	0.91 (0.0062)	0.87 (0.0049)	0.62 (0.0017)	0.47 (0.0010)
ρ	0.18 (0.0011)	0.19 (0.0009)	0.23 (0.0004)	0.19 (0.0004)
Rank ρ	0.18 (0.0008)	0.20 (0.0006)	0.28 (0.0004)	0.25 (0.0004)

Table 1.3: Prediction accuracy measures for family size defined by the two parameters (n_F, λ_F) among the values (2, 0.8), (5, 0.8), (10, 5), (20, 10); n_F indicates the largest number of family members, and λ_F represents the parameter of the Poisson distribution used to generate the additional number of family members apart the main subject.

Also, another limitation of the posterior frailty risk mean is the variability in the prediction. The mean squared prediction error (MSPE) is quite high to provide an accurate estimate of the risk of breast cancer occurrence in a family. We explore the values of the root MSPE conditional on the value of the true frailty risk in Figure 1.2. The MSPE increases as the true frailty risk increases too, regardless of the family size. We observe a possible flattening of the curve with a larger family size in the second plot of Figure 1.2, but the prediction error remained high. This could lead to accurately identify the lower-risk families but not the higher-risk families. This means that we may

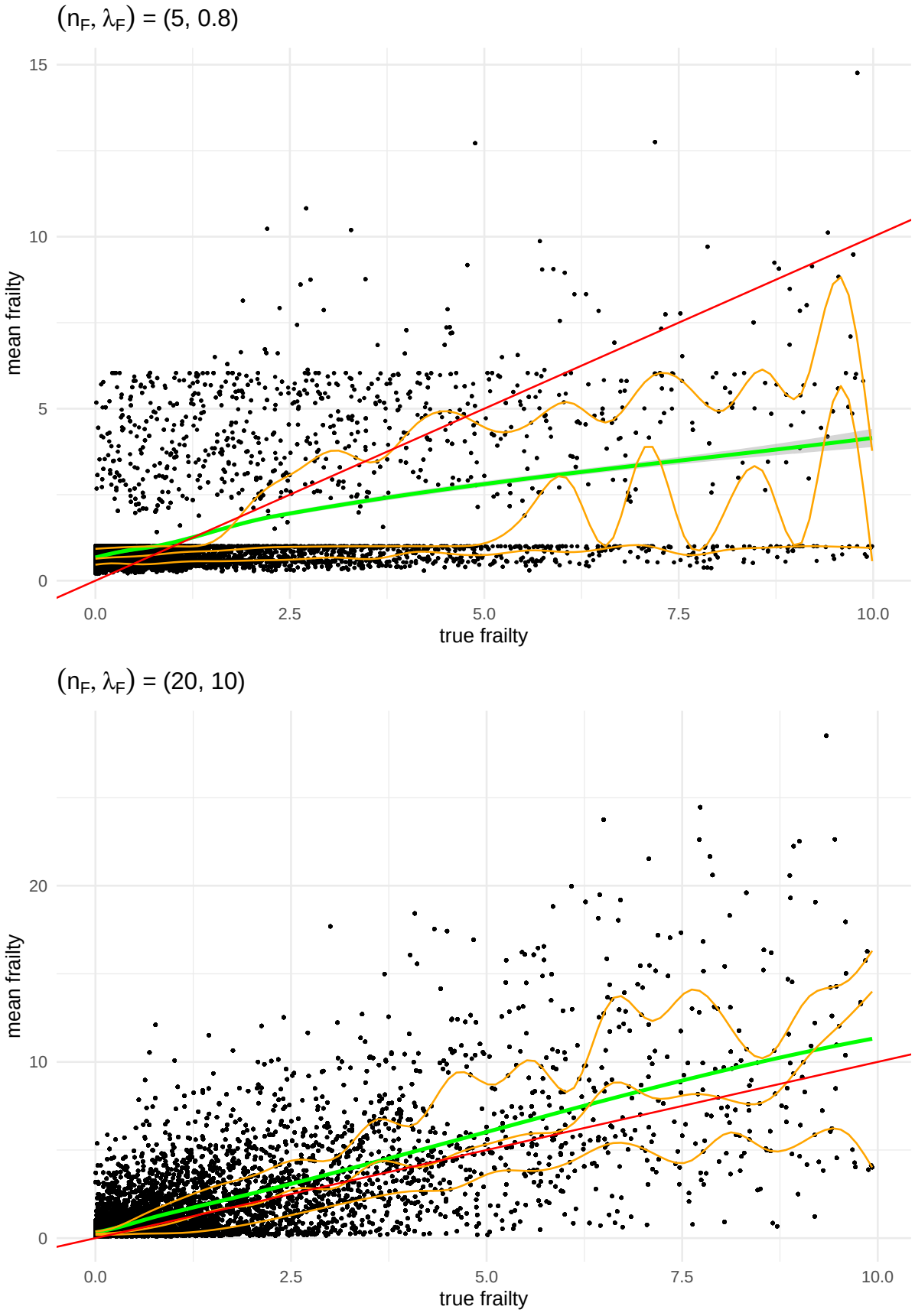

Figure 1.1: True frailty risk versus predicted frailty risk mean with the bisector of the third quadrant (in red), the smoothed regression line (in green with confidence interval) and smoothed quantile regression line on the first, second and third quantile (in orange) for a smaller and higher family size, respectively starting from the top, equal to a maximum of 5 and 20.

encounter a large number of false negative cases. Notice that the family size is not an actionable but it acts directly on the equation of mean, median and family history. We would expect to be more accurate with more information from the family, however it is not guaranteed to happen.

To make the posterior risk mean more meaningful to the end-user, we compute the probability that it exceeds the $(1-\alpha)$ percentile (say 95%) of the prior frailty risk Gamma distribution. In Figure 1.3 the plot of the the true risk percentile versus the predicted probability of belonging to the highest-risk families is reported. This allows us to assess that the relation between the true risk percentile and the predicted probability of belonging to the highest-risk family is coherent, i.e. the (posterior) predicted probability increases when the (prior) true probability increases. Notice that we can distinguish two clouds of points which can be seen as representing the splitting of families into two risk groups. Indeed, a binary splitting can be run fixing the threshold to the 95th quantile in the predicted distribution of the probability of belonging to the highest-risk families, estimated at 0.15.

Therefore, after all of the analyses, we adopt the mean as the best posterior frailty risk summary for Multivariate frailty Cox model.

Prediction accuracy for the posterior mean and the family history, respectively for the Multivariate frailty Cox model and the two Univariate Cox models, is reported in Table 1.4 through the Harrell's Concordance index. Notice that the value of the Harrell's Concordance index depends on the discrete nature of the predicted risk: for example, if a model predicts only two levels of risk, then the numerator of the index will tend to be small, as all tied pairs (i, j) for which $\widehat{r}_i = \widehat{r}_j$ are excluded. This may happen with the two Univariate Cox models which include FH and $FH(t)$. Results show that the Multivariate frailty Cox model is, not surprisingly, the best in prediction performances, as it is used to generate the data. Because the Concordance value from the Multivariate frailty Cox model ranges in between 0.83 and 0.95, it outperforms the Univariate models which are able to coherently predict the risk only for 50 to 62% of pairs (i, j).

Also, the accuracy of the binary classification can be analyzed by reporting the AUC, the positive predicted value PPV, and the negative predictive value NPV. As a result, the AUC increases when the family size is larger, as expected. This can be appreciated also from the ROC curve in Figure 1.4, where the scenario with maximum family size equal to twenty correctly classifies the 97% of families in the right group, the 14% in average more than the scenario with maximum five family members.

In conclusion, the Multivariate Shared Frailty Cox model appears to be an effective approach to the study of family history effects and risk prediction. Once the posterior frailty distribution is obtained for each family, we recommend to report the posterior frailty mean as well as the posterior probability and binary indicator of belonging to the highest-risk families based on a threshold from the population (prior) distribution (for example the top 5% as above).

A sensitivity analysis is additionally run to assess the robustness of the parameter estimates according to the threshold value. The family size is set at maximum 5 with average number of rela-

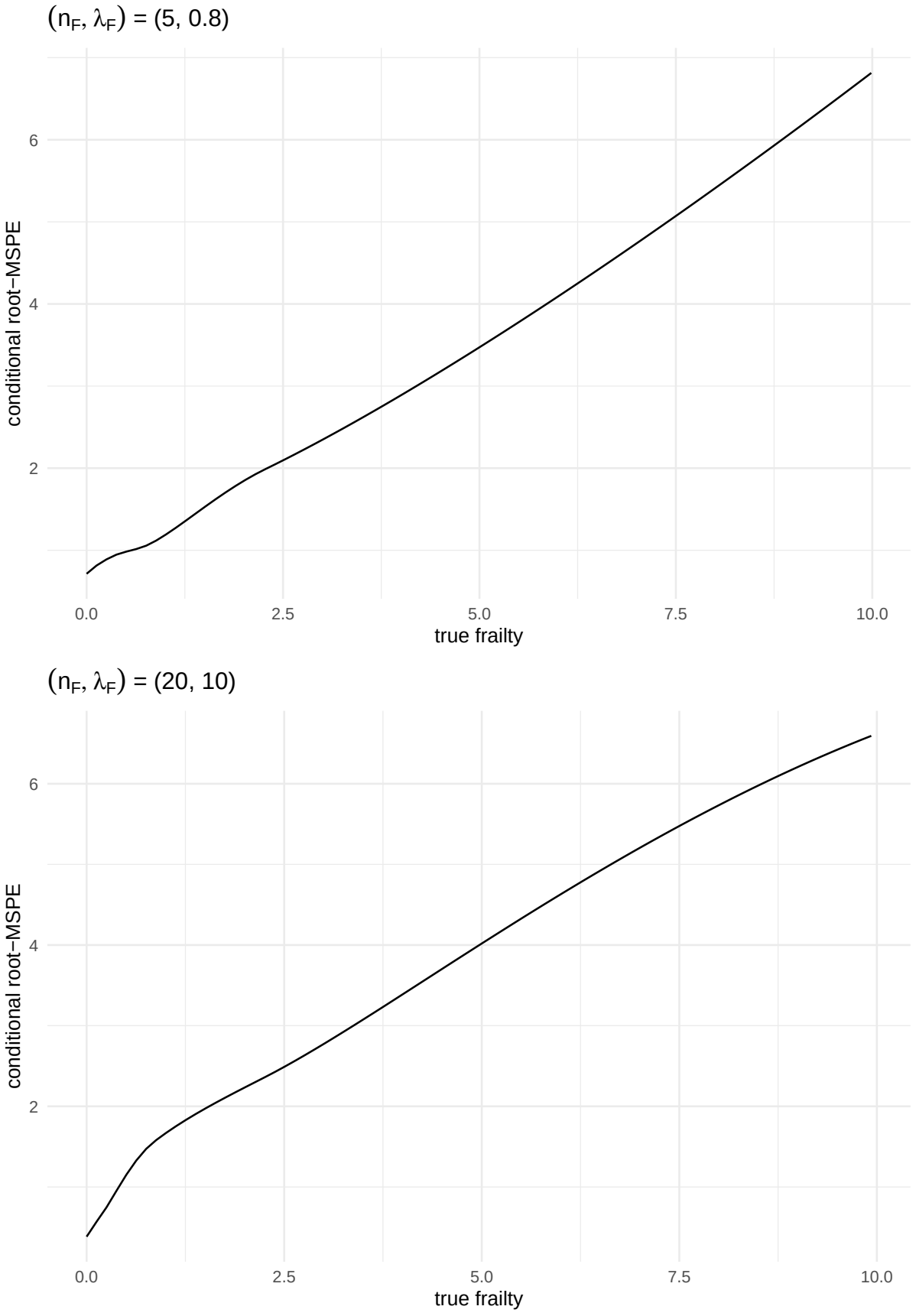

Figure 1.2: True frailty risk versus the conditional MSPE in squared root, for a lower and a higher family size, respectively starting from the top, equal to a maximum of 5 and 20.

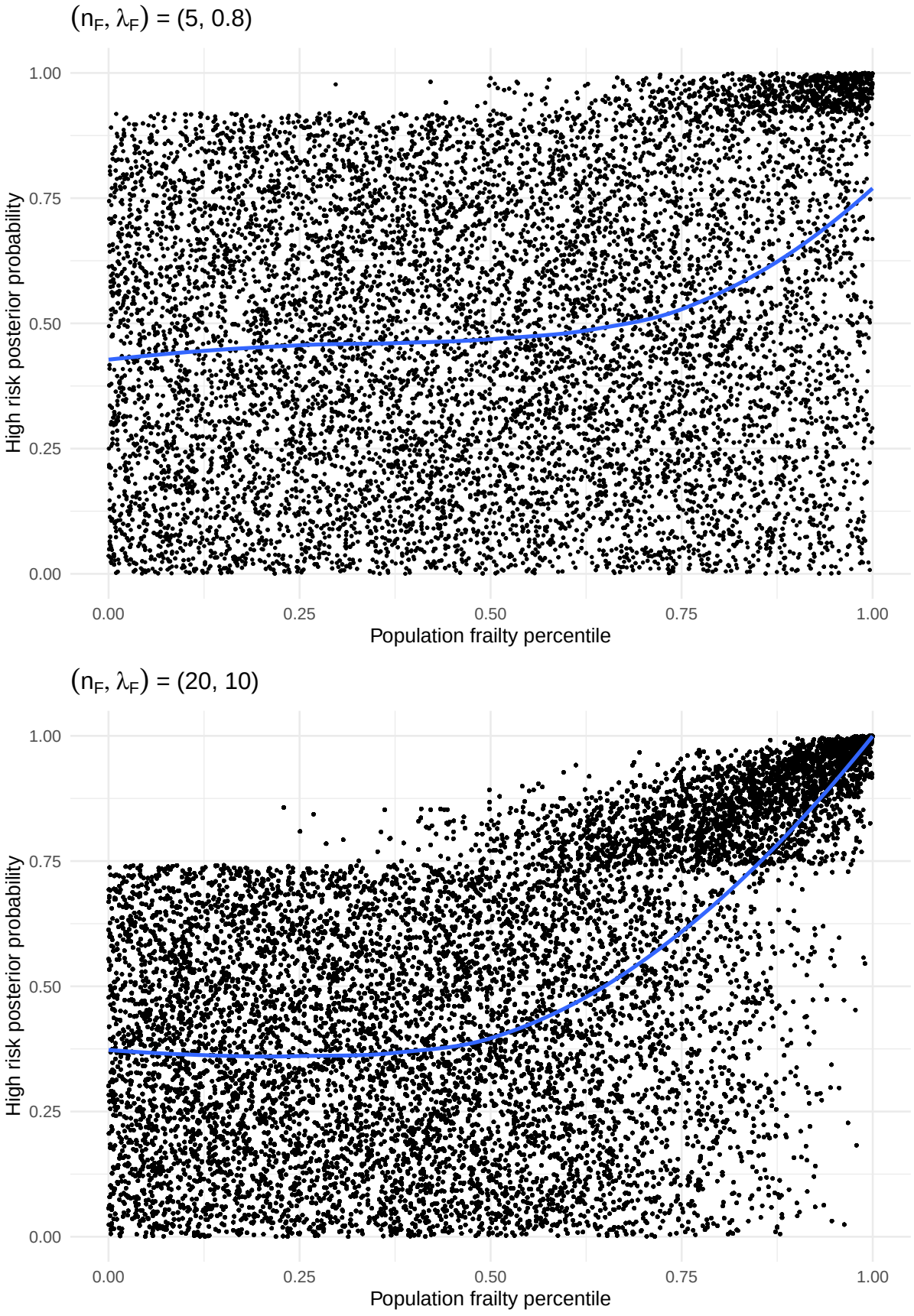

Figure 1.3: True frailty risk percentiles versus the predicted probability of belonging to the highest-risk families, with a smoothed regression line (in blue).

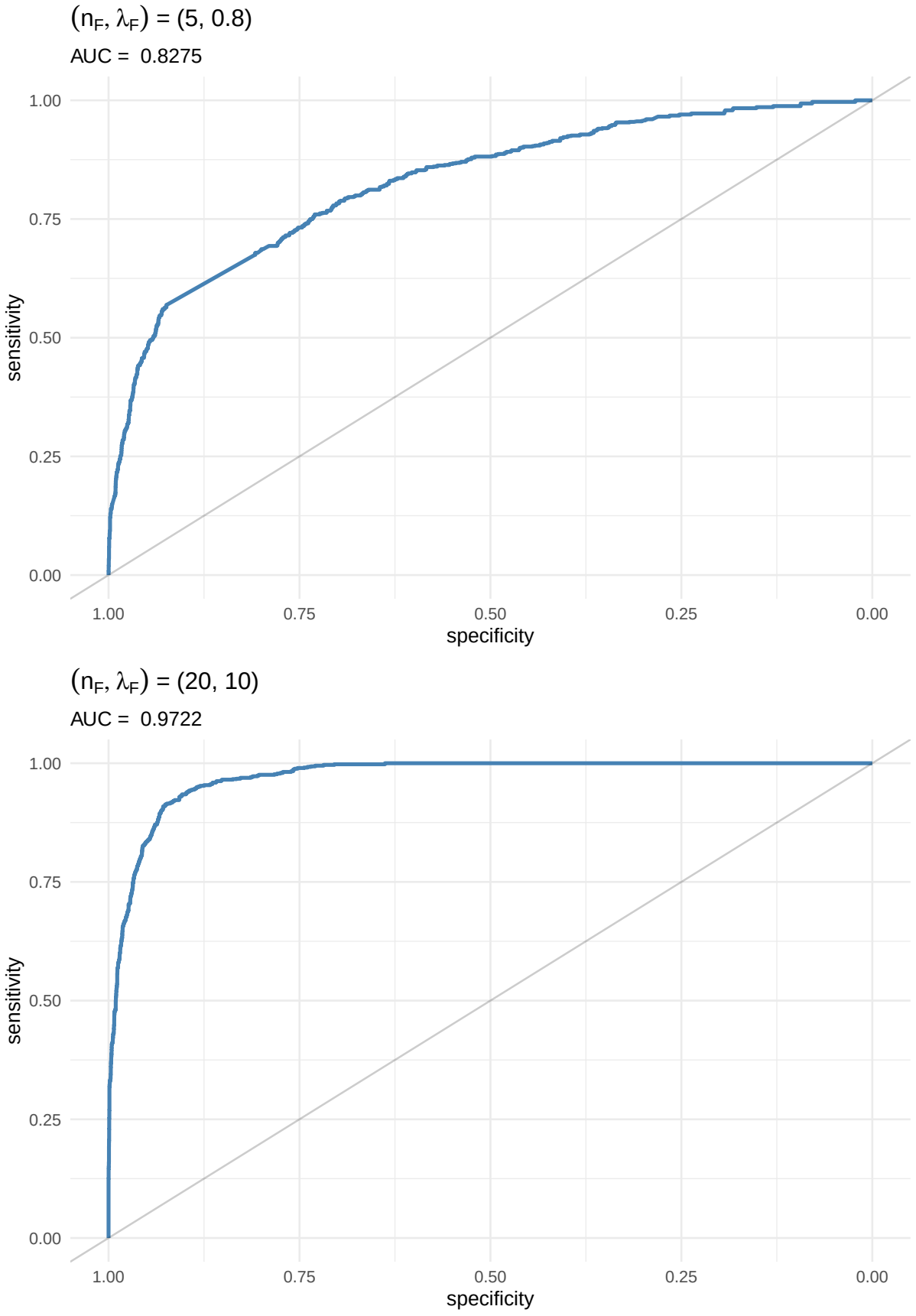

Figure 1.4: ROC curve of the predicted indicator of belonging to the highest-risk families versus the true membership to the highest-risk families (top 5%).

	(n_F, λ_F)			
	(2, 0.8)	(5, 0.8)	(10, 5)	(20, 10)
Multivariate frailty Cox				
AUC	0.73 (0.0001)	0.83 (0.0001)	0.88 (0.0001)	0.97 (<0.0001)
PPV	0.35 (0.0001)	0.37 (0.0001)	0.52 (0.0001)	0.61 (0.0001)
NPV	0.93 (<0.0001)	0.97 (<0.0001)	0.95 (<0.0001)	0.96 (<0.0001)
Concordance	0.96 (<0.0001)	0.94 (<0.0001)	0.86 (<0.0001)	0.83 (<0.0001)
Univariate FH Cox				
AUC	0.56 (0.0001)	0.57 (0.0001)	0.70 (0.0001)	0.69 (0.0001)
PPV	0.28 (0.0002)	0.26 (0.0001)	0.21 (<0.0001)	0.17 (<0.0001)
NPV	0.95 (<0.0001)	0.93 (<0.0001)	0.70 (<0.0001)	0.54 (0.0001)
Concordance	0.50 (<0.0001)	0.50 (<0.0001)	0.51 (0.0001)	0.52 (0.0001)
Univariate $FH(t)$ Cox				
AUC	0.55 (<0.0001)	0.56 (<0.0001)	0.59 (<0.0001)	0.57 (<0.0001)
PPV	0.26 (0.0001)	0.24 (0.0001)	0.16 (<0.0001)	0.14 (<0.0001)
NPV	0.91 (<0.0001)	0.88 (<0.0001)	0.52 (0.0001)	0.36 (0.0001)
Concordance	0.52 (<0.0001)	0.52 (<0.0001)	0.59 (0.0001)	0.62 (0.0001)

Table 1.4: AUC, PPV, and NPV corresponding to the ROC curve on the binary frailty risk with threshold (1 - α) = 0.95, and Harrell's Concordance index.

tives to 0.8. The threshold is varying among 0.85, 0.9, 0.95, 0.99. Results show (Table 1.5) that

	(1-α)			
	0.85	0.9	0.95	0.99
Multivariate frailty Cox				
AUC	0.94 (0.000001)		0.83 (0.0001)	
PPV			0.37 (0.0001)	
NPV	0.93 (<0.0001)	0.97 (<0.0001)	0.95 (<0.0001)	0.96 (<0.0001)
Concordance	0.96 (<0.0001)	0.94 (<0.0001)	0.86 (<0.0001)	0.83 (<0.0001)
Univariate FH Cox				
AUC	0.56 (0.0001)	0.57 (0.0001)	0.70 (0.0001)	0.69 (0.0001)
PPV	0.28 (0.0002)	0.26 (0.0001)	0.21 (<0.0001)	0.17 (<0.0001)
NPV	0.95 (<0.0001)	0.93 (<0.0001)	0.70 (<0.0001)	0.54 (0.0001)
Concordance	0.50 (<0.0001)	0.50 (<0.0001)	0.51 (0.0001)	0.52 (0.0001)
Univariate $FH(t)$ Cox				
AUC	0.55 (<0.0001)	0.56 (<0.0001)	0.59 (<0.0001)	0.57 (<0.0001)
PPV	0.26 (0.0001)	0.24 (0.0001)	0.16 (<0.0001)	0.14 (<0.0001)
NPV	0.91 (<0.0001)	0.88 (<0.0001)	0.52 (0.0001)	0.36 (0.0001)
Concordance	0.52 (<0.0001)	0.52 (<0.0001)	0.59 (0.0001)	0.62 (0.0001)

Table 1.5: AUC, PPV, and NPV corresponding to the ROC curve on the binary frailty risk with varying threshold (1 - α), and Harrell's Concordance index.

On the other side, the Multivariate frailty Cox model is poor in estimating some meaningful measures for explaining the phenomenon of breast cancer. That is why we develop the three aforementioned models into the parametric scenario. This allows us to employee the cure-rate survival function which helps us identifying the cured fraction into the population and estimating the survival curve of breast cancer cases. Let us take a closer look at the development of parametric models in next section.

1.4.3 Parametric setting

Now we explore the Multivariate frailty Cure-Rate, the Univariate frailty Cure-Rate, and the Univariate FH Cure-Rate model in terms of parameter recovery by varying the baseline survival function and accuracy in risk prediction.

Parameter estimation in the parametric case allows for a more detailed description compared to the semiparametric scenario, due to the inclusion of a parametric baseline survival function and the cure-rate structure. We conduct this simulation study on $n = 100,000$ families for 100 iterations to obtain average point estimates and their standard errors across repetitions. As above, to generate the family size we use a parameter to control the maximum of the family size n_F and the rate parameter of a Poisson distribution λ_F which, recall, it can be seen as the average number

of first-degree female relatives that a woman has. For the parameter recovery we set the parameters at $(n_F, \lambda_F) = (5, 0.8)$. The cured fraction p is set at 0.85. The frailty parameter is varying to assess estimate stability across its values. While, the baseline survival distribution has parameters $shape_0 = \mu_0 = 8$, $scale_0 = \sigma_0 = 6$, $\gamma_0 = 15$, according to the distribution it takes among Weibull, Gamma, Lognormal, three-parameter Gamma and three-parameter Lognormal. In Tables 1.6, 1.7, 1.8, 1.9, 1.10 we report the true value of the frailty parameters, varying among 0.2, 0.5, 0.8 in the first column, followed by the recovery of the baseline survival parameters, and the recovery of the frailty parameter in the last column.

	θ	$\widehat{p}_0$ (se)	$\widehat{shape}_0$ (se)	$\widehat{scale}_0$ (se)	$\widehat{\theta}$ (se)
Multivariate frailty Cure-Rate					
Mean (se)	0.2	0.85 (0.0016)	7.99 (0.0498)	5.99 (0.0063)	0.20 (0.0043)
MSE		0.0316	0.2581	0.103	0.0374
Mean (se)	0.5	0.85 (<0.0001)	8.00 (0.0004)	6.00 (0.0003)	0.50 (0.0001)
MSE		0.0316	0.1265	0.0949	0.0316
Mean (se)	0.8	0.85 (0.0001)	8.00 (0.0004)	6.00 (0.0006)	0.80 (0.0003)
MSE		0.0316	0.1265	0.1897	0.0949
Univariate frailty Cure-Rate					
Mean (se)	0.2	0.79 (0.0008)	8.12 (0.0366)	6.37 (0.0374)	0.11 (0.0007)
MSE		0.26	11.5746	11.8327	0.239
Mean (se)	0.5	0.81 (0.0011)	8.59 (0.0241)	6.06 (0.0030)	0.39 (0.0170)
MSE		0.3501	7.6439	0.9506	5.377
Mean (se)	0.8	0.81 (0.0012)	8.36 (0.0247)	6.07 (0.0035)	0.62 (0.0568)
MSE		0.3816	7.8191	1.109	17.9621
Univariate FH Cure-Rate					$\widehat{\beta}_{FH}$ (se)
Mean (se)	0.2	0.89 (0.0014)	5.06 (0.0653)	6.15 (0.0837)	2.88 (0.0587)
MSE		0.4445	20.8579	26.4687	18.755
Mean (se)	0.5	0.88 (0.0016)	4.97 (0.0574)	5.90 (0.0119)	2.02 (0.0505)
MSE		0.5069	18.4026	3.7644	16.0417
Mean (se)	0.8	0.89 (0.0013)	4.72 (0.0633)	5.85 (0.0146)	1.84 (0.0385)
MSE		0.413	20.2842	4.6194	12.2191

Table 1.6: Parameter recovery with a Weibull($shape = 8$, $scale = 6$) baseline survival function.

	θ	$\widehat{p}_0$ (se)	$\widehat{shape}_0$ (se)	$\widehat{scale}_0$ (se)	$\widehat{\theta}$ (se)
Multivariate frailty Cure-Rate					
Mean (se)	0.2	0.85 (<0.0001)	8.24 (0.0003)	5.96 (0.0003)	0.22 (<0.0001)
MSE		0.1897	0.2581	0.103	0.0374
Mean (se)	0.5	0.85 (0.0002)	8.01 (0.0009)	5.99 (0.0008)	0.50 (0.0002)
MSE		0.0632	0.2848	0.2532	0.0632
Mean (se)	0.8	0.85 (<0.0001)	7.99 (0.0003)	6.01 (0.0003)	0.79 (0.0002)
MSE		0.0316	0.0954	0.0954	0.2968
Univariate frailty Cure-Rate					
Mean (se)	0.2	0.79 (0.0006)	9.04 (0.0580)	5.79 (0.0385)	0.13 (0.0021)
MSE		0.199	18.3707	12.1766	0.6678
Mean (se)	0.5	0.79 (0.0009)	8.47 (0.0384)	6.03 (0.0335)	0.17 (0.0039)
MSE		0.2909	12.1522	10.5937	1.2767
Mean (se)	0.8	0.78 (0.0012)	8.08 (0.0563)	6.88 (0.0765)	0.22 (0.0047)
MSE		0.3859	17.8038	24.2074	1.5954
Univariate FH Cure-Rate					$\widehat{\beta}_{FH}$ (se)
Mean (se)	0.2	0.69 (0.0103)	4.51 (0.0776)	53.59 (2.2848)	2.51 (0.0662)
MSE		3.2611	24.7862	724.0828	21.0613
Mean (se)	0.5	0.83 (0.0037)	5.98 (0.0713)	15.05 (0.6704)	1.10 (0.0247)
MSE		1.1702	22.6373	212.1922	7.8338
Mean (se)	0.8	0.85 (0.0012)	6.77 (0.0737)	8.95 (0.2033)	0.79 (0.0247)
MSE		0.3795	23.3384	64.3568	7.8108

Table 1.7: Parameter recovery with a Gamma($shape = 8$, $scale = 6$) baseline survival function.

The findings presented in Tables 1.6, 1.7, 1.8, 1.9, and 1.10 align with our initial expectations: the Multivariate frailty Cure-Rate model is able to accurately identify and estimate the model parameters, given that the data are generated from a multivariate scenario. On the contrary, the Univariate frailty Cure-Rate model loses prediction accuracy by excluding family information; while the Univariate FH Cure-Rate model is not able to recover well the parameters because the family history coefficient is not even part of the data generating process.

However, there are some identification issues regarding the three-parameter distribution. At least, the three-parameter Gamma Multivariate frailty Cure-Rate model can almost accurately identify the true value of the parameters used in the data generating process. Indeed, from the estimated Kaplan-Meier curve in Figure 1.5 we can appreciate how in these scenarios the survival curves remain constant before the time fixed as threshold, that in this case is set at 15, when

	θ	$\widehat{p}_0$	$\widehat{\mu}_0$	$\widehat{\sigma}_0^2$	$\widehat{\theta}$
Multivariate frailty Cure-Rate					
Mean (se)	0.2	0.84 (0.0006)	7.89 (0.0192)	5.93 (0.0060)	0.19 (0.0004)
MSE		0.19	6.0726	1.8987	0.1269
Mean (se)	0.5	0.83 (0.0009)	8.28 (0.0238)	6.01 (0.0068)	0.55 (0.0017)
MSE		0.2853	7.5314	2.1504	0.5399
Mean (se)	0.8	0.84 (0.0008)	8.05 (0.0227)	6.01 (0.0066)	0.95 (0.0039)
MSE		0.2532	7.1785	2.0871	1.2424
Univariate frailty Cure-Rate					
Mean (se)	0.2	0.89 (0.0001)	6.05 (0.0015)	5.42 (0.0015)	0.37 (0.0036)
MSE		0.051	2.0069	0.7493	1.151
Mean (se)	0.5	0.89 (0.0001)	6.16 (0.0017)	5.37 (0.0014)	2.42 (0.0589)
MSE		0.051	1.9169	0.77	18.7245
Mean (se)	0.8	0.89 (0.0001)	6.07 (0.0028)	5.34 (0.0021)	0.65 (0.0052)
MSE		0.051	2.1234	0.9363	1.6512
Univariate FH Cure-Rate					$\widehat{\beta}_{FH}$ (se)
Mean (se)	0.2	0.96 (0.0004)	0.32 (0.0465)	3.75 (0.0472)	0.17 (0.0398)
MSE		0.1676	16.5894	15.0946	12.5859
Mean (se)	0.5	0.95 (0.0007)	0.84 (0.0525)	3.89 (0.0665)	-0.31 (0.0439)
MSE		0.2429	18.0801	21.1347	13.906
Mean (se)	0.8	0.95 (0.0006)	1.09 (0.0555)	3.65 (0.0266)	-0.53 (0.0467)
MSE		0.2145	18.8619	8.7338	14.8276

Table 1.8: Parameter recovery with Lognormal($\mu_0 = 8$, $\sigma_0 = 6$) baseline survival function.

censoring and breast cancer cases begin to show up. On the contrary, the three-parameter Lognormal distribution Multivariate frailty Cure-Rate model shows some problems in recovering the true values of the parameters. We can appreciate a particular behaviour of the estimated Kaplan-Meier curve in Figure 1.6: the jump crossing the threshold is not smooth as it is in the Gamma scenario; and also, it shows that censoring and cases happen after long time. We need to investigate more deeply about this particular behaviour. Perhaps, a higher flexibility into the parameter is not needed to raise the prediction accuracy.

Once the parameters are estimated, we can proceed to run risk prediction by computing the posterior frailty mean, median, and mode by using the parametric equations reported in Section 1.3.2. Notice that in this case we run the simulation with a Weibull($shape_0 = 8$, $scale_0 = 6$), cured fraction $p = 0.85$, and frailty parameter $\theta = 0.2$.

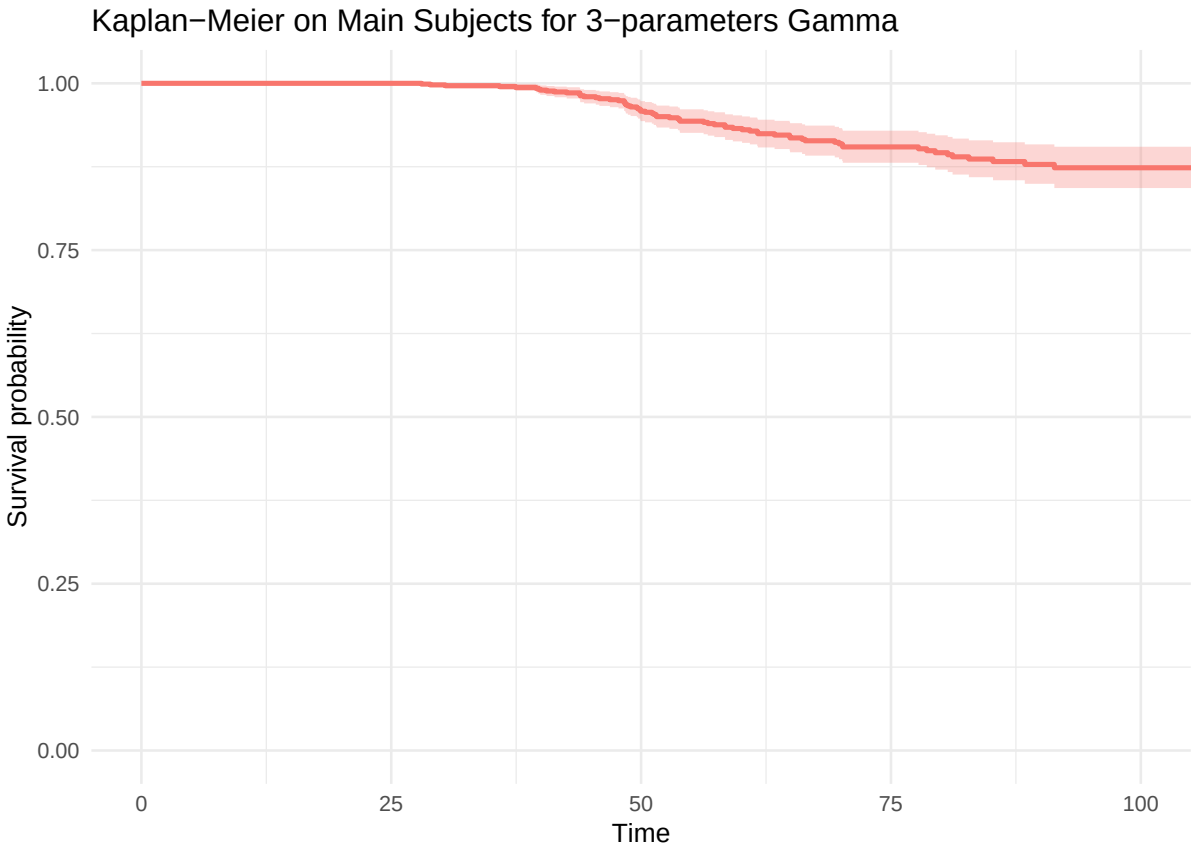

Figure 1.5: Kaplan-Meier estimate with a three-parameter Gamma distribution for the survival function of cases.

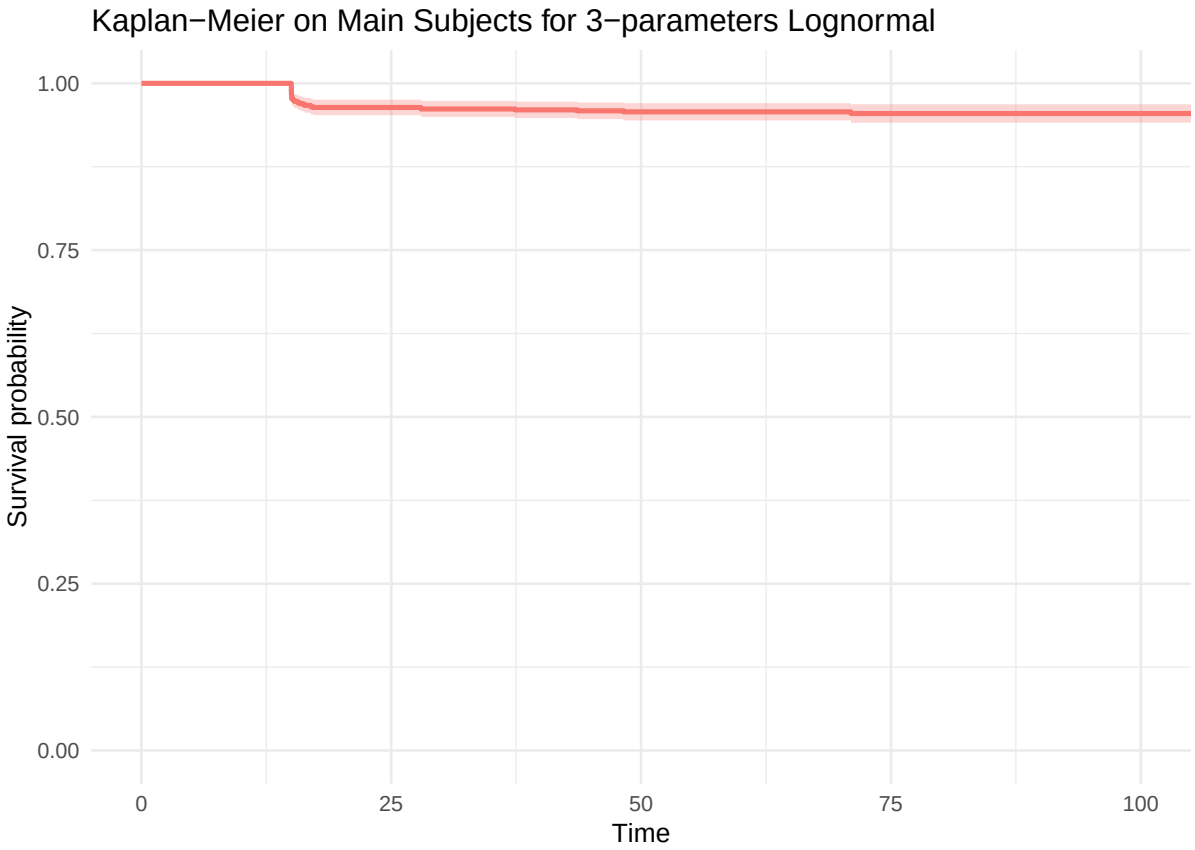

Figure 1.6: Kaplan-Meier estimate with a three-parameter Lognormal distribution for the survival function of cases.

	θ	$\widehat{p}_0$ (se)	$\widehat{shape}_0$ (se)	$\widehat{scale}_0$ (se)	$\widehat{\gamma}_0$ (se)	$\widehat{\theta}$ (se)
MFCR						
Mean (se)	0.2	0.84 (0.0008)	9.34 (0.1751)	6.85 (0.0857)	13.46 (0.3991)	0.19 (0.0021)
MSE		0.2532	55.3877	27.1140	126.2159	0.6642
Mean (se)	0.5	0.84 (0.0004)	7.99 (0.1642)	7.64 (0.1154)	17.32 (0.3211)	0.61 (0.0121)
MSE		0.1269	51.9246	36.5295	101.5672	3.8279
Mean (se)	0.8	0.86 (0.0013)	8.06 (0.1817)	6.48 (0.2611)	17.56 (0.3511)	0.88 (0.0192)
MSE		0.4112	57.4586	82.5685	111.0571	6.0721
UFCR						
Mean (se)	0.2	0.78 (0.0069)	13.18 (0.1567)	4.56 (0.0368)	6.84 (0.3131)	102.81 (2.7683)
MSE		2.1831	49.8229	11.7259	99.3466	881.4064
Mean (se)	0.5	0.76 (0.0071)	13.14 (0.2535)	6.57 (0.1449)	9.14 (0.3799)	98.90 (2.6522)
MSE		2.2470	80.3284	45.8249	120.2778	844.4519
Mean (se)	0.8	0.71 (0.0079)	12.34 (0.2673)	7.74 (0.1807)	11.59 (0.4246)	68.93 (2.7657)
MSE		2.5021	84.6390	57.1688	134.3136	877.2408
UFHCR						$\widehat{\beta}_{FH}$ (se)
Mean (se)	0.2	0.89 (<0.0001)	8.69 (0.0205)	5.49 (0.0084)	14.37 (0.0431)	3.23 (0.0048)
MSE		0.0510	6.5193	2.7048	13.6440	3.3889
Mean (se)	0.5	0.88 (0.0001)	8.23 (0.0189)	5.96 (0.0110)	14.78 (0.0399)	1.64 (0.0022)
MSE		0.0436	5.9811	3.4787	12.6194	1.3355
Mean (se)	0.8	0.87 (0.0001)	6.85 (0.0208)	6.85 (0.0124)	18.34 (0.0413)	1.38 (0.0018)
MSE		0.0374	6.6773	4.0123	13.4805	0.8126

Table 1.9: Parameter recovery with a three-parameter Gamma($shape_0 = 8$, $scale_0 = 6$, $\gamma_0 = 15$) baseline survival function, for the Multivariate frailty Cure-Rate (MFCR), the Univariate frailty Cure-Rate (UFCR), and the Univariate FH Cure-Rate (UFHCR) model.

In Table 1.11 we compare the posterior frailty risk mean, median, and mode, from the posterior frailty risk distribution, by reporting their performances in terms of MSPE, R^2, coefficient of the predicted risk on the true risk, correlation coefficient, and rank correlation coefficient. Family size is varying across the parameter values $(n_f, \lambda_F) = ((2, 0.8), (5, 0.8), (10, 5), (20, 10)^T$. Notice that for the Univariate FH Cure-Rate the predicted risk is the family history indicator itself, and not the mean, median or mode. Let us report the main comments that we can extrapolate from Table 1.11. The MSPE is not getting lower as the family size increases, on the contrary of what we could expect. This can be due to the random generation of the samples every time. However, for the posterior frailty risk mode we can notice that with an increasing family size the MSPE is halved. By comparing the Multivariate frailty Cure-Rate model, the Univariate frailty Cure-Rate model and the Univariate FH Cure-Rate model we can not notice a significative difference in terms of MSPE, even though the best results belong to the performance of the Multivariate frailty Cure-Rate model

	θ	$\widehat{p}_0$ (se)	$\widehat{\mu}_0$ (se)	$\widehat{\sigma}_0^2$ (se)	$\widehat{\gamma}_0$ (se)	$\widehat{\theta}$ (se)
MFCR						
Mean (se)	0.2	0.40 (0.0006)	4.31 (0.0015)	9.84 (0.0055)	14.96 (0.0004)	1.18 (0.0019)
MSE		0.4884	3.7204	4.2155	0.1327	1.1495
Mean (se)	0.5	0.46 (<0.0001)	4.16 (<0.0001)	9.29 (<0.0001)	14.99 (<0.0001)	1.18 (<0.0001)
MSE		0.3913	3.8401	3.2902	0.0332	0.6807
Mean (se)	0.8	0.58 (0.0014)	3.38 (0.0037)	6.82 (0.0118)	14.96 (0.0003)	2.09 (0.0055)
MSE		0.5186	4.7659	3.8205	0.1030	2.1654
UFCR						
Mean (se)	0.2	0.37 (0.0012)	4.06 (0.0063)	9.37 (0.0147)	14.91 (0.0012)	2.51 (0.0173)
MSE		0.6119	4.4150	5.7416	0.3900	5.9384
Mean (se)	0.5	0.56 (0.0017)	3.41 (0.0046)	7.20 (0.0147)	14.93 (0.0010)	2.75 (0.0116)
MSE		0.6108	4.8150	4.8009	0.3239	4.3033
Mean (se)	0.8	0.59 (0.0013)	3.45 (0.0046)	6.94 (0.0134)	14.96 (0.0004)	2.36 (0.0072)
MSE		0.4864	4.7769	4.3405	0.1327	2.9400
UFHCR						$\widehat{\beta}_{FH}$ (se)
Mean (se)	0.2	0.95 (0.0007)	0.89 (0.0083)	3.15 (0.0081)	0.09 (0.0024)	0.17 (0.0045)
MSE		0.2429	7.5790	3.8319	14.9293	1.4233
Mean (se)	0.5	0.94 (0.0012)	0.89 (0.0060)	3.27 (0.0122)	0.17 (0.0051)	0.09 (0.0059)
MSE		0.3900	7.3588	4.7262	14.9174	1.9103
Mean (se)	0.8	0.93 (0.0014)	0.92 (0.0062)	3.30 (0.0131)	0.18 (0.0059)	0.06 (0.0056)
MSE		0.4499	7.3465	4.9448	14.9370	1.9193

Table 1.10: Parameter recovery with a three-parameter Lognormal($\mu_0 = 8, \sigma_0 = 6, \gamma_0 = 15$) baseline survival function, for the Multivariate frailty Cure-Rate (MFCR), the Univariate frailty Cure-Rate (UFCR), and the Univariate FH Cure-Rate (UFHCR) model.

with posterior mean (3.433). The coefficient of determination is instead increasing when the family size tends to a higher number. This is exactly what we expect from before the analysis: a sample with larger families has a greater explainability power of the breast cancer development than a sample with families composed by few subjects. Moreover, here we can appreciate the higher power of the Multivariate frailty Cure-Rate model in comparison to the other two, because in the Univariate models the coefficient of determination is always computed on one subject per family. The best result is when the Multivariate frailty Cure-Rate model with mean or median frailty as covariate, explains around 61-62% of variability of the true risk with maximum twenty family members. For what concerns the coefficient from the linear regression of the true risk on the predicted risk, we would like it to be as closest as possible to one. Here we notice that this value is increasing with the family size for the Multivariate frailty Cure-Rate model, while it performs very badly for the Univariate models. In conclusion, this coefficient is not that meaningful. The

	(n_F, λ_F)			
	(2, 0.8)	(5, 0.8)	(10, 5)	(20, 10)
Multivariate frailty Cure-Rate				
MSPE(mean)	3.873 (0.049)	4.800 (0.074)	3.843 (0.071)	3.433 (0.049)
MSPE(median)	4.420 (0.052)	5.401 (0.078)	4.177 (0.074)	3.643 (0.051)
MSPE(mode)	11.348 (0.061)	12.198 (0.088)	6.009 (0.082)	4.430 (0.052)
R^2(mean)	0.202 (0.002)	0.219 (0.002)	0.478 (0.004)	0.614 (0.002)
R^2(median)	0.207 (0.002)	0.222 (0.002)	0.485 (0.004)	0.619 (0.002)
R^2(mode)	0.127 (0.002)	0.115 (0.002)	0.383 (0.003)	0.583 (0.002)
$\widehat{\beta}_{\text{mean}}$	1.553 (0.008)	1.705 (0.011)	2.401 (0.018)	2.781 (0.018)
$\widehat{\beta}_{\text{median}}$	1.792 (0.010)	1.975 (0.013)	2.608 (0.019)	2.932 (0.019)
$\widehat{\beta}_{\text{mode}}$	0.957 (0.007)	0.927 (0.009)	1.929 (0.015)	2.645 (0.017)
ρ(mean)	0.447 (0.002)	0.465 (0.002)	0.689 (0.007)	0.783 (0.001)
ρ(median)	0.452 (0.002)	0.468 (0.002)	0.694 (0.003)	0.786 (0.001)
ρ(mode)	0.353 (0.002)	0.335 (0.003)	0.616 (0.002)	0.763 (0.001)
Rank ρ(mean)	0.181 (0.002)	0.226 (0.002)	0.438 (0.002)	0.539 (0.001)
Rank ρ(median)	0.181 (0.002)	0.226 (0.002)	0.438 (0.002)	0.539 (0.001)
Rank ρ(mode)	0.151 (0.002)	0.146 (0.002)	0.332 (0.002)	0.490 (0.002)
Univariate frailty Cox				
MSPE(mean)	4.241 (0.053)	5.259 (0.081)	4.738 (0.084)	4.507 (0.059)
MSPE(median)	4.536 (0.055)	5.594 (0.084)	5.040 (0.088)	4.798 (0.062)
MSPE(mode)	6.511 (0.060)	7.666 (0.091)	7.019 (0.098)	6.767 (0.068)
R^2(mean)	0.136 (0.003)	0.157 (0.003)	0.183 (0.002)	0.126 (0.002)
R^2(median)	0.138 (0.003)	0.161 (0.003)	0.191 (0.003)	0.131 (0.002)
R^2(mode)	0.128 (0.003)	0.149 (0.002)	0.190 (0.002)	0.128 (0.002)
$\widehat{\beta}_{\text{mean}}$	2.704 (0.025)	3.158 (0.021)	3.326 (0.037)	2.685 (0.023)
$\widehat{\beta}_{\text{median}}$	2.893 (0.028)	3.369 (0.023)	3.588 (0.038)	2.887 (0.025)
$\widehat{\beta}_{\text{mode}}$	2.621 (0.028)	3.032 (0.022)	3.349 (0.035)	2.658 (0.022)
ρ(mean)	0.356 (0.004)	0.391 (0.003)	0.422 (0.003)	0.351 (0.003)
ρ(median)	0.359 (0.004)	0.395 (0.003)	0.432 (0.003)	0.357 (0.003)
ρ(mode)	0.348 (0.004)	0.381 (0.003)	0.431 (0.003)	0.353 (0.002)
Rank ρ(mean)	0.201 (0.004)	0.206 (0.002)	0.194 (0.002)	0.183 (0.002)
Rank ρ(median)	0.201 (0.004)	0.206 (0.002)	0.194 (0.002)	0.183 (0.002)
Rank ρ(mode)	0.167 (0.003)	0.190 (0.002)	0.203 (0.002)	0.192 (0.001)
Univariate FH Cure-Rate				
MSPE	5.388 (0.057)	6.573 (0.089)	5.625 (0.097)	5.034 (0.062)
R^2	0.043 (0.001)	0.043 (0.001)	0.152 (0.003)	0.186 (0.002)
$\widehat{\beta}_{FH}$	3.584 (0.077)	3.971 (0.077)	3.222 (0.039)	2.806 (0.019)
ρ	0.189 (0.004)	0.195 (0.003)	0.379 (0.004)	0.429 (0.002)
Rank ρ(mean)	0.147 (0.002)	0.159 (0.002)	0.348 (0.001)	0.435 (0.001)

Table 1.11: Prediction accuracy measures.

Pearson correlation and the rank correlation increase when the family size increases, as we expect before the analysis. With maximum twenty family members, the Multivariate frailty Cure-Rate model achieve a correlation in the range 0.76-0.78, which is around the 0.40 points higher that the Univariate frailty Cure-Rate model, and 0.30 points more than the Univariate FH Cure-Rate model. Again for the rank correlation, the Multivariate frailty Cure-Rate model achieve a value in the range 0.49-0.54, which is 0.35 points higher than the Univariate frailty Cure-Rate model. The Univariate FH Cure-Rate model strikes a discrete result with a rank correlation of 0.43. Once again, we conclude that the Multivariate frailty Cure-Rate model outperforms the others.

We can conclude now that the posterior mean as predicted frailty risk through the maximization of the multivariate likelihood yields the most favorable results. Not of less importance is that fact that the mean is an understandable index to the end user, rather than the median and the mode. From Figure 1.7 one can observe the improved accuracy in prediction as the family size increases along with the mean frailty. The observed behavior in the plot with smaller families is attributed to the aforementioned (from semiparametric scenario) shrinkage issue of the mean frailty risk, which is effectively mitigated with a larger family size. No distinction can be observed in the conditional MSPE, plotted against the true risk on the x-axis, in Figure 1.8. This observation highlights the need to explore alternative estimations for frailty risk in order to overcome the various challenges encountered also with the mean estimation even though we adopt it as the best summary of the posterior predicted frailty risk distribution so far.

We thus compare the three parametric models by applying the binary splitting of the predicted risk mean and reporting the AUC, the PPV, the NPV and the Harrell's Concordance index (C) in Table 1.12. The Multivariate frailty Cure-Rate model outperforms the other two Univariate models, with an AUC in the range 70%-95%, increasing according to a higher family size, against an AUC around the 62% for the Univariate frailty Cure-Rate, and in the range 53%-71% for the Univariate FH Cure-Rate model. These additional twenty percentile points in the Multivariate frailty Cure-Rate model allow to reach a strong predictive power that is not reached by any other model. Moreover, the Multivariate frailty Cure-Rate model gives the best result in terms of negative predictive power for maximum family size, with the 98% of families truly not belonging to the highest-risk group in contract to the 96% resulted from the Univariate frailty Cure-Rate model. Similarly to the semiparametric scenario, also here the positive predictive value is weaker than the negative predictive power, meaning that the low-risk families are better identified than the highest-risk families. This allows for a reduction of unnecessary costs, medical treatments and psychological stress for families who do not need to be under surveillance. We refer to Figure 1.9 for an example of ROC curve, one among a hundred repetitions, from the Multivariate frailty Cure-Rate model. One can appreciate how the AUC is significantly increasing from the 88% with a maximum of five members, to 95% with a maximum of twenty members per family.

Lastly, let us say that the Univariate frailty Cure-Rate model can coherently predict the risk according to the time-to-event (see Concordance around the 95%), as well as the Multivariate frailty Cure-Rate model, whose concordance varies among the 93% and the 96%. It is worth noting that for

the Univariate FH Cure-Rate model, the Concordance index cannot be calculated without adjusting for ties. The presence of ties is caused by the family history indicator, which trivially can only take values 0 or 1.

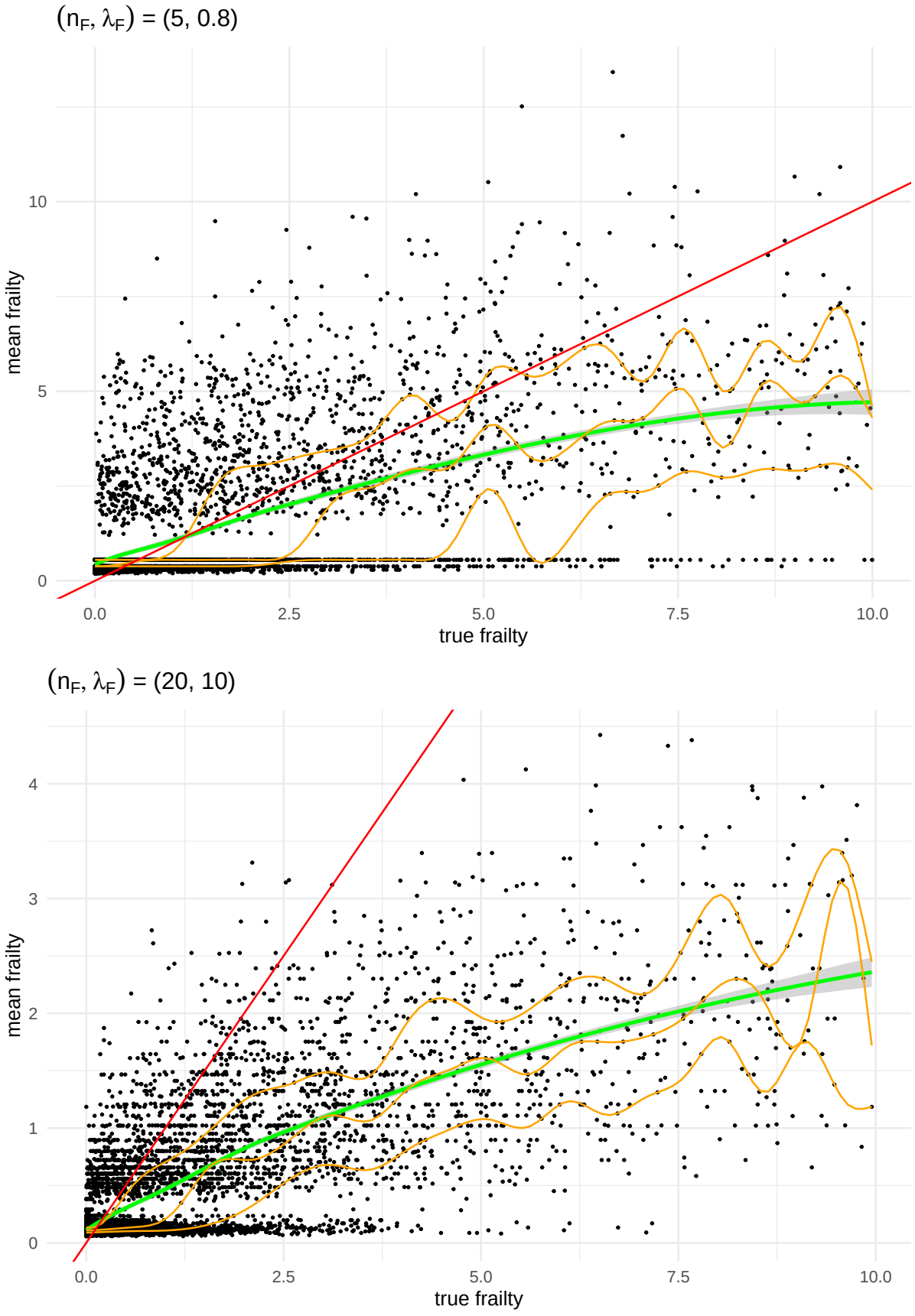

Figure 1.7: True risk versus predicted risk mean with the bisector of the third quadrant (in red), the smoothed regression line, and smoothed quantile regression line on the first, second and third quantile (in orange).

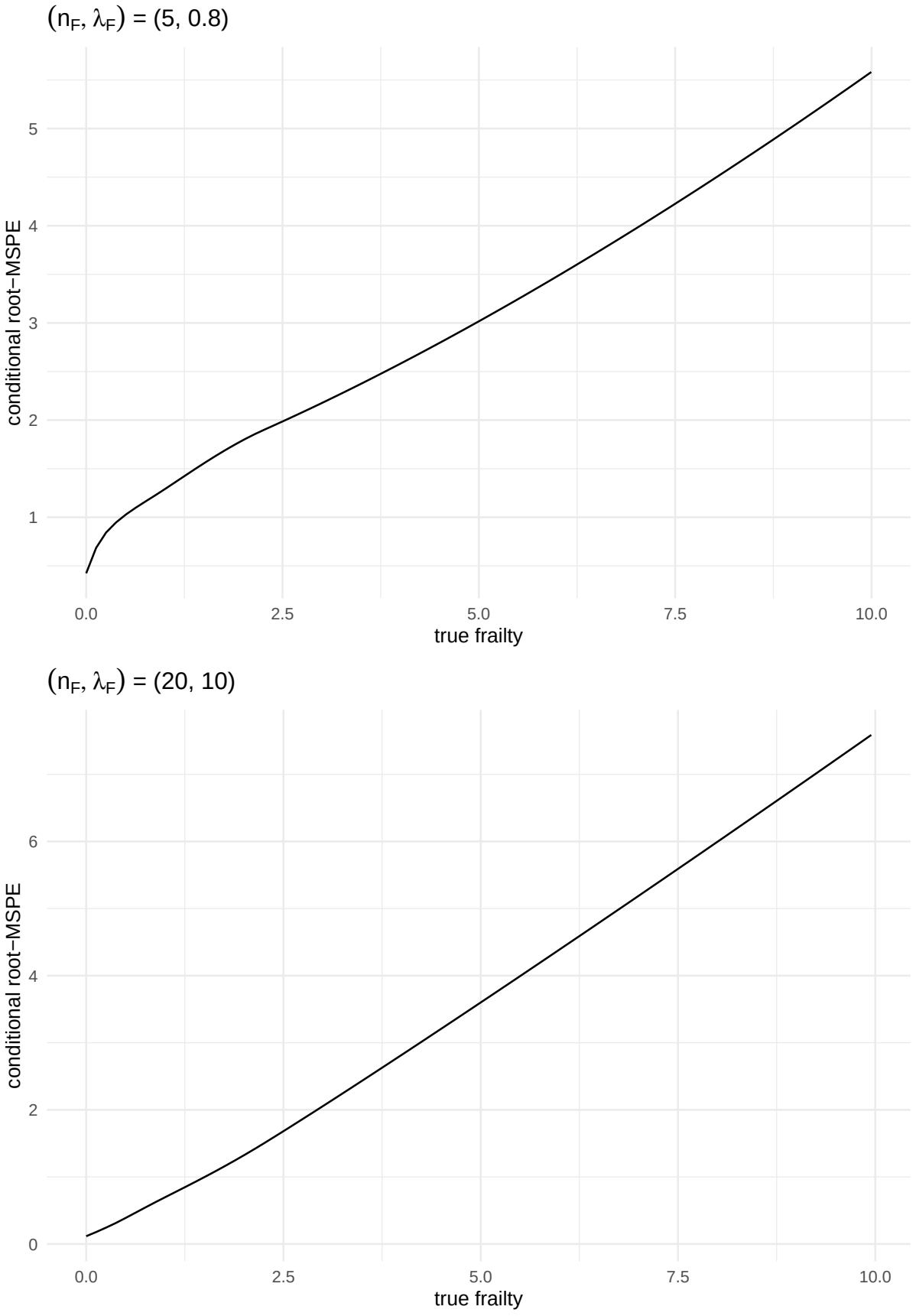

Figure 1.8: True risk versus the conditional MSPE.

	(n_F, λ_F)			
	(2, 0.8)	(5, 0.8)	(10, 5)	(20, 10)
Multivariate frailty Cure-Rate				
AUC	0.697 (0.002)	0.714 (0.002)	0.891 (0.002)	0.948 (0.001)
PPV	0.318 (0.002)	0.352 (0.003)	0.532 (0.003)	0.624 (0.003)
NPV	0.951 (<0.001)	0.959 (<0.001)	0.975 (<0.001)	0.980 (<0.001)
C	0.960 (<0.001)	0.959 (<0.001)	0.943 (0.001)	0.929 (0.001)
Univariate frailty Cure-Rate				
AUC	0.627 (0.002)	0.637 (0.002)	0.632 (0.002)	0.621 (0.001)
PPV	0.286 (0.004)	0.332 (0.003)	0.373(0.004)	0.332 (0.003)
NPV	0.959 (<0.001)	0.961 (<0.001)	0.965 (<0.001)	0.963 (<0.001)
C	0.952 (<0.001)	0.947 (<0.001)	0.951 (<0.001)	0.950 (<0.001)
Univariate FH Cure-Rate				
AUC	0.535 (0.001)	0.533 (0.001)	0.645 (0.002)	0.712 (0.001)
PPV	0.314 (0.009)	0.299 (0.009)	0.288 (0.003)	0.223 (0.001)
NPV	0.991 (<0.001)	0.989 (<0.001)	0.944 (<0.001)	0.897 (<0.001)
C	-	-	-	-

Table 1.12: Mean and standard errors of AUC, PPV, and NPV corresponding to the binary split based on the highest-risk group indicator with threshold $1 - \alpha = 0.95$, and Harrell's concordance index C.

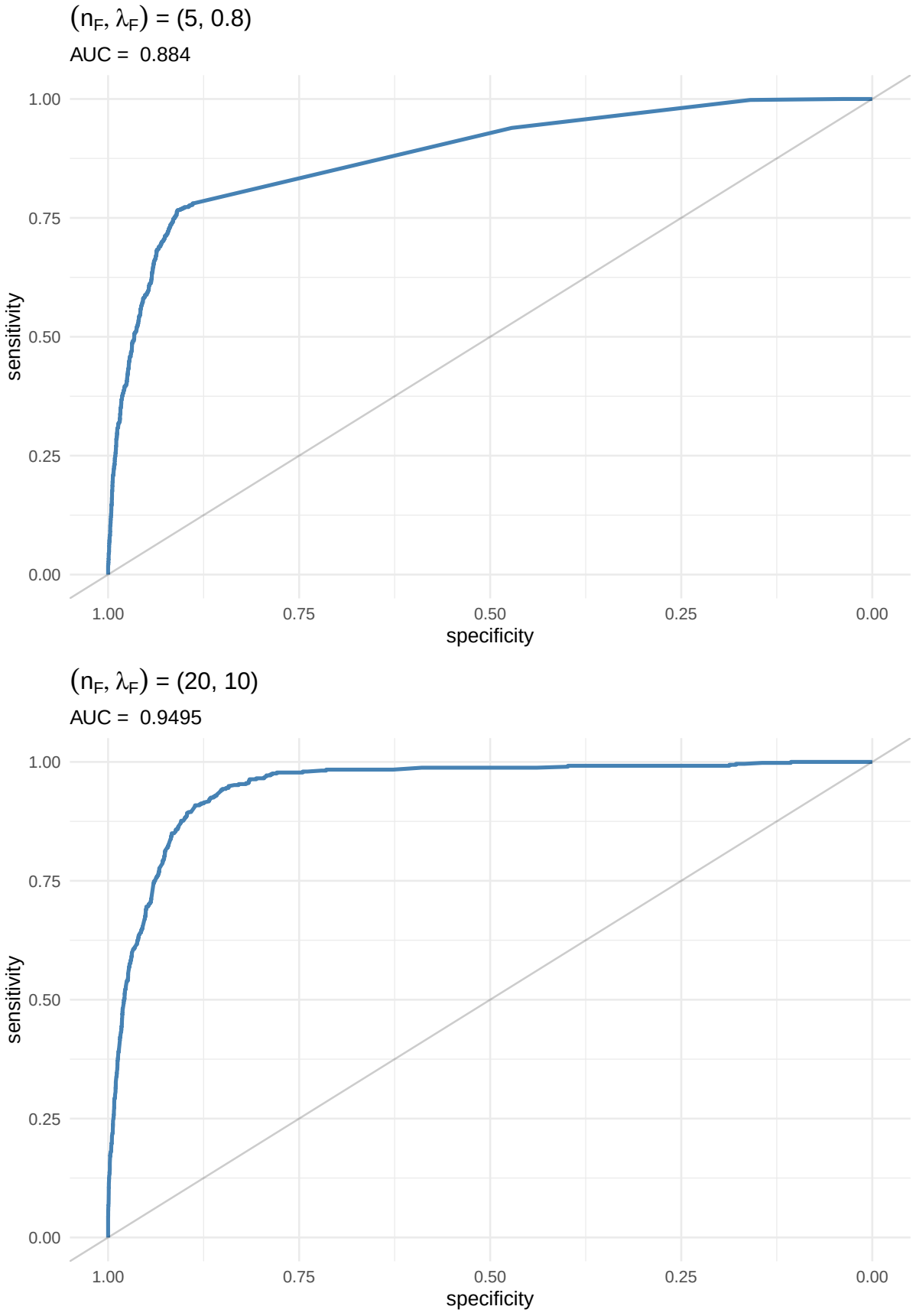

Figure 1.9: ROC curve of the true membership to the highest-risk families (top 5%) versus the predicted risk mean.

1.5 Illustration to the Swedish Multi-Generational Breast Cancer registry

1.5.1 Motivating framework

Breast cancer is the most prevalent form of cancer among females, comprising approximately 30% of all cancer diagnoses [34]. The presence of lesions in secondary locations is a significant risk factor for cancer, as these lesions are known to be highly fatal due to the process of metastasis, which causes over the 90% of cancer-related deaths [28].

In recent decades, a lot has been discovered in cancer research with the goal of improving the survival rates of patients. Despite these progresses, a significant amount of information regarding the development of cancer remains unknown, which presents an opportunity for further improvement. Specifically, areas in which there is potential for improvement include the detection and treatment of primary tumors and metastases. This section will provide an overview of the most recent literature on the etiology, diagnosis, and treatment of cancer. Of particular interest is the advancement of early detection methods, as early detection is crucial for having more treatment options, increasing the probability of a successful recovery, and preventing mortal metastases.

Primary cancers have three distinct causes, namely genetic factors, individual lifestyle, and external environmental factors. In the context of breast cancer, several genes have been identified as crucial for its development. These include breast cancer genes (BRCA1, BRCA2), which are associated with the highest incidence of breast and ovarian cancers in families, as well as the tumor protein mutation (TP53), the single nucleotide polymorphisms (SNPs), and the polygenic risk scores (PRS) [13, 21]. BRCA1 and BRCA2 mutations are highly indicative of cancer detection; however, they only account for 20 – 25% of familial cancer aggregation [28]. As a result, the remaining 75% of cancer familial aggregation is unknown.

Other risk factors for breast cancer, not genetic, include mammography density (MD), age, oral contraceptives, parity and timing of births, breastfeeding, age at menarche and menopause, body mass index (BMI), physical exercise, alcohol and tobacco consumption, and family history [9]. Among these, family history is considered one of the strongest. The term family history refers to the incidence of breast cancer onset among family members. A positive family history is characterized by at least one family member that has experienced breast cancer onset at the time of the analysis. Clearly, the magnitude of the family history effect depends on the number and grade of relatives who have experienced the disease. For example, a positive family history in first-degree relatives, i.e. among mother and sisters, has a greater impact compared to breast cancer cases in second or third-degree relatives, i.e. grandmothers and aunts [7].

Now that we have analysed the possible causes, we explore the three-step process of breast cancer diagnosis. The first step is the clinical analysis, which is the least invasive. If necessary, the process proceeds to the intermediate step of machine imaging. Digital mammography is the most commonly used imaging technique. The final and most invasive step involves a biopsy, where a fine needle is used for cytopathology diagnosis, which studies pathology at the cellular level. After

the identification of a tumor, it needs to be classified in one of the available categories [9].

Tumors can be categorized using different methods, such as primary origin and metastasis stage, histological grade, or molecular subtypes, which are determined by the phenotype, namely the visible expression, of the cells. There are five molecular subtypes of invasive breast cancer: luminal A, luminal B, luminal B-like, HER2-enriched, and triple negative. These categories are mainly determined by the expression of three receptors: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor type 2 (HER2). Luminal cancers are the least dangerous as they need estrogen to grow, thus blocking the estrogen receptor is often sufficient to stop tumor growth. Luminal A is ER positive, PR positive, and HER2 negative, while luminal B cancer is ER positive and HER2 negative, with any result on PR. Luminal B-like cancer is ER positive, PR negative, and HER2 positive. The HER2-enriched is ER and PR negative, but HER2 positive. The triple negative (TN) tumor, which is often referred to as TN-negative, is the most harmful and it is ER, PR, and HER2 negative. These categories are also grouped as HR-positive for luminal tumors, HER2-positive, and TN-negative.

Once that the breast cancer has been categorized, the treatment can be carried out according to the category. Luminal A cancer, which grows slowly and has a better prognosis, can be treated with hormonal therapy, which is the most basic and least invasive therapy. Luminal B and luminal B-like cancers grow faster and have a worse prognosis, but they can typically be treated with appropriate hormonal therapy. HER2-enriched cancer is worse than luminal tumors, and it may require chemotherapy and targeted immunotherapy as primary treatments, followed by hormonal therapy. Triple negative cancer often occurs in individuals with a BRCA1 mutation, young people, and black people. It can be managed through chemotherapy and adjuvant chemotherapy (see e.g. [12]), with surgery as a subsequent option if necessary.

When talking about breast cancer one cannot ignore the process of metastasis. According to recent studies, approximately 30% of breast cancer patients develop metastasis [20]. Therefore, understanding metastasis is crucial for improving patient survival outcomes. Metastasis is a process where cancer cells escape from the primary tumor, which is the original site of cancer, and disseminate to a secondary location. Metastasis can be categorized as regional or distant, depending on the distance the cancer cells have travelled. Regional metastasis occurs when the cancer cells invade the regional lymph nodes. Distant metastasis occurs when cancer cells enter the bloodstream or lymphatic vessels and disseminate to other organs, such as skeleton, lungs, liver, and brain, with the brain being the least frequent but the most harmful [35].

It is still unclear when the metastasis process begins. Researchers have proposed two models of metastatic dissemination: the linear model and the parallel model [28]. The linear model suggests that the metastatic progression occurs right after that the primary tumor grows and becomes malignant. On the other hand, the parallel model suggests that the metastatic progression happens simultaneously with the primary tumor development. This hypothesis is supported by several phenomena, such as early-stage cancer patients found with secondary tumors and second cancer lesions discovered in patients with cancer of unknown origin. The parallel dissemination

is much more dangerous than the linear progression, and both of them do not exclude the other.

Currently, metastatic breast cancer is considered an incurable disease. What it can be done is only about prolonging patient survival and improving their quality of life by implementing a treatment among for example, chemotherapy, hormonal therapy, targeted therapy. Surgical intervention and radiation therapy may also be used in specific cases, according to the progression of the metastatic patient.

As mentioned above, a significant proportion, namely the 75%, of familial aggregation of cancer remains unexplained. It is our goal to investigate this phenomenon and contribute to increase the knowledge about the hereditary aspect of cancer. We believe that the random variable R, namely the frailty risk, can capture the hereditary component and explain the familial aggregation of breast cancer. Performing risk prediction brings to the identification of highest-risk families to target them towards different prevention paths. It indeed helps to modify the intensity of screening schedules, and to facilitate the implementation of tailored preventive strategies [10]. On the other hand, identifying the lowest-risk families may lead to a reduction not only of their psychological stress, but also of unnecessary medical treatments and costs.

1.5.2 The Data

Sweden's centralized information storage process enables the creation of comprehensive and easily accessible registries. Notably, all the information is linked to the patient's unique identification number, known as the "löpnummer" in Swedish. This allows the demographic and medical information to be traced back to each patient and family. The primary focus of this study is on breast cancer data, death, and migration flux, which are available in separated dataset. By merging these dataset, comprehensive information of each female family member can be obtained without losing information about the relationships between members. The process of constructing the clean dataset is detailed in Appendix A.3.

The dataset concerns a cohort of $n = 1,603,920$ Swedish families, consisting of a total of $4,267,803$ women. Among these families, $1,603,920$ women (one per family) were born between the 1947 January 1st, and 1976 December 31st, so that they age between 40 and 70 years old at the end of the follow-up (which coincides with 2016 December 30th), and they are what we call the "main subject" in the family. In this age window the risk of developing breast cancer is at its highest. The reason why we identify a main subject in each family is for moving from the multivariate to the univariate scenario where one subject per family enters the likelihood. Notice that, given a family, a main subject is identified with equal probability among the mother and the group of daughters. If the realization falls into the group of daughters, one among them is randomly sample as the main subject. With this two-steps sample we ensure that also mothers are chosen with probability $1/2$.

Once the main subject has been identified, and consequently her mother and sisters, we can move to obtain the information regarding breast cancer, death, and migration for all of them by merging the Multi-Generational registry, which provides us with the familiar relationships, the

Swedish Cancer registry, the Cause of Death registry, and the Migration registry. From the Cancer registry we select only the invasive cancer as event of interest, by excluding the non-invasive DCIS (ductual carcinoma in-situ) cases. From the Death and Migration registry we obtain the censoring events, i.e. death or emigration before observing the occurrence of breast cancer.

All our analyses were conducted following the approval by the Bocconi Ethics Board.

1.5.3 A preliminary Case-Control analysis

We start by implementing a Nested Case-Control (NCC) study design to estimate the relative risk (RR) of developing breast cancer for women with a positive first-degree family history (mother or sisters) compared to those without a family history. We aim to obtain a value similar to the well-known estimated value from literature, which is RR = 1.8 [14]. To conduct the NCC study, we select subjects based on specific criteria from our initial dataset of 6,633,147 subjects. First, we select women with a birthday within the end of the follow-up and with their mother in the dataset, resulting in 3,798,079 subjects. We further narrow the dataset to women aged between 40 and 70 years old in the time window of 2010-2016, resulting in 1,131,499 subjects. This number is different from the one under analysis because we also admit main subjects without information on the mother.

To perform the NCC we match subjects based on their year of birth and select 19,550 breast cancer cases and 90,833 unique controls (five controls for each case) through a process of sampling with reintroduction. This results in a final dataset of 114,330 subjects, which we call the "Nested Case-Control dataset" (NCCD), whose consort plot [23] is in Figure 1.10 with all the steps from the initial dataset to the NCCD.

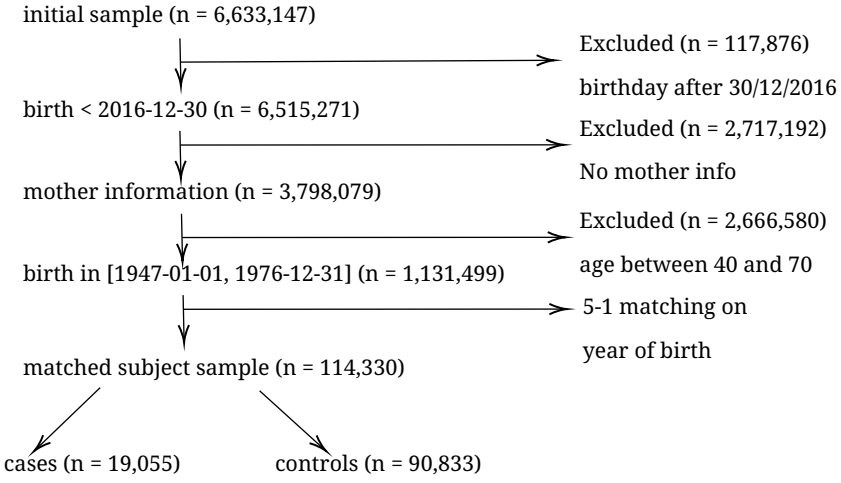

Figure 1.10: Consort flow chart for the construction of the Nested Case-Control dataset.

We carry out a survival Cox model with the time-varying family history covariate $FH(t)$ on the NCCD [42]. The estimated relative risk results $RR = 1.8017$, which is aligned with the value from the literature equal to 1.8 [14, 26]. This result validates the quality of our data.

1.5.4 Descriptive statistics

Table 1.13 reports some descriptive statistics of the clean dataset composed by 4, 267, 803 women, and only $n = 1,603,920$ when the term "Red." accompanies the description.

	Min.	Median	Mean	Max.
Birthday	1853-11-15	1955-03-15	1954-01-15	2018-12-15
Diagnosis	1958-01-15	1994-02-25	1992-03-07	2016-12-30
Death	1947-05-08	1991-08-21	1991-01-18	2018-12-31
Emigration	1948-01-01	1999-09-30	1996-09-12	2018-12-31
Age at diagnosis Red.	11.00	50.00	50.2	69.00
Diagnosis Red.	1961-09-15	2008-08-13	2006-12-27	2016-12-30
Death Red.	1948-10-25	2006-12-26	2000-08-10	2018-12-31
Emigration Red.	1951-05-01	1992-08-24	1991-02-02	2018-12-31
Follow-up (years)	0	53.2922	52.8118	69.9576
Breast cancer onset	Yes	No		
	47,914 (2.99%)	1,556,006 (97.01%)		
EFS	Alive	Dead/emigrated	Diagnosed	
	1,408,072 (87.79%)	147,914 (9.22%)	47,934 (2.99%)	
	Yes	No		
FH at end of FU	105,432 (6.57%)	1,498,488 (93.43%)		
Parity	989,422 (38.31%)	614,498 (61.69%)		
	Min.	Median	Mean	Max.
Number of children	1	1	1.5	10
Age at first child	13.08	27.25	27.67	60.33

Table 1.13: Summaries of the main variables obtained from the Multi-Generational Swedish dataset, where "Red." refers to only the main subject (whose birthday ranges in [1947-01-01, 1976-12-31]); "EFS" means status at the end of the follow-up, "FH" means family history, and "FU" means follow-up.

It is worth noting that the first recorded breast cancer diagnosis occurred in January 1958, with no specific day (we assume the onset date to be the 15th of that month). The final recorded case occurred on December 30th, 2016, providing a 58-year follow-up for analysis. We also consider this date as the end of the follow-up period. The median age at diagnosis is 50 years old, proving that the risk group that we select with ages in-between 40 and 70 is appropriate. This is reflected also in the median follow-up of around 53 years. Around the 3% (47,934 subjects) have been diagnosed with

invasive breast cancer, the 9.22% are censored due to death, emigration or DCIS before the end of the follow-up, and the 87.79% are alive at the end of the follow-up without experiencing neither breast cancer onset nor any other censoring events, so that we can consider this percentage as our cured fractions. Less than the 7% has a positive family history at the end of the follow-up. For what concerns additional covariates that can be included into the models, the majority of women (61.69%) is nulliparous, while, among those who have given birth, the average number of children is 1.5. The mean age at first child (27 years old) is appropriate for Sweden. Notice that all of these variables refer to the most recent follow-up time of the subject. Genetic factors and other details related to breast cancer are not available in the data, limiting the scope of our analysis to familial survival data provided by the women themselves during their visits. No medical tests or blood samples are available.

A majority of families (45.06%) consist of two members, followed by three members (33.7%) and four members (11.37%). To provide complete information on family structure, we also report in Table 1.14 the frequencies of sisters, as they play a critical role in our analysis. The first row of Table 1.14 shows the percentage of families having a specific number of sisters for the "main subjects", where it can be observed that no main subject has precisely eleven sisters. The second row presents the cumulative percentage of sisters for the main subjects, which is equivalent to the non-missing information on the sister columns. For instance, the first row shows that 36% of main subjects have precisely one sister ("Freq." in the first column), while the 52% have at least one sister ("Cum. Freq." in the second column).

	Number of sisters								
	0	1	2	3	4	5	6	7	≥8
Freq.	0.48	0.36	0.12	0.03	0.01	$3.1{\cdot}10^{-3}$	$9{\cdot}10^{-3}$	0.0005	
Cum. Freq.	1	0.52	0.16	0.04	0.01	0.005	0.001	0.0008	0.0002

Table 1.14: Number of sisters of the main subjects.

Once the dataset has been cleaned and merged across families and all necessary subject-specific information is collected, the observed time is determined based on whether the subject has experienced breast cancer onset, death, or emigration out of Sweden before in time. If the individual has not experienced any of these events, her observed time is the last day of the follow-up. The indicator of having observed the event follows immediately. In Figure 1.11, we present the Kaplan-Meier estimator of the survival function for all subjects in the dataset, whose family size ranges from a minimum of one to a maximum of fourteen subjects.

Parameter estimation must be distinguished between the semiparametric and the parametric scenario. Notice that the analysis is conducted both on those main subjects that has a recorded mother in the Multi-Generational dataset and on all main subjects (with or without recorded mother

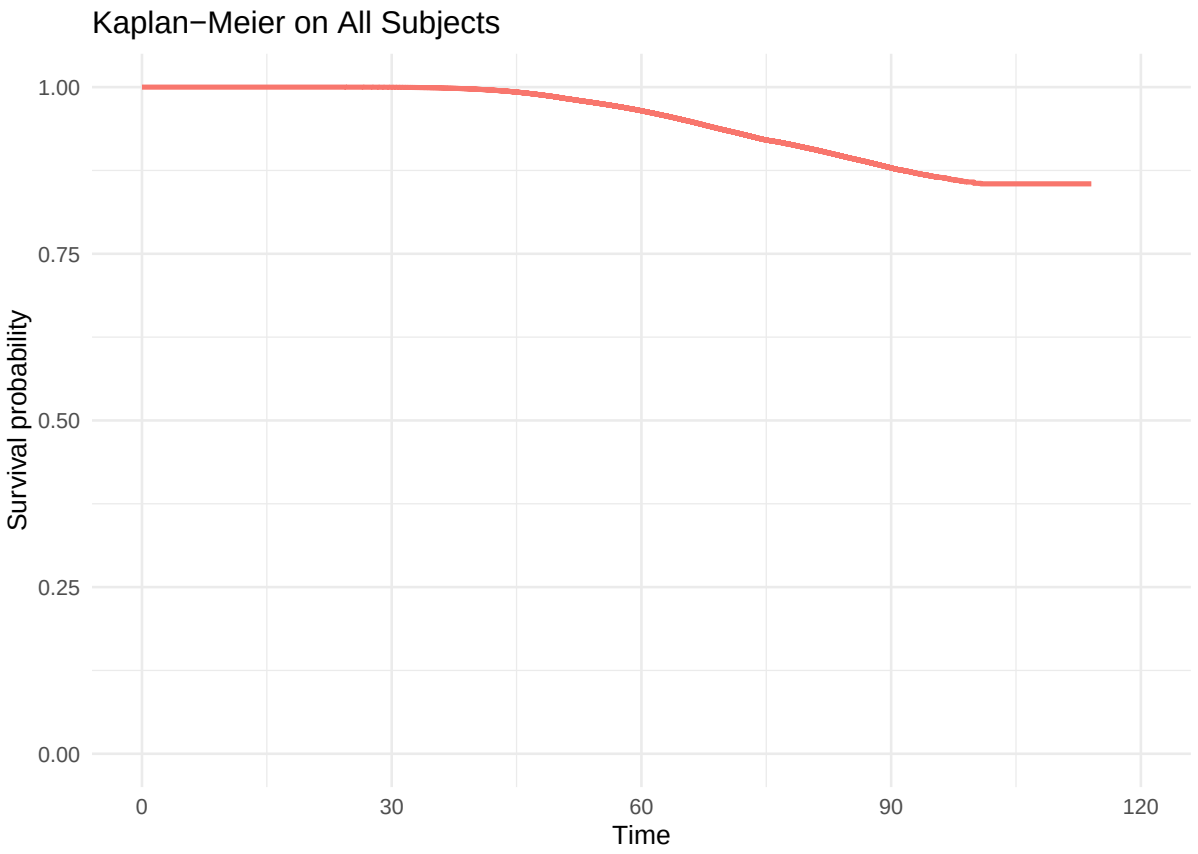

Figure 1.11: Kaplan-Meier survival estimate of age at onset (all subjects).

in the dataset). A preliminary analysis on the covariate extension to parity indicator, number of children, and age at first child from Table 1.16, is also conducted into the semiparametric setting. We plan to extend this preliminary analysis to a proper study of the inclusion of covariates into all models (both from semiparametric and parametric scenario), with a proper simulation study before the real case analysis.

1.5.5 Semiparametric setting

In Section 1.4 the Multivariate frailty Cox, the Univariate FH Cox, and the Univariate $FH(t)$ Cox model have been analysed in simulation studies. Not surprisingly, based on the simulation results, the Multivariate frailty Cox model outperforms the others in terms of predictive accuracy. Similarly to simulation studies, also with the real case data the Multivariate frailty Cox model has a Concordance index of 0.965, which is significantly higher than the Univariate FH Cox model (0.5150) and the Univariate $FH(t)$ Cox model with 0.5036 of concordance. All the results are reported in Table 1.15. By selecting only the main subjects with a recorded mother into the dataset, the Concordance index almost reaches the maximum value, up to the 99% of concordant pairs. This result is outstanding, and the advantage of using the Multivariate model instead of only the

easier Univariate FH model is immediately observed. Indeed, the Univariate FH Cox and Univariate $FH(t)$ Cox models perform poorly, as their prediction concordance is comparable to that of a random flipping of a coin.

	Survival information mother	$\widehat{\theta}$	$\widehat{\beta}_{FH}$	C
Multivariate frailty Cox	yes	1.3624	-	0.9968
Univariate FH Cox		-	1.5372	0.5021
Univariate $FH(t)$ Cox		-	1.8017	0.5042
Multivariate frailty Cox	no	1.327	-	0.9650
Univariate FH Cox		-	1.449	0.5150
Univariate $FH(t)$ Cox		-	1.8405	0.5036

Table 1.15: Estimated parameters and Harrell's concordance index (C).

In Table 1.16 we report results from including the available reproductive covariates as parity, number of children, and age at first child. No significant differences in terms of Concordance index is noticed in comparison to the scenario without covariates.

	Survival information mother	$\widehat{\theta}$	$\widehat{\beta}_{FH}$	C
Multivariate	yes	2	-	0.9335
Univariate FH		-	1.3922	0.6209
Univariate FH(t)		-	2.2254	0.5350
Multivariate	no	2	-	0.9436
Univariate FH		-	1.3922	0.5965
Univariate FH(t)		-	1.8049	0.5184

Table 1.16: Estimated parameters and Harrell's concordance index (C) for the models with parity, number of children, and age at first child.

Notice that the estimated family history coefficient $\widehat{\beta}_{FH}$ has a positive value (around 1.8) meaning that the family history has a positive effect on the breast cancer development, that is indeed coherent with the literature.

We then use the estimated parameters to perform posterior prediction of family-specific frailty risk.

We compute an algorithm which takes as input the familial survival data of the female family members of a new woman. By providing information on her mother and sisters, the algorithm pro-

duces three quantities: the posterior mean frailty risk, the probability of belonging to the highest-risk families (top 5%), and an indicator of whether the woman belongs to the highest-risk families compared to the population (prior) distribution of the probability of belonging to the highest-risk families, based on a fixed threshold. For example, from the Multivariate frailty Cox model we obtain that $\widehat{\theta} = 1.327$, and $\widehat{r}_{(1-\alpha)}$ is 0.1512 for $\alpha = 0.05$. This means that if a woman has a probability of belonging to the highest-risk families over 15.12%, then her indicator will take value one. Risk prediction can be easily done by using the estimated frailty parameter, the (1-α) percentile value, and also the Breslow estimates of the cumulative hazard function obtained by fitting the Multivariate frailty Cox model. These values can be stored for later use, enabling fast prediction without having to recompute the entire process for new women. This algorithm has been developed based on the model chosen in Section 1.4.

In Table 1.17 a summary of the distribution of the cumulative hazard function is provided. The plot of the cumulative hazard function and its density function on the complete population are respectively in Figure 1.12 and 1.13. Interestingly, we can notice how the density function can be split into two regions. It is straightforward to extend this procedure to other percentile levels, although this necessitates refitting the model from the beginning, which can be time-consuming. On the other hand, every computation can be done for once and store it for a later use.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.0000	0.0000	0.0324	0.0752	0.1389	0.2583	360

Table 1.17: Summary of the estimated Breslow cumulative hazard function on a grid of (time) values.

An example of risk prediction into the semiparametric scenario is reported below for a woman with complete family survival information. Observed ages = (45, 90, 60) in years, and indicators of having observed the onset event = (0, 1, 0). Consider the first subject to be the woman who shows up at the hospital at 45 years of age, with the mother who had breast cancer onset at 90, and a sister who has not experienced the onset yet at age 60. The output from the algorithm below means that the woman has a frailty risk with value 1.57 if estimated through the posterior mean; with value 2.78 if estimated through the posterior median. She also has a probability of belonging to the highest-risk families of 12.64% that is slower than the top 5% threshold of 15.12% and thus she does not belong to the highest-risk families. All the quantities can be found in the following output:

```
Posterior mean frailty = 1.5762
Posterior median frailty = 2.7801
Posterior high-risk probability = 0.1264
Posterior high-risk membership = 0
```

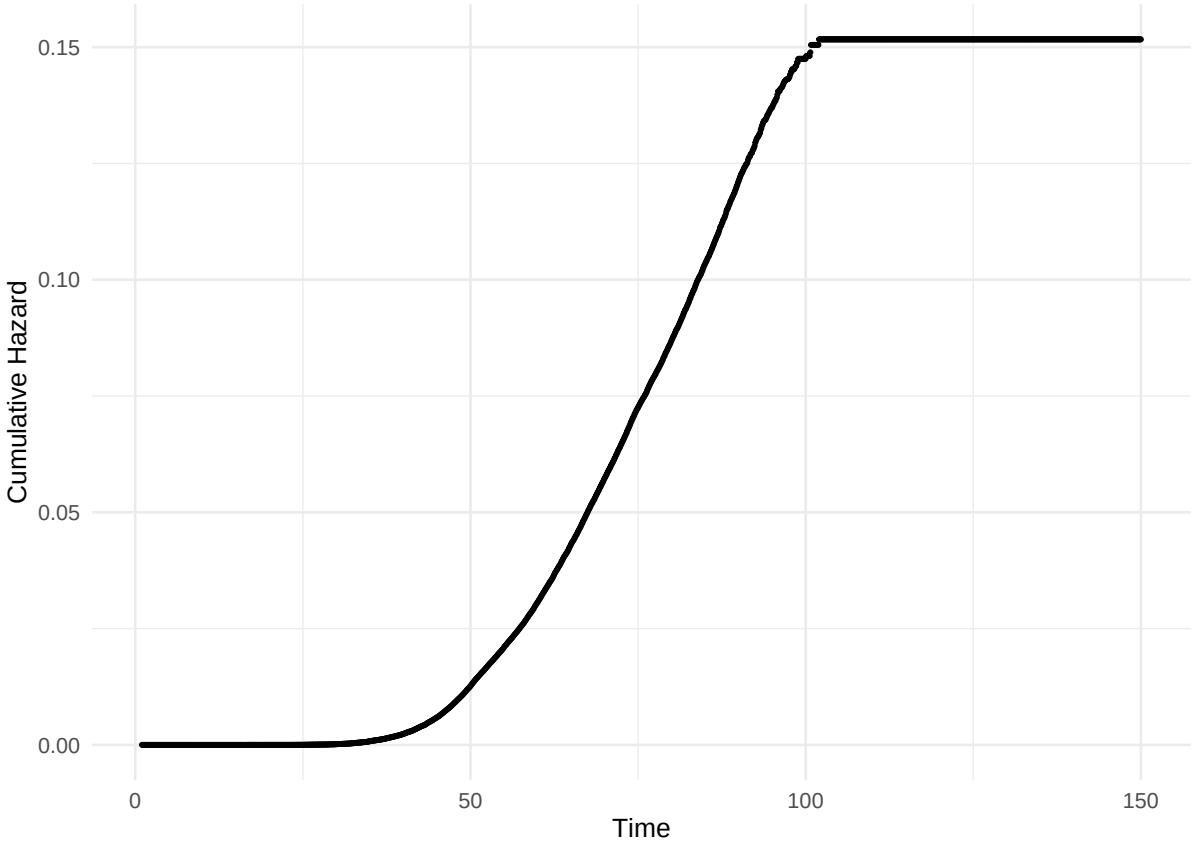

Figure 1.12: Estimated Breslow cumulative hazard function.

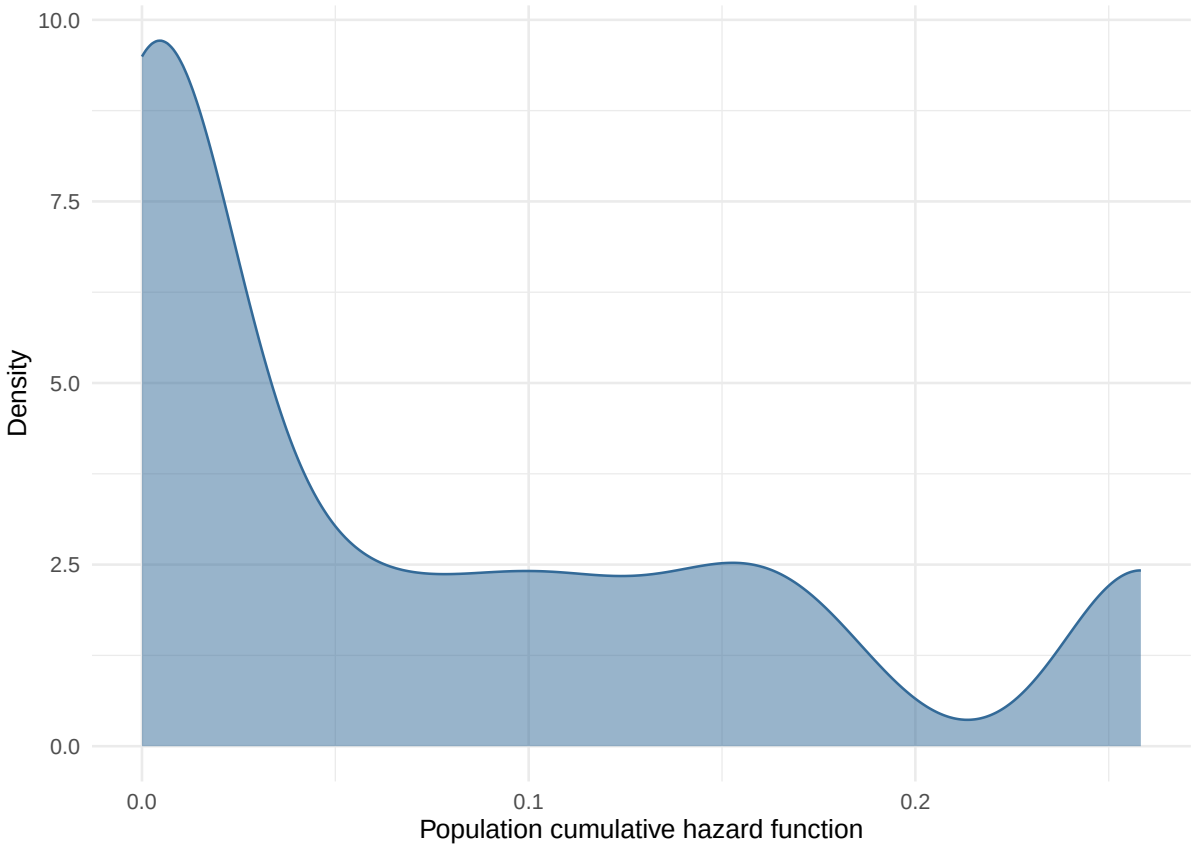

Figure 1.13: Density plot of the Breslow cumulative hazard function estimated on the entire population in object.

It is worth noting that in the parametric scenario, the population parametric distribution provides more information and therefore there is no requirement for computing the nonparametric estimates of the cumulative hazard function and the quantile threshold.

1.5.6 Parametric setting

The Swedish real case dataset also provides us with a precious opportunity to explore the debated issue of whether a cure-rate structure is reasonable in models for breast cancer onset or not.

We first focus on the goodness of fit of Cure-Rate models versus traditional models to the available dataset, to give a qualitative support to the hypothesis that a cured fraction exists and the Cure-Rate models are able to capture it. Notice that, in this preliminary analysis, the frailty quantity has not been involved yet. We employee five parametric distributions: Weibull, Gamma, Lognormal, three-parameter Gamma, and three-parameter Lognormal for the baseline survival function of cases.

Results from the analysis are presented in Table 1.18, where the models are compared using the Akaike Information Criterion (AIC) that is exactly

$$AIC = -2n_\pi - \ln L(\pi),$$

where n_π is the dimension of the parameter collection, and $L(\pi)$ is the likelihood on the parameter collection. It should be noted that the models are not nested. We compare the fit across the several survival function distributions, and also between the cure-rate and non-cure-rate survival structure but always within the multivariate and the univariate cases (which also have very different sample sizes). The Multivariate Cure-Rate three-parameter Lognormal model yields the best result, with an AIC value of 1687555, while the regular Lognormal distribution provides the best performance for the Univariate Cure-Rate model with an AIC of 438368.4. From results, the cure-rate models are always preferred to the non-cure-rate models (except for the case of the Multivariate Lognormal model).

All the curves shown in Figures 1.14, 1.15, and 1.16 support the hypothesis of involving a cure-rate model rather than a (traditional) non-cure model. Indeed, the cure-rate model seems to fit the data (until the end of the follow-up) better than the non-cure models.

It is interesting to notice that support to the cure-rate structure is mainly given by the graphical analysis of the survival function of the mothers, due to their older ages: in Figure 1.17, the Kaplan-Meier curves by subjects (the main subject, the mother, and from the first sister to the last one) show that the tail of the Kaplan-Meier estimator and of the fitted cure-rate models (Figure 1.14, 1.15, 1.16) is mostly attributable by the mothers (in black). A deepening about the reliability of the cure-rate structure and the heavy tail due to the presence of the oldest mothers is in Appendix A.4. We do not find particular graphical differences in curves among the daughters, as it should be expected since the main subject is randomly sampled among all the sisters. One might extend the models to allow for a (say, polynomial) effect of birth cohort on the survival distribution and make distinction between the mother and the sisters.

	Survival function	
	non-cure	cure-rate
Multivariate model		
Weibull	1705353	1692822
Gamma	1698245	1688066
Lognormal	1693854	2334626
three-parameter Gamma	1707589	1687749
three-parameter Lognormal	1698257	1687555
Univariate model		
Weibull	440685.6	439053.4
Gamma	439516	438391.8
Lognormal	438958.8	438368.4
Gamma 3-parameters	444890.7	438386.2
Lognormal 3-parameters	439460.9	438369

Table 1.18: AIC comparison among different survival distributions.

Once proved that the cure-rate model is the most appropriate for describing this real case dataset, we proceed to parameter estimation. We report the values of the estimated parameters in Table 1.19 and 1.20. We run the analysis both on main subjects with a recorded mother in the registry or on main subjects without restrictions. In Table 1.19 the survival function has a Weibull distribution. On the contrary, in Table 1.20 we extend to different baseline survival distributions on all the main subjects. The estimated cured fraction value has a reasonable value for this specific application in the range around 87%-96% for $\widehat{p}$ in the first column of Table 1.19. The tables also report the estimates of the baseline survival distribution parameters, that are $\widehat{shape}_0$, and $\widehat{scale}_0$ for the Weibull and Gamma, with the addition of $\widehat{\gamma}_0$ for the threshold parameter in the three-parameter Gamma; $\widehat{\mu}_0$, and $\widehat{\sigma}_0^2$ for the Lognormal, with the addition of $\widehat{\gamma}_0$ for the threshold parameter in the three-parameters Lognormal. The frailty parameter $\widehat{\theta}$ and the FH coefficient $\widehat{\beta}_{FH}$ are reported in the same column because estimating one of the two exclude the estimation of the other. Notice that only the models with "FH" in the name estimate the family history coefficient β_{FH}.

Similarly to the semiparametric scenario, the Multivariate frailty Cure-Rate model outperforms the other two Univariate models in terms of prediction accuracy through the Harrell's concordance

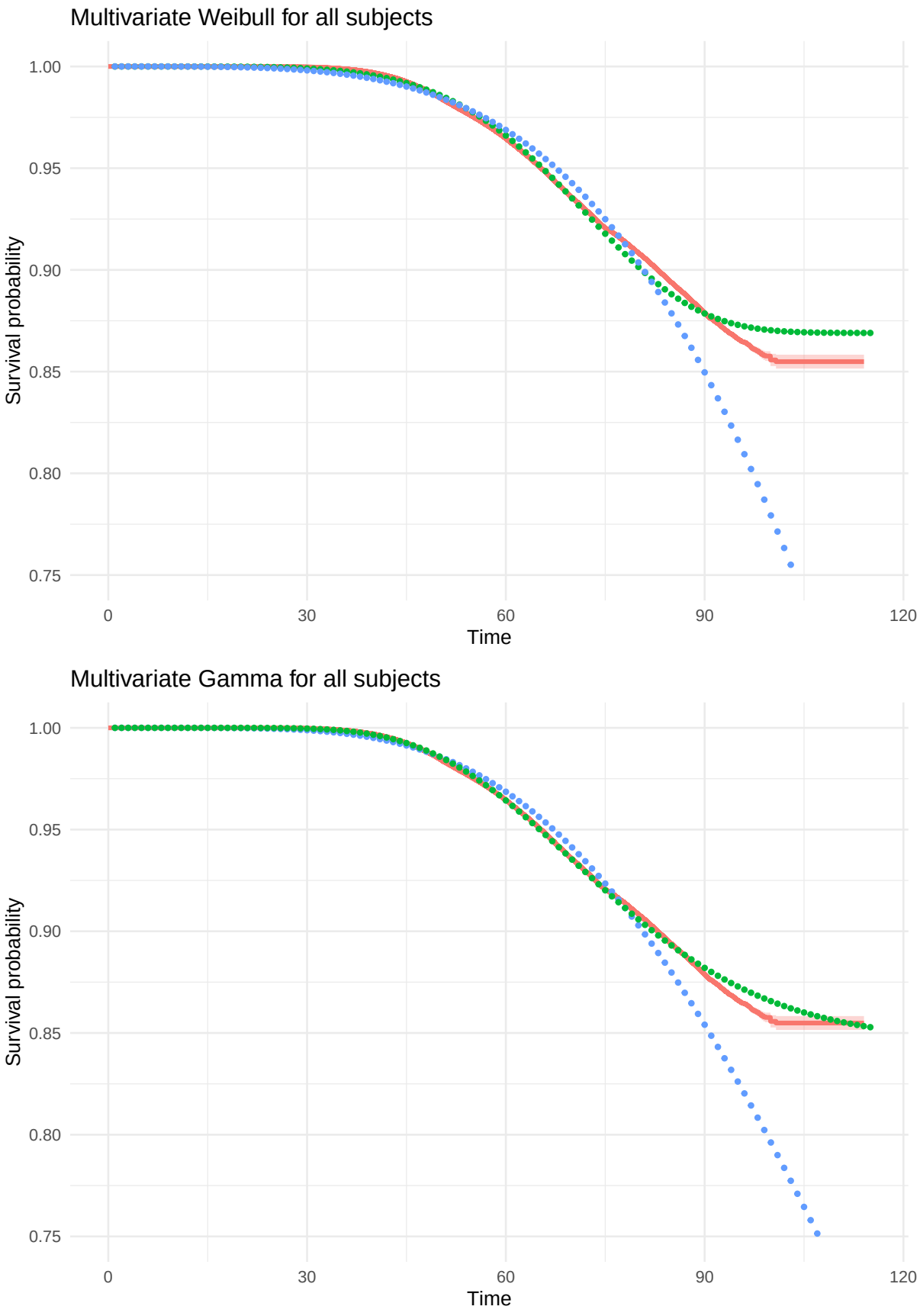

Figure 1.14: Kaplan-Meier for all subjects, in comparison to the estimated survival curves following the cure-rate (green) and the non-cure (blue) survival structure with Weibull and Gamma baseline survival function, on top and below, respectively.

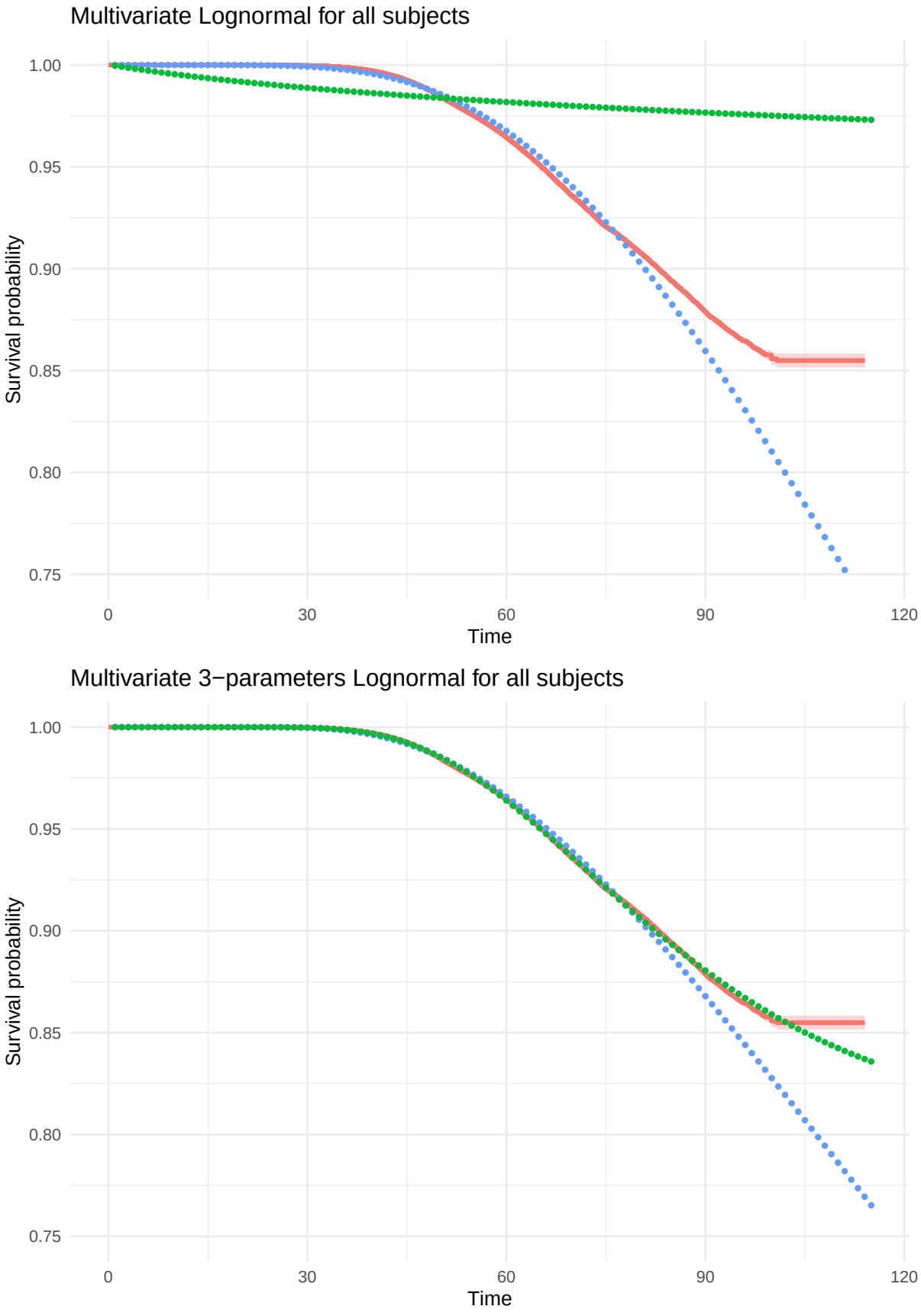

Figure 1.15: Kaplan-Meier for all subjects, in comparison to the estimated survival curves following the cure-rate (green) and the non-cure (blue) survival structure, with Lognormal and three-parameter Lognormal survival function, on top and below, respectively.

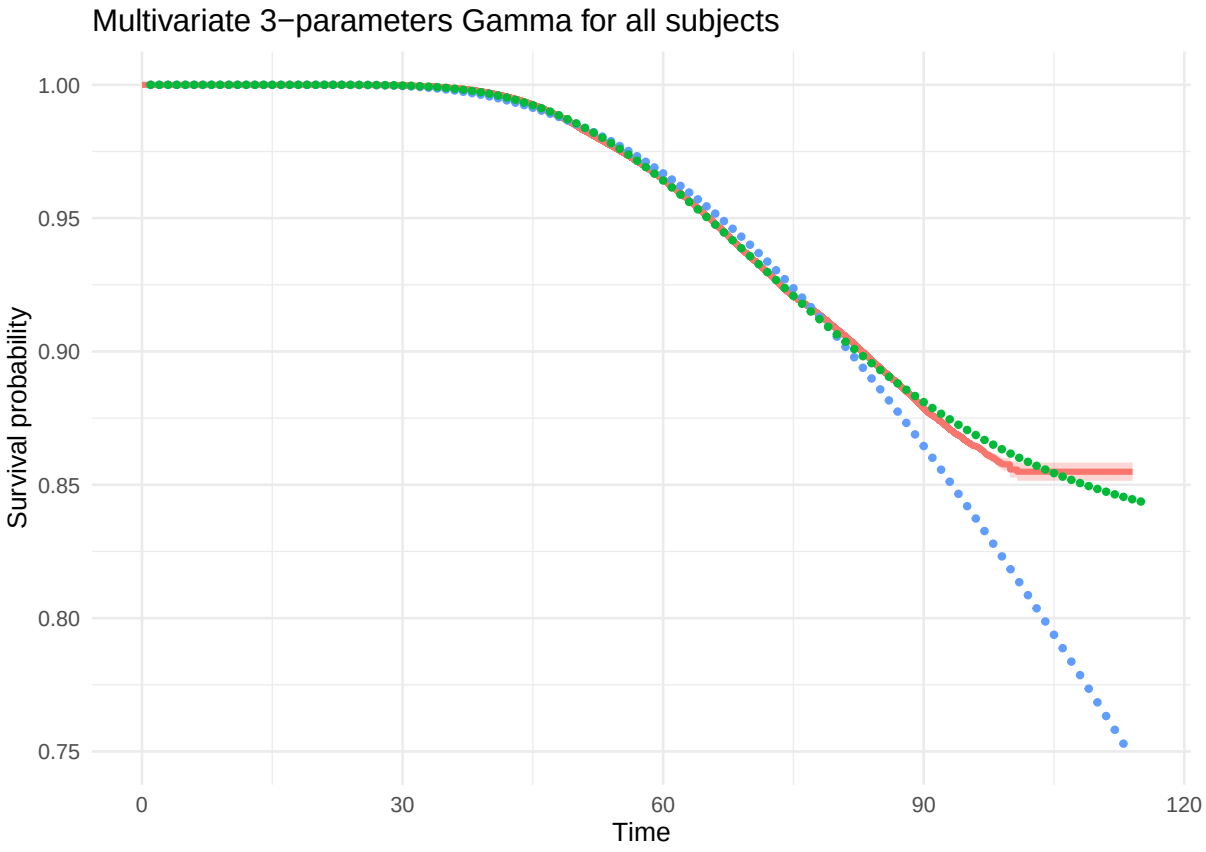

Figure 1.16: Kaplan-Meier for all subjects, in comparison to the estimated survival curves following the cure-rate (green) and the non-cure (blue) survival structure, with a three-parameter Gamma baseline survival function.

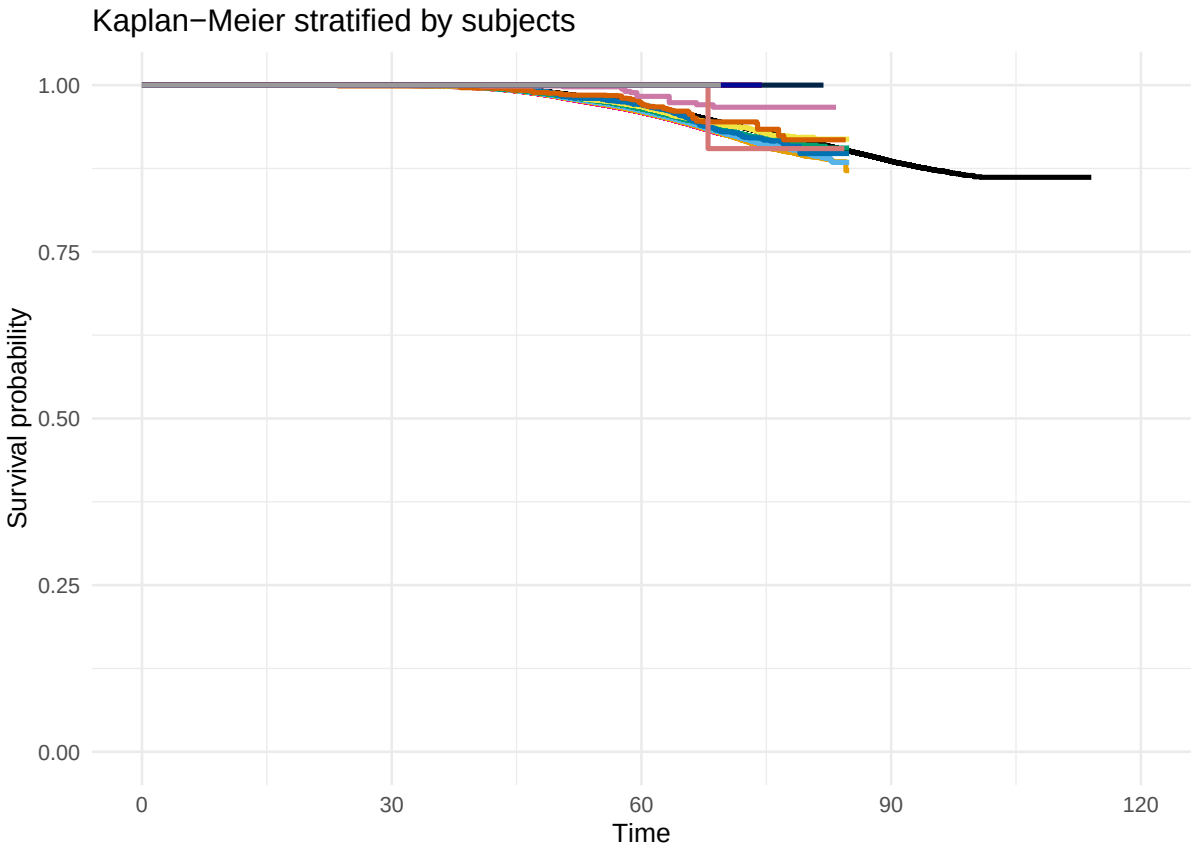

Figure 1.17: Kaplan-Meier of daughters and mother (in black).

	Survival info mother	$\widehat{p}$	$\widehat{shape}_0$	$\widehat{scale}_0$	$\widehat{\theta}/\widehat{\beta}_{FH}$	C
MFCR	yes	0.8710	6.2645	73.2018	4.9053	0.9592
UFCR		0.9594	0.1246	0.1900	1.2150	0.3952
UFHCR		0.9635	0.0926	0.0394	1.2174	-
MFCR	no	0.8703	6.1822	72.9776	4.3569	0.9663
UFCR		0.9750	0.1176	0.1193	1.2614	0.3963
UFHCR		0.9658	0.0965	0.0362	1.1982	-

Table 1.19: Estimated parameters and Concordance index for the Multivariate frailty Cure-Rate (MFCR), the Univariate frailty Cure-Rate (UFCR), and the Univariate FH Cure-Rate (UFHCR) models.

index (C) , reported in the last column of Tables 1.19 and 1.20. The Multivariate frailty Cure-Rate model can indeed coherently predict the risk for the 96% of pairs accordingly to their time-to-event. The concordance increases of around the 40-53% when moving from the Univariate model to the Multivariate model.

In comparison to the semiparametric scenario we can say that the estimated frailty parameter

	$\widehat{p}_0$	$\widehat{\mu}_0/\widehat{shape}_0$	$\widehat{\sigma}_0^2/\widehat{scale}_0$	$\widehat{\gamma}_0$	$\widehat{\theta}/\widehat{\beta}_{FH}$	C
Multivariate frailty Cure-Rate						
Weibull	0.8703	6.1822	72.9776	-	4.3569	0.9663
Gamma	0.8552	18.6981	3.8617	-	5.4790	0.9663
Lognormal	0.8419	4.2911	0.2604	-	5.5265	0.9392
Gamma 3-pars	0.8529	18.6685	3.8172	0.1546	5.8116	0.9679
Lognormal 3-pars	0.8408	4.2746	0.2590	2.7922	5.9126	0.9663
Univariate frailty Cure-Rate						
Weibull	0.9750	0.1176	0.1193	-	1.2614	0.3963
Gamma	0.8626	5.9127	11.5267	-	-5.8713	0.3963
Lognormal	0.6607	4.5317	0.3784	-	1.4119	0.5105
Gamma 3-pars	0.3250	9.5974	11.0534	3.1358	-1.4014	0.4461
Lognormal 3-pars	0.2527	4.8125	0.4431	11.5269	2.0759	0.4900
Univariate FH						
Weibull	0.9635	0.0926	0.0394	-	1.2174	-
Gamma	0.9590	22.0724	2.5169	-	-1.1910	-
Lognormal	0.9774	2.6356	0.9717	-	-3.0954	-
Gamma 3-pars	0.7076	11.8656	6.9557	0.5153	0.7802	-
Lognormal 3-pars	0.5237	4.6639	0.4117	7.9716	2.0892	-

Table 1.20: Estimated parameters for several baseline survival distribution for cases.

has a higher value in the multivariate parametric scenario bringing to assess that the frailty distribution has a higher variance and thus this brings to an easier distinction among low-risk and higher-risk families rather than the semiparametric scenario. Also, the Multivariate frailty Cure-Rate model achieves the equal amount of prediction accuracy than the Multivariate frailty Cox model with additional pros: it is able to estimated the cured fraction and the survival function of breast cancer cases, helping in explaining the phenomenon of breast cancer.

Risk prediction now comes straightforward as shown in Section 1.5. We compute the mean, median and mode of the posterior frailty distribution for each family by employing the posterior frailty risk distribution, given for example by the updated Gamma($shape = 4.3569 + \sum_{j=1}^{n_i} \delta_j$, $rate = 4.3569 - \sum_{j=1}^{n_i} \log\left(0.8703 + (1 - 0.8703)\widetilde{S}(x_j)\right)$), when a Weibull distribution is chosen for the survival function on cases. Once we have the posterior distribution of the risk frailty given the whole detailed family data, one can compute the posterior high-risk probability, and the posterior high-risk membership indicator fixed a frailty mean threshold. All of these measures can be used to assess whether to address her to prevention strategies targeted for the highest-risk families.

1.6 Discussion

This study aims to contribute to the study of risk prediction models for breast cancer. We consider the cure-rate structure as a realistic approach for the Swedish Multi-Generational Breast Cancer registry, where around the 85% of subjects have not experienced yet breast cancer onset within the observed follow-up. We thus extend the traditional proportional hazards assumption in this Lehmann family formulation to cure-rate models. We develop the Multivariate frailty Cure-Rate, the Univariate frailty Cure-Rate and the Univariate FH Cure-Rate parametric models which admit the cure-rate structure of the survival function, in contrast to already developed and known in the literature Cox models which do not admit a cure-rate structure.

Although family information is crucial for risk prediction models for breast cancer, using only a summary of it, like the family history, may not have enough predictive power. Our simulation-based comparison shows that a full multivariate framework induces much better performance in terms of accuracy in risk prediction, when only involving family membership without additional subject-specific covariates. A full assessment of the added value of the Multivariate frailty Cure-Rate model will clearly emerge when additional analyses can be conducted on other dataset. Including family-specific covariates can enhance precision in targeting and improve accuracy in identifying the frailty parameter, as well as classifying families into risk groups. Therefore, incorporating family-specific covariates is an intriguing extension worth exploring.

Our conclusion so far is about the superiority of the Multivariate frailty Cure-Rate model over all the Univariate models and also over its semiparametric counterpart. The Multivariate frailty Cure-Rate model, without losing prediction accuracy, perfectly describe both the fraction of women that won't develop breast cancer during their lifetime and the survival function of cases, thanks to the cure-rate structure involved into the models.

Another point that we want to highlight is the comparison between the Multivariate frailty Cure-Rate model and the BOADICEA model, which is one of the most powerful tool regarding risk prediction seen in the literature so far. The BOADICEA model is based on a multiplicative hazard function that depends on a genetic frailty component. This approach consists of inferring a genetic latent quantity, the polygenic risk score (PRS), for predicting cancer risk based on the family history of the disease and other risk factors. The PRS is a weighted combination of single nucleotide polymorphisms (SNPs), which are genetic mutations, commonly spread into the population, that singularly give a small contribution to increase the risk of breast cancer, but that can be dangerous when combined all together. The BOADICEA model is then univariate and based on the subject-specific hazard function $\lambda(t \mid r) = \lambda_0(t)exp(\beta(r))$, as described in detail in several publications ([1], [3], [2], and [37]). It evaluates the likelihood function through a complex segregation analysis provided by the Mendel software [18].

In contrast, our proposed model differs from the BOADICEA model in several ways. While BOADICEA uses a subject-specific hazard within a univariate framework and infers a genetic latent quantity, our model works in a fully multivariate framework and we infer a generic risk latent quantity, i.e. the subject-specific polygenic score. We incorporate survival information in a family-

specific hazard, while the BOADICEA hazard function is subject-specific and does not account explicitly for the family structure through the estimated genetic frailty component. Our model wins in the simplicity that it has involving only familial breast cancer information and no other factors such as genetic, as BOADICEA does. Nevertheless, it is not difficult to extend our work to the use of additional covariates. The simplicity is also extended to the use of a full likelihood or partial likelihood maximization algorithm for parameter estimation, while BOADICEA relies on a complex segregation analysis to predict the PRS (specifically its variance). Also, the cure-rate structure that we explicitly introduce is not considered into the BOADICEA model. This is a limitation for the BOADICEA model, as we saw how much is crucial relying on the cure-rate structure especially when dealing with breast cancer risk prediction models.

Talking about possible extensions of our work, one could be the use of alternative frailty distributions, such as the Lognormal [8]; or, one could address the problem within the Bayesian framework [17]. One could also compare the prediction accuracy between our Multivariate frailty Cure-Rate model to the BOADICEA model, but it would be necessary the access to the same complete dataset [19] which is so far unavailable.

References

[1] Antoniou, A., Pharoah, P., McMullan, G., Day, N., Stratton, M., Peto, J., Ponder, B., and Easton, D. (2002). A comprehensive model for familial breast cancer incorporating brca1, brca2 and other genes. *British journal of cancer*, 86(1):76–83.

[2] Antoniou, A. C., Cunningham, A., Peto, J., Evans, D., Lalloo, F., Narod, S., Risch, H., Eyfjord, J., Hopper, J., Southey, M., et al. (2008). The boadicea model of genetic susceptibility to breast and ovarian cancers: updates and extensions. *British journal of cancer*, 98(8):1457–1466.

[3] Antoniou, A. C., Pharoah, P., Smith, P., and Easton, D. F. (2004). The boadicea model of genetic susceptibility to breast and ovarian cancer. *British journal of cancer*, 91(8):1580–1590.

[4] Balan, T. A. and Putter, H. (2019). frailtyem: An r package for estimating semiparametric shared frailty models. *Journal of Statistical Software*, 90(7):1–29.

[5] Balan, T. A. and Putter, H. (2020). A tutorial on frailty models. *Statistical methods in medical research*, 29(11):3424–3454.

[6] Bobée, B., Ashkar, F., Roy, R., and Perreault, L. (1991). Risk analysis of hydrological data using gamma family and derived distributions. In *Water Resources Engineering Risk Assessment*, pages 21–41. Springer.

[7] Colditz, G. A., Rosner, B. A., Speizer, F. E., and Group, N. H. S. R. (1996). Risk factors for breast cancer according to family history of breast cancer. *JNCI: Journal of the National Cancer Institute*, 88(6):365–371.

[8] Duchateau, L. and Janssen, P. (2008). *The frailty model*. Springer.

[9] Eriksson, M. (2021). *Risk assessment and prevention of breast cancer*. PhD thesis, Karolinska Institutet (Sweden).

[10] Evans, D. G. R., Kerr, B., Cade, D., Hoare, E., and Hopwood, P. (1996). Fictitious breast cancer family history. *The Lancet*, 348(9033):1034.

[11] Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., and Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute*, 81(24):1879–1886.

[12] Gelber, R. D., Cole, B., Goldhirsch, A., Rose, C., Fisher, B., Osborne, C., Boccardo, F., Gray, R., Gordon, N., Bengtsson, N., et al. (1996). Adjuvant chemotherapy plus tamoxifen compared with tamoxifen alone for postmenopausal breast cancer: meta-analysis of quality-adjusted survival. *The Lancet*, 347(9008):1066–1071.

[13] Government, A. (2022). Risk factors.

[14] Group, C. et al. (2001). Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *The Lancet*, 358(9291):1389–1399.

[15] Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546.

[16] Hougaard, P. (2012). *Analysis of multivariate survival data*. Springer Science & Business Media.

[17] Karamoozian, A., Baneshi, M. R., and Bahrampour, A. (2021). Bayesian mixture cure rate frailty models with an application to gastric cancer data. *Statistical Methods in Medical Research*, 30(3):731–746.

[18] Lee, A., Mavaddat, N., Wilcox, A. N., Cunningham, A. P., Carver, T., Hartley, S., Babb de Villiers, C., Izquierdo, A., Simard, J., Schmidt, M. K., et al. (2019). Boadicea: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genetics in Medicine*, 21(8):1708–1718.

[19] Lee, M., Reilly, M., Lindström, L. S., and Czene, K. (2017). Differences in survival for patients with familial and sporadic cancer. *International Journal of Cancer*, 140(3):581–590.

[20] Lee, V. (2022). Metastatic breast cancer.

[21] Loibl, S., Poortmans, P., and Morrow, M. (2021). Breast cancer. *The Lancet*.

[22] Meijering, E. (2002). A chronology of interpolation: from ancient astronomy to modern signal and image processing. *Proceedings of the IEEE*, 90(3):319–342.

[23] Moher, D., Schulz, K. F., Altman, D. G., and Group*, C. (2001). The consort statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Annals of internal medicine*, 134(8):657–662.

[24] Mota, A., Milani, E. A., Leão, J., Ramos, P. L., Ferreira, P. H., Junior, O. G., Tomazella, V. L., and Louzada, F. (2022). A new cure rate frailty regression model based on a weighted lindley distribution applied to stomach cancer data. *Statistical Methods & Applications*, pages 1–27.

[25] Ozenne, B., Sørensen, A. L., Scheike, T., Torp-Pedersen, C., and Gerds, T. A. (2017). riskregression: predicting the risk of an event using cox regression models. *The R Journal*, 9(2):440–460.

[26] Pharoah, P. D., Day, N. E., Duffy, S., Easton, D. F., and Ponder, B. A. (1997). Family history and the risk of breast cancer: a systematic review and meta-analysis. *International journal of cancer*, 71(5):800–809.

[27] Rahman, M. S., Ambler, G., Choodari-Oskooei, B., and Omar, R. Z. (2017). Review and evaluation of performance measures for survival prediction models in external validation settings. *BMC medical research methodology*, 17(1):1–15.

[28] Riggio, A. I., Varley, K. E., and Welm, A. L. (2021). The lingering mysteries of metastatic recurrence in breast cancer. *British journal of cancer*, 124(1):13–26.

[29] Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56(4):1016–1022.

[30] Rodrıguez, G. (2010). Multivariate survival models.

[31] Rosner, B., Colditz, G. A., Iglehart, J. D., and Hankinson, S. E. (2008). Risk prediction models with incomplete data with application to prediction of estrogen receptor-positive breast cancer: prospective data from the nurses' health study. *Breast Cancer Research*, 10(4):1–11.

[32] Scheike, T. H., Hjelmborg, J. B., and Holst, K. K. (2015). Estimating twin pair concordance for age of onset. *Behavior genetics*, 45:573–580.

[33] Schmid, M., Wright, M. N., and Ziegler, A. (2016). On the use of harrell's c for clinical risk prediction via random survival forests. *Expert Systems with Applications*, 63:450–459.

[34] Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(1):7–34.

[35] Strandberg, R. (2022). *Breast cancer natural history models and risk prediction in mammography screening cohorts*. Karolinska Institutet (Sweden).

[36] Taroni, P., Quarto, G., Pifferi, A., Ieva, F., Paganoni, A. M., Abbate, F., Balestreri, N., Menna, S., Cassano, E., and Cubeddu, R. (2013). Optical identification of subjects at high risk for developing breast cancer. *Journal of biomedical optics*, 18(6):060507–060507.

[37] Thomas, D. C. et al. (2004). *Statistical methods in genetic epidemiology*. Oxford University Press.

[38] Tillander, A. (2022). Biostat3: Utility functions, datasets and extended examples for survival analysis. r package version 0.1.8.

[39] Tokatli, Z. F., Türe, M., Ömürlü, İ. K., Alas, R. Ç., and Uzal, M. C. (2011). Developing and comparing two different prognostic indexes for predicting disease-free survival of nonmetastatic breast cancer patients. *Turkish Journal of Medical Sciences*, 41(5):769–780.

[40] Tyrer, J., Duffy, S. W., and Cuzick, J. (2004). A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in medicine*, 23(7):1111–1130.

[41] Weir, A. J. (1973). *Lebesgue integration and measure*, volume 1. Cambridge University Press.

[42] Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E., and Groothuis-Oudshoorn, C. G. (2018). Time-varying covariates and coefficients in cox regression models. *Annals of translational medicine*, 6(7).

Chapter 2

Two-latent-class Lehmann Cure-Rate models for age at disease onset - a simulation study

We are interested in a specific aspect of time-to-onset modelling, namely the investigation of family-specific risk of disease onset with particular emphasis on the risk component present from birth, as opposed to the environmental component. We assume that the true family ("frailty") risk is latent and remains constant from birth. We focus on breast cancer development, even though this work can be extended to a variety of diseases. Our goal is to estimate the true family risk, as assessing the risk is crucial in suggesting tailored screening and prevention strategies based on and individual's family risk level. In particular, we employ a univariate or multivariate frailty model on the time to breast cancer onset, using a binary risk classification to stratify families into a low-risk group and a high-risk group. We compare this model to one that uses as a strong risk factor the observed family history indicator, as a covariate to replace the unkown latent binary risk group. Indeed, the family history indicator should be expected to be a weak indicator of the complete detailed breast cancer familial history.

keywords: breast cancer, family history, frailty models, survival analysis

2.1 Introduction

In 2020, female breast cancer overtook lung cancer as the most common cancer in the world. An estimated 2.3 million new cases of breast cancer were reported, constituting 12% of all new cancer cases and 25% of female cancer cases. Breast cancer ranked 5th in mortality, with 6.9% among all cancer deaths, remaining the most common cause of cancer death in women (16%). For example, in the USA the 12% of women are estimated to experience breast cancer in their lifetime (see [20]), while in Sweden is the 9.4% before age 75 [17].

To tackle this problem, we want to use risk prediction models for breast cancer to identify families with the highest risk of developing breast cancer and provide them with targeted and more intensive screening and prevention strategies. Indeed, classifying subjects into risk groups allows for the modification of screening schedules (more/less intensive) depending on the risk of breast cancer for a given woman, for the implementation of additional prevention efforts, and for the reduction of unnecessary medical treatments, costs and psychological stress [12, 15].

Remarkable contributions to the breast cancer risk modelling field include the Gail model (refer to [7]), which employs logistic regression to integrate risk factors, such as the number of first-degree relatives with breast cancer, with the aim to compute the long-term probability of developing breast cancer. The Tyrer and Cuzick (TC) model, which (refer to [19]) integrates personal risk factors and complete genetic analysis (involving also BRCA1 and BRCA2 gene mutations) to model the risk of developing breast cancer by combining the genetic and familial components. The Rosner and Colditz model, which (refer to [14]) is based on a logistic model for incidence that is affected by reproductive risk factors, including age at menarche, age at menopause, and age at childbirth. These models have been implemented in various studies, such as those described in [4] and [6]. Models for disease onset also include the popular two-hit Moolgavkar-Venson-Knudson (MVK) cell-splitting model, which has all subjects eventually experience the disease, if right censoring does not intervene to end the observation of the time-to-event [10].

As we can observe from literature the inclusion of strong risk factors associated to breast cancer are commonly used into risk prediction models. One among the strongest risk factors (such as BRCA1, BRCA2, TP53, and SNPs, mammography density (MD) and body mass index (BMI) [2], [9], [16], [18]), the family history, is still involved in risk prediction models although we believe it is a weak indicator since it only provides a summary of the clinical history experienced by a family. Specifically, it is defined as the collection of breast cancer experiences within a family and is represented as a binary variable that takes a value of one when at least one family member has experienced breast cancer onset, and zero if none has. For comprehensive and complex data, family history may not fully capture the familial aggregation of breast cancer development.

On the other hand, the family history indicator motivates the binary nature of the breast cancer risk which leads to the split of families into a low-risk group and a highest-risk group. Thus, our objective is to develop a risk prediction model for age at breast cancer onset, say the beginning of the disease, which involves a family-specific risk assumed to be latent and unchanged from birth. Drawing inspiration from the family history indicator, we allow for this latent risk, namely

the frailty risk, to be discrete and comprise two risk levels (low and high), which we denote as 0 and 1, respectively. In the following, we explore a Univariate *FH* Cure-Rate model, a Univariate frailty Cure-Rate model and a Multivariate frailty Cure-Rate model, where frailty is referred to the latent risk of breast cancer development, "Cure-Rate" refers to the peculiar survival function which allows for a fraction of the population to not experience breast cancer onset eventually, and the difference between "Multivariate" and "Univariate" stands in jointly modelling all the time-to-events of a family, in opposition to model only the time-to-event of one subject per family. We refer to "main subject" the member that we randomly select when moving from the multivariate to the univariate scenario, and also the one we compute the family history on, since the family history is a subject-specific characteristic and the model that comprises it is univariate as well.

We seek to illustrate and quantify how family data can be better used to learn about family-specific risk of developing diseases by using such a Multivariate frailty Cure-Rate model for disease onset instead of summary-based methods (see family history), as usually is easier and used in the literature. Lastly, to provide a comprehensive and complete assessment, we implement the Univariate frailty Cure-Rate model to determine the significant loss of information incurred when subjects are viewed as not part of a family sharing the same risk of breast cancer development (as it is in the Multivariate frailty Cure-Rate model).

The chapter is outlined as follows: we introduce the univariate and multivariate background of the frailty Cure-Rate model in Section 2.2, the methods about the Multivariate frailty Cure-Rate model in Section 2.2.5. A comparison among the Univariate *FH* Cure-Rate model, the Univariate frailty Cure-Rate model, and the Multivariate frailty Cure-Rate model is run in Section 2.3, and we close with some discussion in Section 2.4.

2.2 Models for age at disease onset

2.2.1 Introduction to the Cure-Rate frailty models

We explain how step by step we can develop a univariate or multivariate Cure-Rate frailty models. We start from introducing frailty models and then we incorporate the Cure-Rate structure to them.

Univariate frailty models

The Univariate Frailty model [5] on the time-to-event $T = t$ allows the hazard function $\lambda_r(t) = \lambda(t \mid r)$ to have a particular form including the frailty risk R which captures the unobserved heterogeneity among subjects. The hazard is given by

$$\lambda_r(t) = \alpha(r)\lambda_0(t),$$

where $\lambda_0(t)$ is the baseline hazard function that can assume a parametric distribution with parameter collection θ or a semiparametric form. The quantity $\alpha(r)$ is a general function of the risk, that we may use in the linear form $\lambda_r(t) = r\lambda_0(t)$.

The model can be extended to the inclusion of subject-specific covariates. In this case the frailty quantity explains the unobserved heterogeneity that the covariates are not able to capture. The hazard function is given by:

$$\lambda_r(t) = r\lambda_0(t; x), \tag{2.1}$$

where x are the subject-specific covariates. Notice that in the multivariate setting, the shared frailty hazard function allows to define the frailty as a family-specific quantity. Thus, this specification is used with clustered data, as it is our case where we see families as clusters. The hazard function has the same equation as in 2.1, but the frailty risk r is seen this time as a familial characteristic.

In the binary case, the latent quantity can be represented as $R = (0, 1)$, where typically the relation between the hazard functions is $\lambda_1(t) = \alpha\lambda_0(t)$. This assumption allows the hazard and survival functions of group "0" to coincide with the baseline functions, while $\alpha < 1$ ensures coherence with the assumption of a highest-risk group in the population and thus we can rely on the assumption of proportional hazard. Therefore, we obtain:

$$\begin{aligned}
&\lambda_0(t) = \lambda(t \mid R = 0) = \lambda_0(t; x), \qquad S_0(t) = S(t \mid R = 0) = S_0(t; x),\\
&\lambda_1(t) = \lambda(t \mid R = 1) = \alpha\lambda_0(t; x), \qquad S_1(t) = S(t \mid R = 1) = [S_0(t; x)]^{\alpha}.
\end{aligned}$$

Multivariate frailty models

Now, recall the Multivariate frailty survival model [8] to describe modelling jointly the time-to-events in a family [13]. We handle multiple time-to-event data by leveraging the assumption of conditional independence. For instance, consider the case where two women belong to the same family, resulting in dependent time-to-events. However, assuming conditional independence given the family (i.e. given the shared frailty risk), and letting $T_1 = t_1$ and $T_2 = t_2$ be the time-to-events of the two women in the family, and let r denote the risk value, the joint survival function factorizes given the risk, so that

$$S_{12}(t_1, t_2 \mid R) \overset{T_1 \perp T_2 | R}{=} S_1(t_1 \mid R) S_2(t_2 \mid R),$$

where one therefore assumes conditional independence given the frailty risk term R. Recall that, if $R = 0$, we have

$$S_{12}(t_1, t_2) = S_0(t_1) S_0(t_2),$$

while, if $R = 1$, we have

$$S_{12}(t_1, t_2) = S_1(t_1) S_1(t_2) = [S_0(t_1) S_0(t_2)]^{\alpha}.$$

It is important to notice that this case can be immediately extended to more than two survival times per cluster sharing the same risk R. More generally, the marginal survival function for n_i subjects per family is given by

$$S_{1\ldots n_i}(t_1, \ldots, t_{n_i}) = h \prod_{j=1}^{n_i} S_0(t_j) + (1 - h) \prod_{j=1}^{n_i} \left[S_0(t_j)\right]^{\alpha},$$

with h the probability of belonging to the low-risk group of families.

Incorporating the Cure-Rate structure

Now, given the nature of the phenomenon, not all women will experience breast cancer onset, regardless of how long they will live. Therefore, we rely on the cure-rate survival function [11], which can be considered as a mixture of a proper survival function, which models the fraction of individuals who will experience the event: the "cases", and a degenerate distribution, which models the fraction of individuals who will not experience the event: the "non-cases". We define a "proper" survival function in the case it tends to 0 when the time-to-events tends to $+\infty$, such as $\lim_{t\to+\infty} S(t) = 0$, and has probability equal to one that the time-to-event can not assumes value $+\infty$, such as $P(T < +\infty) = 1$.

Let T indicate a non-negative time-to-event random variable, the survival function that defines a cure-rate model takes the form:

$$S_0(t) = p_0 + (1 - p_0)\widetilde{S}_0(t)$$

(see, e.g., [1]) with $\widetilde{S}_0(t)$ the proper survival function. Indeed, let the survival random variable T be such that, conditionally on being a case, it is absolutely continuous, and let $\widetilde{f}_0(t)$ indicate the conditional density function of the cases corresponding to the proper survival function $\widetilde{S}_0(t)$. In contrast, the fraction p_0 is defined the "cured fraction", i.e. the fraction of the subjects who will never experience the event of interest, so that $T = +\infty$ with probability p_0. Figure 2.1 shows the difference between a traditional (in blue) and a cure-rate (in red) survival model on randomly generated data.

The question of whether a cure-rate model is appropriate for a given phenomenon can be addressed by noting that a traditional proper survival function can be seen as a special case of a cure-rate model with $p_0 = 0$. In other words, allowing for a cure-rate simply enlarges the set of available models, within which traditional survival functions are nested through such constraint. We believe that implementing a cure-rate model is the right way to address the problem of modelling breast cancer development.

Thus, we assume that there exist two latent risk classes: low (or "general") risk (R=0) and high-risk (R=1). Let $h = P(R = 1)$. For the two risk classes one has $S_r(t) = p_r + (1 - p_r)\widetilde{S}_r(t)$, with $r \in \{0, 1\}$ (see Figure 2.2 from a trivial simulation study in R), such that

$$S_0(t) = p_0 + (1 - p_0)\widetilde{S}_0(t) \tag{2.2}$$

$$S_1(t) = [p_0 + (1 - p_0)\widetilde{S}_0(t)]^{\alpha} = p_1 + (1 - p_1)\widetilde{S}_1(t) \tag{2.3}$$

After this introduction on frailty cure-rate models, let us give an insight on the reason why we develop models with a two-latent-class approach. The family history model may be a valid indicator which splits families into two risk groups of developing breast cancer. A section regarding the family history model follows.

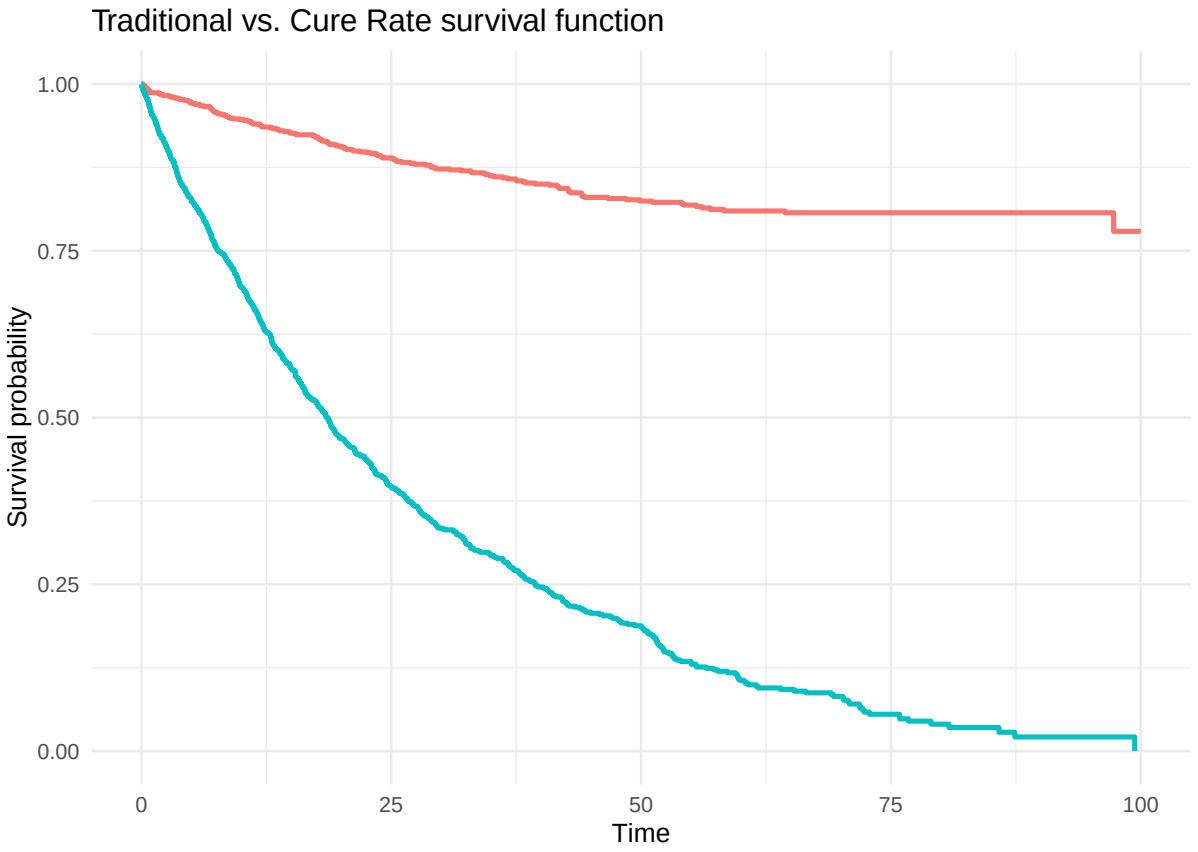

Figure 2.1: traditional survival function (in blue) vs. a cure-rate survival function (in red).

2.2.2 The Univariate FH Lehmann Cure-Rate model

The family history (FH) model can be a valid approach for modelling the familial risk of breast cancer. We develop a Univariate model involving the family history indicator as commonly done in the literature.

We consider one main subject i per family, and define $FH(u)$ as the indicator function that takes value 1 if one or more relatives of the subject have experienced the disease by the subject age u. For a family with four members (subject, sister, mother, and grandmother), we have for example $FH(u) = 1 - \mathbb{I}(bg + tg \geq b + t)\mathbb{I}(bm + tm \geq b + t)\mathbb{I}(bs_1 + ts_1 \geq b + t) = 1 - \mathbb{I}(tg \geq t + 60)\mathbb{I}(tm \geq t + 30)\mathbb{I}(ts_1 \geq t)$, assuming each generation is 30 years apart one from the other (so that the grandmother is 60 years old, and the mother 30 years old when the subject and sister are born). Let us apply the same Lehmann family and cure-rate structure as in the frailty setting, to obtain the general form of the Lehmann survival function depending on $FH(u)$ given by:

$$S_F(x_i) = [S_0(x_i)]^{\beta_F FH(x_i)}.$$

Notice that, the same cure-rate baseline survival function and conditional density function for

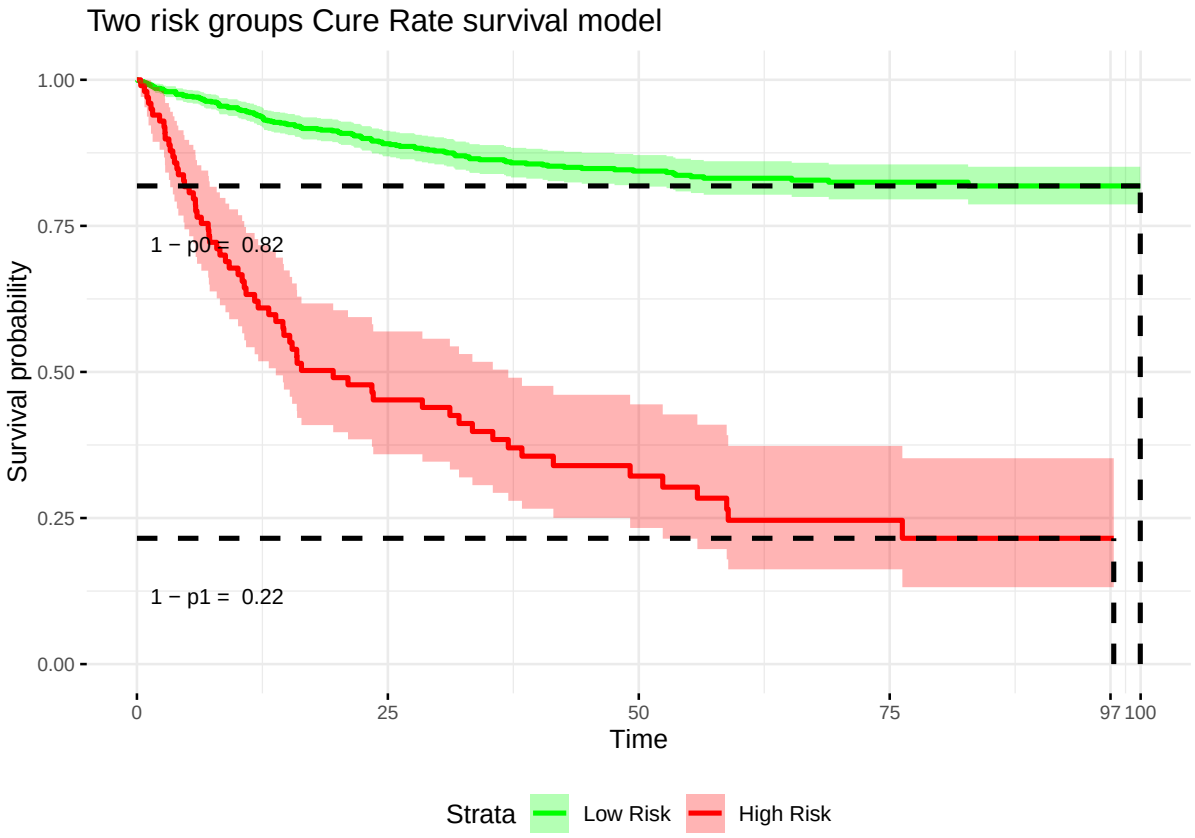

Figure 2.2: cure-rate model with two latent risk groups.

the cases are involved for the low-risk group. Consequently, for the high-risk group we have

$$S_1(x_i) = [S_0(x_i)]^{\beta_F} = [p_0 + (1-p_0)\widetilde{S}_0(x_i)]^{\beta_F} = p_1 + (1-p_1)\widetilde{S}_1(x_i)$$
$$f_1(x_i) = (1-p_1)\left(\frac{1-p_0}{1-p_1}\right)\beta_F \widetilde{f}_1(x_i)\left(p_0 + (1-p_0)\widetilde{S}_0(x_i)\right)^{\beta_F - 1}$$
$$\text{with, } p_1 = p^{\beta_F} \text{ and, } \widetilde{S}_1(x_i) = \frac{(p_0 + (1-p_0)\widetilde{S}_0(x_i))^{\beta_F} - p_1}{1-p_1}$$

The parameter β_F is the observed family history risk modifier, and it is typically used to account for the increased family risk for subjects that have a positive family history of breast cancer. In other words, $FH(t)$ is meant to estimate the risk of breast cancer development from the observed onset histories at time of the analysis t.

We build the closed form of the family history likelihood, without frailty quantity involved. The parameter collection is $\underline{\theta}_{FH} = \{p_0, \underline{\lambda}^T, \beta_F\}^T$, where recall p_0 is the cured fraction, and $\underline{\lambda}^T$ is the (vector) parameter collection of the baseline survival function, whose dimension depends on the

chosen distribution. The univariate likelihood involving the family history indicator is given by

$$
\begin{aligned}
L_{FH}(\underline{\theta}_{FH}; \text{subject data}) &= \prod_{i=1}^{n} f_X(\underline{FH}_i, \underline{x}_i; \underline{\theta}_{FH}) \qquad (2.4)\\
&= \prod_{i=1}^{n} f_X(\underline{x}_i \mid FH_i = 0; \underline{\theta}_{FH})^{(1-FH_i)} f_X(\underline{x}_i \mid FH_i = 1; \underline{\theta}_{FH})^{FH_i}\\
&= \prod_{i=1}^{n} \left[f_0(x_i)^{\delta_i} S_0(x_i)^{1-\delta_i} \right]^{(1-FH_i)} \left[f_1(x_i)^{\delta_i} S_1(x_i)^{1-\delta_i} \right]^{FH_i},
\end{aligned}
$$

where $\underline{x}_i = (x_i, \delta_i)^T$, with observed time $x_i = \min(t_i, c_i)$, t_i the time-to-event, and c_i the right-censoring time, both measured from the same origin, and $\delta_i = \mathbb{I}(t_i \leq c_i)$ the indicator of having observed the event, for subject i.

This model motivates the development of two-latent-class models because it splits families into those who has at least one breast cancer case into the family and those who has not. Thus we move to the development of two-latent-class models in the univariate and multivariate setting to prove that they can outperforms this too simplistic family history model.

Let us highlight few differences between the family history indicator and the binary frailty risk. The risk R takes value zero or one from birth and does not change over time, $FH(t)$ is a counting process that takes value one as soon as the first onset occurs among any of the other family members. Replacing the true unknown risk group R with the proxy $FH(t)$ leads to measurement error in the unknown value of R for the family. A detailed comparison of FH vs. R in terms of probability of agreement is illustrated in Appendix B.3, and an alternative building of the FH indicator is developed in B.4.

2.2.3 A note of non-identifiability

Before developing the final univariate model that we will use into the simulation studies, we deeply run an analysis of the non-identifiability of frailty models. This is very interesting and not easy to manage in the case of two risk groups.

Let us take again the two conditional distributions $\widetilde{S}_0(t)$ and $\widetilde{S}_1(t)$ from 2.2: they can be chosen freely, and to them correspond two given density functions $\widetilde{f}_0(t)$ and $\widetilde{f}_1(t)$ with (possibly vector) parameters θ_0 and θ_1, respectively. Thus the complete (vector) parameter for the model is $\underline{\theta} = (p_0, p_1, \theta_0^T, \theta_1^T, h)^T$.

Recall that the complete observed data ($\underline{x} = (x_1, x_2, \ldots, x_n)^T$, $\underline{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)^T$) is an i.i.d. sample of independently right-censored observed survival times from the population, where for the generic subject i, $x_i = \min(t_i, c_i)$, and $\delta_i = \mathbb{I}(t_i \leq c_i)$. Without additional constraints, from the observed data $(\underline{x}, \underline{\delta})$ one may not identify the parameter vector $\underline{\theta}$.

We explore different scenarios in the following lines as: (I) identifiability of the classical survival function $S_0(t) = \widetilde{S}_0(t)$; (II) identifiability of the cure-rate survival function $S_0(t) = p_0 + (1 - p_0)\widetilde{S}_0(t)$, with $\widetilde{S}_0(t)$ a proper survival function; (III) identifiability of the Lehmann structure $S_1(t) = [S_0(t)]^{\alpha(z)}$; (IV) identifiability of the cure-rate Lehmann structure $S_1(t) = [p_0 + (1 - p_0)\widetilde{S}_0(t)]^{\alpha(z)}$;

(V) identifiability of the marginal cure-rate survival function $S(t) = (1-h)S_0(t) + hS_1(t)$. Notice that we generalize the form of $\alpha(z)$ to be in function of covariates z, but it may be also a constant.

Trivially cases (I), (II), and (III) can be proved. The proof of the tricky case (IV) follows.

Proof.

$$
\begin{aligned}
&S_1(t) = [p_0 + (1-p_0)\widetilde{S}_0(t)]^{\alpha(z)} = p_0^{\alpha(z)} + \left(1 - p_0^{\alpha(z)}\right)\widetilde{S}_1(t;\alpha(z)) \\
&[p_0 + (1-p_0)\widetilde{S}_0(t;\theta)]^{\alpha(z)} = [p_0 + (1-p_0)\widetilde{S}_0(t;\theta')]^{\alpha'(z)} \quad \forall z \\
&\alpha(z)\log[p_0 + (1-p_0)\widetilde{S}_0(t;\theta)] = \alpha'(z)\log[p_0 + (1-p_0)\widetilde{S}_0(t;\theta')] \\
&\underbrace{\frac{\alpha(z)}{\alpha'(z)}}_{\text{not in function of t}} = \underbrace{\frac{\log[p_0 + (1-p_0)\widetilde{S}_0(t;\theta)]}{\log[p_0 + (1-p_0)\widetilde{S}_0(t;\theta')]}}_{\text{not in function of z}} = c.
\end{aligned}
$$

Since,

$$
\begin{aligned}
&\lim_{t\to\infty}\widetilde{S}_0(t;\theta) = 0,\ \lim_{t\to\infty}\log(p_0 + (1-p_0)\widetilde{S}_0(t;\theta)) = \log(p_0) \Rightarrow c = 1 \\
&\Rightarrow \alpha(z) = \alpha'(z).
\end{aligned}
$$

Then also,

$$
\begin{aligned}
&\log[p_0 + (1-p_0)\widetilde{S}_0(t;\theta)] = \log[p_0 + (1-p_0)\widetilde{S}_0(t;\theta')] \\
&\Rightarrow p_0 + (1-p_0)\widetilde{S}_0(t;\theta) = p_0 + (1-p_0)\widetilde{S}_0(t;\theta') \\
&\Rightarrow \widetilde{S}_0(t;\theta) = \widetilde{S}_0(t;\theta') \\
&\Rightarrow \theta = \theta'
\end{aligned}
$$

□

This proves the identifiability of the cure-rate Lehmann survival function, in the case where the baseline survival function $S_0(t;\theta)$ has a parametric distribution.

The case (V) seems trickier than the other cases. The marginal survival function $S(t)$ can be expressed in terms of both the baseline $S_0(t)$ and the distribution of the frailty risk R. This relationship is determined through the moment generating function (MGF) of R evaluated at the argument $\log(S_0(t))$. Thus, the marginal survival function is given by

$$
S(t) \quad = \quad \mathbb{E}_R\left[S_0(t)^R\right] = \mathbb{E}_R\left[\mathrm{e}^{R\log(S_0(t))}\right] = MGF_R\left(\log\left(S_0(t)\right)\right).
$$

As long as the integral converges, this form applies to many multiplicative frailty models. Recall that if $P(R \geq 0) = 1$, the MGF coincides with the Laplace transform of the random variable R, evaluated at minus the argument.

Again, we structure the binary frailty model that has $R \in \{0,1\}$ as a binary multiplicative frailty model, since under proportional hazards assumption the two hazard functions $\lambda_0(t) = \lambda(t \mid R=0)$ and $\lambda_1(t) = \lambda(t \mid R=1)$ are such that $\lambda_1(t) = \alpha\,\lambda_0(t)$ for the constant $\alpha = \lambda_1(t)/\lambda_0(t)$ for any

t. As a consequence, the survival time T has the two conditional survival distributions $S_0(t) = P(T \leq t \mid R = 0)$ and $S_1(t) = P(T \leq t \mid R = 1)$, and its distribution can be described as a multiplicative frailty model with frailty random variable R such that $R = 0$ w.p. $P(R = 0) = 1 - h$ and $R = 1 \iff \alpha = \lambda_1(t)/\lambda_0(t)$ w.p. $P(R = 1) = h$. For such a random variable the MGF is

$$MGF_R(u) = \mathbb{E}_R(e^{ur}) = \mathrm{e}^u(1-h) + \mathrm{e}^{u\alpha}h = \mathrm{e}^u + \left(\mathrm{e}^{u\alpha} - \mathrm{e}^u\right)h = e^u(1-h) + e^{u\alpha}$$

and as a consequence the marginal survival distribution of T is equal to

$$S(t) = MGF_R\left(\log(S_0(t))\right) = \mathrm{e}^{\log(S_0(t))}(1-h) + \mathrm{e}^{\log(S_0(t))\alpha} = S_0(t)(1-h) + S_1(t)h.$$

where, recall, the two survival functions follow a cure-rate structure, such that $S_0(t) = p_0 + (1 - p_0)\widetilde{S}_0(t)$ and $S_1(t) = p_1 + (1 - p_1)\widetilde{S}_1(t)$. At this point we obtain the complete form of the marginal survival distribution of T, without specifying the baseline survival function that can be fixed later. The model, which may be seen as a double mixture of survival functions, is not identifiable unless some constraint is set.

Let us ignore the presence of administrative right censoring, and therefore assume that all censored observations are all (and the only) "non-cases." This is a case in which more information is available on the model parameters, since additional right censoring would reduce the information available on the "cases." The generic contribution l_i to the likelihood function by subject i with observed data (x_i, δ_i) is

$$L_i(\underline{\theta}; (x_i, \delta_i)) = \left[(1-h)(1-p_0)\widetilde{f}_0(x_i) + h(1-p_1)\widetilde{f}_1(x_i)\right]^{\delta_i} \left[(1-h)\,p_0 + h\,p_1\right]^{1-\delta_i},$$

so that the likelihood function is equal to

$$\begin{aligned}
L(\underline{\theta}; (\underline{x}, \underline{\delta})) &= \prod_{i=1}^{n} L_i(\underline{\theta}; (x_i, \delta_i)) \\
&= \prod_{i \in cases} \left[(1-h)(1-p_0)\widetilde{f}_0(x_i) + h(1-p_1)\widetilde{f}_1(x_i)\right] \prod_{i \in non-cases} \left[(1-h)\,p_0 + h\,p_1\right] \\
&= \left\{ \prod_{i \in cases} \left[(1-h)(1-p_0)\widetilde{f}_0(x_i) + h(1-p_1)\widetilde{f}_1(x_i)\right] \right\} \left[(1-h)\,p_0 + h\,p_1\right]^{n_\infty},
\end{aligned}$$

with n_∞ the number of non-cases in the data (and $n - n_\infty$ the number of cases).

Now, let $\beta_1 = (1-h)(1-p_0)$; $\beta_2 = h(1-p_1)$, and $\beta_3 = (1-h)\,p_0 + h\,p_1$. The likelihood function can be re-written as

$$L(\underline{\theta}; (\underline{x}, \underline{\delta})) = \left\{ \prod_{i \in Cases} \left[\beta_1 \widetilde{f}_0(x_i) + \beta_2 \widetilde{f}_1(x_i)\right] \right\} \beta_3^{n_\infty},$$

where one can easily check that $\beta_1 + \beta_2 + \beta_3 = 1$, with all three terms positive.

The proportion n_∞/n of non-cases can estimate non parametrically the parameter β_3, and from it the quantity $1-\beta_3 = \beta_1+\beta_2$. As a consequence, the term $\beta_1+\beta_2$ is identified. If one then multiplies and divides the likelihood by the term $(\beta_1 + \beta_2)^{n-n_\infty}$, it seems clear that the quantity $\beta_1/(\beta_1 + \beta_2)$ (and thus also the quantity $\beta_2/(\beta_1 + \beta_2)$) is also identified from the mixture terms in the curly

bracket, which is based on the cases, together with the parameters θ_0 and θ_1 of the two density functions $\widetilde{f}_0$ and $\widetilde{f}_1$.

Therefore, the two parameters θ_0 and θ_1, as well as the two quantities β_1 and β_2, are identified. On the other hand, in general the individual parameters h, p_0, p_1 are not identified from the observed data.

Given the constraints $p_0 \in (0,1)$ and $p_1 \in (0,1)$, and the assumption that $p_0 > p_1$ (which is without loss of generality given the freedom of deciding which group is "0" and which is "1"), from knowledge of the values of β_1 and β_2 one may rule out some regions of $(0,1)$ as possible values for h. Indeed, since $(1-p_0) = \beta_1/(1-h)$ and $(1-p_1) = \beta_2/h$, and noting that $p_0 > p_1 \iff 1-p_0 < 1-p_1$, simple algebra shows that h must fall in the interval $[\beta_2, \beta_2/(\beta_1+\beta_2)]$. This in turn restricts the possible values that the pair (p_0, p_1) can take, since $p_0 = 1-\beta_1/(1-h)$ and $p_1 = 1-\beta_2/h$. □

As a consequence of this fact, one may try to place some constraints on the parameters to create identifiability. One example is the following restriction, associated with the hazard functions $\widetilde{\lambda}_0(t)$ and $\widetilde{\lambda}_1(t)$ for the two time-to-event distributions for the cases in the two groups:

$$\frac{\widetilde{\lambda}_1(t)}{\widetilde{\lambda}_0(t)} = \frac{p_0}{p_1} = \frac{1}{\alpha}, \tag{2.5}$$

thus imposing the PH structure on the distributions of the cases in the two groups, plus the assumption that the factor $\alpha \in (0,1)$ that relates $\widetilde{\lambda}_0(t) = \alpha\,\widetilde{\lambda}_1(t)$ is the same that relates p_0 to $p_1 = \alpha\, p_0$.

We call such model the Proportional Hazards Constrained Cure-Rate (PHCCR) model. Notably, in the PHCRR model the higher-risk group is associated with both a larger fraction of cases and earlier age at onset for their disease.

Note that to achieve identifiability one may also try to impose prior distributions on the parameters. Or, one may perform a sensitivity analysis that replaces this restriction with a fixed value for $p_1/p_0 = \rho$.

Example 1.

Let the two conditional distributions of the survival times of the cases be distributed as $\text{Exp}(\lambda_0)$ and $\text{Exp}(\lambda_1)$ respectively for the two risk groups, with $\lambda_1 > \lambda_0$, i.e. such that $\lambda_0 = \alpha\,\lambda_1$ with $\alpha \in (0,1)$. Note that we also have $p_1 = \alpha\, p_0$.

The following output illustrates the PHCCR model with two exponential CR survival sub-models. The simulations are based on 1,000 simulated dataset of size n=100,000 individuals each.

In Table 2.2 is reported the parameter recovery in mean and standard error of the estimated parameter values across the 1,000 repetitions. The square root of the mean square error $\sqrt{MSE}$ represents a measure of the absolute distance between the true value from the data generating process and the estimated value from observed data. The MSE is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{r}_i - r_i)^2.$$

	p_0	l_0	α	h
True value	0.8	0.1	0.3333	0.8
Mean	0.7995	0.1001	0.3335	0.7994
Standard error	0.0005	0.0002	0.0004	0.0002
$\sqrt{MSE}$	0.0220	0.0051	0.0152	0.0059
95% C.I. Lower	0.7981	0.0998	0.3326	0.7991
95% C.I. Upper	0.8008	0.1004	0.3345	0.7998

Table 2.1: Results on the identifiability of a two risk groups Cure-Rate model with Exponential survival function.

Also, the lower-bound and the upper-bound of the 95% confidence interval (C.I.) are reported.

Example 2.

Let the two conditional distributions of the survival times of the cases be Weibull($shape_0$, $scale_0$) and Weibull($shape_1, scale_1$) respectively for the two risk groups, with $shape_0 = shape_1$. To implement the conditional proportional hazards model $\widetilde{\lambda}_0(t) = \alpha\,\widetilde{\lambda}_1(t)$ one simply sets $scale_1 = scale_0\ (\alpha^{1/shape_0})$. Again, $p_1 = \alpha\, p_0$ (easy to check).

These two small examples confirm that the parameter values that are used to generate the data are recovered correctly by the maximum likelihood estimators, with only small residual biases for the estimators.

	$shape_0$	$scale_0$	$shape_1$	α	h
True value	20	65	20	0.70	0.80
Mean	20.2413	64.9588	20.1350	0.7006	0.8000
Se	0.0360	0.0088	0.0185	0.0001	0.0004
$\sqrt{MSE}$	1.1626	0.2814	0.6014	0.0032	0.0131
95% C.I. Lower	20.0184	64.9042	20.0201	0.7000	0.7974
95% C.I. Upper	20.4642	65.0133	20.2498	0.7013	0.8025

Table 2.2: Results on the identifiability of a two risk groups Cure-Rate model with Weibull survival functions.

2.2.4 The Univariate frailty Lehmann Cure-Rate model

As an alternative to the PHCCR model, we now extend the definition of the Lehmann family of distributions to the case of cure-rate models, still within the latent class framework.

Recall the Lehmann family of distributions is equivalent to the definition of the proportional hazards (PH) structure for proper survival distributions:

$$\left\{S_\alpha(t) = [S_0(t)]^\alpha,\ \ \alpha > 0\right\},$$

where $S_0(t)$ is the (proper) baseline survival function and the parameter α modifies it to become the (also proper) survival function $S_\alpha(t)$. Let all random variables in the family be absolutely continuous random variables. It is then easy to check that $\lambda_\alpha(t) = \alpha\,\lambda_0(t)$ for any choice of (positive and finite) α. Indeed, when α is modelled through a regression structure one has the celebrated semiparametric Cox proportional hazards (PH) survival model [3].

Here, we suggest extending the PH model to the model defined by the more general Lehmann cure-rate family obtained by applying the Lehmann power transformation to a baseline cure-rate model:

$$\left\{S_\alpha(t) = \left[p_0 + (1-p_0)\widetilde{S}_0(t)\right]^\alpha,\ \ \alpha > 0\right\}.$$

For a fixed value α, the survival function $S_\alpha(t)$ also defines a cure-rate model. Indeed, $\lim_{t\to\infty} S_\alpha(t) = p_0^\alpha$, and $S_\alpha(t)$ can be written as

$$S_\alpha(t) = p_0^\alpha + \left(1 - p_0^\alpha\right)\widetilde{S}_\alpha(t),$$

with conditional (proper) survival function for the cases equal to

$$\widetilde{S}_\alpha(t) = \frac{\left[p_0 + (1-p_0)\widetilde{S}_0(t)\right]^\alpha - p_0^\alpha}{1 - p_0^\alpha},$$

whose conditional density function is

$$\widetilde{f}_\alpha(t) = -\frac{d}{dt}\widetilde{S}_\alpha(t) = \frac{1-p_0}{1-p_0^\alpha}\alpha\left[p_0 + (1-p_0)\widetilde{S}_0(t)\right]^{(\alpha-1)}\widetilde{f}_0(t).$$

We note that here, too, a regression model with $\alpha = \alpha(z)$ can also be constructed for a vector z of observed covariates if they are available.

A two (or indeed more) latent class parametric Lehmann Cure-Rate model can now be easily defined. Recall the Lehmann structure on the survival function characterizing the risk group $S_r(t) = [S_0(t)]^{\alpha(r)}$, such that we have:

$$S_0(t) = p_0 + (1-p_0)\widetilde{S}_0(t), \qquad\qquad S_1(t) = [p_0 + (1-p_0)\widetilde{S}_0(t)]^\alpha.$$

For a fixed α, also the high-risk survival function $S_1(t)$ defines a cure-rate model. It is easy to check that $\lim_{t\to\infty} S_1(t) = p_0^\alpha = p_1$, and that $S_1(t)$ can be written as

$$S_1(t) = p_1 + (1-p_1)\,\widetilde{S}_1(t),$$

with conditional (proper) survival function for the cases equal to:

$$\widetilde{S}_1(t) = \frac{\left[p_0 + (1-p_0)\widetilde{S}_0(t)\right]^\alpha - p_0^\alpha}{1 - p_0^\alpha},$$

and conditional density function

$$\widetilde{f}_1(t) = -\frac{d}{dt}\widetilde{S}_1(t) = \frac{1-p_0}{1-p_0^\alpha}\alpha\left[p_0 + (1-p_0)\widetilde{S}_0(t)\right]^{(\alpha-1)}\widetilde{f}_0(t).$$

Since the cure-rate survival function is not proper, the density function associated with the cure-rate model is also not proper. Note that, without loss of generality, for $\alpha > 1$ one has $S_1(t) < S_0(t)\ \forall t > 0$ and $p_1 < p_0$. Indeed, we may reparametrize $\alpha_1 = 1/\alpha \in (0,1)$ to impose $\alpha > 1$.

For example, if one assumes a survival function distributed according to the Exponential distribution $\widetilde{S}_0(t) = \mathrm{e}^{-\lambda_0 t}$ for the distribution of $(T \mid R = 0, case)$, then

$$\widetilde{S}_1(t) = \frac{\left[p_0 + (1-p_0)\mathrm{e}^{-\lambda_0 t}\right]^\alpha - p_0^\alpha}{1 - p_0^\alpha},$$

and

$$\widetilde{f}_1(t) = \frac{\alpha(1-p_0)\lambda_0}{1-p_0^\alpha}\left[p_0 + (1-p_0)\mathrm{e}^{-\lambda_0 t}\right]^{\alpha-1}\mathrm{e}^{-\lambda_0 t}.$$

Interesting comments about the two-latent-class Lehmann cure-rate model are in Appendix B.1.

The Univariate likelihood

Recall that in the univariate setting only one subject per family contributes to the likelihood. Hence, the observed data univariate likelihood on the parameter collection $\underline{\theta} = \{p_0, \underline{\lambda}^T, \alpha, h\}^T$ is given by

$$L_u(\underline{\theta}; \text{subject data}) = \prod_{i=1}^{n}\left[f_X(\underline{x}_i|R_i = 0)(1-h) + f_X(\underline{x}_i|R_i = 1)\,h\right],$$

where subscript "u" stays for univariate likelihood. One has just $\underline{x}_i = (x_i, \delta_i)^T$ for $i = 1, \ldots, n$, with

$$f(\underline{x}|R=1) = f_1(x)^\delta S_1(x)^{(1-\delta)} = \left[(1-p_1)\widetilde{f}_1(x)\right]^\delta\left[p_1 + (1-p_1)\widetilde{S}_1(x)\right]^{1-\delta}$$

The goal of parameter estimation is to determine the risk difference α between the low-risk and high-risk groups, along with the other parameters. The extended likelihood function can be derived. Notice that the survival function and density function for the low and high-risk groups

are given by

$$
\begin{aligned}
S_0(t) &= p_0 + (1-p_0)\widetilde{S}_0(t) \\
f_0(t) &= (1-p_0)\widetilde{f}_0(t) \\
S_1(t) &= [S_0(t)]^{\alpha} = [p_0 + (1-p_0)\widetilde{S}_0(t)]^{\alpha} = p_1 + (1-p_1)\widetilde{S}_1(t) \\
\widetilde{S}_1(t) &= \frac{(p_0 + (1-p_0)\widetilde{S}_0(t))^{\alpha} - p_1}{1-p_1} \\
f_1(t) &= (1-p_1)\widetilde{f}_1(t) \\
\widetilde{f}_1(t) &= \left(\frac{1-p_0}{1-p_1}\right)\alpha\widetilde{f}_0(t)\left(p_0 + (1-p_0)\widetilde{S}_0(t)\right)^{\alpha-1} \\
&\text{with, } p_1 = p^{\alpha}.
\end{aligned}
$$

Thus, the likelihood is given by

$$
\begin{aligned}
L_u(\underline{\theta}; \text{subject data}) = \prod_{i=1}^{n} f_X(\underline{x}_i; \underline{\theta}) &= \prod_{i=1}^{n} [f_X(\underline{x}_i \mid R_i = 0; \theta)P(R_i = 0) \\
&+ f_X(\underline{x}_i \mid R_i = 1; \underline{\theta})P(R_i = 1)] \\
&= \prod_{i=1}^{n} \left[f_0(x_i)^{\delta_i} S_0(x_i)^{1-\delta_i}\right](1-h) + \left[f_1(x_i)^{\delta_i} S_1(x_i)^{1-\delta_i}\right] h.
\end{aligned} \tag{2.6}
$$

We move now from the univariate to the multivariate setting by computing the multivariate likelihood after a description of the data which contributes to the likelihood of the model.

2.2.5 The Multivariate frailty Lehmann Cure-Rate model

Family data

Consider the family cluster formed by main subject, sister, mother, and grandmother.

Figure 2.3 shows a depiction of the calendar times of birth (b) and of the times to onset (t) for a group of four family members. Notice that in the figure all family members experience the breast cancer onset, so that the cure-rate structure is not considered here. However, recall that the Cure-Rate model also allows for one or more of the times t, ts, tm, or tg to be equal to $+\infty$.

The data generating process produces the family time-to-event data

$$(B, Bg, Bm, Bs, T, Tg, Tm, Ts)^T,$$

for families indexed by $i = 1, \ldots, n$. We observe a realization of the multivariate random variable $(B, Bg, Bm, Bs, \underline{X}, \underline{X}g, \underline{X}m, \underline{X}s)^T$, where $\underline{X} = (min(T, C), \Delta)^T$, i.e. we observe the value $\underline{X} = \underline{x} = (x, \delta)^T$. The notation for the other family members is obtained by having x, t, c, b be followed by g, m, and s (meaning respectively, "granmother", "mother", and "sister"). The distinction between grandmother, mother and sister is not strictly needed here, it will make the extension to a more complex model easier. One may, for example, specify a relative-specific survival function to capture the generational differences among family members.

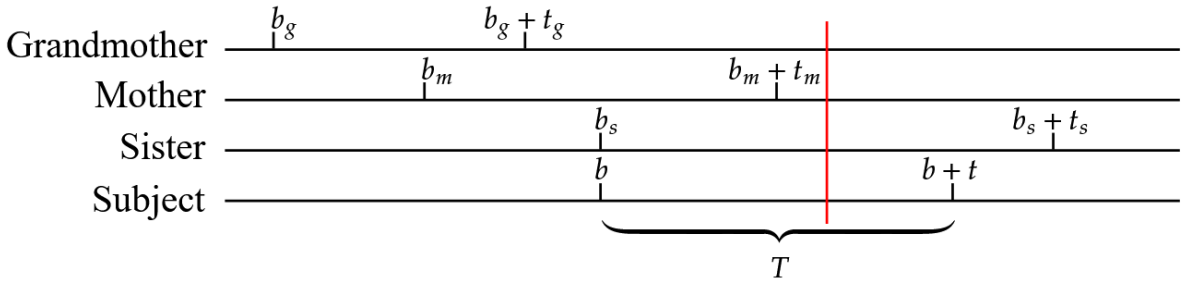

Figure 2.3: birth calendar times and times to disease onset for a family.

The Multivariate likelihood

The assumptions we rely on to deal with the multivariate aspect of this model are the conditional independence among survival times within the same family and the assumption of shared frailty risk within a family which allows the time-to-events to be i.i.d..

In this model, the observed data likelihood function incorporates common family memberships by grouping their contributions to the likelihood within each risk group. A deepening about the observed data likelihood is in Appendix B.2. Indeed, let $\underline{\theta} = \{p_0, \underline{\lambda}^T, \alpha, h_0\}^T$ be the whole parameter vector of the model. The observed data likelihood is given by

$$L(\underline{\theta}; \text{all data}) = \prod_{i=1}^{n} \left[f_{\underline{\mathbf{X}}}(\underline{\mathbf{x}}_i | R_i = 0; \theta)(1-h) + f_{\underline{\mathbf{X}}}(\underline{\mathbf{x}}_i | R_i = 1; \theta)\, h \right],$$

where “all data” is composed by the observed time and the indicator of having observed the event, respectively $\underline{\mathbf{x}} = (\underline{x} = (x, \delta)^T, \underline{x}s = (xs, \delta s)^T, \underline{x}m = (xm, \delta m)^T, \underline{x}g = (xg, \delta g)^T)^T$. Here, α is the target parameter for inference because of its crucial meaning. Indeed, it is the risk difference between the low-risk and the high-risk group of developing breast cancer in the two-latent-class setting involved in the PH structure which leads to $\lambda_1(t) = \alpha\lambda_0(t)$.

Let us compute the closed form of the likelihood. For ease of notation we drop writing the baseline survival parameter collection $\underline{\lambda}$. The first component is obtained, under the assumption

of conditional independence of the survival times within each family, as

$$
\begin{aligned}
&f_{\underline{X}}(\underline{x}_i \mid R_i = 0) = [f_T(x_i \mid R_i = 0) S_C(x_i)]^{\delta_i} \, [S_T(x_i \mid R_i = 0) f_C(x_i)]^{1-\delta_i} \\
&\propto f_T(x_i \mid R_i = 0)^{\delta_i} S_T(x_i \mid R_i = 0)^{1-\delta_i} = \left[((1-p_0)\widetilde{f}_0(x_i))^{\delta_i} (p_0 + (1-p_0)\widetilde{S}_0(x_i))^{(1-\delta_i)} \right] \\
&f_{\underline{X}}(\underline{xs}_i \mid R_i = 0) = f_T(xs_i \mid R_i = 0)^{\delta s_i} S_T(xs_i \mid R_i = 0)^{1-\delta s_i} = \\
&= \left[((1-p_0)\widetilde{f}_0(xs_i))^{\delta s_i} (p_0 + (1-p_0)\widetilde{S}_0(xs_i))^{(1-\delta s_i)} \right] \\
&f_{\underline{X}}(\underline{xm}_i \mid R_i = 0) = f_T(xm_i \mid R_i = 0)^{\delta m_i} S_T(xm_i \mid R_i = 0)^{1-\delta m_i} = \\
&= \left[((1-p_0)\widetilde{f}_0(xm_i))^{\delta m_i} (p_0 + (1-p_0)\widetilde{S}_0(xm_i))^{(1-\delta m_i)} \right] \\
&f_{\underline{X}}(\underline{xg}_i \mid R_i = 0) = f_T(xg_i \mid R_i = 0)^{\delta g_i} S_T(xg_i \mid R_i = 0)^{1-\delta g_i} = \\
&= \left[((1-p_0)\widetilde{f}_0(xg_i))^{\delta g_i} (p_0 + (1-p_0)\widetilde{S}_0(xg_i))^{(1-\delta g_i)} \right] \\
&f_{\underline{\mathbf{X}}}(\underline{\mathbf{x}}_i \mid R_i = 0; \theta) \overset{\perp | R}{=} f_{\underline{X}}(\underline{x}_i \mid R_i = 0) f_{\underline{X}}(\underline{xs}_i \mid R_i = 0) f_{\underline{X}}(\underline{xm}_i \mid R_i = 0) f_{\underline{X}}(\underline{xg}_i \mid R_i = 0) \\
&\qquad = \left[((1-p_0)\widetilde{f}_0(x_i))^{\delta_i} (p_0 + (1-p_0)\widetilde{S}_0(x_i))^{(1-\delta_i)} \right] \cdot \\
&\qquad \cdot \left[((1-p_0)\widetilde{f}_0(xs_i))^{\delta s_i} (p_0 + (1-p_0)\widetilde{S}_0(xs_i))^{(1-\delta s_i)} \right] \cdot \\
&\qquad \cdot \left[((1-p_0)\widetilde{f}_0(xm_i))^{\delta m_i} (p_0 + (1-p_0)\widetilde{S}_0(xm_i))^{(1-\delta m_i)} \right] \cdot \\
&\qquad \cdot \left[((1-p_0)\widetilde{f}_0(xg_i))^{\delta g_i} (p_0 + (1-p_0)\widetilde{S}_0(xg_i))^{(1-\delta g_i)} \right].
\end{aligned}
$$

Similarly for the second component, given the Lehmann survival function and density function for the high-risk group

$$
\begin{aligned}
S_1(t) &= \left[p_0 + (1-p_0)\widetilde{S}_0(t) \right]^{\alpha} = p_0^{\alpha} + (1-p_0^{\alpha})\widetilde{S}_1(t), \\
\widetilde{S}_1(t) &= \frac{\left[p_0 + (1-p_0)\widetilde{S}_0(t) \right]^{\alpha} - p_0^{\alpha}}{1-p_0^{\alpha}}, \\
f_1(t) &= (1-p_0^{\alpha})\widetilde{f}_1(t), \\
\widetilde{f}_1(t) &= \frac{\alpha(1-p_0)}{1-p_0^{\alpha}} \left[p_0 + (1-p_0)\widetilde{S}_0(t) \right]^{\alpha-1} \widetilde{f}_0(t),
\end{aligned}
$$

we have

$$
\begin{aligned}
&f_{\underline{X}}(\underline{x}_i \mid R_i = 1) = [f_T(x_i \mid R_i = 1)S_C(x_i)]^{\delta_i}\,[S_T(x_i \mid R_i = 1)f_C(x_i)]^{1-\delta_i} \\
&\propto f_T(x_i \mid R_i = 1)^{\delta_i} S_T(x_i \mid R_i = 1)^{1-\delta_i} = f_1(x_i)^{\delta_i} S_1(x_i)^{1-\delta_i} \\
&= \left[\frac{r\widetilde{f}_0(x_i)}{(1-p_0)^{-1}}\left(p_0 + (1-p_0)\widetilde{S}_0(x_i)\right)^{r-1}\right]^{\delta_i} \cdot \left[p_1 + (1-p_1)\widetilde{S}_1(x_i)\right]^{1-\delta_i} \\
&f_{\underline{X}}(\underline{xs}_i \mid R_i = 1) = f_1(xs_i)^{\delta s_i} S_1(xs_i)^{1-\delta s_i} = \\
&= \left[\frac{r\widetilde{f}_0(xs_i)}{(1-p_0)^{-1}}\left(p_0 + (1-p_0)\widetilde{S}_0(xs_i)\right)^{r-1}\right]^{\delta s_i} \left[p_1 + (1-p_1)\widetilde{S}_1(xs_i)\right]^{1-\delta s_i} \\
&f_{\underline{X}}(\underline{xm}_i \mid R_i = 1) = f_1(xm_i)^{\delta m_i} S_1(xm_i)^{1-\delta m_i} = \\
&= \left[\frac{r\widetilde{f}_0(xm_i)}{(1-p_0)^{-1}}\left(p_0 + (1-p_0)\widetilde{S}_0(xm_i)\right)^{r-1}\right]^{\delta m_i} \left[p_1 + (1-p_1)\widetilde{S}_1(xm_i)\right]^{1-\delta m_i} \\
&f_{\underline{X}}(\underline{xg}_i \mid R_i = 1) = f_1(xg_i)^{\delta g_i} S_1(xg_i)^{1-\delta g_i} = \\
&= \left[\frac{r\widetilde{f}_0(xg_i)}{(1-p_0)^{-1}}\left(p_0 + (1-p_0)\widetilde{S}_0(xg_i)\right)^{r-1}\right]^{\delta g_i} \left[p_1 + (1-p_1)\widetilde{S}_1(xg_i)\right]^{1-\delta g_i} \\
&f_{\underline{\mathbf{X}}}(\underline{\mathbf{x}}_i \mid R_i = 1;\theta) \overset{\perp|R}{=} f_{\underline{X}}(\underline{x}_i \mid R_i = 1) f_{\underline{X}}(\underline{xs}_i \mid R_i = 1) f_{\underline{X}}(\underline{xm}_i \mid R_i = 1) f_{\underline{X}}(\underline{xg}_i \mid R_i = 1)
\end{aligned}
$$

For simplicity we can see the expression as composed by the quantity

$$
\begin{aligned}
f_{\underline{\mathbf{X}}}(\underline{\mathbf{x}}|R=1) &\overset{\perp|R}{=} f(\underline{x}|R=1) f(\underline{xs}|R=1) f(\underline{xm}|R=1) f(\underline{xg}|R=1) \\
&= \left[f_1(x)^{\delta} S_1(x)^{(1-\delta)}\right]\left[f_1(xs)^{\delta s} S_1(xs)^{(1-\delta s)}\right]\left[f_1(xm)^{\delta m} S_1(xm)^{(1-\delta m)}\right]\left[f_1(xg)^{\delta g} S_1(xg)^{(1-\delta g)}\right],
\end{aligned}
$$

and similarly for the other family members, and for the $R = 0$ terms.

The specific mathematical calculations for the most common baseline survival distributions, i.e., the Exponential and the Weibull distributions, are presented in the following lines.

Exponential case

The high-risk group survival function with a Lehman structure follows a cure-rate model as well as the low-risk survival function, that is: $S_0(t) = p + (1-p)\widetilde{S}(t)$, with $\widetilde{S}(t)$ a proper survival function that converges to zero over time. We can easily prove that $S_1(t)$ follows a cure-rate structure with a different fraction of the population that will never experience the event:

$$
\begin{aligned}
&S_1(t) = S_T(t \mid R = 1) = [S_T(t \mid R_i = 0)]^{\alpha} = [S_0(t)]^{\alpha} = [p + (1-p)\widetilde{S}(t)]^{\alpha} = \widetilde{p} + (1-\widetilde{p})\widetilde{S}_1(t), \\
&\widetilde{S}_1(t) = \frac{(p + (1-p)\widetilde{S}(t))^{\alpha} - \widetilde{p}}{1-\widetilde{p}} = \frac{(p + (1-p)\mathrm{e}^{-\lambda t})^{\alpha} - \widetilde{p}}{1-\widetilde{p}} \\
&\widetilde{p} = p^{\alpha},
\end{aligned}
$$

for the baseline survival function distributed according to an Exponential(λ) distribution.

In the data generating process, we will derive the formulas for generating the time-to-event based on the group membership of individuals. We begin by considering the most straightforward example, that is the Exponential distribution of time-to-event. Generating data for the low-risk group is as follows:

1. we sample a Bernoulli random variable with probability p of not experiencing the event. Those who will have a positive value, will be assigned the time-to-event value $t = +\infty$;

2. the other individuals will be assigned a time-to-event value obtained from the inverse survival function:

$$\widetilde{S}(t) = \mathrm{e}^{-\lambda t} = y \sim \mathrm{U}[0,1] \iff t = -\frac{1}{\lambda}\log(y).$$

Similarly, for the high-risk group, the data generation process is as follows:

1. we sample a Bernoulli random variable with probability $\widetilde{p}$ of not experiencing the event. Those who will have a positive value, will be assigned the time-to-event value $t = +\infty$;

2. the other individuals will be assigned a time-to-event value obtained from the inverse survival function:

$$\widetilde{S}_1(t) = \frac{(p + (1-p)\mathrm{e}^{-\lambda t})^{\alpha} - \widetilde{p}}{1-\widetilde{p}} = y \sim \mathrm{U}[0,1] \iff t = -\frac{1}{\lambda}\log\left(\frac{(y(1-\widetilde{p}) + \widetilde{p})^{1/\alpha} - p)}{1-p}\right)$$

Weibull case

Equivalently for the case the Exponential distribution in Appendix **??**, the low and high-risk group survival functions follow a cure-rate structure. This structure is defined as $S(t) = p + (1-p)\widetilde{S}(t)$, where $\widetilde{S}(t)$ represents a proper survival function and p denotes the cured fraction.

Additionally, there is a Lehman structure between the survival functions, given by $S_1(t) = [S_0(t)]^{\alpha}$. We assume that $S_0(t)$ follows a Weibull distribution with shape parameter k and scale parameter λ.

Now, let us outline the crucial formulas for this scenario:

$$S_0(t) = p + (1-p)\widetilde{S}(t), \text{ with } \widetilde{S}(t) = e^{-\left(\frac{t}{\lambda}\right)^k}, \text{ and } \widetilde{f}(t) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} e^{-\left(\frac{t}{\lambda}\right)^k}, \ \forall t \geq 0$$

$$S_1(t) = \widetilde{p} + (1-\widetilde{p})\widetilde{S}_1(t), \text{ with } \widetilde{S}_1(t) = \frac{(p + (1-p)\widetilde{S}(t))^{\alpha} - \widetilde{p}}{1-\widetilde{p}} \text{ and } \widetilde{p} = p^{\alpha}$$

$$\widetilde{f}_1(t) = \left(\frac{1-p}{1-\widetilde{p}}\right)\alpha\left(p + (1-p)e^{-\left(\frac{t}{\lambda}\right)^k}\right)^{\alpha-1} e^{-\left(\frac{t}{\lambda}\right)^k}\frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1}$$

In the low-risk group, the data generation process follows these steps:

- we sample a Bernoulli random variable with probability p of not experiencing the event. Those who will have a positive value, will be assigned the time-to-event value $t = +\infty$;
- the other individuals will be assigned a time-to-event value obtained from the inverse survival function:

$$\widetilde{S}(t) = \mathrm{e}^{-(t/\lambda)^k} = y \sim \mathrm{U}[0,1] \iff t = \lambda \log\left(\frac{1}{y}\right)^{1/k}$$

Similarly, the generation for the high-risk group is given by the following steps:

- we sample a Bernoulli random variable with probability $\widetilde{p}$ of not experiencing the event. Those who will have a positive value, will be assigned the time-to-event value $t = +\infty$;
- the other individuals will be assigned a time-to-event value obtained from the inverse survival function:

$$\widetilde{S}_1(t) = \frac{(p + (1-p)\mathrm{e}^{-\left(\frac{t}{\lambda}\right)^k})^\alpha - \widetilde{p}}{1-\widetilde{p}} = y \sim \mathrm{U}[0,1] \iff t = -\lambda \left[\log\left(\frac{(y(1-\widetilde{p}) + \widetilde{p})^{1/\alpha} - p}{1-p}\right)\right]^{1/k}.$$

However, when the distribution of $S_0(t)$ is not as trivial as the, typically not useful, Exponential or Weibull distribution, the generation of observations from the high-risk group requires the inversion of the survival function $S_1(t)$. This may not be easy to do analytically, and typically requires numerical integration from the conditional density function of the cases in the high-risk group $f_1(t)$. The numerical inversion would then be needed to produce samples from the distribution. However, a much faster (and more precise) algorithm can be obtained by recalling the form of the marginal survival function for the high-risk group. Indeed, one may generate values u from the $U(0,1)$ distribution and invert $S_1(t)$ directly to produce the value $t = S_1^{-1}(u)$. Given the nature of the random variable T, however, one should produce the value $T = 1$ whenever $u < p_1 = p^\alpha$, and solve $u = S_1(t) = [S_0(t)]^\alpha$ for t. It is easy to check that

$$t = \widehat{S}_0^{-1}\left(\frac{u^{1/\alpha} - p}{1-p}\right),$$

which can be easily computed from the quantile function, available in most software packages for a large number of distributions (one just needs to make sure that the quantile function is never invoked for $u < p^\alpha$). As an example, the data generation from the high-risk group of a Lehmann Cure-Rate model based on the Weibull distribution for the times to onset for the cases in the high-risk group may be implemented as

- $T = +\infty$ if $u < p + \alpha$,
- $T = S_1^{-1}\left(\min\left(0.9, \frac{1-u^{1/\alpha}}{1-p}\right)\right)$ if $u > p^\alpha$.

We have temporarily implemented this latest version of the data generation process to address the computational challenges we are currently facing.

Example 3.

An example of implementation of the Multivariate frailty Lehmann Cure-Rate model based on a Weibull baseline distribution for the cases in the low-risk group produces the output in Table 2.3 (recall that we reparametrize $\alpha_1 = 1/\alpha \in (0,1)$ to force $\alpha > 1$). Parameters value are perfectly recovered by the multivariate likelihood estimation process.

	p_0	$shape_0$	$scale_0$	α_1	h
True value	0.8	10	70	0.4	0.2
Mean	0.7902	10.1515	69.8595	0.3678	0.1373
Se	0.0002	0.0004	0.0004	0.0004	0.0006
$\sqrt{MSE}$	0.0111	0.1521	0.1415	0.0370	0.0700
95% C.I. Lower	0.7899	10.1507	69.8584	0.3667	0.1354
95% C.I. Upper	0.7905	10.1523	69.8605	0.3689	0.1393

Table 2.3: Parameter recovery for the Multivariate frailty Lehmann Cure-Rate model.

Now that we outlined all of the components and the likelihood of the Univariate FH Lehmann Cure-Rate model, the Univariate Lehmann frailty Cure-Rate, and the Multivariate Lehmann frailty Cure-Rate model we can move to a comparative analysis through a simulation study.

2.3 Comparison of univariate vs. multivariate estimation

This section is outlined as follows: a description of a fast algorithm for sampling data is reported. Right after that, parameters are estimated through the maximization of the univariate and multivariate likelihoods. The numerical method of Nelder-Mead is chosen to achieve likelihood maximization. With the parameter estimation we want to internally validate the models. Once one has the estimated parameters, the models can be compared in terms of risk classification. The risk can be predicted both for for each woman that constitutes the available dataset and for a new woman whose family is not part of the available data.

We expect that fitting the multivariate likelihood allows for: (i) more accurate estimation of the model parameters; (ii) exploring of the dependence structure within families (goodness of fit); and (iii) more accurate risk prediction.

2.3.1 Data generation from the two-latent-class Lehmann Cure-Rate model

A fast algorithm for data generation

In simpler models (ex. the Exponential and Weibull seen above), the generation of observations from the high-risk group can be easily achieved by applying the analytical inversion of $\widetilde{S}_1(t)$ (for

cases), after having generated the case/non-case status.

Splitting subjects into two risk groups, we have the closed form of the survival function on the observed time for cases into the low-risk and high-risk group. For the Exponential baseline survival function $\widetilde{S}_0(t) = \mathrm{e}^{-\lambda t}$, and the Weibull survival function $\widetilde{S}_0(t) = \mathrm{e}^{-(t/\lambda)^k}$, with scale λ and shape k, inverting the survival function brings to generating the time-to-events from, respectively,

$$t = -\frac{1}{\lambda}\log(u), \qquad\qquad t = \lambda \log\left(\frac{1}{u}\right)^{1/k}.$$

where $u \sim \mathrm{Unif}[0,1]$. Similarly, we obtain for an observation from $\widetilde{S}_1(t)$, where the time-to-event generation formula for Exponential and Weibull distribution is given by, respectively,

$$t = -\frac{1}{\lambda}\log\left(\frac{\left[(1-p_0^\alpha)u + p_0^\alpha\right]^{1/\alpha} - p_0}{1-p_0}\right), \qquad t = -\lambda\left[\log\left(\frac{\left[(1-p_0^\alpha)u + p_0^\alpha\right]^{1/\alpha} - p_0}{1-p_0}\right)\right]^{1/k}.$$

While this model is clearly appealing in its interpretation, it is difficult to identify its parameters. Thus, fitting of such frailty model requires that one has additional external information on the value of some of the parameters.

When the distribution $\widetilde{S}_0(t)$ is not as trivial as the – typically not very useful – Exponential distribution, generation of observations from the high-risk group requires inversion of the survival function $\widetilde{S}_1(t)$ through numerical integration from $\widetilde{f}_1(t)$, the conditional density function of the cases in the high-risk group. This allows to produce samples from the distribution.

However, a much faster and precise algorithm exists from the form of the marginal survival function for the high-risk group. One may generate values u from the $U(0,1)$ distribution and invert $S_1(t) = [S_0(t)]^\alpha$ directly to produce the value $t = S_1^{-1}(u)$. One should produce the value $T = +\infty$ whenever $u < p_1 = p_0^\alpha$, and solve $u = S_1(t)$ for t when $u \geq p_1$. It is easy to check that this yields

$$t = \widetilde{S}_0^{-1}\left(\frac{u-p_0}{1-p_0}\right) = \widetilde{F}_0^{-1}\left(\frac{1-u}{1-p_0}\right), \text{ and } t = \widetilde{S}_0^{-1}\left(\frac{u^{1/\alpha}-p_0}{1-p_0}\right) = \widetilde{F}_0^{-1}\left(\frac{1-u^{1/\alpha}}{1-p_0}\right),$$

respectively for the low-risk and high-risk groups. These can be easily computed from the quantile function available in most software packages for a large number of distributions (one just needs to make sure that the quantile function is never invoked for $u < p_0^\alpha$).

Data generation

According to the specific structure of the model, it is necessary to generate data separately for cases and non-cases belonging to the two risk groups. Recall that cases refer to subjects who have a non-zero probability of developing disease onset, while non-cases are individuals who will never develop disease onset, regardless of their lifespan. Notice that non-cases represent the cured fraction of the population.

Families, composed by a (main) subject, her sister, her mother, and her grandmother, are generated with uniformly distributed birth calendar times, with uniformly distributed distance between

grandmothers, mother, and daughter (the sister of the main subject) between 25 and 35 years:

$$
\begin{aligned}
Bg &\sim Unif(min = 1880, max = 1910),\\
Bm &= Bg + Unif(min = 25, max = 35),\\
Bs &= Bm + Unif(min = 25, max = 35),\\
Bval &= Bm + Unif(min = 25, max = 35),
\end{aligned}
$$

so that births are as late as 2000.

Data are right-censored by the end of follow up or a the event of death, whose generation is given by

$$
\begin{aligned}
Deathg &= Bg + Unif(min = 60, max = 105),\\
Deathm &= Bm + Unif(min = 60, max = 105),\\
Deaths &= Bs + Unif(min = 60, max = 105),\\
Death &= Bval + Unif(min = 60, max = 105).
\end{aligned}
$$

We set the end of the study at the year 2020, so that the censored observation for each subject is

$$
\begin{aligned}
Censg &= pmin(Deathg, 2020),\\
Censm &= pmin(Deathm, 2020),\\
Censs &= pmin(Deaths, 2020),\\
Cens &= pmin(Death, 2020).
\end{aligned}
$$

Most importantly, the times-to-event is the crucial point of the data generation process. The faster algorithm steps (which in part has been described above in Paragraph 2.3.1) are given by

1. Fixing a parametric distribution of the proper survival function $\widetilde{S}_0(t)$, so that $S_0(t) = p_0 + (1 - p_0)\widetilde{S}_0(t)$ and $f_0(t) = (1 - p_0)\widetilde{f}_0(t)$;
2. generating the time-to-event $t \sim \widetilde{f}_0(t)$ for low-risk cases;
3. generating the time-to-event $t = \widetilde{S}_0^{-1}\left(\frac{u^\alpha - p_0}{1 - p_0}\right)$, $u \sim \mathrm{U}[0,1]$ for high-risk cases with the approximate method.

In the following scenario, we simulated $n = 100,000$ families, each consisting of 4 members, and repeated this simulation 100 times. Within the algorithm, a reparametrization process was implemented for the parameters to guarantee adherence to the non-negative constraint.

We aim to obtain the parameter value used in data generation by maximizing the likelihood. At each iteration, given n_p number of parameters, we fix the $(n_p - 1)$ parameters and vary the risk parameter α over a few values. The cured fraction and the proportion of high-risk families into the population are set at $p = 0.8$, and $h = 0.3$. Results are presented in Table 2.4 for

a baseline survival function distributed according to an Exponential($\lambda = 0.3$), Table 2.5 for the Weibull($shape = 10, scale = 70$) case, Tables 2.6 for the Gamma($shape = 10, scale = 2$) case, using the univariate vs. the multivariate likelihood of the frailty Lehmann Cure-Rate model, and the univariate likelihood of the FH Lehmann Cure-Rate model. Specifically, in the first column in reported the true value of the risk difference $\alpha = (1/2, 1/4, 1/5)$, fixed at the data generation step. All the other columns report the mean, standard error and mean square error ($\sqrt{\text{MSE}}$) associated to the estimated parameter values across the repeated simulations.

The slight discrepancies observed between the estimated and true values can be attributed to the approximation algorithm used for generating data. One way to overcome this issue is by increasing the sample size, particularly the average family size into the sample, as it would result in a more accurate estimate of the true value. Additionally, this would help in reducing the variance around the mean for sure. Nevertheless, on the contrary to the univariate models, the multivariate model is capable of accurately recovering the parameter values.

We notice also that the estimates of the model parameters for the univariate likelihood have standard errors that are much larger than those of the estimates based on the multivariate likelihood. We take as example the first scenario with the Weibull baseline survival function from Table 2.5. Specifically, their standard errors are between 1.75 and 12.18 times larger than their multivariate counterparts, as we can assess from Table 2.7

Such increase is noteworthy in particular because one may expect the effective sample size of the multivariate estimator to represent a four-fold increase from the univariate estimator, given the use of information from not one but four relatives for each family. Indeed, we illustrate such aspect for the univariate model in the following lines.

	True value of α	$\widehat{p}_0$	$\widehat{\alpha}$	$\widehat{\lambda}_0$	$\widehat{h}$
Multivariate					
Mean	0.5	0.8174	0.4918	0.0316	0.3651
Se		0.0168	0.0223	3e-04	0.13
$\sqrt{\text{MSE}}$		0.7842	0.309	0.4684	0.2101
Mean	0.25	0.8	0.2441	0.0315	0.212
Se		0.0016	0.0025	2e-04	0.0055
$\sqrt{\text{MSE}}$		0.7667	0.5559	0.2185	0.0132
Mean	0.2	0.7995	0.1947	0.0316	0.2078
Se		0.0014	0.0016	2e-04	0.0039
$\sqrt{\text{MSE}}$		0.0018	0.0186	6e-04	7e-04
Univariate					
Mean	0.5	0.8432	0.3516	0.0336	0.2225
Se		0.0875	2.1274	0.0021	0.2227
$\sqrt{\text{MSE}}$		0.0976	2.2886	0.0022	0.2542
Mean	0.25	0.8004	0.2501	0.0335	0.0997
Se		0.002	0.0214	7e-04	0.0016
$\sqrt{\text{MSE}}$		0.002	0.0214	7e-04	0.0017
Mean	0.2	0.8004	0.1999	0.0335	0.0999
Se		0.0017	0.0186	6e-04	7e-04
$\sqrt{\text{MSE}}$		0.0186	0.0018	6e-04	7e-04
Univariate FH			$\widehat{\beta}_F$		
Mean	0.5	0.7860	2.1491	0.0388	
Se		0.0021	0.0129	5e-04	
$\sqrt{\text{MSE}}$		0.0142	1.5347	0.0055	
Mean	0.25	0.828	0.9928	0.0655	
Se		0.002	0.0136	7e-04	
$\sqrt{\text{MSE}}$		0.0281	2.9928	0.0321	
Mean	0.2	0.8342	0.9314	0.0731	
Se		0.0018	0.0149	9e-04	
$\sqrt{\text{MSE}}$		0.0342	3.9264	0.0397	

Table 2.4: Parameter identifiability for α varying, with Exponential baseline survival function.

	True value of α	$\widehat{p}_0$	$\widehat{\alpha}$	$\widehat{shape}_0$	$\widehat{scale}_0$	$\widehat{h}$
Multivariate						
Mean	0.5	0.8134	0.5369	10.1756	69.8799	0.3452
Se		0.0003	0.0008	0.0012	0.0011	0.0017
$\sqrt{\text{MSE}}$		0.0256	0.1068	0.1907	0.1876	0.1745
Mean	0.25	0.8014	0.2479	10.0886	69.8911	0.2008
Se		0.0001	0.0001	0.0009	0.0011	0.0003
$\sqrt{\text{MSE}}$		0.0071	0.0150	0.1275	0.1575	0.0261
Mean	0.2	0.8015	0.1974	10.1489	69.9141	0.2031
Se		0.0001	0.0001	0.0007	0.0008	0.0002
$\sqrt{\text{MSE}}$		0.0074	0.0064	0.1662	0.1177	0.0156
Univariate						
Mean	0.5	0.8690	0.2230	10.2890	70.9132	0.3201
Se		0.0007	0.0019	0.0021	0.0134	0.0031
$\sqrt{\text{MSE}}$		0.1011	0.3334	0.3553	1.6226	0.3304
Mean	0.25	0.7494	0.5183	9.7783	69.8134	0.1557
Se		0.0002	0.0025	0.0025	0.0051	0.0008
$\sqrt{\text{MSE}}$		0.0541	0.3641	0.3337	0.5463	0.0884
Mean	0.2	0.7662	0.3012	9.9325	69.9320	0.1903
Se		0.0003	0.0015	0.0030	0.0036	0.0007
$\sqrt{\text{MSE}}$		0.0476	0.1812	0.3090	0.3655	0.0723
Univariate FH			$\widehat{\beta}_F$			
Mean	0.5	0.7895	0.4855	9.995	69.9747	
Se		0.0006	0.0028	0.0010	0.0023	
$\sqrt{\text{MSE}}$		0.0658	0.2855	0.1045	0.235	
Mean	0.25	0.7827	0.4218	9.8695	69.5787	
Se		0.0009	0.0026	0.0010	0.0033	
$\sqrt{\text{MSE}}$		0.0963	0.3083	0.1664	0.5342	
Mean	0.2	0.7895	0.3374	9.7897	69.3433	
Se		0.0012	0.0019	0.0010	0.0036	
$\sqrt{\text{MSE}}$		0.1169	0.2419	0.2347	0.7512	

Table 2.5: Parameter identifiability for α varying, with Weibull baseline survival function.

	True value of α	$\widehat{p}_0$	$\widehat{\alpha}$	$\widehat{shape}_0$	$\widehat{scale}_0$	$\widehat{h}$
Multivariate						
Mean	0.5	0.7875	0.5281	10.7928	1.8435	0.1333
Se		0.0166	0.2010	0.3618	0.0560	0.0900
$\sqrt{\text{MSE}}$		0.0208	0.2029	0.8715	0.1662	0.1121
Mean	0.25	0.8030	0.2400	10.5283	1.8873	0.1992
Se		0.0131	0.0189	0.4718	0.0952	0.0464
$\sqrt{\text{MSE}}$		0.0134	0.0214	0.7083	0.1475	0.0464
Mean	0.2	0.7977	0.1958	10.2873	1.9442	0.1967
Se		0.0087	0.0189	0.4019	0.0913	0.0185
$\sqrt{\text{MSE}}$		0.0090	0.0194	0.4940	0.1070	0.0188
Univariate						
Mean	0.5	0.7445	0.7568	10.3722	1.9144	0.0483
Se		0.0130	0.2419	0.6032	0.1254	0.1205
$\sqrt{\text{MSE}}$		0.0570	0.4877	0.7088	0.1518	0.1938
Mean	0.25	0.6973	0.5938	10.1955	1.8798	0.0739
Se		0.0396	0.2878	0.6771	0.1198	0.1000
$\sqrt{\text{MSE}}$		0.1101	0.5349	0.7048	0.1697	0.1610
Mean	0.2	0.7459	0.1870	10.6457	1.8082	0.1833
Se		0.1006	1.3575	0.1561	0.1845	0.2938
$\sqrt{\text{MSE}}$		0.1142	0.1787	1.5033	0.2662	0.2942
Univariate FH			$\widehat{\beta}_{FH}$			
Mean	0.5	0.9908	0.1124	52.5408	20.6369	
Se		<0.0001	0.0031	0.1846	0.0725	
$\sqrt{\text{MSE}}$		0.1908	0.4975	46.3738	19.9978	
Mean	0.25	0.9949	0.5069	29.1894	11.4650	
Se		0.0001	0.0052	0.3077	0.1209	
$\sqrt{\text{MSE}}$		0.1950	0.5798	36.2618	15.3504	
Mean	0.2	0.9939	0.4083	35.0272	13.7579	
Se		0.0001	0.0051	0.3015	0.1184	
$\sqrt{\text{MSE}}$		0.1939	0.5502	39.1814	16.6870	

Table 2.6: Parameter identifiability for α varying, with Gamma baseline survival function.

	$\widehat{p}_0$	$\widehat{\alpha}$	$\widehat{shape}_0$	$\widehat{scale}_0$	$\widehat{h}$
Multivariate Se	0.0003	0.0008	0.0012	0.0011	0.0017
Univariate Se	0.0007	0.0019	0.0021	0.0134	0.0031
Univariate Se / Multivariate Se	2.3333	2.3750	1.7500	12.1818	1.8235

Table 2.7: comparison between the SE from the Multivariate model vs. Univariate model.

A quick check of the effect of using a larger sample size in the univariate model

We compare the estimated standard deviation of the parameter estimators from the univariate likelihood for a sample of size $n = 1,000,000$ in Table 2.8 to that of the parameter estimators from the univariate likelihood for a sample of size $n = 500,000$ in Table 2.9. The study is repeated for 100 simulations.

Results from the Univariate estimation, with n=1, 000, 000 families, and thus subjects are collected in Table 2.8.

	p_0	$shape_0$	$scale_0$	α	h	AUC
True value	0.8	10	70	0.4	0.2	
Mean	0.8110	10.0309	70.1812	0.3813	0.2157	0.5725
Sd	0.0342	0.0530	0.2126	0.1685	0.1101	0.0007
$\sqrt{MSE}$	0.0360	0.0614	0.2794	0.1696	0.1113	
95% C.I. Lower	0.0360	0.0614	0.2794	0.1696	0.1113	
95% C.I. Upper	0.8131	10.0342	70.1944	0.3918	0.2225	

Table 2.8: parameter estimation for the univariate likelihood for a sample size of $n = 1,000,000$.

While, the univariate estimation with $n = 500,000$ families, and thus subjects is collected in Table 2.9.

	p_0	$shape_0$	$scale_0$	α	h	AUC
True value	0.8	10	70	0.4	0.2	
Mean	0.8122	10.0245	70.1663	0.4073	0.2245	0.5726
Sd	0.0494	0.0656	0.2788	0.2047	0.1616	0.0010
$\sqrt{MSE}$	0.0508	0.0700	0.3247	0.2049	0.1635	
95% C.I. Lower	0.80914	10.02039	70.14906	0.39465	0.21452	
95% C.I. Upper	0.81526	10.02853	70.18363	0.42003	0.23456	

Table 2.9: parameter estimation for the univariate likelihood for a sample size of $n = 500,000$.

The comparison between the standard errors is in Table 2.10.

As anticipated, the ratio of the estimated standard deviations of the estimators closely approximates the expected value of $\sqrt{2} = 1.1442$. This implies that when the sample size doubled, the standard deviation increased by approximately the square root of two. Hence, we can conclude that there is a proportional relationship between the sample size and the standard deviation, with the standard deviation increasing by a factor close to the square root of two when the sample size is doubled.

The comparison of the root MSE (RMSE) for the estimated parameters from the univariate vs. the multivariate likelihood yields, as we can assess from Table 2.11 showing that the RMSEs of the univariate estimators are larger than those of the multivariate estimators.

	p_0	$shape_0$	$scale_0$	α	h
Sd 1M	0.0342	0.0530	0.2126	0.1685	0.1101
Sd 500k	0.0494	0.0656	0.2788	0.2047	0.1616
Sd 1M / Sd 500k	0.6923077	0.8079268	0.7625538	0.8231558	0.6813119
Sd 500k / Sd 1M	1.444444	1.237736	1.311383	1.214837	1.467757

Table 2.10: comparison between standard errors for double sample size, where “1M” and “500k” mean n = 1,000,000 and n = 500,000 from Tables 2.8, and 2.9.

	$\widehat{p}_0$	$\widehat{\alpha}$	$\widehat{shape}_0$	$\widehat{scale}_0$	$\widehat{h}$
Multivariate RMSE	0.1011	0.3334	0.3553	1.6226	0.3304
Univariate RMSE	0.0256	0.1068	0.1907	0.1876	0.1745
Univariate RMSE / Multivariate RMSE	3.9492	3.1217	1.8631	8.6492	1.8934

Table 2.11: Comparison between the RMSE from the Multivariate model vs. Univariate model.

On the other hand, given the presence of bias in the estimators, such bias becomes more visible in the multivariate case, so that if one compares the standardized RMSE (SRMSE) obtained by dividing it by each estimator’s estimated standard deviation, obtains the results in Table 2.12 which shows that, but the scale baseline parameter, the relative RMSE is larger for the univariate estimators.

	$\widehat{p}_0$	$\widehat{\alpha}$	$\widehat{shape}_0$	$\widehat{scale}_0$	$\widehat{h}$
Multivariate SRMSE	85.3333	133.5000	158.9167	170.5454	102.6471
Univariate SRMSE	144.4286	175.4737	169.1905	121.0896	106.5806
Univariate / Multivariate SRMSE	1.6925	1.3144	1.0646	0.7100	1.0383

Table 2.12: Comparison between the standardized RMSE from the Multivariate model vs. Univariate model.

The increase in precision achieved by the estimators obtained from the multivariate likelihood is possibly due in part to the fact that the added relatives (grandmother and mother in particular), having been born earlier than the main subjects that appear in the univariate likelihood, are less likely to have their survival times be (administratively) censored.

However, the effect of the shared frailty risk component of the model is also possibly contributing to the increase in precision. Such effect is however not easy to quantify, as estimating the parameters of a model that does not include the shared frailty component would be such that either a different data generating model should be used, or a misspecified model is being used.

2.3.2 Risk group prediction for univariate vs. multivariate models

Our primary goal is to predict the risk for an woman whose family is not part of the data used to fit the model. Subsequently, risk prediction and related metrics are calculated and presented, followed by a final evaluation to determine the most informative and effective approach among the three studied. Notice that we can run risk prediction only for the frailty Lehmann Cure-Rate model because in the case of the *FH* Lehmann Cure-Rate model, the family history is used as the predicted risk.

The complete shape of the low-risk and high-risk survival functions are

$$\begin{aligned} S_0(t) &= \widehat{p}_0 + (1-\widehat{p}_0)\widetilde{S}(t) \\ S_1(t) &= \widehat{p}_0^{\widehat{\alpha}} + (1-\widehat{p}_0^{\widehat{\alpha}})\widetilde{S}_1(t) \\ \widetilde{S}_1(t) &= \frac{(\widehat{p}_0 + (1-\widehat{p}_0)\widetilde{S}_0(t))^{\widehat{\alpha}} - \widehat{p}_0^{\widehat{\alpha}}}{1-\widehat{p}_0^{\widehat{\alpha}}} \end{aligned}$$

where $\widetilde{S}_0(t)$ has a defined distribution, previously fixed.

Risk prediction is achieved through calculation of the conditional probability for each family:

$$P(R=1 \mid \underline{x}; \widehat{\theta}) = \frac{f(\underline{x} \mid R=1; \widehat{\theta})P(R=1)}{f(\underline{x})} = \frac{f(\underline{x} \mid R=1; \widehat{\theta})P(R=1)}{f(\underline{x} \mid R=1; \widehat{\theta})P(R=1) + f(\underline{x} \mid R=0; \widehat{\theta})P(R=0)}.$$

Recall from above that $\underline{x} = (x, \delta)$ indicate the univariate survival data couple. The vector collection $\underline{\mathbf{x}} = ((x,\delta)^T, (xs, \delta s_1)^T, (xm, \delta m)^T, (xg, \delta g)^T)^T$ represents the whole family data, always for a four members family. The purpose to define the family with at least three members, i.e. the grandmother, the mother, and the first sister is to cover at least the second-degree-generational heritability of breast cancer. Notice again that this is easily extendable to a higher number of sisters, or other degree members of the family (see e.g. father's mother, aunts, female cousins). Hence, the multivariate model is a generalization of the formula above, involving here all survival information from all family members and accounting for the conditional independence assumption. The multivariate posterior probability of belonging to the high-risk group is given by:

$$P(R=1 \mid \underline{\mathbf{x}}; \widehat{\theta}) = \frac{f(\underline{\mathbf{x}} \mid R=1; \widehat{\theta})P(R=1)}{f(\underline{\mathbf{x}} \mid R=1; \widehat{\theta})P(R=1) + f(\underline{\mathbf{x}} \mid R=0; \widehat{\theta})P(R=0)}. \tag{2.7}$$

Where, recall that for the two risk groups the familial density function is given by:

$$\begin{aligned} f(\underline{\mathbf{x}} \mid R=1) &\overset{\perp | R}{=} f(\underline{x} \mid R=1) f(\underline{xg} \mid R=1) f(\underline{xm} \mid R=1) f(\underline{xs} \mid R=1), \\ f(\underline{\mathbf{x}} \mid R=0) &\overset{\perp | R}{=} f(\underline{x} \mid R=0) f(\underline{xg} \mid R=0) f(\underline{xm} \mid R=0) f(\underline{xs} \mid R=0), \end{aligned}$$

where the univariate density function, split for the two risk groups, for the subject is

$$\begin{aligned} f(\underline{x} \mid R=1) &= f_1(x)^{\delta} S_1(x)^{(1-\delta)}, \\ f(\underline{x} \mid R=0) &= f_0(x)^{\delta} S_0(x)^{(1-\delta)}, \end{aligned}$$

with the following quantities of interest:

$$
\begin{aligned}
&S_0(x) = \widehat{p}_0 + (1-\widehat{p}_0)\widetilde{S}_0(x),\\
&f_0(x) = (1-\widehat{p}_0)\widetilde{f}_0(x),\\
&S_1(x) = [S_0(x)]^{\widehat{\alpha}} = [\widehat{p}_0 + (1-\widehat{p}_0)\widetilde{S}_0(x)]^{\widehat{\alpha}} = \widehat{p}_0^{\widehat{\alpha}} + (1-\widehat{p}_0^{\widehat{\alpha}})\widetilde{S}_1(x),\\
&f_1(x) = (1-\widehat{p}_0^{\widehat{\alpha}})\left(\frac{1-\widehat{p}_0}{1-\widehat{p}_0^{\widehat{\alpha}}}\right)\widehat{\alpha}\widetilde{f}_0(x)\left(\widehat{p}_0 + (1-\widehat{p}_0)\widetilde{S}_0(x)\right)^{\widehat{\alpha}-1},\\
&\text{with } \widetilde{S}_1(x) = \frac{(\widehat{p}_0 + (1-\widehat{p}_0)\widetilde{S}_0(x))^{\widehat{\alpha}} - \widehat{p}_0^{\widehat{\alpha}}}{1-\widehat{p}_0^{\widehat{\alpha}}}.
\end{aligned}
$$

Very useful is the probability of surviving within the next k years, for those women who have not already experienced the onset. The probability is estimated as:

$$S(x+k \mid \underline{\mathbf{x}}; \widehat{\theta}) = S(x+k \mid R=1; \widehat{\theta})P(R=1 \mid \underline{\mathbf{x}}; \widehat{\theta}) + S(x+k \mid R=0; \widehat{\theta})P(R=0 \mid \underline{\mathbf{x}}; \widehat{\theta}),$$

with

$$
\begin{aligned}
S(x+k \mid R=1; \widehat{\theta}) &= S_1(x+k; \widehat{\theta}),\\
S(x+k \mid R=0; \widehat{\theta}) &= S_0(x+k; \widehat{\theta}),
\end{aligned}
$$

where $P(R=1 \mid \underline{\mathbf{x}}; \widehat{\theta})$ is obtained in the previous step in Formula 2.7. Notice that for such women, who have not experienced the disease onset yet, the observed time x always corresponds to the censoring time $x = c$. Specifically, this is achieved for the univariate estimators through calculation, for each family, of the conditional probability

$$P(R_i = 1 \mid (x_i, \delta_i); \widehat{\theta}) = \frac{h\,\widetilde{f}_1(x_i)^{\delta_i}\,\widetilde{S}_1(x_i)^{1-\delta_i}}{h\,\widetilde{f}_1(x_i)^{\delta_i}\,\widetilde{S}_1(x_i)^{1-\delta_i} + (1-h)\,\widetilde{f}_0(x_i)^{\delta_i}\,\widetilde{S}_0(x_i)^{1-\delta_i}}, \tag{2.8}$$

where $\widehat{\theta}$ is the vector of the estimated model parameters, such that in the aforementioned Formula 2.8 each survival and density function are of the type $f(x) = f(x; \widehat{\theta})$.

The formula for the multivariate model is similar, but it involves the survival information from all family members, taking into account the conditional independence assumption within each family:

$$P(R_i = 1 \mid (\mathbf{x}_i, \boldsymbol{\delta}_i); \widehat{\theta}) = \frac{h\,\widetilde{f}_1(\mathbf{x}_i)^{\boldsymbol{\delta}_i}\,\widetilde{S}_1(\mathbf{x}_i)^{1-\boldsymbol{\delta}_i}}{h\,\widetilde{f}_1(\mathbf{x}_i)^{\boldsymbol{\delta}_i}\,\widetilde{S}_1(\mathbf{x}_i)^{1-\boldsymbol{\delta}_i} + (1-h)\,\widetilde{f}_0(\mathbf{x}_i)^{\boldsymbol{\delta}_i}\,\widetilde{S}_0(\mathbf{x}_i)^{1-\boldsymbol{\delta}_i}}.$$

To simplify the expressions, above we have used the notation

$$\widetilde{f}_r(\mathbf{x}_i)^{\boldsymbol{\delta}_i} = \widetilde{f}_r(xg_i)^{\delta g_i}\,\widetilde{f}_r(xm_i)^{\delta m_i}\,\widetilde{f}_r(xs_i)^{\delta s_i}\,\widetilde{f}_r(x_i)^{\delta_i}$$

for $r = 0, 1$. The notation for $\widetilde{S}_r(\mathbf{x}_i)^{1-\boldsymbol{\delta}_i}$ is analogous.

2.3.3 ROC and AUC: univariate vs. multivariate model

The overall performance of the risk group classifier as a function of the cutoff for assignment to the groups can be assessed through the ROC curve [12]. The ROC shows the plot of the points (1-specificity, sensitivity) for all values of p in $(0, 1)$. Figure 2.4 is an example of the output of a shiny-app that we developed to illustrate the functioning of the ROC curve and the AUC measure for the general setting of diagnostic tests. The link https://marcobonetti.shinyapps.io/shinyapp gives access to the shiny-app.

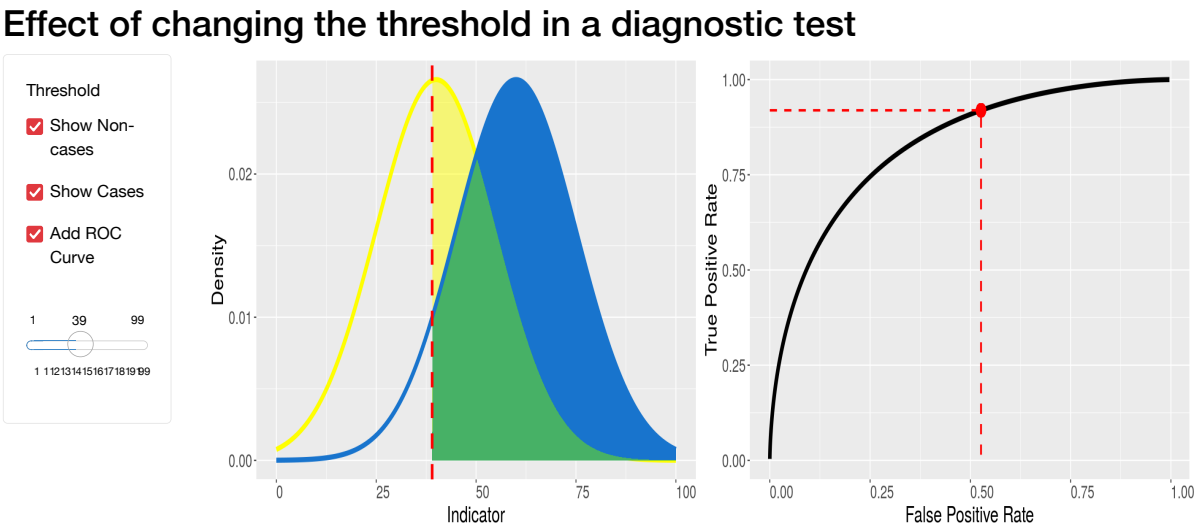

Figure 2.4: Sample output from shiny-app illustrating ROC curves.

From the ROC curve an overall measure of the performance of the classifier, the Area Under the Curve (AUC), can be computed [12]. The AUC estimates are constructed from 100 simulated multivariate samples by comparing the known true value R, which we generate at the data generation step, to its posterior family-specific expected value obtained through likelihood estimation, with either one subject in the univariate case or all family members in the multivariate case. We consider the baseline survival distribution and the family sample size, which is one in the univariate case and four in the multivariate case. The posterior expected value of the latent risk is given by

$$\mathbb{E}(R \mid \text{family data}; \widehat{\theta}) = \frac{P(R_i = 1) f_{\mathbf{X}}(\underline{\mathbf{x}}_i \mid R_i = 1; \widehat{\theta})}{P(R_i = 0) f_{\mathbf{X}}(\underline{\mathbf{x}}_i \mid R_i = 0; \widehat{\theta}) + P(R_i = 1) f_{\mathbf{X}}(\underline{\mathbf{x}}_i \mid R_i = 1; \widehat{\theta})}.$$

This quantity is thus compared to the real risk group. Straightforwardly, one can obtain the univariate counterpart of the posterior expected value of the latent risk.

The results showing the average AUC over 100 simulated samples are in Table 2.13. In this analysis, the sample size is allowed to vary over three values: $n = 10^2, 10^3, 10^4$ to appreciate possible changes in increasing the sample size.

	Number of families		
	10^2	10^3	10^4
Multivariate			
Exponential	0.6917 (0.0062)	0.6902 (0.0019)	0.6907 (0.0006)
Weibull	0.6564 (0.0070)	0.6559 (0.0023)	0.6558 (0.0007)
Gamma	**0.6923 (0.0046)**	**0.6988 (0.0008)**	**0.6957 (<0.0001)**
Univariate			
Exponential	0.5393 (0.0139)	0.5379 (0.0101)	0.5409 (0.0008)
Weibull	0.5492 (0.0068)	0.5498 (0.0030)	0.5503 (0.0007)
Gamma	**0.5927 (0.0079)**	**0.5798 (0.0010)**	**0.5816 (0.0001)**
Univariate FH			
Exponential	0.4492 (0.0905)	0.4137 (0.0231)	0.4137 (0.0079)
Weibull	**0.5183 (0.0143)**	**0.5213 (0.0026)**	**0.5211 (0.0007)**
Gamma	0.5140 (0.0892)	0.4969 (0.0892)	0.4698 (0.0008)

Table 2.13: AUC results from the Multivariate frailty Lehmann Cure-Rate model, the Univariate frailty Lehmann Cure-Rate model, and the Univariate FH Lehmann Cure-Rate model.

It is noteworthy that, for each model, the variance of the AUC values decreases as the sample size increases, as expected. The models that perform the best are highlighted in bold. Notably, the multivariate model outperforms the other models for each sample size and distribution. Moving from the univariate to the multivariate likelihood shows an increase of 10% and more in the AUC, and considering that 0.5 coincides with a classification procedure by following the flipping of a coin, such improvement in classification performance is indeed significant. Computing the AUC with R vs. its expected value $\mathbb{E}(R)$ has no meaning with the observed family history model because the information of the frailty is not involved in this model. Due to this fact, a comparison between FH and R both obtained in the data simulation process is applied to replace $\mathbb{E}(R)$ (results always in Table 2.13). Results are quite poor for the univariate FH model because the AUC, in some cases (≈ 0.4), has a lower value than having accuracy in classification with the flipping of a coin.

Notice that by increasing the sample size there is no significant change in the value of the AUC. ROC curves from one of the 100 dataset are reported in Figures 2.5, 2.6, 2.7, 2.8, and 2.9 as a graphical example. We report the number of the families into the sample, but no the number of subjects involved. Consider that this last is greater or equal to the number of families.

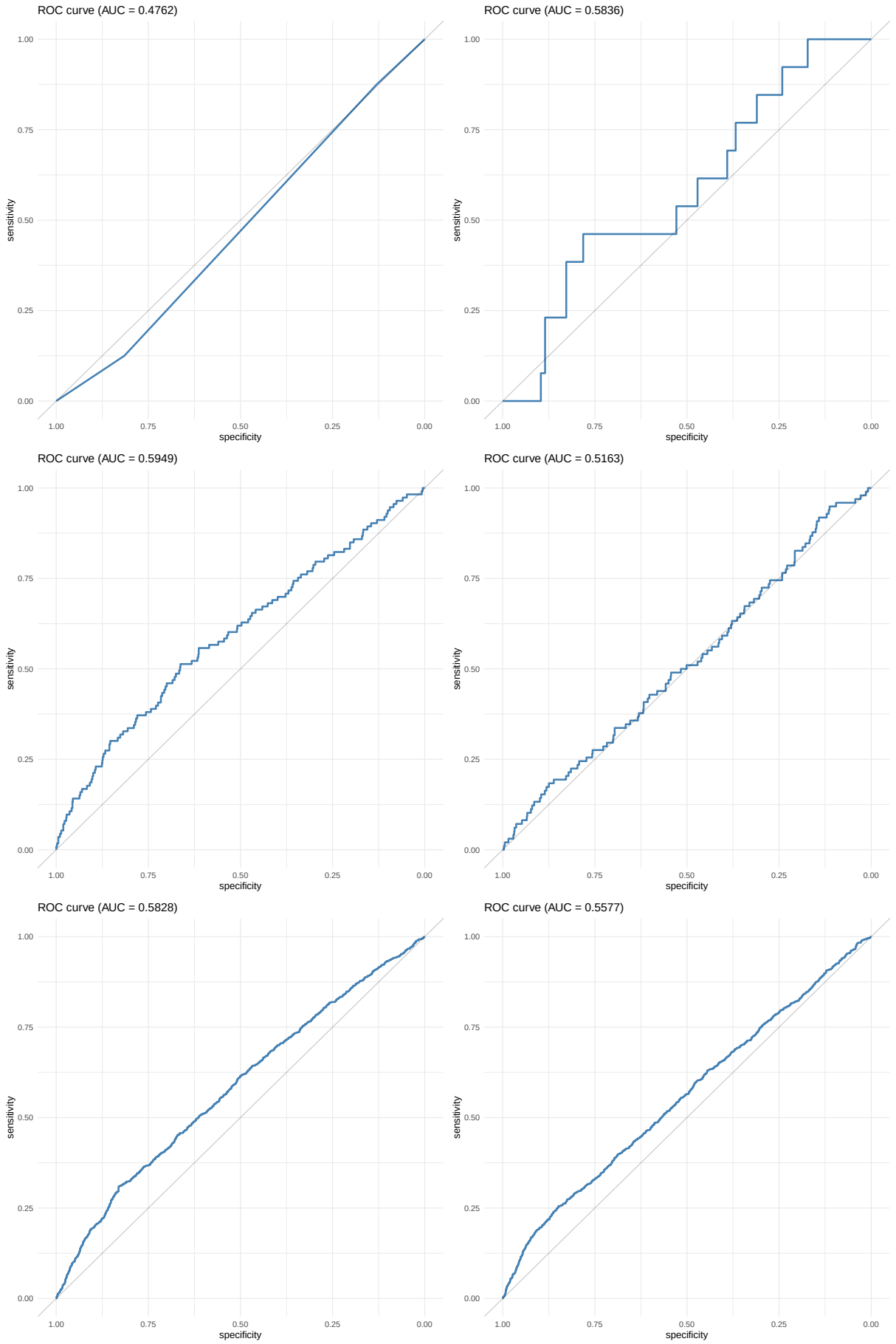

Figure 2.5: Exponential (left) and Weibull (right) ROC curves with n = 10^2 (top), 10^3 (middle), and 10^4 (bottom) for the Univariate model.

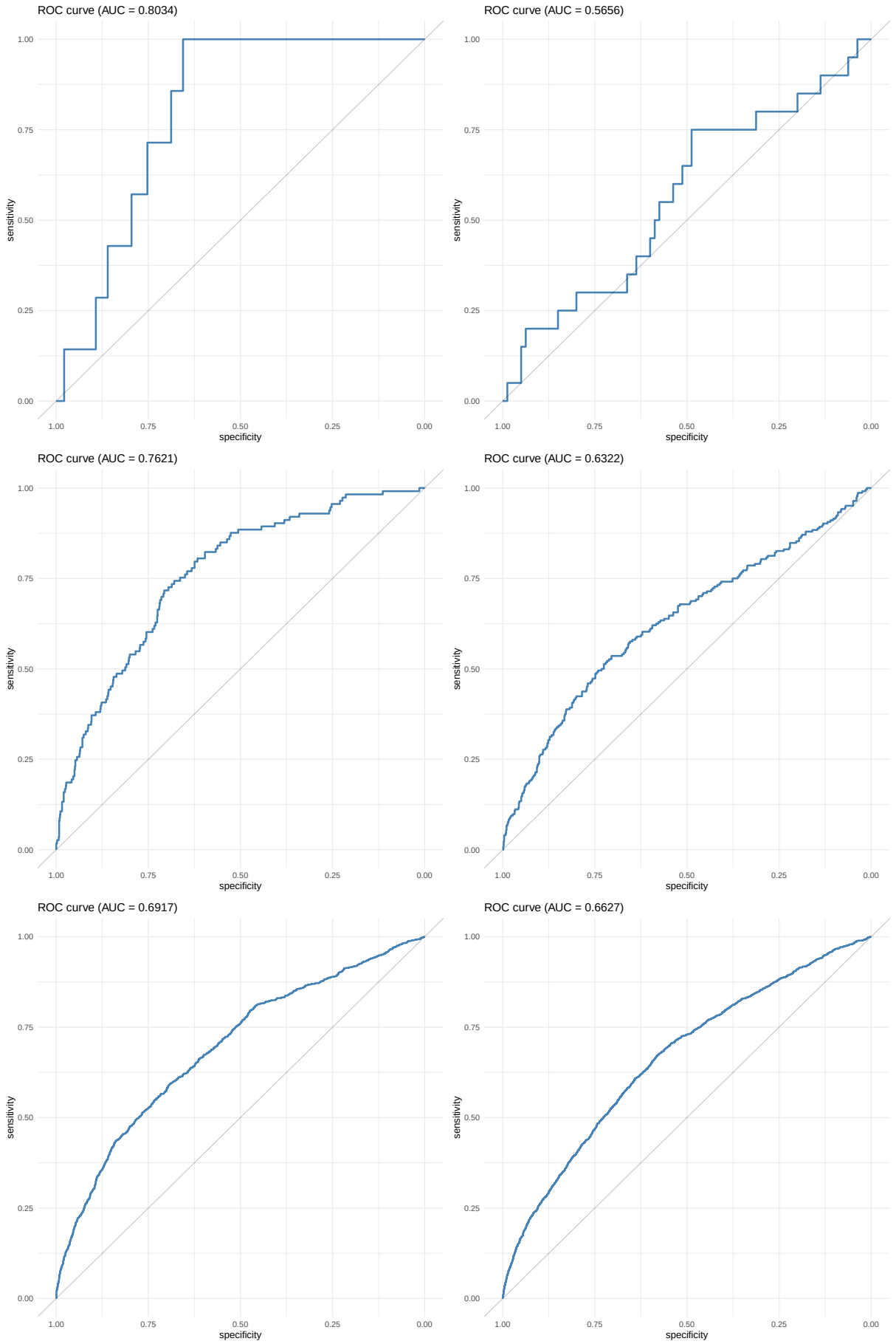

Figure 2.6: Exponential (left) and Weibull (right) ROC curves n = 10^2 (top), 10^3 (middle), and 10^4 (bottom) for the Multivariate model.

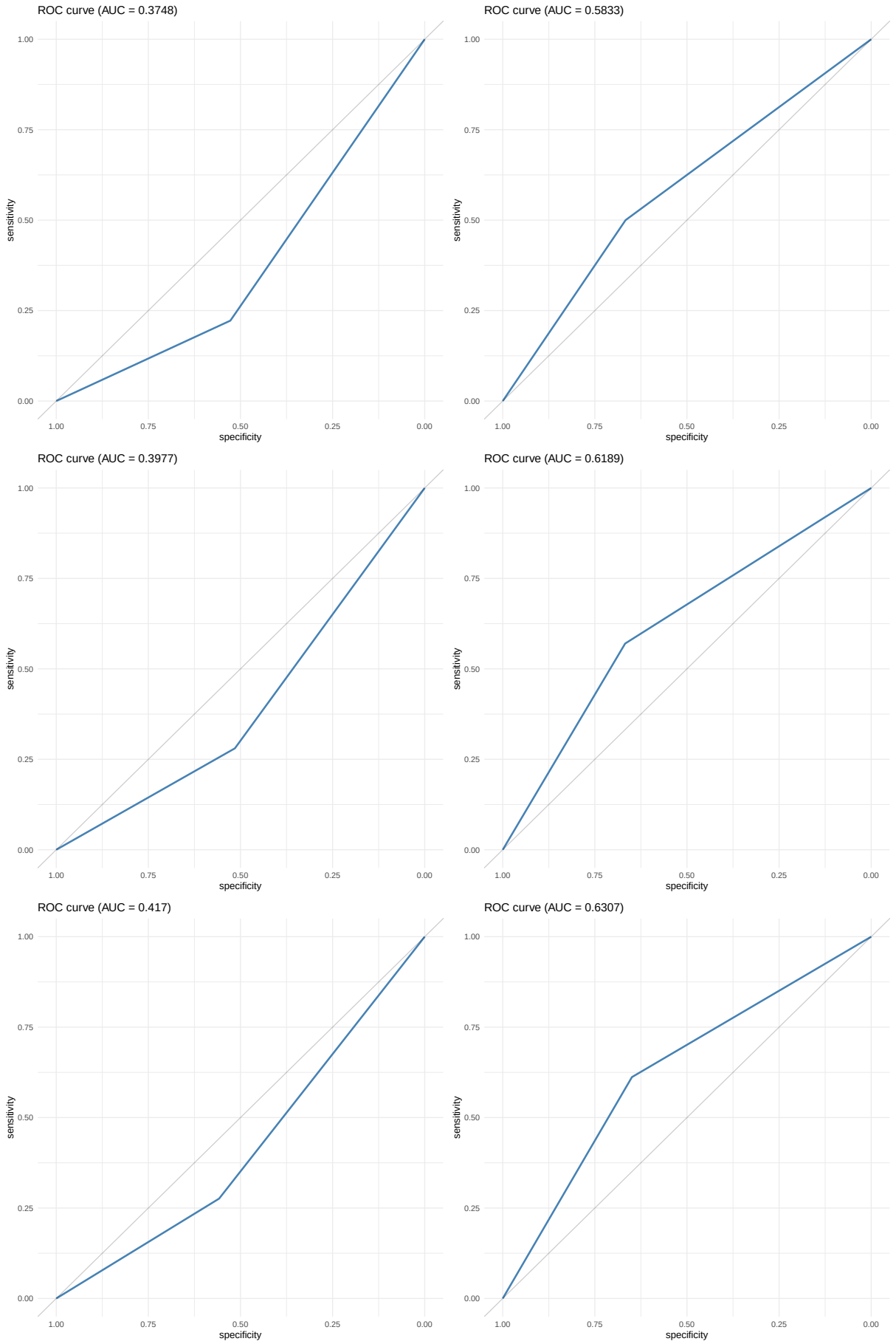

Figure 2.7: Exponential (left) and Weibull (right) ROC curves with n = 10^2 (top), 10^3 (middle), and 10^4 (bottom), for the Univariate FH model.

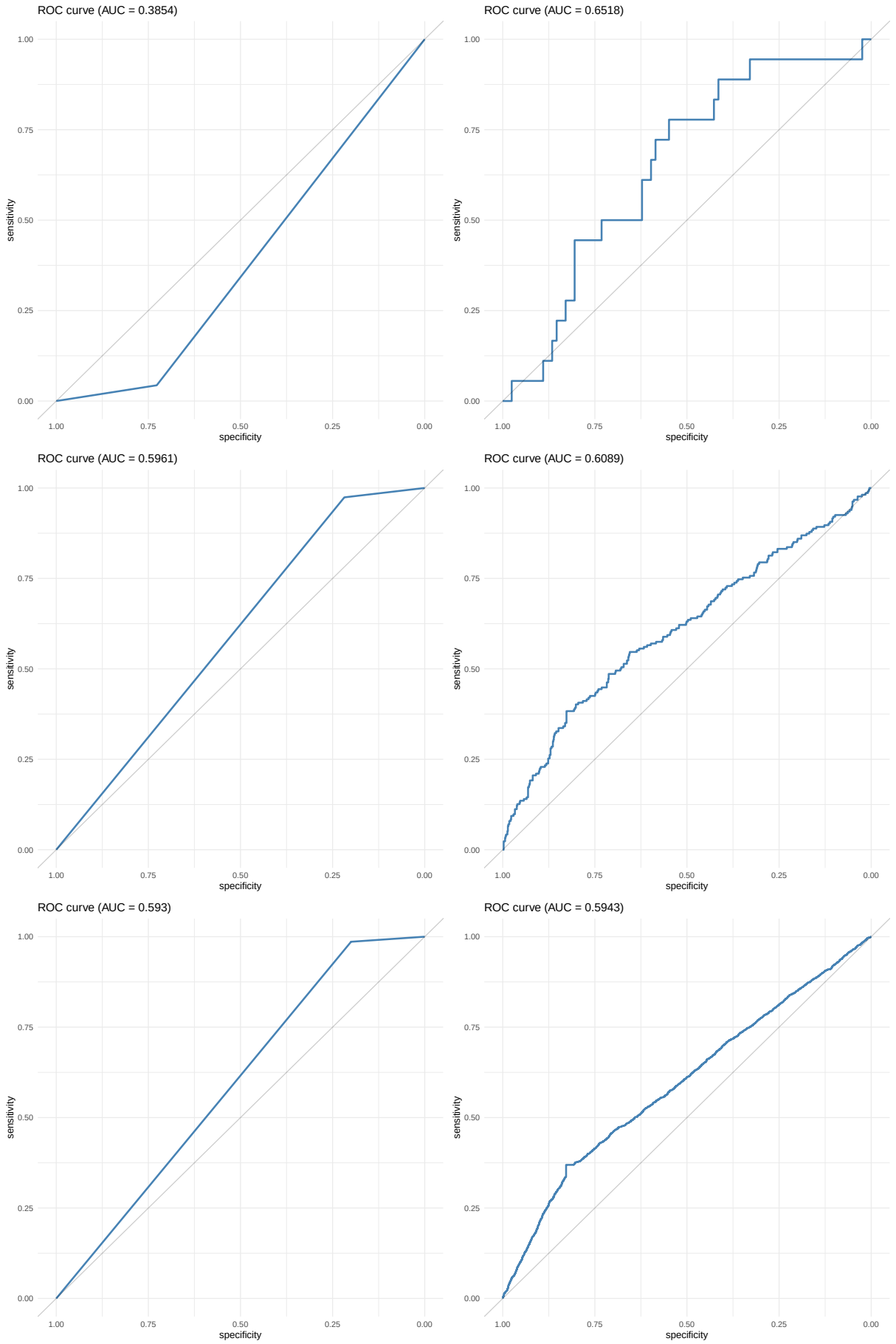

Figure 2.8: Gamma Univariate FH (left) and Univariate model (right) ROC curves with n = 10^2 (top), 10^3 (middle), and 10^4 (bottom).

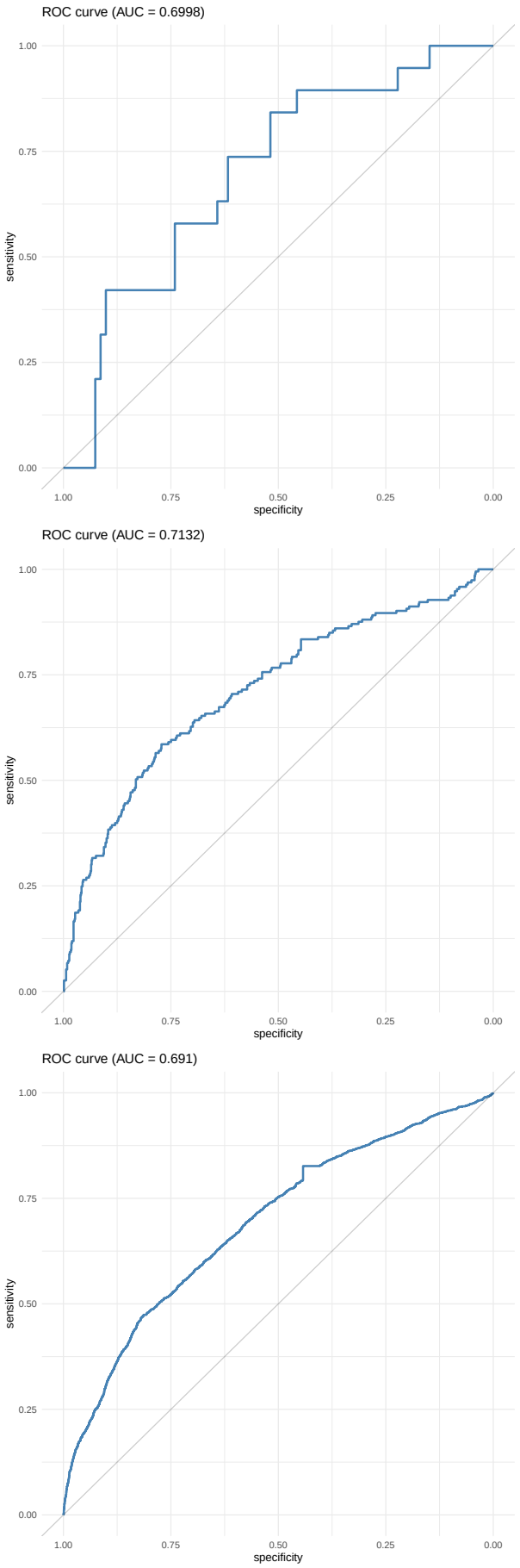

Figure 2.9: Gamma Multivariate model ROC curves with n = 10^2 (top), 10^3 (middle), and 10^4 (bottom).

We report some illustrative results based on one sample of varying number of families into the sample: n = 100, 1,000, 10,000, exploring the three baseline survival distributions, Exponential, Weibull and Gamma. Figures 2.10, 2.11, 2.12, 2.13, 2.14, 2.15, 2.16, 2.17, 2.18 show the histograms of the family-specific posterior frailty risk probability $P(R_i \mid \underline{x}_i, \underline{\delta}_i; \widehat{\theta})$ as estimated from the simulated dataset using the univariate likelihood and the multivariate likelihood with overlapping densities, for the two risk groups and distinguished by distributions. Consider that the lighter area with dashed borders represents the results of the distribution from the Multivariate frailty Lehmann Cure-Rate model, while the darker area without borders is from the Univariate frailty Lehmann Cure-Rate model. We can generally notice from these figures that the multivariate distribution of the probability of belonging to the high-risk group is more distributed over all the range [0,1], contrarily to the univariate distribution. This allows to better identify the highest-risk families in order to accurately address them to more intensive prevention strategies.

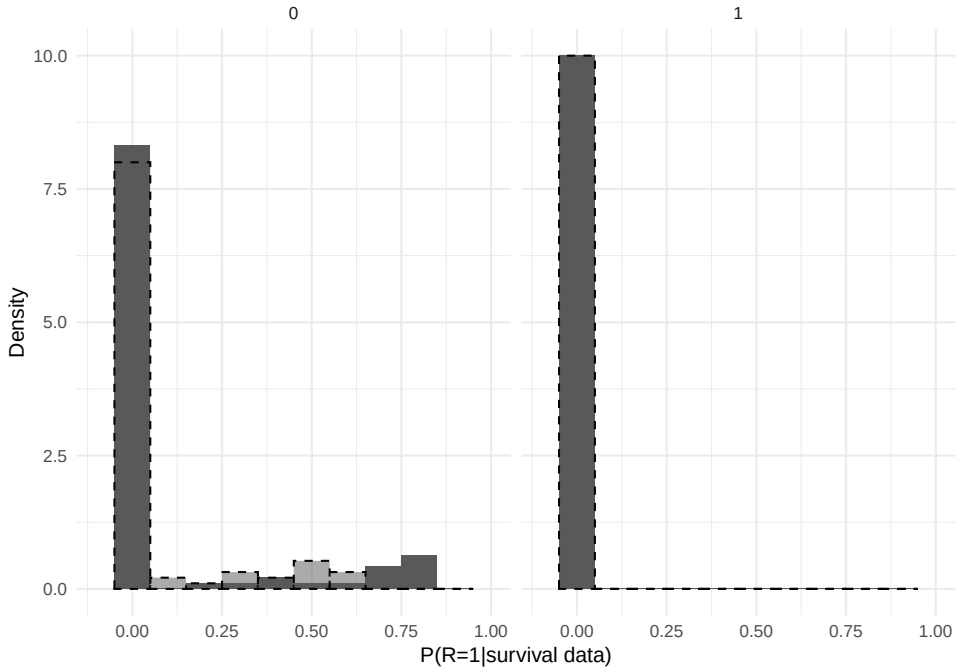

Figure 2.10: Family-specific estimated probability of belonging to the high-risk group, through maximization of the Exponential univariate likelihood (in grey) vs. the Exponential multivariate likelihood (lighter with dashed borders) grouped by true risk $R = 0/1$ for $n = 100$.

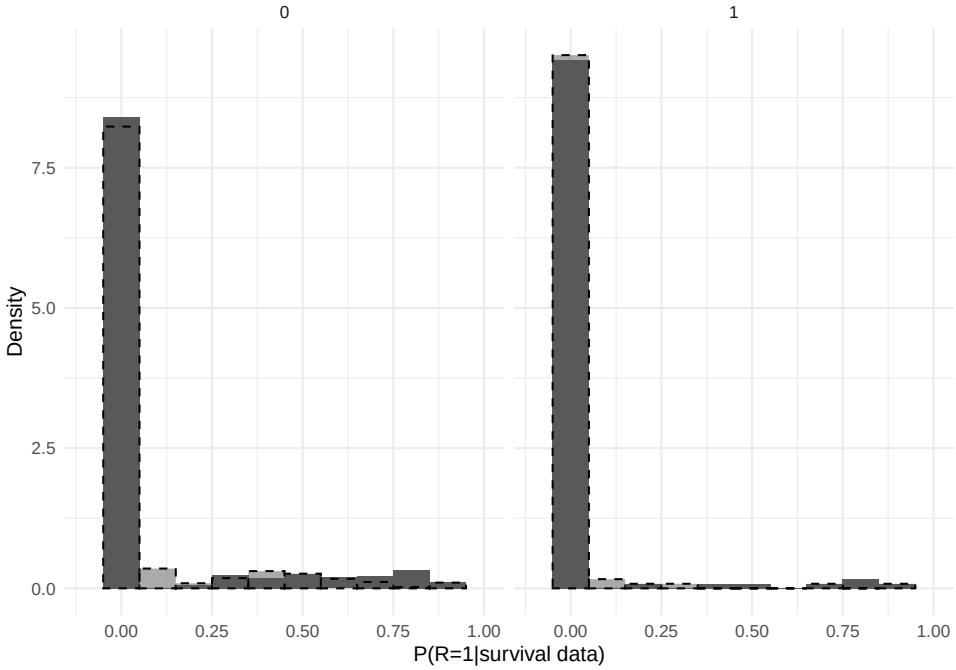

Figure 2.11: Same as Figure 2.10 with $n = 1,000$.

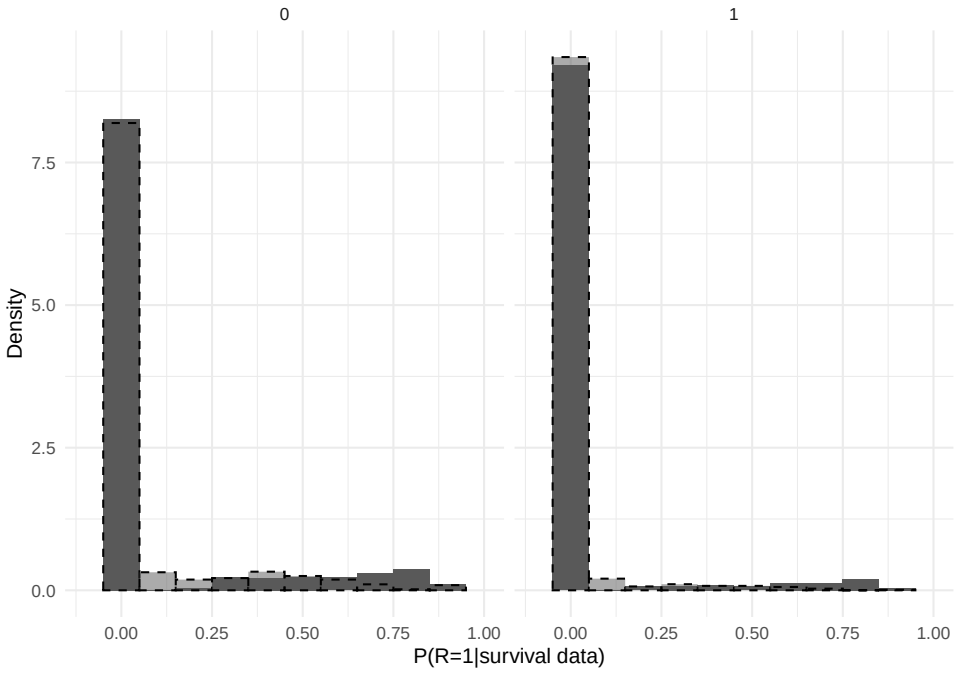

Figure 2.12: Same as Figure 2.10 with $n = 10,000$.

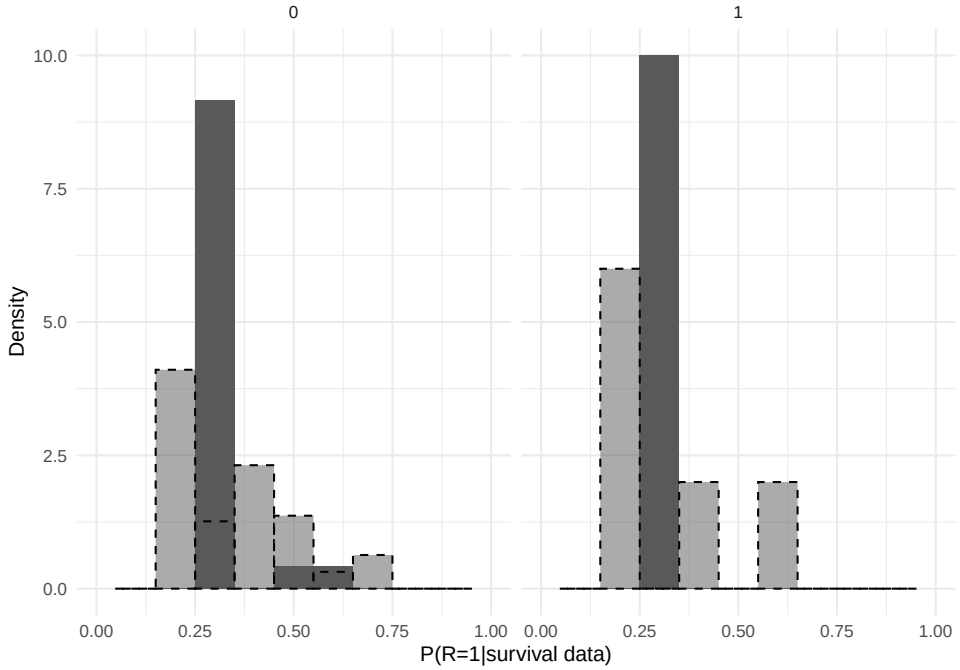

Figure 2.13: Family-specific estimated probability of belonging to the high-risk group through maximization of the Weibull univariate likelihood (in grey) vs. the Weibull multivariate likelihood (lighter with dashed borders) grouped by true risk $R = 0/1$ for $n = 100$.

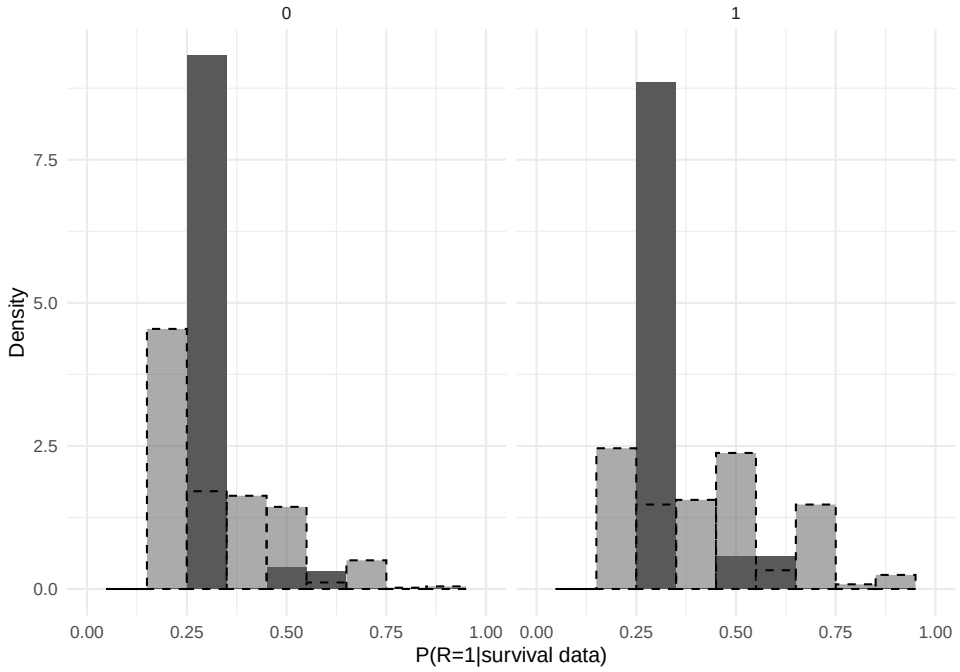

Figure 2.14: Same as Figure 2.13 with $n = 1,000$.

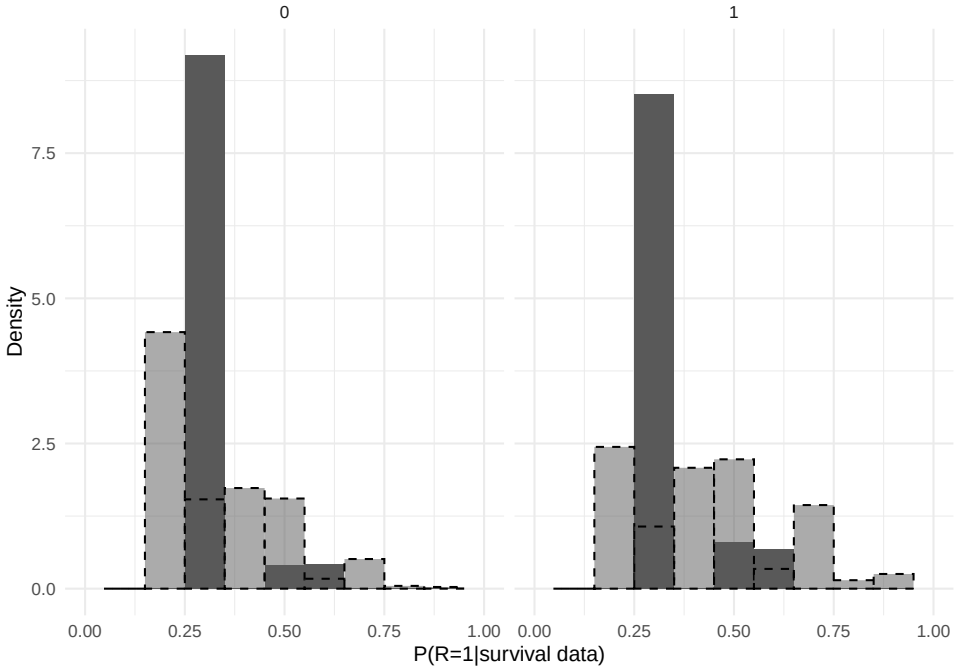

Figure 2.15: Same as Figure 2.13 with $n = 10,000$.

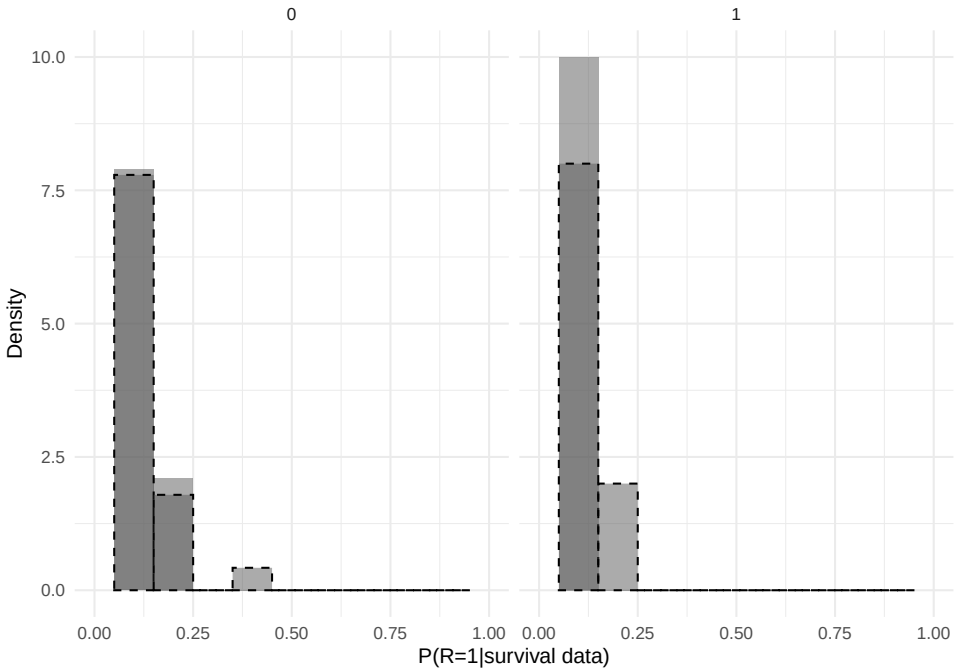

Figure 2.16: Family-specific estimated probability of belonging to the high-risk group through maximization of the Gamma univariate likelihood (in grey) vs. the Gamma multivariate likelihood (in grey with dashed borders) grouped by true risk $R = 0/1$ for $n = 100$.

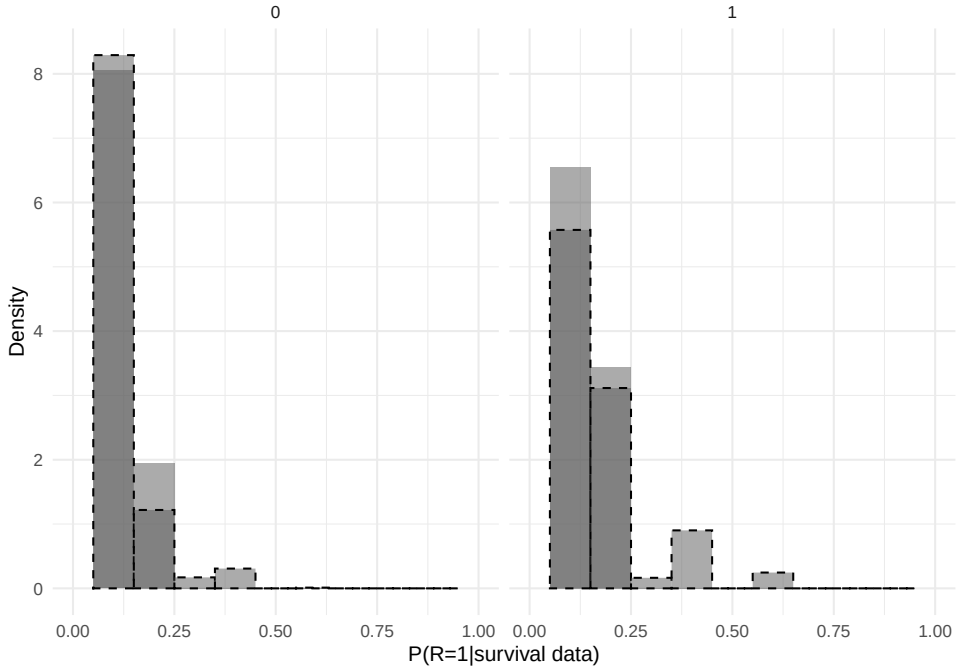

Figure 2.17: Same as Figure 2.16 with $n = 1,000$.

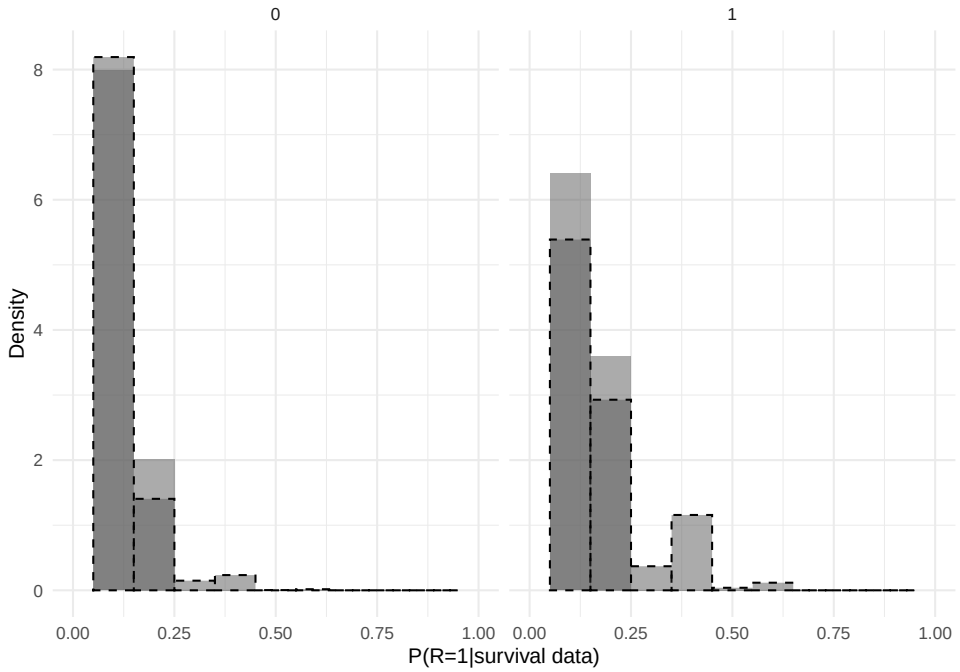

Figure 2.18: Same as Figure 2.16 with $n = 10,000$.

2.4 Discussion

Breast cancer is a significant global health concern, and despite some progress in recent years, there is still wide room for improvement. Prior studies have explored conventional survival models and the use of family history indicator. We believe that solely involving the family history into a risk prediction model can be inadequate to capture the dynamic of breast cancer development. By relying on frailty models, in this work we showed that a two-latent-class approach can outperforms a traditional family history model in terms of precision and prediction. We focused on frailty models and cure-rate survival functions as we believe that a portion of subjects may never experience disease onset, no matter how long their lifespan will be. In the hypothetical scenario where every person develops breast cancer, the cure-rate model relies on the usual survival model. In particular, we developed a Univariate FH which allows for the extension of the Lehmann family model to a Cure-Rate model. Drawing inspiration from the family history models which motivates the two-latent-class approach, we then developed a Univariate frailty Lehmann Cure-Rate model and a Multivariate frailty Lehmann Cure-Rate model.

We discussed the identifiability of univariate models, and performed our comparative analysis, which shows that, as expected, the Multivariate frailty Lehmann Cure-Rate model should be preferred over the univariate models. This shows that detailed family-level data plays a critical role in elucidating the clustering of breast cancer cases, and that merely relying on a raw summary of familial information, such as a family history indicator, may prove inadequate in capturing the full complexity of the phenomenon.

Although the family history model justifies the two-latent-class approach, an extension to more than two risk groups may bring further improvement in risk prediction.

Additional improvements may come from including relatives over the first-degree relationship. This could involve the grandmother from both the maternal and paternal sides. Clearly, and very importantly, while here we have focused only on the latent family risk, covariates may be incorporated to tailor the risk as done in traditional risk models. Some ideas in that direction are described in Appendix B.5. Moreover, moving to the analysis on available data would be a crucial point to validate our work in a real case setting.

References

[1] Bondi, L., Bonetti, M., Grigorova, D., and Russo, A. (2023). Approximate bayesian computation for the natural history of breast cancer, with application to data from a milan cohort study. *Statistics in Medicine*, 42(18):3093–3113.

[2] Bravi, F., Decarli, A., and Russo, A. G. (2018). Risk factors for breast cancer in a cohort of mammographic screening program: a nested case–control study within the fr icam study. *Cancer medicine*, 7(5):2145–2152.

[3] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

[4] Darabi, H., Czene, K., Zhao, W., Liu, J., Hall, P., and Humphreys, K. (2012). Breast cancer risk prediction and individualised screening based on common genetic variation and breast density measurement. *Breast Cancer Research*, 14(1):1–11.

[5] Duchateau, L. and Janssen, P. (2008). *The frailty model*. Springer.

[6] Evans, D. G. R., Kerr, B., Cade, D., Hoare, E., and Hopwood, P. (1996). Fictitious breast cancer family history. *The Lancet*, 348(9033):1034.

[7] Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., and Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute*, 81(24):1879–1886.

[8] Hougaard, P. (2012). *Analysis of multivariate survival data*. Springer Science & Business Media.

[9] Lee, M., Reilly, M., Lindström, L. S., and Czene, K. (2017). Differences in survival for patients with familial and sporadic cancer. *International Journal of Cancer*, 140(3):581–590.

[10] Moolgavkar, S. H. and Venzon, D. J. (1979). Two-event models for carcinogenesis: incidence curves for childhood and adult tumors. *Mathematical biosciences*, 47(1-2):55–77.

[11] Mota, A., Milani, E. A., Leão, J., Ramos, P. L., Ferreira, P. H., Junior, O. G., Tomazella, V. L., and Louzada, F. (2022). A new cure rate frailty regression model based on a weighted lindley distribution applied to stomach cancer data. *Statistical Methods & Applications*, pages 1–27.

[12] Pepe, M. S. et al. (2003). *The statistical evaluation of medical tests for classification and prediction*. Medicine.

[13] Rodrıguez, G. (2010). Multivariate survival models.

[14] Rosner, B., Colditz, G. A., Iglehart, J. D., and Hankinson, S. E. (2008). Risk prediction models with incomplete data with application to prediction of estrogen receptor-positive breast cancer: prospective data from the nurses' health study. *Breast Cancer Research*, 10(4):1–11.

[15] Skates, S. J., Pauler, D. K., and Jacobs, I. J. (2001). Screening based on the risk of cancer calculation from bayesian hierarchical changepoint and mixture models of longitudinal markers. *Journal of the American Statistical Association*, 96(454):429–439.

[16] Strandberg, J. R. and Humphreys, K. (2019). Statistical models of tumour onset and growth for modern breast cancer screening cohorts. *Mathematical Biosciences*, 318:108270.

[17] Strandberg, R. (2022). *Breast cancer natural history models and risk prediction in mammography screening cohorts*. Karolinska Institutet (Sweden).

[18] Strandberg, R., Czene, K., Eriksson, M., Hall, P., and Humphreys, K. (2022). Estimating distributions of breast cancer onset and growth in a swedish mammography screening cohortestimating distributions of breast cancer onset and growth. *Cancer Epidemiology, Biomarkers & Prevention*.

[19] Tyrer, J., Duffy, S. W., and Cuzick, J. (2004). A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in medicine*, 23(7):1111–1130.

[20] Waks, A. G. and Winer, E. P. (2019). Breast cancer treatment. *Jama*, 321(3):316–316.

Chapter 3

Familial mortality risk - a simulation study

Published in the SIS 2022 | Book of Short Papers

We study the family-level risk associated with longevity, as we assume that families can be categorized in different groups of mortality risk and that family members share a common risk that is latent and unchanged from birth. We develop a classification algorithm that operates by computing a chosen family-specific risk posterior quantile. This algorithm is applied to scenarios involving both discrete k-level risk and continuous risk. By conducting this analysis we aim to contribute to the fundamental task of quantifying the mortality risk of subjects from information on the survival of their family members, thus allowing the development of prevention strategies that may protect individuals belonging to frail families, which is crucial in enhancing the survival chances of individuals.

keywords: heritability of longevity, risk prediction, shared frailty, survival analysis

3.1 Introduction

We are interested in studying the effect of (family-specific) risk on longevity conditionally to the genetic make-up of the family. Longevity is known to be in part hereditary, so the risk of dying has a familial component. The survival family history is crucial to involve in the analysis, to assess the subject's (being part of a family) risk of survival. The definition of a positive family history refers to the collection of the survival experiences of the other family members. The significance of a family history increases with the number of deaths, their ages of death, and the closeness of the genetic relationship with the subject [5]. The significance of family history varies according to other aspects also (see e.g. [5, 10]). We may divide families into different clusters, within which they share the same risk of mortality. The interest is in classifying a subject's family to one of a set of risk groups or, more generally, to estimate the family-specific risk from the available data.

First consider the effect of family risk on longevity. A recent model in this direction is called heritability of longevity [6]. The key components of this model are outlined as follows. For the jth individual, we let t_j be the longevity, and $s_j \in \{\text{male, female}\}$ the sex. Longevity is defined as the difference between the age at death of the subject and the expected age of death based on temporal and environmental factors. Below we use the subscripts m (mother), f (father) and p (generic parent) to identify the corresponding family members of subject j. The simplest model is linearly based on the mid-parent heritability:

$$t_j = \gamma_0 \frac{(t_{m_j} + t_{f_j})}{2} + \delta.$$

A second model comes from considering the heritability to be different based on whether the parent has concordant sex with the individual j or not. Accordingly, the models become two, one for concordant-sex and one for discordant-sex, i.e.:

$$t_j = \delta_0 + \gamma_0 t_{p_j} + \gamma_1 \mathbb{I}(s_j, s_{p_j}) + \gamma_2 (t_{p_j} \times \mathbb{I}(s_j, s_{p_j})),$$

where the indicator of concordant sex is $\mathbb{I}(s_j, s_{p_j}) = 1 \iff s_j = s_{p_j}$, $p \in \{m, f\}$. Estimation of such models can be performed by the least square method [6] from data consisting of individual medical and biological information in population-scale family trees.

We investigate an alternative model that can be used to address the effect of family history on survival from a fully multivariate perspective. Different from what we have seen so far, the quantity of interest t_i is the time-to-event, where the event is death. We therefore develop survival analysis models and methods (see, e.g. [7]). We assume that families are split into groups with different hazard functions and characterized by different survival curves.

This structure recalls a mixture for survival models, where the family risk is the mixing quantity. Family risk is therefore treated as a latent family feature on survival, at the family level (where the family is seen as the cluster). So this means that all the family members have the same risk of mortality from birth and that is not directly observable.

Frailty models offer a viable approach for constructing a mixture in the context of survival models. The frailty quantity is a random effect that captures the unobserved heterogeneity among

groups, given different distributions to subjects who belong to different groups. We refer our proposal to the general conditional model (see [4]) called the univariate frailty model. To fix the idea, conditionally to the frailty quantity of interest, i.e. the unobserved family risk, the hazard function has a multiplicative form involving the baseline hazard and the risk. Notice that the distribution of the risk can be seen as the mixture distribution. The very first set to develop the univariate frailty model is based on splitting families into two groups: a low-risk and a high-risk group. The latent family risk, which is called "R", assumes the value "$R = low/high$". The model for the risk of death is represented as follows. The hazard function for the survival times of all family members in the two groups can be defined as $\lambda_0(t)$ and λ_1(t) = $\alpha\lambda_0$(t) for $R = 0$ and 1, respectively. Notice that the hazard function in the low-risk group $\lambda_0(t)$ is taken as the baseline hazard. While the hazard of the high-risk group $\lambda_1(t)$ is proportional to the baseline hazard up to some constant α. The corresponding survival functions are the baseline survival function and the high-risk group survival:

$$S_0(t) = e^{-\int_0^t \lambda_0(u)\mathrm{d}u}$$
$$S_1(t) = e^{-\int_0^t \lambda_1(u)\mathrm{d}u} = e^{-\int_0^t \alpha\lambda_0(u)\mathrm{d}u} = \left[e^{-\int_0^t \lambda_0(u)\mathrm{d}u}\right]^{\alpha} = [S_0(t)]^{\alpha},$$

following the typical Lehmann survival structure [8]. We assume $\alpha > 0$ because, by definition, the high-risk survival function should always be lower than the low-risk survival. This assures that in the high-risk group subjects die earlier and in a higher number.

Notice that, according to the value assumed by the baseline hazard, we can observe different scenarios. When the baseline hazard is not "too small" then the two hazards are different and the high-risk group produces events earlier. Then, inferring the latent group for the jth subject should be relatively easy at the beginning of the calendar time axis because one will already observe some deaths mainly in high-risk families. On the other hand, when the baseline hazard is small, inferring the latent group for the jth subject should be easy later when more events occur in the high-risk families compared to a few events in the low-risk families. However, learning about R will be difficult at the beginning when the risk of death is low and none or very few events occur in both groups.

We begin by extending the univariate frailty model to accommodate multiple time-to-event observations. In this framework, we incorporate the birth cohort effect for each family (cluster), as described in [12] and other relevant studies. An interesting fact about this model is that the conditional independence assumption holds. For example, consider families made of two subjects, say, mother and daughter. We thus have a bivariate frailty model, where R is again the family risk parameter, so that $T_1 \perp T_2 | R$. Also, the pairs (T_1, T_2) within each risk group are independent. The frailty (random) effect R has a multiplicative effect on the hazard function as described above.

In Section 3.2 we explore the methods, in Section 3.3 we implement the risk classification algorithm, and in Section 3.4 we show some results from simulation studies, as long as estimation is possible only if all the family survival data are available. We conclude the analysis with some comments in Section 3.5.

3.2 Methods

Recall that R is the continuous frailty variable that follows a parametric distribution characterized by θ. Within such a framework, we use i to identify the family (out of n) and j to identify its n_i members. Following [4] we develop the complete likelihood $L(\underline{R};\underline{X})$ for the problem, where $\underline{X} = \{\underline{x}_i, i = 1,\dots,n\}$, and $\underline{x}_i = (x_{ij}, \delta_{ij})^T$, $x_{ij} = \min(t_{ij}, c_{ij})$, $\delta_{ij} = \mathbb{I}(t_{ij} \leq c_{ij})$ follow the usual notation, that has t indicate the survival time and c indicate the (independent) censoring time. X_{ij} indicates the baseline covariate vector for subject j in family i. The complete likelihood $L(\underline{R};\underline{X})$ can be written in terms of the frailty parameter θ and the survival parameters, i.e. the vector coefficient β for the covariate effects and the baseline hazard function λ_0. So, following the shared frailty hazards structure, we have $\lambda_{ij}(t|z_{ij}, R_i) = R_i \cdot \lambda_{0ij}(t|z_{ij})$, and $\lambda_{0ij}(t|z_{ij}) = \lambda_0(t)\exp(z'_{ij}\beta)$ for family i. The full likelihood is composed by two quantities: $\mathcal{L}(\beta, \lambda_0, \theta) = \mathcal{L}_1(\theta)\mathcal{L}_2(\beta, \lambda_0)$. The estimation procedure follows the approach from [4] with the notation from [13]. The frailty R can be taken to be distributed as a Gamma with shape θ and rate $1/\theta$:

$$\mathcal{L}_1(\theta) = \prod_i \frac{1}{\Gamma(1/\theta)\theta^\theta} R_i^{\theta-1} e^{-R_i/\theta},$$

$$L_1 = \log\mathcal{L}_1(\theta) = \sum_i \left[-\log(\Gamma(\theta)) - \theta\log(\theta) + (\theta-1)\log(R_i) - \frac{R_i}{\theta}\right].$$

We compute also the survival component of the likelihood:

$$\mathcal{L}_2(\beta, \lambda_0) = \prod_{i=1}^{n}\prod_{j=1}^{n_i} \lambda_{ij}(x_{ij})^{\delta_{ij}} S_{ij}(x_{ij}) = \prod_{i=1}^{n}\prod_{j=1}^{n_i} \left(R_i \cdot \lambda_{0ij}(x_{ij}|z_{ij})\right)^{\delta_{ij}} \exp(-R_i \cdot \Lambda_{0ij}(x_{ij}|z_{ij})),$$

$$L_2 = \log\mathcal{L}_2(\beta, \lambda_0) = \sum_{i=1}^{n}\sum_{j=1}^{n_i} \delta_{ij}\log(R_i \cdot \lambda_{0ij}(x_{ij}|z_{ij})) - R_i \cdot \Lambda_{0ij}(x_{ij}|z_{ij}),$$

where $\Lambda(t)$ indicates the cumulative hazard function. The full log-likelihood is then $L(\theta, \beta, \lambda_0) = L_1(\theta) + L_2(\beta, \lambda_0)$.

To estimate the model parameters, we may specify the form of the baseline hazard function. Indeed, the baseline hazard can assume a parametric form or it can be left unspecified (this corresponds to the semiparametric case) [3].

For example, for the parametric specification a common model for the time-to-event variable is the Weibull distribution $T \sim \text{Weibull}(\text{shape}=\gamma, \text{scale}=\mu)$ with the corresponding hazard functions. Given the multiplicative frailty structure, one can reparametrize the conditional (on R) survival distribution as $T \sim \text{Weibull}(\text{shape} = \delta, \text{scale} = \mu/R^{1/\delta})$. Parameter estimation is then achieved by maximizing the log-likelihood function [9] through the Expectation Maximization (EM) algorithm [3], [1]. In both parametric and semiparametric cases, all parameters can be estimated and used to perform classification. We may compute some summary measures for the estimated parameter.

The variance of $\widehat{\theta}$ is computed following the procedure:

$$\widehat{\theta} \pm 1.96\widehat{\sigma}_{\widehat{\theta}}$$
$$U - L = 4\widehat{\sigma}_{\widehat{\theta}}$$
$$\widehat{\sigma}_{\widehat{\theta}} = \frac{U - L}{4}.$$

We can assess the identifiability of the parameter estimation through a comparison between the empirical variance and the one computed here above.

Notice that this survival method can be extended to the framework of disease development. This extension coincides with the previous Chapter 2.

3.3 Risk classification

We implement a risk classification algorithm for k-latent discrete risk classes, that can be also used in the continuous frailty risk setting. To fix ideas, we discretize the continuous frailty assuming infinitely many classes of risk. In this way, the risk can be considered discrete. Further, we will extend to a proper classification algorithm for the continuous framework. The detailed algorithm is described in the following lines.

We start from the distribution $f_{R|Z}(r \mid z, \theta)$, where R is the frailty quantity and z are the covariates. We assume the frailty distribution $R \sim \text{Gamma}(shape = \theta, scale = \theta)$. We integrate out the covariate from the posterior distribution between frailty and covariates in order to obtain the Gamma distribution again. We wish to prove that

$$\int_{\underline{Z}} f_{R|\underline{Z}}(r, z; \widehat{\theta}) f_{\underline{Z}}(z) \mathrm{d}z \stackrel{?}{=} f_R(r; \theta) = \mathbb{E}_{\underline{Z}}[f_{R|Z}(r \mid z; \widehat{\theta})]$$

We use the grid method $\{r_1, \ldots, r_k\}$, as described in the text in Chapter 3, to obtain the prediction of the frailty quantity $\widehat{R}$. We may choose a Weibull distribution for $Z \mid R \sim$ Weibull. We want to assess whether $f(r \mid z)$ has the equivalent behaviour of r_i. From the distribution of $Z \mid r = r_i$ we may compute analytically and compare the density function $f(r \mid z_i)$ to r_i. Additionally, we want to assess whether the median of $f_R(u \mid z) = R$.

$$r_i \qquad \text{vs} \qquad \left\{ f_{R|Z}(r \mid z);\ z \sim f(z \mid r_i) \right\}$$

We can see r as a parameter: for all the ith family, r_i is the "parameter" that governs $f_{(Z|R)}(z \mid r_i)$. We need to estimate optimally, through the maximum likelihood estimators (MLE) or the variances of the estimators, the unknown parameter r_i from the n=1 sample $z = z_i$. In the frequentist context the procedure may be based on $\mathbb{L}(r; z_i) = f(z_i; r)$ and

$$\widehat{r_i} = \underset{r \in \mathbb{R}^+}{\arg\max} f(z_i \mid r).$$

The MLE of the frailty risk is computed through the following procedure:

$$\log \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[(\lambda_0(x_{ij}) r)^{\widehat{\delta}_{ij}} S_0(x_{ij})^r \right] = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left[\widehat{\delta}_{ij} (\log \lambda_0(x_{ij} + \log r) + r S_0(x_{ij}) \right]$$

$$\frac{\partial \left(\sum_{i=1}^{n} \sum_{j=1}^{n_i} \left[\widehat{\delta}_{ij} (\log \lambda_0(x_{ij} + \log r) + r S_0(x_{ij}) \right] \right)}{\partial r} = \sum_{j} \left[\frac{\delta_{ij}}{r} + \log S_0(x_{ij}) \right] = 0$$

$$\iff \widehat{r} = -\left[\left(\frac{\sum_j \log S_0(x_{ij})}{\sum_j \delta_{ij}} \right) \right]^{-1} = -\frac{\sum_j \delta_{ij}}{\sum_j \log S_0(x_{ij})}.$$

where $\sum_{ij}(\cdot) = \sum_{i=1}^{n} \sum_{j=1}^{n_i}(\cdot)$.

The parameter r depends only on the observed events or censored observations through the numerator, indeed no indicator of event is involved in S_0. This means that all the observed survival times are included in the computation of r. This causes an issue that we are required to explore. We check the second derivative to assess the direction of the first derivative:

$$\frac{\partial^2 \left(\sum_{ij} \left[\frac{\delta_{ij}}{r} + \log S_0(x_{ij}) \right] \right)}{\partial r} = -\frac{1}{r^2} \sum_{ij} \delta_{ij} < 0.$$

The fact that the second derivative is always negative indicates a computational stability issue when there are no events (i.e., no occurrence of the event of interest) in a particular family. This scenario poses challenges because it results in a lack of variability and can lead to numerical instability in calculations. However, there are no such problems when the sum of δ_{ij} (indicating event occurrences) for a given individual i is non-zero. In this case, the presence of at least one event provides the necessary variability to avoid computational stability issues.

Indeed, the estimated risk can be comouted as $\widehat{r} = -\sum_{ij} \delta_{ij} / \sum_{ij} \log \widehat{S}_0(x_{ij})$ where $\widehat{S}_0(\cdot)$ can be replaced with a consistent estimator, for example Breslow. We can rewrite r as

$$\widehat{r} = \frac{\sum_{ij} \delta_{ij}}{-\log \prod_{j=1}^{n_i} \widehat{S}_0(x_{ij})}.$$

Notice again that this new computation is possible only when we have at least one event in each family. The problem remains when there exists a family with only censored observations. Lastly, notice that if we know the original distribution of $f(r \mid z)$ than we can compare that distribution with $\widehat{r}$.

An alternative estimation procedure is explored. We wish to find a estimator of R_i that we call δ_i such that the expected value of the absolute measurement error between the true frailty and the estimator is the minimum, as

$$\delta(z) : \min \int_{\mathbb{R}^+} \int (\delta(z) - r)^2 f_{Z|R}(z \mid r) \mathrm{d}z f_R(r) \mathrm{d}r = \int \left[\int_{\mathbb{R}^+} (\delta(z) - r)^2 f_{R|Z}(r \mid z) \mathrm{d}r \right] f_Z(z) \mathrm{d}z,$$

with R unknown. We can then minimize $\int_{\mathbb{R}^+} (\delta(z) - r)^2 f_{R|Z}(r \mid z) \mathrm{d}r$. This justifies using $\delta(z) = \mathbb{E}(R \mid z)$ as an estimator for R. There is a problem of underestimation and there is a scaling problem for

the frailty quantity. We are wondering which function would be appropriate, since the squared function is not. We may apply a logarithmic transformation in order to solve the problem. This approach leads to

$$
\begin{aligned}
& r \to \widetilde{r} = \log(r) \iff r = e^{\widetilde{r}}, \text{ and } \frac{\partial r}{\partial \widetilde{r}} = e^{\widetilde{r}} \\
& f_{\widetilde{R}}(\widetilde{r}) = f_R(e^{\widetilde{r}})e^{\widetilde{r}} \\
& f_{Z|\widetilde{R}}(z \mid \widetilde{r}) = \frac{f_{Z,\widetilde{R}}(z,\widetilde{r})}{f_{\widetilde{R}}(\widetilde{r})} = f_{Z|R}(z \mid e^{\widetilde{r}}) \\
& z, \widetilde{r} : \min \int_Z \int_{\mathbb{R}^+} (\delta(z) - \widetilde{r})^2 f_{\widetilde{R}|Z}(\widetilde{r} \mid z) \mathrm{d}\widetilde{r} f_R(z) \mathrm{d}z \\
& \quad \iff \int_{\mathbb{R}^+} (\delta(z) - \widetilde{r})^2 f_{\widetilde{R}|Z}(\widetilde{r} \mid z) e^{\widetilde{r}} \mathrm{d}\widetilde{r} = \int_{\mathbb{R}^+} (\delta(z) - \widetilde{r})^2 f_{R|Z}(e^{\widetilde{r}} \mid z) e^{\widetilde{r}} \mathrm{d}\widetilde{r} = \\
& \int_{\mathbb{R}^+} \left[e^{\widetilde{r}} (\delta(z) - \widetilde{r})^2 \right] f_{R|Z}(e^{\widetilde{r}} \mid z) \mathrm{d}\widetilde{r} = \int_{\mathbb{R}^+} \left[r(\delta(z) - \log(r))^2 \right] f_{R|Z}(r \mid z) \frac{1}{r} \mathrm{d}r = \mathbb{E}\left[(\delta(z) - \log(r))^2 \right] \\
& \quad \Rightarrow \widehat{\delta} = \mathbb{E}(\log r \mid z),
\end{aligned}
$$

as estimator of $\log(r_i)$, so then we have $\mathrm{e}^{\widetilde{\delta}}$ as estimator for r_i.

Beyond the few past ideas to carry out risk classification, we now explain the procedure used here. The first step consists of predicting the membership distribution of the subjects to risk groups, involving the estimated parameters at the step before:

$$
\begin{aligned}
& f(R_i \mid \underline{x}_i; \widehat{\theta}) = \frac{f(R_i, \underline{x}_i; \widehat{\theta})}{f(\underline{x}_i; \widehat{\theta})} = \frac{f(\underline{x}_i \mid R_i) f(R_i; \widehat{\theta})}{f(\underline{x}_i; \widehat{\theta})} = \frac{f(\underline{x}_{i1} \mid R_i) f(\underline{x}_{i2} \mid R_i) \ldots f(\underline{x}_{in_i} \mid R_i) f(R_i; \widehat{\theta})}{\int_{\mathbb{R}^+} f(\underline{x}_i \mid R_i) f(R_i; \widehat{\theta}) \mathrm{d}R_i} \\
& \propto \prod_{j=1}^{n_i} \left[f_{T|R}(x_{ij} \mid R_i)^{\delta_{ij}} S_{T|R}(x_{ij} \mid R_i)^{1-\delta_{ij}} \right] f(R_i \mid \widehat{\theta}) = \prod_{j=1}^{n_i} \left[\lambda_{T|R}(x_{ij} \mid R_i)^{\delta_{ij}} S_{T|R}(x_{ij} \mid R_i) \right] f(R_i \mid \widehat{\theta}),
\end{aligned}
$$

up to the denominator. The notation above indicates the data $\underline{x}_i = (\underline{x}_{i1}, \ldots, \underline{x}_{in_i})^T$. The density function is written as $f(\underline{x}_i \mid R_i, \widehat{\theta}) = f(\underline{x}_i \mid R_i)$ because $\widehat{\theta}$ is irrelevant once R_i is estimated through $f(R_i; \widehat{\theta}) \sim \text{Gamma}(shape = \widehat{\theta}, scale = s1/\widehat{\theta})$. We can obtain $f(\underline{x}_i \mid R_i)$ from the Breslow estimator usually available in software packages, such as

$$
\widehat{f}(\underline{x}_i \mid R_i) = \widehat{\lambda}_0(\underline{x}_i) \mathrm{e}^{\beta' z_{ij}} R_i [\widehat{S}_0(\underline{x}_i)]^{R_i \mathrm{e}^{\beta' z_{ij}}}
$$

where the baseline survival function $\widehat{S}_0(t) = \mathrm{e}^{-\widehat{\Lambda}_0(t)}$, with $\Lambda_0(t)$ the baseline cumulative hazard function estimated through Breslow. The distribution of the frailty conditionally to the family is

$$
f(R_i \mid \underline{x}_i; \widehat{\theta}) \propto \prod_{j=1}^{n_i} \left[\widehat{\lambda}_{T|R}(x_{ij} \mid R_i)^{\delta_{ij}} \widehat{S}_{T|R}(x_{ij} \mid R_i) \right] f(R_i \mid \widehat{\theta}) = \prod_{j=1}^{n_i} \left[(\widehat{\lambda}_0(x_{ij}) R_i)^{\delta_{ij}} [\widehat{S}_0(x_{ij})]^{R_i} \right] f(R_i \mid \widehat{\theta}).
$$

This density function can be seen as the posterior predictive distribution of the risk. Hence, we can predict the risk which assumes one of the values in a grid that we set, according to the density distribution.

3.3.1 Classification procedure

In the parametric approach, the prediction follows a method based on the expectation step of the expectation-maximization (EM) algorithm [9]. Our contribution refers to the semiparametric approach instead. We suggest performing risk prediction by fixing a grid of values $\{r_1, \ldots, r_K\}$ for the frailty quantity and implementing the following steps:

(i) obtain $\widehat{S}_0(x_{ij})$ and $\widehat{\lambda}_0(x_{ij})$ from the Breslow estimator. Details are collected in Appendix C.1;

(ii) compute $\widehat{f}(\underline{x}_i|r_k) = \prod_{j=1}^{n_i}\left[(\widehat{\lambda}_0(x_{ij})r_k)^{\delta_{ij}}[\widehat{S}_0(x_{ij})]^{r_k}\right]$;

(iii) compute $\widehat{f}(\underline{x}_i, r_k; \widehat{\theta}) = \prod_{j=1}^{n_i}\left[(\widehat{\lambda}_0(x_{ij})r_k)^{\delta_{ij}}[\widehat{S}_0(x_{ij})]^{r_k}\right] f(r_k|\widehat{\theta})$, $\forall r_k$ in the grid;

(iv) compute the integral $\widehat{f}(\underline{x}_i; \widehat{\theta}) = \int_{\mathbb{R}^+} f(\underline{x}_i|r_i)f(r_i; \widehat{\theta})\mathrm{d}r_i = \sum_k \Delta(r_k)\widehat{f}(\underline{x}_i, r_k; \widehat{\theta})$, where $\Delta(r_k) = r_{k+1} - r_k$;

(v) compute $\widehat{f}(r_k|\underline{x}_i; \widehat{\theta}) = \widehat{f}(\underline{x}_i, r_k; \widehat{\theta})/\sum_k \Delta(r_k)\widehat{f}(\underline{x}_i, r_k; \widehat{\theta})$.

The predicted continuous shared frailty value $\widehat{R}_i$ for each family i is computed with the rule below, i.e. it takes the value corresponding to summing up the estimated (as above) density function on the grid values until the desired threshold quantile $q \in [0, 1]$:

$$\widehat{R}_i = r_k : \sum_{j:r_j \le r_k} \widehat{f}(r_j|\underline{x}_i; \widehat{\theta})\Delta(r_j) \le q. \tag{3.1}$$

Indeed we choose the posterior percentile so that it minimizes the misclassification rate.

To explore the discrete splitting of families into, say, two risk groups, we may transform the continuous frailty into a binary two-group variable. We can then carry out the (actionable) classification according to the rule:

$$\widehat{RB}_i = \begin{cases} \text{low} & P(r_i < \widehat{\eta}_r|\underline{x}_i; \widehat{\theta}) \ge q \\ \text{high} & P(r_i < \widehat{\eta}_r|\underline{x}_i; \widehat{\theta}) < q \end{cases} \Leftrightarrow \begin{cases} \text{low} & \text{Quantile}(r_i|\underline{x}_i; \widehat{\theta}) \le \widehat{\eta}_r \\ \text{high} & \text{Quantile}(r_i|\underline{x}_i; \widehat{\theta}) > \widehat{\eta}_r \end{cases} \tag{3.2}$$

where $\widehat{\eta}_r = \text{Quantile}\left(\text{Gamma}(\widehat{\theta}, 1/\widehat{\theta})\right) \in [0, 1]$, $\text{Quantile}(r_k|\underline{x}_i; \widehat{\theta}) = r_k : P(r_k|\underline{x}_i; \widehat{\theta}) \le q$ with $q \in [0, 1]$ as above. And, $P(r_k|\underline{x}_i; \widehat{\theta}) = \sum_{j:r_j \le r_k} \widehat{f}(r_j|\underline{x}_i; \widehat{\theta})\Delta(r_j)$.

The idea is represented in Figure 3.1. For simplicity of interpretation and convenience, we first fix $q = 0.5$ to obtain the median as the threshold. So $\widehat{\eta}_z$ is the estimated frailty median as well. If we rely on binary splitting we can then predict the risk groups for each woman in the sample. Especially, we may compute $P(R_i = 1|\underline{x}_i, \widehat{\theta})$; hence the (actionable) classification into a risk group can be carried out with a threshold, say q:

$$\widetilde{R}_i = \begin{cases} 1 & P(R_i = 1|\underline{x}_i, \widehat{\theta}) > q \\ 0 & P(R_i = 1|\underline{x}_i, \widehat{\theta}) < q \end{cases}$$

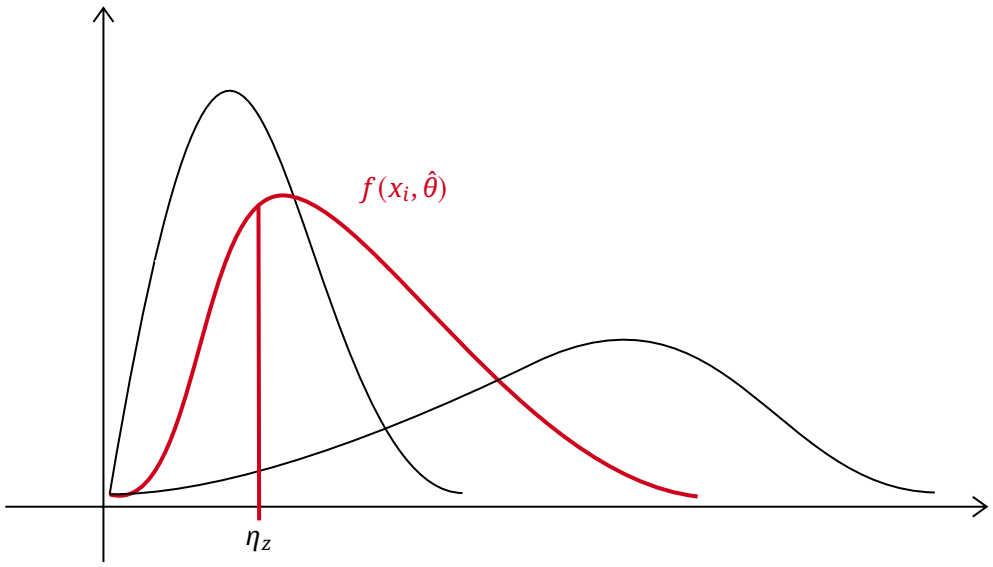

Figure 3.1: Classification method for two risk groups.

where $\widehat{\theta} = \widehat{\theta}(\underline{z}_i)$. The true conditional probability of $R_i = 0$ or $R_i = 1$ is

$$P(R_i = 1|\underline{x}_i, \widehat{\theta}) = \frac{P(R_i = 1 \cap \underline{x}_i; \widehat{\theta})}{P(\underline{x}_i; \widehat{\theta})} = \frac{P(R_i = 1)P(\underline{x}_i|R_i = 1; \widehat{\theta})}{P(\underline{x}_i; \widehat{\theta})}$$

where $P(R_i = 1) = h$ is an element of $\widehat{\underline{\zeta}}$. Then

$$P(R_i = 1|\underline{x}_i, \widehat{\theta}) = \frac{\widehat{P}(R_i = 1) f_{X_i}(\underline{x}_i|R_i = 1; \widehat{\theta})}{\widehat{P}(R_i = 0) f_{X_i}(\underline{x}_i|R_i = 0; \widehat{\theta}) + \widehat{P}(R_i = 1) f_{X_i}(\underline{x}_i|R_i = 1; \widehat{\theta})}$$

where the density function $f_{X_i}(\underline{x}_i|R_i = 1; \widehat{\theta})$ is computed as described above.[1] Alternative procedures can be implemented but are not treated here.

One can then apply some diagnostic tools to analyse the goodness in classification, such as the scatter-plot of R versus $\widehat{R}$ (see 3.1) to obtain a visual analysis of the classification accuracy, and the confusion matrix (see Table 3.1) between the median-based risk group $RB = \mathbb{I}(R \leq \text{Median}(R))$ and the estimated risk group $\widehat{RB}$ obtained as in 3.2 for a fixed q.

		$\widehat{R}$	
		0	1
R	0	TN	FP
	1	FN	TP

Table 3.1: Confusion matrix between R vs. $\widehat{R}$.

We can use the agreement index Cohen's kappa [2] to have a summary of the binary classification results, whose equation follows as

$$k = \frac{2(TP \cdot TN - FN \cdot FP)}{(TP + FP) \cdot (TN + FP) + (TP + FN) \cdot (TN + FN)},$$

with TP, TN, FP, FN indicate the true positive, true negative, false positive, and false negative proportions, where positive (negative) stands for $\widehat{RB}$=high-risk (low-risk). We can compute also the classic agreement index which is given by

$$\text{agreement index} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Both indices vary in the range $[0, 1]$ where 0 means no agreement and 1 means perfect agreement. Also, sensitivity and specificity can be computed as additional classification accuracy measures.

[1] If $f_R(r)$ is a continuous frailty instead of the two-group discrete mixing variable discussed so far, then we could use $f_R(r \mid \underline{x}_i; \widehat{\theta}) \Rightarrow \mathbb{E}(R \mid \underline{x}_i; \widehat{\theta})$ for classification. Some possibilities would be: (i) $\tilde{R}_i = \mathbb{I}(\mathbb{E}(R \mid \underline{x}_i; \widehat{\theta}) \geq 1)$ assuming a log Weibull-Gamma frailty model [12, 4]; (ii) $\tilde{R}_i = \mathbb{I}(P(R > 1 \mid \underline{x}_i; \widehat{\theta}) \geq 0.5)$.

3.4 A preliminary simulation scenario

One can generate some family structures and survival times, and implement the two-group risk classification. The number of families in the dataset is fixed. Each woman has a mother and a grandmother. Instead, the number of sisters and aunts varies. They can be distributed as a Poisson($\lambda = \lambda_s$) and a Poisson($\lambda = \lambda_a$) respectively. We fix $\lambda_s = 1$, $\lambda_a = 0.5$, so that the resulting family size is

$$\text{family size} = 3 + \text{Poisson}(1) + \text{Poisson}(0.5).$$

The expected value of the sample size N is therefore $\mathbb{E}(N) = 3 + 1 + 0.5 = 4.5$. We have the family structure for each family, and for each family member, we build the day of birth (DOB), the observed time (X), and the observed event indicator (δ). The data is visualized as:

	DOB	X	δ
subject			
mother			
grandmother			
sister 1?			
sister 2?			
...			
sister k_1?			
aunt 1?			
...			
aunt k_2?			

where X=min(today-DOB, diagnosis-DOB, death-DOB), and δ takes value 0/1 according to the value of X.

$$\delta = \begin{cases} 0 & \text{X=today-DOB or X=death-DOB;} \\ 1 & \text{X=diagnosis-DOB.} \end{cases}$$

We directly generate the time-to-event T from the Weibull distribution $T \sim \text{Weibull}(\text{shape} = \gamma, \text{scale} = \mu/R^{1/\gamma})$ conditional to the frailty, with $R \sim \text{Gamma}(1,1)$, $\mu = 1$, $\gamma = 5$. The censoring time are generated from a Uniform distribution $C \sim \text{U}(0,12)$, and for each subject we generate $X = \min(T, C)$ and $\delta = \mathbb{I}(X = T)$.

Thus, we sample three thousands families and explore the classification accuracy in three different scenarios: (1) parametric hazard and binary classification with median as threshold; (2) semiparametric case with $q = 0.5$; (3) semiparametric case with $q = 0.25$. We extend to $q = 0.25$ so that we keep low and realistic the posterior high-risk families proportion (see text below). We carry out this analysis stratified by family size and overall. The results for family size are not reported because irrelevant, while the summary of the overall results is in Table 3.2. The posterior high-risk families proportion in the semiparametric case, varying q, is: 0.19 ($q = 0.25$); 0.24 ($q = 0.5$)

reported in Table 3.2 in the last column "HR". Notice that we expect that involving the true value of parameter θ, the true survival function $S_0(x_{ij})$ and hazard function $\lambda_0(x_{ij})$ we are able to reach the best performance in classification. This is the best scenario, and we intend to compare this to the already seen scenarios above.

	Cohen's kappa [2]	Accuracy [11]	Sensitivity	Specificity	HR
1	0.32(0.16)	0.66(0.08)	67.19(9.35)	64.92(6.89)	
2	0.86(0.02)	0.93(0.01)	**98.03(0.53)**	90.64(1.07)	0.24(0.01)
3	**0.90(0.01)**	**0.95(0.01)**	95.41(1.01)	**95.53(0.92)**	**0.19(0.01)**

Table 3.2: Mean and standard deviation of some diagnostics measures for the three models under study.

In Table 3.2 in bold, there are the best results in each category. Notice that the scenario with the first quantile as the threshold has the best performance overall, with the 90% of concordance between the true group and the predicted group of risk of the families; a 95% of accuracy in prediction, with 95% of sensitivity and specificity. These results outperform all the others, but the higher sensitivity at the 98% reached by using the median threshold.

3.5 Discussion

Preliminary results suggest the absence of important differences in classification accuracy across different family sizes. Important is to notice a substantial difference in classification accuracy between the parametric and the semiparametric setting, particularly when employing the median as the classification threshold. Moreover, the use of the first quantile appears to be favorable in terms of both classification accuracy and posterior proportion of high-risk families. These conclusions can be further supported by additional examinations by plotting some figures and being validated also through a real case dataset.

The results are thus so far promising and may help clinicians in identifying high-risk families to target them to intensive prevention paths according to different health problems. Thus, identifying the families with the lowest longevity is directly connected to improve their survival by carrying out disease screening for early detection. We also believe that knowing to be at high-risk of mortality can increase a family's awareness and lead the family's members to live a life with better habits and attention on their health.

As a next step, we wish to complete the simulation studies by, for example, exploring several distributions of the survival baseline function. This further exploration is driven by the fact that the time-to-event variable in observational studies is unlikely to be distributed according to a simple distribution, as the Exponential or the Weibull can be. We may try to analyse a three-parameters distribution to gain higher flexibility to explain the phenomenon of heritability of longevity.

References

[1] Balan, T. A. and Putter, H. (2019). frailtyem: An r package for estimating semiparametric shared frailty models. *Journal of Statistical Software*, 90(7):1–29.

[2] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

[3] Duchateau, L. and Janssen, P. (2007). *The frailty model*. Springer Science & Business Media.

[4] Hougaard, P. (2012). *Analysis of multivariate survival data*. Springer Science & Business Media.

[5] Jenkins, M. et al. (2018). Colorectal cancer risk according to family history. `https://wiki.cancer.org.au/australia/Clinical_question:Family_history_and_CRC_risk`.

[6] Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., Gershovits, M., Markus, B., Sheikh, M., Gymrek, M., et al. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science*, 360(6385):171–175.

[7] Lawless, J. F. (2011). *Statistical models and methods for lifetime data*, volume 362. John Wiley & Sons.

[8] Lehmann, E. L., Romano, J. P., and Casella, G. (2005). *Testing statistical hypotheses*, volume 3. Springer.

[9] Munda, M., Rotolo, F., Legrand, C., et al. (2012). parfm: Parametric frailty models in r. *Journal of Statistical Software*, 51(11):1–20.

[10] Nadeem, Q. (2009). Family history and improving health.

[11] Powers, D. M. W. (2012). The problem with kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 345–355.

[12] Rodrıguez, G. (2005). Multivariate survival models.

[13] Yu, B. (2006). Estimation of shared gamma frailty models by a modified em algorithm. *Computational statistics & data analysis*, 50(2):463–474.

Chapter 4

Exploration of most powerful tests for right-censored survival data

Published in the SEAS IN Book of short papers 2023

We explore some ideas on the most powerful tests for survival data with right censoring. This goal is to carry out a test based on the proportional hazards assumption, aiming to evaluate whether a population exhibits survival times that are governed by a known survival function denoted as $S_0(t)$, rather than by $S_1(t) = [S_0(t)]^{\beta^*}$. In terms of hazard functions, the test compares $\lambda_0(t)$ to $\lambda_1(t) = \beta^* \lambda_0(t)$ and we test the hypothesis that β^* is equal to one, with a two-tailed alternative hypothesis.

We begin by discussing the test without censoring, initially exploring the scenario of a sample with size equal to one and subsequently extending our analysis to a sample size greater than one. Subsequently, we derive an explicit formulation for the most powerful test in the case of a single sample element subject to independent right censoring. The determination of the most powerful test in situations involving independent right censoring for sample sizes greater than one remains an open problem.

keywords: most powerful test, proportional hazards, survival analysis

4.1 Introduction

This chapter presents a comprehensive examination of the application of Most Powerful (MP) tests for survival data, specifically focusing on scenarios that adhere to the proportional hazards (PH) assumption, both in the presence and absence of independent right censoring. An event of interest is defined, where each subject is associated with a time-to-event random variable, denoted as T, and an indicator random variable, denoted as Δ, which signifies whether the event onset has been observed or not. Typically, events such as death or disease onset are studied in this context. In cases where the event is not observed ($\Delta = 0$) because another event happened before, the subject is classified as a censored case, with a censoring time denoted with the random variable C. Consequently, the observed time for each subject is determined as the minimum value between the time-to-event and the censoring time, i.e., $X = \min(T, C)$, and the indicator variable is represented by $\Delta = \mathbb{I}(T \leq C)$.

We aim to build a hypothesis test to determine whether an independently and identically distributed (i.i.d.) sample is composed by survival times that are generated by a known survival function denoted as $S_0(t)$, or by an alternative survival function $S_1(t) = [S_0(t)]^{\beta^*}$, where β^* is an unknown parameter of interest. To provide insight into this test, one can think of a situation where clinicians need to determine whether one or more individuals, perhaps from the same cluster (e.g. a family), come from a higher or lower survival, in order to apply how and when different clinical interventions regarding mere mortality or a disease diagnosis. One may think, for example, at the identification of low-survival (high-risk) subjects for a specific disease, say hepatitis C patients for liver cancer [2], in order to target them towards intensive screening and personalised prevention strategies to enhance their chances of survival.

In Section 4.2, we delve into the examination of MP tests for survival data without censoring events. Subsequently, in Section 4.3, we extend our analysis to admit right-censoring. These tests have been developed for both sample size equal to one and sample size greater than one, with the exception of the latter case, which is currently an ongoing work. Finally, in Section 4.4, we provide a comprehensive discussion on the findings and implications of this analysis.

4.2 MP test with no censoring

4.2.1 Sample size equal to one

Within the context of survival analysis without censoring, the observed time coincides with the value t of the time-to-event random variable T.

Under the assumption of proportional hazards (PH), we define $\lambda_1(t) = \beta^* \lambda_0(t)$, which is equivalent to $S_1(t) = [S_0(t)]^{\beta^*}$. In light of this, the hypothesis system can be expressed as follows:

$$\begin{cases} H_0 : \text{The survival times of the i.i.d. sample are generated by } S_0(t). \\ H_1 : \text{The survival times of the i.i.d. sample are generated by } S_1(t) = [S_0(t)]^{\beta^*}. \end{cases}$$

We wish to assess the statistical evidence supporting one hypothesis over the other, particularly investigating whether the survival times of the i.i.d. sample conform to $S_0(t)$ or deviate by following $S_1(t)$. This evaluation entails the comparison of the hazard functions $\lambda_0(t)$ and $\lambda_1(t)$, or equivalently, examining the relationship between the survival functions $S_0(t)$ and $S_1(t)$, where the latter is in function of the survival function $S_0(t)$ and β^*. In essence, the research question at hand revolves around determining whether the true value of β^* significantly differs from one, so that the hypothesis system can be given by both:

$$\begin{cases} H_0 : S(t) = S_0(t) \\ H_1 : S(t) = S_0(t)^{\beta^*} \end{cases} \iff \begin{cases} H_0 : \beta = 1 \\ H_1 : \beta = \beta^*, \end{cases}$$

with $\beta^* \neq 1$, assuming $S_0(t)$ known. The Neyman-Pearson level α Most Powerful (MP) test, as described in the work of Lehmann on hypothesis testing ([1]), is employed in the context of survival analysis to make inference based on a single observation, denoted as t. This MP test is designed to achieve the highest statistical power among all tests at a given significance level α.

The rejection rule for this MP test is defined as follows: the test rejects the null hypothesis if and only if

$$\Lambda(t) = \frac{f_1(t)}{f_0(t)} \geq k_\alpha, \text{ with } k_\alpha : P\left(\frac{f_1(T)}{f_0(T)} \geq k_\alpha; H_0\right) = \alpha.$$

Considering the proportional hazards (PH) assumption, we explore the implications of this assumption within the context of survival analysis.

Under the PH assumption, we consider the hazard rate at any given time t as the product of a baseline hazard function $\lambda_0(t)$ and a time-independent function β^*, denoted as $\lambda_1(t) = \beta^*\lambda_0(t)$. This formulation suggests that the hazard rates for different individuals are proportional to each other over time, with the parameter β^* representing the proportional change in hazard rates. Indeed, we consider

$$\lambda_0(t) = \frac{f_0(t)}{S_0(t)}, \ \lambda_1(t) = \beta^*\lambda_0 = \beta^*\frac{f_0(t)}{S_0(t)}.$$

Also, $\lambda_1(t) = f_1(t)/S_1(t) = f_1(t)/[S_0(t)]^{\beta^*}$. As a result, it is crucial for the two forms to coincide:

$$\beta^*\frac{f_0(t)}{S_0(t)} = \frac{f_1(t)}{[S_0(t)]^{\beta^*}} \iff \frac{f_1(t)}{f_0(t)} = \beta^*\frac{[S_0(t)]^{\beta^*}}{S_0(t)} = \beta^*[S_0(t)]^{\beta^*-1}.$$

Indeed,

$$f_1(t) = -\frac{\mathrm{d}S_1(t)}{\mathrm{d}t} = -\frac{\mathrm{d}[S_0(t)]^{\beta^*}}{\mathrm{d}t} = -\beta^*[S_0(t)]^{\beta^*-1}(-f_0(t)) = \beta^*[S_0(t)]^{\beta^*-1}f_0(t),$$

that trivially gives $f_1(t)/f_0(t) = \beta^*[S_0(t)]^{\beta^*-1}$.

In the context of the Neyman-Pearson testing problem with sample size equal to one, the test statistics can be expressed as the ratio of the densities of the alternative hypothesis and the null hypothesis, denoted as $f_1(t)/f_0(t)$. Remarkably, this ratio can be further simplified as $\beta^*S_0(t)^{\beta^*-1}$.

Consequently, the rejection rule for the test can be reformulated as $\beta^* S_0(t)^{\beta^*-1} \geq k_\alpha$, where k_α represents the critical value or threshold corresponding to the chosen significance level α.

Depending on the value of the parameter β^*, we can distinguish between two distinct scenarios, indeed by considering the specific value of β^* and evaluating the test statistic against the critical value, we can determine the appropriate rejection rule and draw conclusions about the relationship between the survival times and the hypothesized survival functions. Hence, depending on whether $\beta^* > 1$ or $\beta^* < 1$, we have:

- if $\beta^* < 1$, then $\beta^* S_0(t)^{\beta^*-1} \geq k_\alpha \iff S_0(t) \leq \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} = \widetilde{k}_\alpha$, such that
  $P(\text{Reject } H_0; H_0) = P(S_0(T) \leq \tilde{k}_\alpha; H_0) = \alpha$. Since $S_0(T) \sim \text{Unif}[0,1]$, $\widetilde{k}_\alpha = \alpha$. Then the rejection region in terms of the time-to-event is $T \geq S_0^{-1}(\alpha)$.

- if $\beta^* > 1$, then $\beta^* S_0(t)^{\beta^*-1} \geq k_\alpha \iff S_0(t) \geq \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} = \widetilde{k}_\alpha$, such that
  $P(\text{Reject } H_0; H_0) = P(S_0(T) \geq \widetilde{k}_\alpha; H_0) = \alpha$. Again, by $S_0(T) \sim \text{Unif}[0,1]$, $\widetilde{k}_\alpha = 1-\alpha$ follows immediately. Again, the rejection region in terms of the time-to-event is $T \leq S_0^{-1}(1-\alpha)$.

Since $\widetilde{k}_\alpha$ does not depend on β^* except for the sign, it implies that the two tests, based on the rejection rule $\beta^* S_0(t)^{\beta^*-1} \geq \widetilde{k}_\alpha$, are uniformly most powerful (UMP) at level α for the two broader hypothesis testing scenarios, and this is explained deeper in the following lines.

The first scenario involves testing the null hypothesis $H_0 : \beta = 1$ against the alternative hypothesis $H_1 : \beta < 1$. In this case, the alternative hypothesis suggests a proportional decrease in hazard rates relative to the null hypothesis. By employing the rejection rule $\beta^* S_0(t)^{\beta^*-1} \geq \widetilde{k}_\alpha$, the test is UMP at level α for this problem. This means that among all possible tests at significance level α, this test has the highest statistical power to detect a true alternative hypothesis of $\beta < 1$.

Similarly, the second scenario involves testing the null hypothesis $H_0 : \beta = 1$ against the alternative hypothesis $H_1 : \beta > 1$. Here, the alternative hypothesis implies a proportional increase in hazard rates compared to the null hypothesis. The rejection rule $\beta^* S_0(t)^{\beta^*-1} \geq \widetilde{k}_\alpha$ provides a UMP test at level α for this problem. This indicates that among all possible tests at significance level α, this particular test has the highest statistical power to detect a true alternative hypothesis of $\beta > 1$.

By establishing the UMP property for these two wider hypothesis testing scenarios, we ensure that the respective tests are optimal in terms of statistical power, consistently achieving the highest level of sensitivity in detecting the specified alternative hypotheses.

4.2.2 Sample size greater than one

Recall the previously mentioned hypothesis system within the context of survival data without censoring:

$$\begin{cases} H_0 : S(t) = S_0(t) \\ H_1 : S(t) = S_0(t)^{\beta^*} \end{cases} \iff \begin{cases} H_0 : \lambda(t) = \lambda_0(t) \\ H_1 : \lambda(t) = \beta^* \lambda_0(t) \end{cases} \iff \begin{cases} H_0 : \beta = 1 \\ H_1 : \beta = \beta^*, \end{cases}$$

with $\beta^* \neq 1$, assuming $S_0(t)$, and equivalently $\lambda_0(t)$, known. Previously, we obtain that

$$\beta^* S_0(t)^{\beta^*-1} = \frac{f_1(t)}{f_0(t)}, \text{ recalling that } S_0(t) = e^{-\Lambda_0(t)} = e^{-\int_0^t \lambda_0(u)\mathrm{d}u}.$$

In the absence of censoring, we revisit the formulation of the two tests based on this hypothesis system. Notably, these tests focus on assessing the relationship between the survival times and the hypothesized survival functions. This result holds significant importance for the subsequent analysis. In the case of an independent and identically distributed (i.i.d.) sample denoted as $(t_1, t_2, \ldots, t_n)$, where the sample size n is greater than 1, the Neyman-Pearson MP test can be expressed as follows:

$$\Lambda(t_1, \ldots, t_n) = \prod_{i=1}^{n} \left[\frac{f_1(t_i)}{f_0(t_i)}\right] \geq k_\alpha, \text{ with } k_\alpha : P\left(\prod_{i=1}^{n} \left[\frac{f_1(T_i)}{f_0(T_i)}\right] \geq k_\alpha; H_0\right) = \alpha.$$

Once again, after performing a few calculations, we can rewrite the rejection rule way simpler. The test statistic can be expressed as $(\beta^*)^n \prod_{i=1}^{n} [S_0(t_i)]^{\beta^*-1} \geq k_\alpha$, which quantifies the combined effect of the individual survival times on the test outcome. The computations are given by:

$$\begin{aligned}
&\prod_{i=1}^{n} f_0(t_i) = \prod_{i=1}^{n} \left[\frac{f_1(t_i)}{f_0(t_i)}\right] = \prod_{i=1}^{n} \left[\beta^* S_0(t_i)^{\beta^*-1}\right] = (\beta^*)^n \prod_{i=1}^{n} [S_0(t_i)]^{\beta^*-1} \geq k_\alpha \\
&\iff \prod_{i=1}^{n} [S_0(t_i)]^{\beta^*-1} \geq \frac{k_\alpha}{(\beta^*)^n} \iff \prod_{i=1}^{n} S_0(t_i) \geq \left[\frac{k_\alpha}{(\beta^*)^n}\right]^{1/(\beta^*-1)} \\
&\iff \sum_{i=1}^{n} [\log(S_0(t_i))] \geq \log\left(\frac{k_\alpha}{(\beta^*)^n}\right)^{1/(\beta^*-1)} \\
&\iff -\sum_{i=1}^{n} [\log(S_0(t_i))] \leq -\log\left(\frac{k_\alpha}{(\beta^*)^n}\right)^{1/(\beta^*-1)} = -\frac{1}{1-\beta^*} \log\left(\frac{k_\alpha}{(\beta^*)^n}\right)
\end{aligned}$$

Depending on the value of β^*, specifically whether $\beta^* > 1$ or $\beta^* < 1$, we can distinguish between two distinct scenarios:

- if $\beta^* < 1$, then $-\sum_{i=1}^{n} [\log(S_0(t_i))] \geq -\log\left(\frac{k_\alpha}{(\beta^*)^n}\right)^{1/(\beta^*-1)} = \widetilde{k}_\alpha$,
- if $\beta^* > 1$, then $-\sum_{i=1}^{n} [\log(S_0(t_i))] \leq -\log\left(\frac{k_\alpha}{(\beta^*)^n}\right)^{1/(\beta^*-1)} = \widetilde{k}_\alpha$,

with the appropriate (different) values $\widetilde{k}_\alpha$. We define the statistic $W = W(T_1, T_2, \ldots, T_n) = -\sum_{i=1}^{n} [\log(S_0(T_i))]$, where T_i represents the i-th survival time from the i.i.d. sample. It is worth noting that under the null hypothesis, $-\log(S_0(T_i))$ follows an Exponential distribution with parameter 1, denoted as $\mathrm{Exp}(1)$. Consequently, we can establish that the statistic W, since it is the sum of exponential random variables, follows a Gamma distribution with shape parameter n and rate parameter 1, i.e., $W \sim \mathrm{Gamma}(n, 1)$.

By leveraging the known distribution of W under the null hypothesis, we can determine the threshold for rejection, denoted as $\widetilde{k}_\alpha$. Remarkably, this threshold can be easily obtained. Furthermore, since $\widetilde{k}_\alpha$ does not depend on β^* except for its sign, the rejection regions for the uniformly

most powerful (UMP) tests can be defined as, for the hypothesis testing problem $H_0 : \beta = 1$ versus $H_1 : \beta < 1$, the rejection region is given by $W \geq \text{Ga}(n,1)_{1-\alpha}$, where $\text{Ga}(n,1)_{1-\alpha}$ represents the $(1-\alpha)$th percentile of the Gamma distribution with shape parameter n and rate parameter 1. In this case, rejecting the null hypothesis indicates that the observed sample provides strong evidence supporting the alternative hypothesis of $\beta < 1$.

Conversely, for the hypothesis testing problem $H_0 : \beta = 1$ versus $H_1 : \beta > 1$, the rejection region is defined as $W \geq \text{Ga}(n,1)_{\alpha}$. Here, $\text{Ga}(n,1)_{\alpha}$ represents the α-th percentile of the gamma distribution with shape parameter n and rate parameter 1. Rejecting the null hypothesis in this case implies compelling evidence in favor of the alternative hypothesis of $\beta > 1$.

By employing the respective rejection regions based on the computed percentiles of the Gamma distribution, we can effectively determine the rejection or acceptance of the null hypothesis. These rejection regions play a crucial role in the uniformly most powerful tests, enabling us to draw robust conclusions regarding the relationship between the survival times and the hypothesized survival functions under different alternative hypotheses.

4.3 MP test for independently right-censored data

4.3.1 Sample size equal to one

In the context of right-censored survival analysis, we consider the generic subject i and observe the pair $\underline{x}_i = (x_i, \delta_i)^T$, where $x_i = \min(t_i, c_i)$ and $\delta_i = \mathbf{I}(t_i \leq c_i)$. Here, x_i represents the observed time, which is the minimum of the actual survival time t_i and the censoring time c_i. Additionally, δ_i serves as an indicator variable, taking the value of 1 if the event is observed ($t_i \leq c_i$) and 0 otherwise.

It is customary to use the notation t_i to denote the survival time and c_i to denote the independent censoring time, both measured from the same origin. This notation aids in distinguishing between the actual survival time and the censoring time for each subject.

Under the proportional hazards (PH) assumption, we can establish the same hypothesis system as before. Recall that

$$\begin{cases} H_0 : S(t) = S_0(t) \\ H_1 : S(t) = S_0(t)^{\beta^*} \end{cases} \iff \begin{cases} H_0 : \lambda(t) = \lambda_0(t) \\ H_1 : \lambda(t) = \beta^* \lambda_0(t) \end{cases} \iff \begin{cases} H_0 : \beta = \beta_0 = 1 \\ H_1 : \beta = \beta^*, \end{cases}$$

with $\beta^* \neq 1$, where we assume the known survival function $S_0(t)$, or equivalently, the known hazard function $\lambda_0(t)$.

Under the PH assumption, the Neyman-Pearson test statistics, when only one observation (x, δ) is available, is given by

$$\frac{f_1(x,\Delta)}{f_0(x,\delta)} = \frac{f_{(X,\Delta)}(x,\delta,\beta^*)}{f_{(X,\delta)}(x,\delta,\beta_0)} = \frac{f_1(x)^{\delta} S_1(x)^{1-\delta}}{f_0(x)^{\delta} S_0(x)^{1-\delta}},$$

with β_0 denote the specific value of β under the null hypothesis, which in our case coincides with one. Then, since we are in the right-censored survival setting, we have the observed time x, which

represents the minimum value between the time-to-event t and the censoring time c. Additionally, we have the indicator variable δ, which indicates whether the event has been observed.

The MP test, when only one observation (x, δ) is available, rejects the null hypothesis if and only if:

$$\Lambda(x,\delta) = \frac{f_1(x)^{\delta} S_1(x)^{1-\delta}}{f_0(x)^{\delta} S_0(x)^{1-\delta}} \geq k_\alpha, \text{ with } k_\alpha : P(\Lambda(X,\Delta) \geq k_\alpha; H_0) = \alpha.$$

By performing some simple computations, we can establish that the rejection rule can be expressed as $(\beta^*)^{\delta} S_0(x)^{\beta^*-1} \geq k_\alpha$, or equivalently, $S(x)^{\beta^*-1} \geq \left(\frac{k_\alpha}{(\beta^*)^{\delta}}\right)$. The entire computation is given by

$$\frac{f_1(x)}{f_0(x)} = \beta^* [S_0(x)]^{\beta^*-1} \Rightarrow \left[\frac{f_1(x)}{f_0(x)}\right]^{\delta} = (\beta^*)^{\delta} [S_0(x)]^{\delta(\beta^*-1)},$$

and,

$$\left[\frac{S_1(x)}{S_0(x)}\right]^{1-\delta} = \left[\frac{[S_0(x)]^{\beta^*}}{S_0(x)}\right]^{1-\delta} = \left[S_0(x)^{\beta^*-1}\right]^{1-\delta}.$$

Then,

$$\Lambda(x,\delta) = (\beta^*)^{\delta} [S_0(x)]^{\delta(\beta^*-1)} \cdot \left[S_0(x)^{\beta^*-1}\right]^{1-\delta} = (\beta^*)^{\delta} [S_0(x)]^{\beta^*-1}.$$

Building upon previous results, we can observe that the quantity $\frac{k_\alpha}{(\beta^*)^{\delta}}$ serves as a threshold for the rejection rule. Depending on the values of β^* and δ, this threshold determines the critical region in which we reject the null hypothesis.

The value of the threshold must be selected in a manner that ensures the desired significance level for the hypothesis test. Specifically, the threshold should be chosen such that the probability of observing a test statistic greater than or equal to the threshold, under the null hypothesis, is equal to the fixed significance level α. This can be expressed formally as:

$$k_\alpha : P((\beta^*)^{\Delta} [S_0(X)]^{(\beta^*-1)} \geq k_\alpha; H_0) = \alpha.$$

In other words, the threshold should be determined to achieve a desired level of significance, ensuring that the probability of falsely rejecting the null hypothesis is controlled at the specified significance level. Thus, we explore different paths.

MP test for independently right censoring for $\alpha^* < \alpha$

Firstly, we aim to solve the MP test for a significance level $\alpha^* < \alpha$, utilizing the threshold k_α. We seek the threshold k_α such that:

- if $\beta^* > 1, k_\alpha :\ S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right) = S_C^{-1}\left((1-\alpha)\cdot\left(\frac{k_\alpha}{\beta^*}\right)^{\beta^*-1}\right)$,
- if $\beta^* < 1, k_\alpha :\ S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right) = S_C^{-1}\left(\alpha \cdot k_\alpha^{\beta^*-1}\right)$,

where $S_C(c)$ is the survival function associated to the censored observations.

Since the value of β^* is fixed, we need to solve the equations (potentially using numerical methods) to determine the thresholds $k_\alpha^* = k_\alpha^*(\beta^*)$. The rejection regions for the MP test are defined as follows: for $\beta^* > 1$, the rejection region is given by $\{X \leq \gamma_1, T \leq C\} \cup \{X \leq \gamma_2, T > C\}$; while for $\beta^* < 1$, the rejection region is given by $\{X \geq \gamma_1, T \leq C\} \cup \{X \geq \gamma_2, T > C\}$. Here, we have

$$\gamma_1 = S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{\frac{1}{\beta^*-1}}\right), \text{ and } \gamma_2 = S_0^{-1}\left(k_\alpha^{\frac{1}{\beta^*-1}}\right).$$

The proof for the determination of these rejection regions is provided below.

Let us first consider the case where $\beta^* > 1$. Since $\beta^* > 1$, we have

$$\frac{k_\alpha}{\beta^*} < k_\alpha, \text{ and} \frac{1}{\beta^*-1} > 0.$$

Therefore, we can deduce that

$$\left(\frac{k_\alpha}{\beta^*}\right)^{\frac{1}{\beta^*-1}} < k_\alpha^{\frac{1}{\beta^*-1}}.$$

By applying the inverse of the baseline survival function to both sides, we obtain

$$S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{\frac{1}{\beta^*-1}}\right) > S_0^{-1}\left(k_\alpha^{\frac{1}{\beta^*-1}}\right).$$

Consequently, we can conclude that $\gamma_1 > \gamma_2$, where

$$\gamma_1 = S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{\frac{1}{\beta^*-1}}\right), \text{ and} \gamma_2 = S_0^{-1}\left(k_\alpha^{\frac{1}{\beta^*-1}}\right).$$

The computation continues fixing the probability of no rejection with threshold k_α such that

$$P_{(X,\Delta)}\left((\beta^*)^{\Delta} S_0(X)^{\beta^*-1} \leq k_\alpha; H_0\right) = 1 - \alpha.$$

This is given by

$$
\begin{aligned}
&P_{(X,\Delta)}\left((\beta^*)^{\Delta}S_0(X)^{\beta^*-1} \le k_\alpha; H_0\right) = \\
&= P_{(X,\Delta)}\left((\beta^*)^{\delta}S_0(X)^{\beta^*-1} \le k_\alpha, \delta = 1; H_0\right) + P_{(X,\Delta)}\left((\beta^*)^{\delta}S_0(X)^{\beta^*-1} \le k_\alpha, \delta = 0; H_0\right) \\
&= P\left(\beta^* S_0(X)^{\beta^*-1} \le k_\alpha, T \le C; H_0\right) + P\left(S_0(X)^{\beta^*-1} \le k_\alpha, T > C; H_0\right) \\
&= P\left(S_0(X) \le \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}, T \le C; H_0\right) + P\left(S_0(X) \le (k_\alpha)^{1/(\beta^*-1)}, T > C; H_0\right) \\
&= P\left(X \ge S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right), T \le C; H_0\right) + P\left(X \ge S_0^{-1}\left((k_\alpha)^{1/(\beta^*-1)}\right), T > C; H_0\right) \\
&= P\left(X \ge \gamma_1, T \le C; H_0\right) + P\left(X \ge \gamma_2, T > C; H_0\right) \\
&= P\left(T \ge \gamma_1, C \ge \gamma_1, T \le C; H_0\right) + P\left(T \ge \gamma_2, C \ge \gamma_2, T > C; H_0\right) \\
&= \int_{\gamma_1}^{\infty}\int_{t}^{\infty} f_T(t) f_C(c)\mathrm{d}c\mathrm{d}t + \int_{\gamma_2}^{\infty}\int_{c}^{\infty} f_C(c) f_T(t)\mathrm{d}t\mathrm{d}c \\
&= \int_{\gamma_1}^{\infty} f_T(t) S_C(t)\mathrm{d}t + \int_{\gamma_2}^{\infty} f_C(c) S_T(c)\mathrm{d}c.
\end{aligned}
$$

Since $\gamma_1 > \gamma_2$, we have that

$$
\begin{aligned}
&\int_{\gamma_1}^{\infty} f_T(t) S_C(t)\mathrm{d}t + \int_{\gamma_2}^{\infty} f_C(c) S_T(c)\mathrm{d}c > \int_{\gamma_1}^{\infty} f_T(t) S_C(t)\mathrm{d}t + \int_{\gamma_1}^{\infty} f_C(c) S_T(c)\mathrm{d}c \\
&= \int_{\gamma_1}^{\infty} \left[f_T(u) S_C(u) + f_C(u) S_T(u)\right]\mathrm{d}u = -S_T(u) S_C(u)\big|_{\gamma_1}^{\infty} = S_T(\gamma_1) S_C(\gamma_1).
\end{aligned}
$$

We set $k_\alpha^* : S_T(\gamma_1) S_C(\gamma_1) = 1 - \alpha$, so that $P(\text{Reject } H_0; H_0) = \alpha^* \le \alpha$, i.e. we control the type I error probability. For that true probability of type I error P(type I error), the test is, therefore, MP with level α^*. It is important to note that k_α^* will depend on the survival function S_C. Thus, the value of k_α^* for a given β^* can be obtained by solving the equation $S_T(\gamma_1) S_C(\gamma_1) = 1 - \alpha$, where $\gamma_1 = \gamma_1(k_\alpha)$, or

$$
\begin{aligned}
&S_0\left(S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)\right) S_C\left(S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)\right) = 1 - \alpha \\
&\iff \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} S_C\left(S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)\right) = 1 - \alpha \\
&\iff S_C\left(S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)\right) = (1-\alpha)\left(\frac{k_\alpha}{\beta^*}\right)^{\beta^*-1} \\
&\iff S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right) = S_C^{-1}\left((1-\alpha)\left(\frac{k_\alpha}{\beta^*}\right)^{\beta^*-1}\right)
\end{aligned}
$$

Now, consider the case $\beta^* < 1$. Similarly to the first case, we have now $\frac{k_\alpha}{\beta^*} > k_\alpha$, and $\frac{1}{\beta^*-1} < 0 \Rightarrow \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} > k_\alpha^{1/(\beta^*-1)} \Rightarrow \gamma_1 > \gamma_2$, like before. The rejection region is however different. Indeed,

we reject the null hypothesis in the case $H_0 \iff (\beta^*)^{\delta} S_0(x)^{\beta^*-1} \geq k_\alpha$, again with threshold $k_\alpha : P_{(X,\Delta)}\left((\beta^*)^{\Delta} S_0(X)^{\beta^*-1} \geq k_\alpha; H_0\right) = \alpha$. We now split the probability of rejection as

$$
\begin{aligned}
& P_{(X,\Delta)}\left((\beta^*)^{\Delta} S_0(X)^{\beta^*-1} \geq k_\alpha; H_0\right) = \\
& = P\left(S_0(X) \leq \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}, T \leq C; H_0\right) + P\left(S_0(X) \leq k_\alpha^{1/(\beta^*-1)}, T > C; H_0\right) \\
& = P\left(X \geq S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right), T \leq C; H_0\right) + P\left(X \geq S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right), T > C; H_0\right) \\
& = P(X \geq \gamma_1, T \leq C; H_0) + P(X \geq \gamma_2, T > C; H_0) \\
& = P(T \geq \gamma_1, C \geq \gamma_1, T \leq C; H_0) + P(T \geq \gamma_2, C \geq \gamma_2, T > C; H_0) \\
& \leq \int_{\gamma_2}^{\infty} f_T(t) S_C(t) \mathrm{d}t + \int_{\gamma_2}^{\infty} f_C(c) S_T(c) \mathrm{d}c = S_T(\gamma_2) S_C(\gamma_2).
\end{aligned}
$$

Again, set the threshold $k_\alpha^* : S_T(\gamma_2) S_C(\gamma_2) = \alpha$, such that the probability of rejection P(Reject H_0;H_0)= $\alpha^* \leq \alpha$, and again the test is MP with level α^*. Hence, the value of k_α^* is found by setting $S_T(\gamma_2) S_C(\gamma_2) = \alpha$, such that

$$
\begin{aligned}
& S_0\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) S_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) = \alpha \iff k_\alpha^{1/(\beta^*-1)} S_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) = \alpha \\
& \iff S_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) = \frac{\alpha}{k_\alpha^{1/(\beta^*-1)}} \iff S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right) = S_C^{-1}\left(\alpha \cdot k_\alpha^{\beta^*-1}\right).
\end{aligned}
$$

Alternative solution

Recall that the rejection region is $S(x)^{\beta^*-1} \geq \left(\frac{k_\alpha}{(\beta^*)^{\delta}}\right)$, under the null hypothesis. Depending on whether the indicator of having observed the event $\delta = 0$ or $\delta = 1$ the cases are split in two, $S(x) \geq (k_\alpha)^{1/(\beta^*-1)}$, and $S(x) \geq \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}$, respectively. Now, also depending on whether the value of β^* is $\beta^* > 1$ or $\beta^* < 1$, we obtain four different cases for the rejection region

- if $\beta^* > 1$ and $\delta = 1$, then $S_0(x) \geq (k_\alpha/\beta^*)^{1/(\beta^*-1)}$
- if $\beta^* > 1$ and $\delta = 0$, then $S_0(x) \geq k_\alpha^{1/(\beta^*-1)}$;
- if $\beta^* < 1$ and $\delta = 1$, then $S_0(x) \leq (k_\alpha/\beta^*)^{1/(\beta^*-1)}$
- if $\beta^* < 1$ and $\delta = 0$, then $S_0(x) \leq k_\alpha^{1/(\beta^*-1)}$;

Since β^* is fixed, these should be solved for the threshold $k_\alpha^* = k_\alpha^*(\beta^*)$. In the end, we will find out that the approximated form of threshold is $k_\alpha = \left(\frac{1-\alpha}{2}\right)^{\beta^*-1}$, and consequently the rejection regions are, depending on whether the event has been observed or not $\delta = 0$ or $\delta = 1$, the following:

- if $\delta = 0$, then $S_0(x) \geq (1-\alpha)/2 \iff x \leq S_0^{-1}((1-\alpha)/2)$;
- if $\delta = 1$, then $S_0(x) \geq (1-\alpha)/(2(\beta^*)^{1/(\beta^*-1)}) \iff x \leq S_0^{-1}\left((1-\alpha)/2(\beta^*)^{1/(1-\beta^*)}\right)$.

It should be noted that there are no additional cases or splits for different values of β^*. This will become evident as we proceed with the proof.

The rejection regions can be summarized as $\{X \leq \gamma_1, T \leq C\} \cup \{X \leq \gamma_2, T > C\}$ and $\{X \geq \gamma_1, T \leq C\} \cup \{X \geq \gamma_2, T > C\}$, respectively for $\beta^* > 1$, and $\beta^* < 1$, with $\gamma_1 = S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)$, and $\gamma_2 = S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)$. Therefore, the presence of the indicator δ leads to a splitting of the probability $P(\Lambda(X, \Delta) \geq k_\alpha)$ into two distinct regions, as illustrated in Figure 4.1.

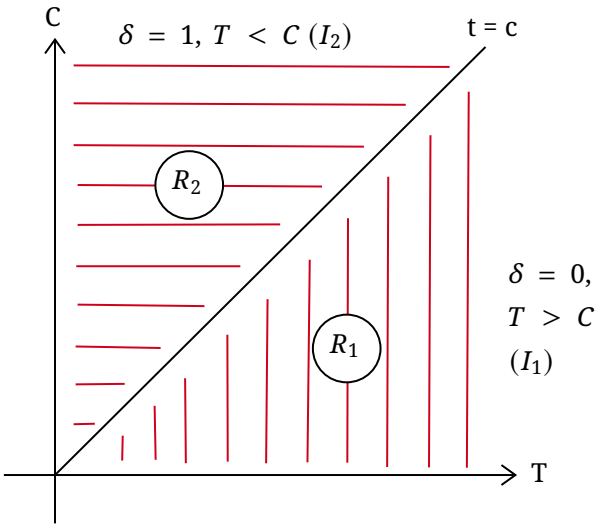

Figure 4.1: splitting into two regions the probability of rejection.

The rejection region is the union of two disjoint areas. We thus have:

$$P(S_0(X)^{\beta^*-1}(\beta^*)^{\Delta} \leq k_\alpha; H_0) = P(S_0(X)(\beta^*)^{\Delta/(\beta^*-1)} \leq k_\alpha^{1/(\beta^*-1)}; H_0)$$
$$= P(S_0(X)^{\beta^*-1}(\beta^*)^{\delta} \leq k_\alpha; H_0, \delta = 1) + P(S_0(X)^{\beta^*-1}(\beta^*)^{\delta} \leq k_\alpha; H_0, \delta = 0) = 1 - \alpha$$

that can be rewritten in terms of integrals

$$\int_0^\infty \int_0^\infty \mathbb{I}\left[S_0(\min(t,c))(\beta^*)^{\mathbb{I}(t\leq c)/(\beta^*-1)} \leq k_\alpha^{1/(\beta^*-1)}\right] f_T(t) f_C(c) \mathrm{d}t\mathrm{d}c$$
$$= \int_{R_1} \int \mathbb{I}\left[S_0(c) \leq k_\alpha^{1/(\beta^*-1)}\right] f_T(t) f_C(c) \mathrm{d}t\mathrm{d}c +$$
$$+ \int_{R_2} \int \mathbb{I}\left[S_0(t)(\beta^*)^{(\beta^*-1)} \leq k_\alpha^{1/(\beta^*-1)}\right] f_T(t) f_C(c) \mathrm{d}t\mathrm{d}c$$
$$= I_1 + I_2 = 1 - \alpha$$

where we can refer to the two integral components as I_1 and I_2, which can be solved separately.

The first integral, denoted as I_1, is expressed as follows

$$\begin{aligned}
I_1 : &\int_0^\infty \int_0^t \left[\mathbb{I}\left(S_0(c) \le k_\alpha^{1/(\beta^*-1)}\right) f_C(c)\mathrm{d}c\right] f_T(t)\mathrm{d}t \\
&= \int_0^\infty \int_0^t \left[\mathbb{I}\left(c \ge S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) f_C(c)\mathrm{d}c\right] f_T(t)\mathrm{d}t \\
&= \int_0^\infty \mathbb{I}\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right) \le c \le t\right) f_T(t)\mathrm{d}t \\
&= \int_{S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)}^\infty P\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right) \le C \le T\right) f_T(t)\mathrm{d}t \\
&= \int_{S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)}^\infty F_C(t) f_T(t)\mathrm{d}t - F_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) S_0\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) \\
&= \int_{S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)}^\infty F_C(t) f_T(t)\mathrm{d}t - F_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) k_\alpha^{1/(\beta^*-1)}.
\end{aligned}$$

Similarly the second integral is given by

$$\begin{aligned}
I_2 : &\int_0^\infty \int_0^c \left[\mathbb{I}\left(S_0(t)(\beta^*)^{1/(\beta^*-1)} \le k_\alpha^{1/(\beta^*-1)}\right) f_T(t)\mathrm{d}t\right] f_C(c)\mathrm{d}c \\
&= \int_0^\infty \int_0^c \left[\mathbb{I}\left(t \ge S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)\right) f_T(t)\mathrm{d}t\right] f_C(c)\mathrm{d}c \\
&= \int_0^\infty \int_{S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)}^c P\left(S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right) \le T \le C\right) f_T(t) f_C(c)\mathrm{d}t\mathrm{d}c \\
&= \int_0^\infty \left[F_T(c) - F_T\left(S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)\right)\right] f_C(c)\mathrm{d}c \\
&= \int_0^\infty F_T(c) f_C(c)\mathrm{d}c - F_T\left(S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)\right).
\end{aligned}$$

Therefore, the sum of the two components, denoted as I_1 and I_2, can be expressed as follows:

$$\begin{aligned}
1-\alpha = I_1 + I_2 = &\int_{S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)}^\infty F_C(t) f_T(t)\mathrm{d}t - F_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) k_\alpha^{1/(\beta^*-1)} \\
&+ \int_0^\infty F_T(c) f_C(c)\mathrm{d}c - F_T\left(S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)\right).
\end{aligned}$$

Alternatively, we can express the second region as follows:

$$\begin{aligned}
I_2 : &\int_0^\infty \mathbb{I}\left[t \ge S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)\right] \int_t^\infty f_C(c)\mathrm{d}c f_T(t)\mathrm{d}t \\
&= S_0\left(S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)\right) - \int_{S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)}^\infty F_C(t) f_T(t)\mathrm{d}t,
\end{aligned}$$

By combining the expressions for I_1 and I_2, the sum of the two regions can be expressed as follows

$$1-\alpha = I_1 + I_2 = \int_{S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)}^{\infty} F_C(t) f_T(t)\mathrm{d}t - \int_{S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)}^{\infty} F_C(t) f_T(t)\mathrm{d}t + \\ + F_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) k_\alpha^{1/(\beta^*-1)} + \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}.$$

To obtain the precise form of k_α, we aim to derive a closed-form solution. For this purpose, we rely on an approximation of the censoring distribution function approaching to one, which allows us to simplify the expression. Hence, we extract the approximate form of the threshold k_α by noting that if the distribution function $\lim_{t\to\infty} F_C(t) = 1$, also the distribution function $F_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right)$ can be approximated by one since $S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)$ is large. Indeed, for $\beta^* > 1$, $k_\alpha > k_\alpha/\beta^* \Rightarrow S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right) > S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)$, and

$$\begin{aligned} 1-\alpha &= S_0\left[S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right] - S_0\left[S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)\right] + k_\alpha^{1/(\beta^*-1)} + \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} \\ &= k_\alpha^{1/(\beta^*-1)} - \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} + k_\alpha^{1/(\beta^*-1)} + \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} = 2k_\alpha^{1/(\beta^*-1)} \\ \iff k_\alpha &= \left(\frac{1-\alpha}{2}\right)^{(\beta^*-1)} \end{aligned} \tag{4.1}$$

We can conclude that the probability of :

$$P\left(S_0(X)(\beta^*)^{\delta/(\beta^*-1)} \le \frac{1-\alpha}{2}; H_0\right) = 1-\alpha$$
$$\iff P\left(S_0(X)(\beta^*)^{\delta/(\beta^*-1)} \ge \frac{1-\alpha}{2}; H_0\right) = \alpha.$$

Therefore, the approximated rule to reject the null hypothesis for the MP test with $\beta^* > 1$ is given by:

$$S_0(x)(\beta^*)^{\delta/(\beta^*-1)} \ge k_\alpha^{1/(\beta^*-1)} = \left[\left(\frac{1-\alpha}{2}\right)^{(\beta^*-1)}\right]^{1/(\beta^*-1)} = \frac{1-\alpha}{2}.$$

Equivalently we have for the case with $\beta^* < 1$, $k_\alpha < k_\alpha/\beta^* \Rightarrow k_\alpha^{1/(\beta^*-1)} > \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} \Rightarrow S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right) > S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)$, and

$$\begin{aligned} 1-\alpha =& S_0\left[S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right] - S_0\left[S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)\right] + k_\alpha^{1/(\beta^*-1)} + \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} \\ =& k_\alpha^{1/(\beta^*-1)} - \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} + k_\alpha^{1/(\beta^*-1)} + \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} \\ =& 2k_\alpha^{1/(\beta^*-1)} \Rightarrow k_\alpha = \left(\frac{1-\alpha}{2}\right)^{(\beta^*-1)} \end{aligned}$$

Hence, the threshold k_α has the same approximated value also for the case $\beta^* < 1$.

We provide few comments on this peculiar rejection regions. When $\delta = 0$, the observed time coincides with the censoring time, i.e., $X = C$. In this case, the rejection region is defined as $C \leq S_0^{-1}\left(\frac{1-\alpha}{2}\right)$, which is referred to as the region R_1 in Figure 4.2. This region is determined under the regularity condition that the inverse of the survival function exists, ensuring the existence of $S_0^{-1}\left(\frac{1-\alpha}{2}\right)$. Similarly, when $\delta = 1$, it implies that the observed time coincides with the time-to-event, i.e., $X = T$. In this scenario, the rejection region is given by $T \leq \frac{S_0^{-1}(1-\alpha)}{(2(\beta^*)^{(\beta^*-1)})}$, which is referred to as the region R_2 in Figure 4.2. It is worth noting that when $\beta^* > 1$, we have $\frac{(1-\alpha)}{(2(\beta^*)^{(\beta^*-1)})} > \frac{(1-\alpha)}{2}$, which implies $S_0^{-1}\left(\frac{1-\alpha}{2(\beta^*)^{(\beta^*-1)}}\right) < S_0^{-1}\left((1-\alpha)/2\right)$. Consequently, we could rewrite the rejection region as $T \leq \frac{S_0^{-1}(1-\alpha)}{2}$, as this region includes the original one. The peculiar shape of the rejection region may be attributed to the approximation of the cumulative distribution function in the presence of censoring. Specifically, the shape of the rejection region R_2 might not be immediately intuitive. Additionally, it is noteworthy that the rejection region under censoring is wider compared to the case without censoring, indicating that the presence of censored cases leads to easier rejection of the null hypothesis.

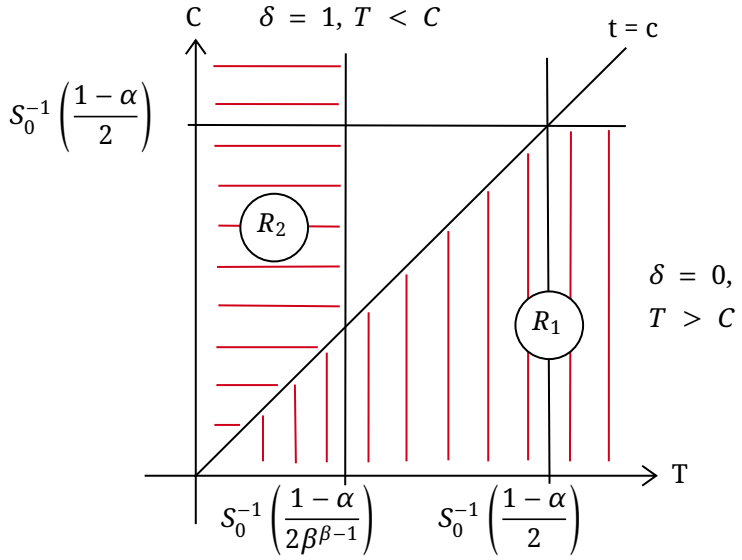

Figure 4.2: the approximated rejection region of the MP test.

An approximation of β^* to one

Recall the previously computed approximation of the censoring distribution function to one, as denoted in formula 4.1. Now, we focus on the specific case where $\beta^* > 1$ and β^* is approximately equal to 1. We call $\gamma_1 = S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)$, and $\gamma_2 = S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)$. If $\beta^* > 1$, we have

$$
\begin{aligned}
&S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right) - S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right) \geq 0 \\
&\int_{S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)}^{S_0^{-1}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right)} F_C(t) f_T(t)\mathrm{d}t \simeq (\gamma_1 - \gamma_2) F_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) f_T\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) \\
&= \gamma_3.
\end{aligned}
$$

Notice that we can see these approximation also as the integral on the area given by the base multiplied by the height in the easiest point, which coincides to the point $S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)$. The probability of rejection, following the aforementioned approximation where $F_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) \approx 1$, is given by

$$
\begin{aligned}
&I_1 + I_2 = \gamma_3 + F_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) k_\alpha^{1/(\beta^*-1)} + \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} = \\
&= F_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right)\left((\gamma_1 - \gamma_2) f_T\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) + k_\alpha^{1/(\beta^*-1)}\right) + \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}
\end{aligned}
$$

where we can approximate $(\gamma_1 - \gamma_2)$ as the derivative of the baseline survival function $\frac{\partial}{\partial x} S_0^{-1}(x)$, in the point $x = k_\alpha^{1/(\beta^*-1)}$, using a Taylor expansion. This approximation holds when β^* is close to 1, as it causes γ_1 to approach γ_2. Therefore, we obtain the following expression:

$$
\begin{aligned}
&\left(\frac{\partial}{\partial x} S_0^{-1}(x)\Big|_{k_\alpha^{1/(\beta^*-1)}}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} - k_\alpha^{1/(\beta^*-1)}\right) f_T\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) + k_\alpha^{1/(\beta^*-1)}\right) \cdot \\
&\cdot F_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) + \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} \simeq 1 - \alpha
\end{aligned}
$$

where

$$
\frac{\partial}{\partial x} S_0^{-1}(x)\Big|_{t:S_0(t)=x} = \frac{\partial}{\partial t}(1 - F_0)^{-1}(t)\Big|_{t:F_0(t)=1-x} = -\frac{1}{f_0(t)}
$$

and,

$$
\frac{\partial}{\partial x} S_0^{-1}(x)\Big|_{t:S_0(t)=x} \simeq \frac{-1}{f_0\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right)} = \frac{-1}{f_T\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right)}.
$$

given that $\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} - k_\alpha^{1/(\beta^*-1)} \leq 0$. Thus, we have

$$\left(\frac{-1}{f_T\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right)}\left(\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} - k_\alpha^{1/(\beta^*-1)}\right) f_T\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) + k_\alpha^{1/(\beta^*-1)}\right) \cdot$$
$$\cdot F_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) + \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} =$$
$$= F_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right)\left[2k_\alpha^{1/(\beta^*-1)} - \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)}\right] + \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} = 1-\alpha$$

Furthermore, approximating the censoring distribution function to one, we recall that if $F_C(\cdot) = 1$ at a singular point, then $F_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) \approx 1$. In the case of $\beta^* > 1$ with $\beta^* \approx 1$, we have:

$$F_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right)\left(2 - \left(\frac{1}{\beta^*}\right)^{1/(\beta^*-1)}\right) + \left(\frac{1}{\beta^*}\right)^{1/(\beta^*-1)} = \left(\frac{1-\alpha}{k_\alpha}\right)^{1/(\beta^*-1)}$$
$$\text{If } \beta^* \downarrow 1^+ \iff (1-\beta^*) \downarrow 0^+ \iff \frac{1}{\beta^*-1} \uparrow +\infty \iff k_\alpha^{1/(\beta^*-1)} \in [0,1] \to 0$$
$$\iff S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right) \uparrow +\infty \iff F_C\left(S_0^{-1}\left(k_\alpha^{1/(\beta^*-1)}\right)\right) \approx 1.$$

Hence again,

$$2k_\alpha^{1/(\beta^*-1)} - \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} + \left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} = 2k_\alpha^{1/(\beta^*-1)} = 1-\alpha \iff k_\alpha = \left(\frac{1-\alpha}{2}\right)^{\beta^*-1}$$

Thus, the conclusion is that for $\beta^* \approx 1$ and $\beta^* > 1$, we have $\left(\frac{k_\alpha}{\beta^*}\right)^{1/(\beta^*-1)} \approx k_\alpha^{1/(\beta^*-1)}$, which gives us a form with β^* entering only in the exponent.

Third approximation

Following the binary split of the rejection region, we have two thresholds that now we denote as k_0 and k_1, which define the region according to whether the observed time coincides to the censoring time c or the survival time t. The rejection region can be rewritten as

$$\alpha = \pi \int_0^{k_0} \frac{f_C(x)S_T(x)}{\pi}\mathrm{d}x + (1-\pi)\int_0^{k_1} \frac{f_T(x)S_C(x)}{(1-\pi)}\mathrm{d}x.$$

Thus, the probability of type I error is given by

$$\begin{aligned}
&P(\text{Reject}; H_0) = P((X,\Delta) \in R; H_0) = \\
&P(X \in R_0, \Delta = 0; H_0) + P(X \in R_1, \Delta = 1; H_0) = \\
&P(\Delta = 0; H_0)P(X \in R_0 | \Delta = 0; H_0) + P(\Delta = 1; H_0)P(X \in R_1 | \Delta = 1; H_0) = \\
&\pi P(X \in R_0 | \Delta = 0; H_0) + (1-\pi)P(X \in R_1 | \Delta = 1; H_0) = \alpha
\end{aligned} \tag{4.2}$$

with the probability of not experiencing the event under null hypothesis $\pi = P(\Delta = 0; H_0)$, and the rejection region split into two different regions $R = (X \in R_0; \Delta = 0) \cup (X \in R_1; \Delta = 1)$. We call $p_0 = P(X \in R_0 | \Delta = 0; H_0)$, and $p_1 = P(X \in R_1 | \Delta = 1; H_0)$ the probabilities to belong to the two rejection regions conditional to one of the two value of the indicator of having observed the event. Formula 4.2 can then be written as $(1-\pi)p_0 + \pi p_1$, so that $\pi p_0 + (1-\pi)p_1 = \alpha \iff \pi p_0 + p_1 - \pi p_1 = \alpha$. A closed form of the MP test can be achieved restricting to the case where the two probabilities of rejection coincide $p_0 = p_1 \iff p_0 = \alpha$. Then we have

$$P(X \leq k_0 | \Delta = 0; H_0) = \int_0^{k_0} \frac{f_C(x) S_T(x)}{\pi} \mathrm{d}x = \alpha.$$

Hence,

$$\int_0^{k_0} \frac{f_C(x) S_T(x)}{\pi} \mathrm{d}x = \alpha \iff k_0 : \alpha P(\Delta = 0)$$

We give an example about the two threshold MP test: let T follow an Exponential distribution with parameter λ_0, and C follow an exponential distribution with parameter λ_C. We have:

$$\begin{aligned}
&\alpha P(\Delta = 0) = \int_0^{k_0} \lambda_C \mathrm{e}^{(-x\lambda_C)} \mathrm{e}^{(-\lambda_0 x)} \mathrm{d}x \\
&\iff \alpha P(\Delta = 0) = \lambda_C \int_0^{k_0} \mathrm{e}^{(-(\lambda_0 + \lambda_C)x)} \mathrm{d}x \\
&\iff \alpha P(\Delta = 0) = \frac{\lambda_C}{\lambda_C + \lambda_0} \int_0^{k_0} (\lambda_C + \lambda_0) \mathrm{e}^{(-(\lambda_0 + \lambda_C)x)} \mathrm{d}x
\end{aligned}$$

where we call the density function $f_Q(q) = (\lambda_C + \lambda_0)\exp(-(\lambda_0 + \lambda_C)q)$ thus the random variable is distributed as an Exponential $Q \sim \mathrm{Exp}(\lambda_0 + \lambda_C)$. Thus we have

$$\begin{aligned}
&\alpha P(\Delta = 0) = \frac{\lambda_C}{\lambda_C + \lambda_0} F_Q(k_0) \iff \alpha P(\Delta = 0) = \frac{\lambda_C}{\lambda_C + \lambda_0} \left(1 - \mathrm{e}^{-k_0(\lambda_0 + \lambda_C)}\right) \\
&\iff \frac{\alpha P(\Delta = 0)(\lambda_C + \lambda_0)}{\lambda_C} = 1 - \mathrm{e}^{-k_0(\lambda_0 + \lambda_C)} \\
&\iff \mathrm{e}^{-k_0(\lambda_0 + \lambda_C)} = 1 - \alpha P(\Delta = 0) \frac{(\lambda_C + \lambda_0)}{\lambda_C} \\
&\iff -k_0(\lambda_C + \lambda_0) = \log\left(1 - \alpha P(\Delta = 0) \frac{(\lambda_C + \lambda_0)}{\lambda_C}\right) \\
&\iff k_0 = \frac{\log\left(1 - \alpha P(\Delta = 0) \frac{(\lambda_C + \lambda_0)}{\lambda_C}\right)}{\lambda_C + \lambda_0}.
\end{aligned}$$

Similarly, the second threshold is given by

$$k_1 = \frac{\log\left(1 - \alpha P(\Delta = 1) \frac{(\lambda_C + \lambda_0)}{\lambda_C}\right)}{\lambda_C + \lambda_0}.$$

With the two probabilities of rejection not equal $p_0 \neq p_1$ the two rejection thresholds are

$$k_0 = \frac{\log\left(1 - p_0 P(\Delta = 0)\frac{(\lambda_C + \lambda_0)}{\lambda_C}\right)}{\lambda_C + \lambda_0}$$
$$k_1 = \frac{\log\left(1 - p_1 P(\Delta = 1)\frac{(\lambda_C + \lambda_0)}{\lambda_C}\right)}{\lambda_C + \lambda_0}$$

where the values of $k_0,\ k_1$ are such that all the condition set at the beginning yield.

4.3.2 Sample size greater than one

Following the same hypothesis system

$$\begin{cases} H_0 : S(t) = S_0(t) \\ H_1 : S(t) = S_0(t)^{\beta^*} \end{cases} \iff \begin{cases} H_0 : \lambda(t) = \lambda_0(t) \\ H_1 : \lambda(t) = \beta^* \lambda_0(t) \end{cases} \iff \begin{cases} H_0 : \beta = \beta_0 = 1 \\ H_1 : \beta = \beta^*, \end{cases}$$

with $\beta^* \neq 1$, assuming $S_0(t)$, and equivalently $\lambda_0(t)$, known. The rejection rule of the MP test, when the sample size is $n > 1$, is given by:

$$\Lambda(\underline{x}, \underline{\delta}) = \frac{L_{\underline{X},\underline{\Delta}}(\underline{x}, \underline{\delta}|H_1)}{L_{\underline{X},\underline{\Delta}}(\underline{x}, \underline{\delta}|H_0)} = \prod_{i=1}^{n} \left[(\beta^*)^{\delta_i} S_0(x_i)^{\beta^* - 1}\right] \geq k_\alpha;$$
$$k_\alpha : P(\Lambda(\underline{X}, \underline{\Delta}) \geq k_\alpha; H_0) = \alpha,$$

with $\underline{x} = (x_1, \ldots, x_n)^T$ and $\underline{\delta} = (\delta_1, \ldots, \delta_n)^T$. No simpler description of the rejection region is currently available for this case. However, the MP test statistics reduces to

$$\begin{aligned} \frac{L_{\underline{X},\underline{\Delta}}(\underline{x}, \underline{\delta}|H_1)}{L_{\underline{X},\underline{\Delta}}(\underline{x}, \underline{\delta}|H_0)} &= \frac{\prod_{i=1}^{n} \left\{f_1(x_i)^{\delta_i} S_C(x_i)^{\delta_i} f_C(x_i)^{1-\delta_i} S_1(x_i)^{1-\delta_i}\right\}}{\prod_{i=1}^{n} \left\{f_T(x_i)^{\delta_i} S_C(x_i)^{\delta_i} f_C(x_i)^{1-\delta_i} S_T(x_i)^{1-\delta_i}\right\}} \\ &= \frac{\prod_{i=1}^{n} \left[\frac{f_1(x_i)}{S_1(x_i)}\right]^{\delta_i} S_1(x_i)}{\prod_{i=1}^{n} \left[\frac{f_T(x_i)}{S_T(x_i)}\right]^{\delta_i} S_T(x_i)} \\ &= \frac{\prod_{i=1}^{n} [\beta^* \lambda_0(x_i)]^{\delta_i} S_0(x_i)^{\beta^*}}{\prod_{i=1}^{n} \lambda_0(x_i)^{\delta_i} S_0(x_i)} = \prod_{i=1}^{n} \left[(\beta^*)^{\delta_i} S_0(x_i)^{\beta^* - 1}\right]. \end{aligned}$$

It is easily seen that the test statistics depends on $(\underline{X}, \underline{\Delta})$ only through the quantities $(\sum_{i=1}^{n} \log(S_0(X_i)), \sum_{i=1}^{n} \Delta_i)^T$. The joint distribution of the bivariate test statistics is needed to obtain the threshold k_α, and thus the implementable form of the test. Also, note that the test statistics reduces to $(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} \Delta_i)^T$ up to a known constant, the sufficient statistic that appears in the maximum likelihood estimator of the parameter λ when the sample of time-to-event is distributed following an Exponential, i.e. $T_1, \ldots, T_n \sim \text{Exp}(\lambda)$. Indeed, for the exponential model, the survival function has shape $S_0(t; \lambda) = \mathrm{e}^{-\lambda_0 t}$, thus we have $\sum_{i=1}^{n} \log(S_0(x_i)) = -\lambda_0 \sum_{i=1}^{n} x_i$, with the baseline hazard function λ_0 known.

4.4 Discussion

As discussed in Section 4.3, the bivariate test statistics for the case of right-censored data depends only on the joint distribution of $(\underline{X}, \underline{\Delta})$ through the transformed variables $(\sum_{i=1}^{n} \log(S_0(X_i)), \sum_{i=1}^{n} \Delta_i)^T$. Obtaining the implementable form of the test requires knowledge of this joint distribution.

The fact that both $\sum_{i=1}^{n} \log(S_0(X_i))$ and $\sum_{i=1}^{n} \Delta_i$ are essential in the context of censored data is not surprising. The value of X_i depends on the realization of Δ_i, and the observation of an event is directly related to the value of X_i. This interdependence indicates that the mechanism of partial observation in censored data cannot be ignored.

To illustrate this point, let us consider the non-parametric estimation of the survival function. It is not appropriate to simply discard the observations for which $\Delta_i = 0$ and estimate $S(t)$ using the remaining observed data with the empirical survival function $\widehat{S}^*(t) = \frac{\sum_{i=1}^{n} \Delta_i \cdot \mathbb{I}(X_i \geq t)}{\sum_{i=1}^{n} \Delta_i}$ without additional adjustments. The Kaplan-Meier estimator, for example, accounts for this dependence and provides a proper estimation method. It is crucial to acknowledge that the consistency of such an estimator for $S(t)$ holds only in the case of a censoring random variable that is degenerate to infinity with probability one. The proof for this statement can be found in Appendix D.1.

Although the test so far can be applied in a real case scenario to identify the survival of one subject with or without censoring, or of a sample without censoring, meaning that we observe the event of interest of all subject, our intention is to investigate the application of the MP test to survival data with right censoring. Later we would like to run simulation studies, and subsequently, extend the analysis to available data.

References

[1] Lehmann, E. L., Romano, J. P., and Casella, G. (2005). *Testing statistical hypotheses*, volume 3. Springer.

[2] Tayob, N., Stingo, F., Do, K.-A., Lok, A. S., and Feng, Z. (2018). A bayesian screening approach for hepatocellular carcinoma using multiple longitudinal biomarkers. *Biometrics*, 74(1):249–259.

Appendix A

A.1 Lehmann family of cure-rate models and proportional hazards

We want to explore the meaning of the proportional hazard (PH) assumption when $P(T = +\infty) = p > 0$. Indeed, in the cure-rate model we have $S_T(t) = p \cdot 1 + (1-p)S_0(t)$. This means that there is a proportion p of the population that will never experience the event, no matter how long they will live, while the proportion $(1-p)$ of the population experience the event according to the survival function $S_0(t)$.

Let us compute the hazard function of a cure-rate model:

$$\begin{aligned}
\lambda(t) &\stackrel{\Delta}{=} \lim_{\Delta_t \downarrow 0} \frac{1}{\Delta_t} P(T \in [t, t+\Delta_t] \mid T \geq t) = \lim_{\Delta_t \downarrow 0} \frac{1}{\Delta_t} \frac{P(T \in [t, t+\Delta_t])}{P(T \geq t)} \\
&= \lim_{\Delta_t \downarrow 0} \frac{1}{\Delta_t} \frac{1}{S_T(t)} [F_T(t+\Delta_t) - F_T(t)] = \frac{1}{S_T(t)} \lim_{\Delta_t \downarrow 0} \frac{1}{\Delta_t} [S_T(t) - S_T(t+\Delta_t)] \\
&= \frac{1}{S_T(t)} \lim_{\Delta_t \downarrow 0} \frac{1}{\Delta_t} [p + (1-p)S_0(t) - (p + (1-p)S_0(t+\Delta_t))] \\
&= \frac{(1-p)}{S_T(t)} \lim_{\Delta_t \downarrow 0} \frac{1}{\Delta_t} [F_T(t+\Delta_t) - F_T(t)] = \frac{(1-p)}{S_T(t)} f_0(t) = \frac{(1-p)f_0(t)}{p + (1-p)S_0(t)}.
\end{aligned}$$

The PH assumption would require that for two groups A and B one had

$$\frac{(1-p_B)f_{0B}(t)}{p_B + (1-p_B)S_{0B}(t)} = \lambda_B(t) = \beta\lambda_A(t) = \beta \cdot \frac{(1-p_A)f_{0A}(t)}{p_A + (1-p_A)S_{0A}(t)}. \tag{A}$$

Notice that this model assumption is different from the traditional PH assumption on cases: $\lambda_{0B}(t) = \beta \cdot \lambda_{0A}(t)$. Let us assume that $p_A > 0$, $p_B > 0$, and we study the following cases:

- (I) if $p_A = p_B = 0$ then we recover the usual PH: $\lambda_{0B}(t) = \beta \cdot \lambda_{0A}(t)$;

- (II) if $p_A = p_B) p > 0$ and $S_{0A}(t) \neq S_{0B}(t)$ we obtain

$$\frac{\cancel{(1-p)}f_{0B}(t)}{p + (1-p)S_{0B}(t)} = \beta \cdot \frac{\cancel{(1-p)}f_{0A}(t)}{p + (1-p)S_{0A}(t)} \tag{B}$$

- (III) if $p_A, p_B > 0$, $p_A \neq p_B$, and $S_{0A}(t) = S_{0B}(t)$ we obtain

$$\frac{(1-p_B)f_0(t)}{p_B + (1-p_B)S_0(t)} = \beta \cdot \frac{(1-p_A)f_0(t)}{p_A + (1-p_A)S_0(t)}$$
$$(1-p_B)\cdot(p_A + (1-p_A)S_0(t)) = \beta(1-p_A)(p_B + (1-p_B)S_0(t))$$
$$(1-p_B)(1-p_A)S_0(t) + p_A(1-p_B) = \beta(1-p_A)(1-p_B)S_0(t) + \beta p_B(1-p_A)$$
$$S_0(t) = \frac{\beta(1-p_A)p_B - p_A(1-p_B)}{(1-\beta)(1-p_A)(1-p_B)},$$

  and the only case in which the survival function is a constant is the degenerate case $S_0(t) = 1\ \forall t$, or $P(T = +\infty) = 1$;

- (IV) if $p_A = p_B = p > 0$ and $S_{0A}(t) = S_{0B}(t)$ then, $\beta \equiv 1$;

- (V) if $p_A = p_B = p > 0$, $p_A \neq p_B$ and $S_{0A}(t) \neq S_{0B}(t)$ then, we obtain the general form in (A) above.

Thus, the interpretation of the PH assumption for cure-rate models is more complicated as we must distinguish all the different cases and respect the conditions.

A.2 Harrell's index for one-dimensional multiplicative Gamma frailty models

Harrell's c-index is used to quantify the concordance between a risk index and right censored survival times. Here, we study what the reference population values for the index are when the multiplicative frailty random variable R is distributed as a Gamma(θ, θ) random variable using the shape-rate parametrization, i.e. with density

$$g_R(r;\theta) = \frac{1}{\Gamma(\theta)}\theta^\theta r^{\theta-1}\mathrm{e}^{-\theta r},\ \ r \geq 0,\ \theta > 0$$

so that $\mathbb{E}(R) = \theta/\theta = 1$, $\mathrm{var}(R) = \theta/\theta^2 = 1/\theta$, and

$$\mathrm{MGF}_R(t) = E\left(\mathrm{e}^{tR}\right) = \left(1 - \frac{t}{\theta}\right)^{-\theta},\ \text{for } t < \theta.$$

Let us recall here that in the multiplicative frailty survival model the conditional (on the frailty) hazard function has the form $\lambda(t \mid r) = r\,\lambda_0(t)$ for some baseline hazard function $\lambda_0(t)$. This coincides with the proportional hazards assumption, or equivalent with assuming the Lehmann structure $S(t \mid r) = [S_0(t)]^r$, with $S_0(t)$ the baseline survival function, and with corresponding conditional density function $f_{T|R}(t \mid r)$.

We now define the population Harrell's index as the probability of concordance between the observed survival times and the frailty terms of the subjects. The term "population" here refers to the fact that we consider the true frailty terms r, and refer to the survival distribution without reference to right censoring.

The population Harrell's index C can be defined as

$$\begin{aligned} C &= P\left(\{R_1 < R_2\} \cap \{T_1 > T_2\}\right) + P(\{R_2 < X_1\} \cap \{T_2 > T_1\}) \\ &= 2P(\{R_1 < R_2\} \cap \{T_1 > T_2\}), \end{aligned}$$

where (R_1, T_1) and (R_2, T_2) are i.i.d. with the same joint bivariate density function $f_{(R,T)}(r,t) = g_R(r) f_{T|R}(t \mid r)$.

Proposition For the gamma multiplicative frailty model, the value of the population Harrell's index $C(\theta)$ does not depend on the baseline survival function $S_0(t)$. Depending on the value of θ, its value varies in the range $[0.25, 0.5]$. Lastly, the exact value of $C(\theta)$ is given by

$$C(\theta) = \mathbb{E}(Y \mid Y > 0.5) = S_{Y^*}\left(\frac{1}{2}\right),$$

with $S_{Y^*}(t)$ the survival function of the random variable Y^*, which is distributed as $\text{Beta}(\theta + 1, \theta)$.

Proof We have

$$\begin{aligned} C &= E\left[\mathbb{I}(R_1 < R_2)\mathbb{I}(T_1 > T_2)\right] \qquad \text{(A.1)} \\ &= \int_{\mathbb{R}^+}\int_{\mathbb{R}^+}\int_{\mathbb{R}^+}\int_{\mathbb{R}^+} \mathbb{I}(r_1 < r_2)\mathbb{I}(t_1 > t_2) f_{R,T}(r_1, t_1) f_{R,T}(r_2, t_2) \mathrm{d}t_1 \mathrm{d}t_2 \mathrm{d}r_1 \mathrm{d}r_2 \\ &= \int_{\mathbb{R}^+}\int_{\mathbb{R}^+} \mathbb{I}(r_1 < r_2) \left[\int_{\mathbb{R}^+}\int_{\mathbb{R}^+} \mathbb{I}(t_1 > t_2) f_{T|R}(t_1 \mid r_1) f_{T|R}(t_2 \mid r_2) \mathrm{d}t_1 \mathrm{d}t_2\right] f_R(r_1) f_R(r_2) \mathrm{d}r_1 \mathrm{d}r_2. \end{aligned}$$

Now, the inner double integral can be written as

$$\begin{aligned} &\int_{\mathbb{R}^+}\int_{\mathbb{R}^+} \mathbb{I}(t_1 > t_2) f_{T|R}(t_1 \mid r_1) f_{T|R}(t_2 \mid r_2) \mathrm{d}t_1 \mathrm{d}t_2 = \int_{\mathbb{R}^+} \left[\int_{\mathbb{R}^+} \mathbb{I}(t_1 > t_2) f_{T|R}(t_1 \mid r_1) \mathrm{d}t_1\right] f_{T|R}(t_2 \mid r_2) \mathrm{d}t_2 \\ &= \int_{\mathbb{R}^+} \left[\int_{t_2}^{\infty} f_{T|R}(t_1 \mid r_1) \mathrm{d}t_1\right] f_{T|R}(t_2 \mid r_2) \mathrm{d}t_2 = \int_{\mathbb{R}^+} S_{T|R}(t_2 \mid r_1) f_{T|R}(t_2 \mid r_2) \mathrm{d}t_2 \\ &= \int_{\mathbb{R}^+} [S_0(t_2)]^{r_1} f_{T|R}(t_2 \mid r_2) \mathrm{d}t_2 = \mathbb{E}\left\{[S_0(V)]^{r_1}\right\} \qquad \text{(A.2)} \end{aligned}$$

with $V \sim F_{T|R}(t \mid r_2)$. Now, it is well known that for any absolutely continuous random variable V, the transformed random variables $F_V(V)$ and $S_V(V)$ are both $\text{Unif}[0,1]$-distributed. In this case, the survival function of the random variable V is $[S_0(t)]^{r_2}$, and therefore $[S_0(V)]^{r_2} \sim \text{Unif}[0,1]$. Since

$$[S_0(t)]^{r_1} = \left\{[S_0(t)]^{r_2}\right\}^{\frac{r_1}{r_2}} \iff [S_0(V)]^{r_1} = \left\{[S_0(t)]^{r_2}\right\}^{\frac{r_1}{r_2}} \sim [U]^{\frac{r_1}{r_2}},$$

where, $U \sim \text{Unif}[0,1]$. The expression in (A.2) is then equal to

$$\mathbb{E}\left(U^{\frac{r_1}{r_2}}\right) = \mathbb{E}\left[\mathrm{e}^{\log\left(U^{\frac{r_1}{r_2}}\right)}\right] = \mathbb{E}\left[\mathrm{e}^{\left(\frac{r_1}{r_2}\right)\log(U)}\right] = \mathbb{E}\left[\mathrm{e}^{\left(-\frac{r_1}{r_2}\right)(-\log(U))}\right]. \qquad \text{(A.3)}$$

It is well known that if $U \sim \text{Unif}[0,1]$, then $-\log(U) \sim \text{Exp}(1)$. Hence the expected value in (A.3) coincides with the moment generating function of the Exp(1) random variable evaluated as the (negative) value $-r_1/r_2$. Since $MGF_{\text{Exp}(1)}(t) = (1-t)^{-1}$, and it is defined for $t < 1$, it is always defined at $-r_1/r_2$ for any positive r_1 and r_2. Hence the inner integral in (A.1) is equal to

$$MGF_{\text{Exp}(1)}\left(-\frac{r_1}{r_2}\right) = \frac{1}{1+-\frac{r_1}{r_2}} = \frac{r_2}{r_1+r_2}.$$

Putting this all together, the Population Harrell's index C is therefore equal to

$$\begin{aligned} C(\theta) &= 2\int_{\mathbb{R}^+}\int_{\mathbb{R}^+} \mathbb{I}(r_1 < r_2)\frac{r_2}{r_1+r_2} f_R(r_1) f_R(r_2)\mathrm{d}r_1\mathrm{d}r_2 \\ &= 2P(R_1 < R_2)\int_{\mathbb{R}^+}\int_{\mathbb{R}^+} \frac{r_2}{r_1+r_2} f_{(R_1,R_2)|R_1<R_2}(r_1,r_2) r_1 \mathrm{d}r_2 = \mathbb{E}\left(\frac{R_2}{R_1+R_2} \mid R_1 < R_2\right) \end{aligned} \tag{A.4}$$

where the 0.5 term in the front comes from the fact that R_1 and R_2 are i.i.d.
Let us now focus on the random variable $Y = \dfrac{R_2}{R_1+R_2}$. Without conditioning, it is easy to check that $Y \sim \text{Beta}(\theta,\theta)$, i.e.

$$f_Y(y;\theta) = \frac{1}{Be(\theta,\theta)} y^{\theta-1}(1-y)^{\theta-1}\mathbb{I}(0<y<1), \text{ with } Be(\theta,\theta) = \frac{\Gamma(\theta)\Gamma(\theta)}{\Gamma(2\theta)}.$$

For any positive value of θ, the density function $f_Y(y;\theta)$ is trivially symmetric in $(0,1)$ around 0.5, and it is such that $\mathbb{E}(Y) = 0.5$.

Let us now turn to the conditioning in the expected value in (A.4). Easily, $R_1 < R_2 \Leftrightarrow Y > 0.5$, so that (A.4) becomes

$$C(\theta) \quad = \quad \mathbb{E}(Y \mid Y > 0.5). \tag{A.5}$$

This allows one to conclude immediately that, since Y takes values in $[0,1]$, conditionally on $Y > 0.5$ it takes values in $[0.5,1]$, and therefore the index C must take values in $[0.25, 0.5]$, regardless of the value of θ.

Recall the density function of Y. By its noted symmetry around 0.5 one can write

$$\begin{aligned} f_{Y|Y>0.5}(y) &= \frac{f_Y(y)\mathbb{I}\left(y > \frac{1}{2}\right)}{\frac{1}{2}} = \frac{2}{Be(\theta,\theta)} y^{\theta-1}(1-y)^{\theta-1}\,\mathbb{I}(0<y<1)\,\mathbb{I}\left(y>\frac{1}{2}\right) \\ &= \frac{2}{Be(\theta,\theta)} y^{\theta-1}(1-y)^{\theta-1}\,\mathbb{I}\left(\frac{1}{2}<y<1\right). \end{aligned}$$

Then,

$$\begin{aligned} C(\theta) =& \mathbb{E}\left(Y \mid Y > \frac{1}{2}\right) = \int_{\mathbb{R}^+} y f_Y(y;\theta)\mathbb{I}\mathrm{d}y = \int_{\frac{1}{2}}^1 \left[y\frac{2}{Be(\theta,\theta)} y^{\theta-1}(1-y)^{\theta-1}\right]\mathrm{d}y \\ =& \frac{2}{Be(\theta,\theta)}\int_{\frac{1}{2}}^1 \left[y^{(\theta+1)-1}(1-y)^{\theta-1}\right]\mathrm{d}y = 2\frac{Be(\theta+1,\theta)}{Be(\theta,\theta)}\int_{\frac{1}{2}}^1 f_{Y*}(y;\theta)\mathrm{d}y \end{aligned}$$

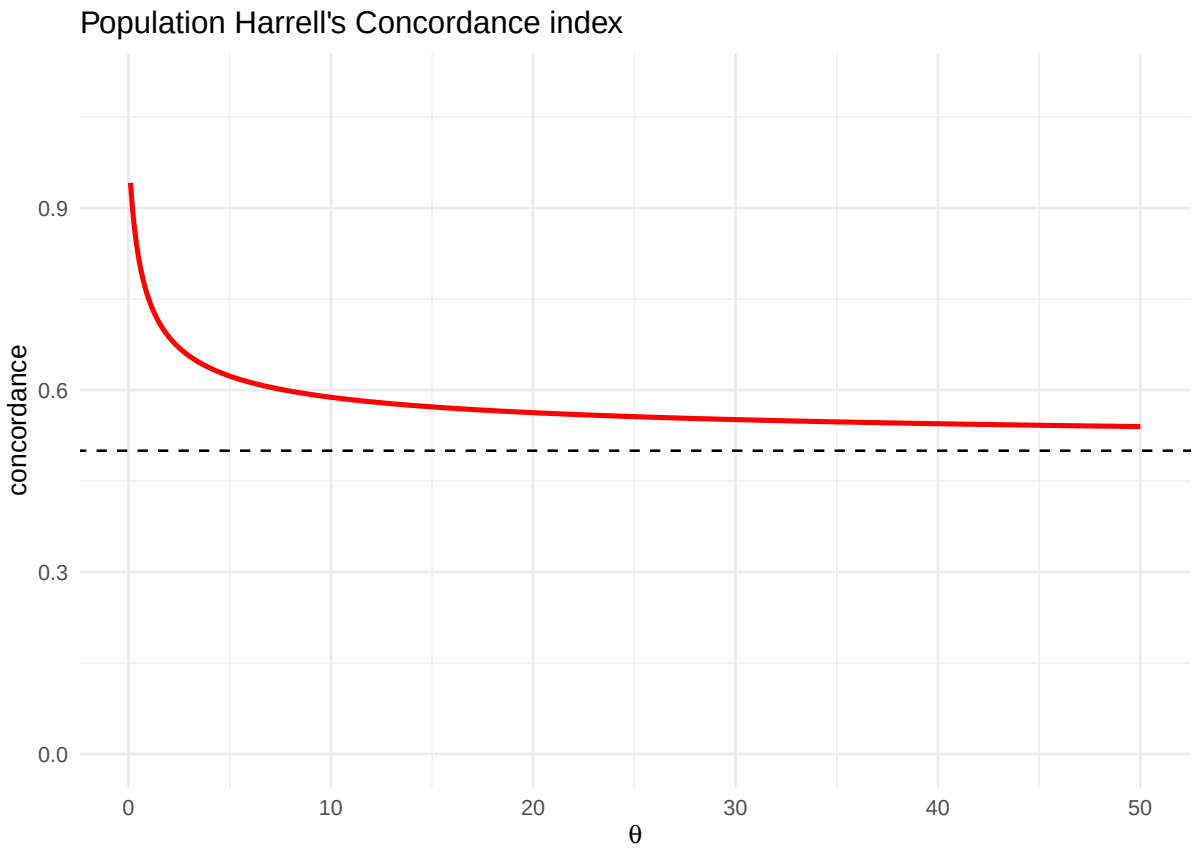

Figure A.1: population Harrell's Index in the multiplicative gamma frailty model, for varying θ.

where $Y^* \sim \text{Beta}(\theta+1, \theta)$. Given the definition of the Beta function, this is finally

$$\begin{aligned} C(\theta) &= 2\frac{\Gamma(\theta+1)\Gamma(\theta)}{\Gamma(2\theta+1)}\frac{\Gamma(2\theta)}{\Gamma(\theta)\Gamma(\theta)}\int_{\frac{1}{2}}^{1} f_{Y*}(y;\theta)\mathrm{d}y \\ &= 2\frac{\theta\,\Gamma(\theta)}{2\theta\,\Gamma(2\theta)}\frac{\Gamma(2\theta)}{\Gamma(\theta)}\int_{\frac{1}{2}}^{1} f_{Y*}(y;\theta)\mathrm{d}y = \frac{1}{2}\int_{\frac{1}{2}}^{1} f_{Y*}(y;\theta)\mathrm{d}y = P\left(Y^* > \frac{1}{2}\right) = S_{Y^*}\left(\frac{1}{2}\right). \end{aligned}$$

Figure A.1 shows the value of the index for varying θ.

Note 1 Clearly, $C(\theta)$ does not depend on the baseline survival function $S_0(t)$. Depending on the value of θ, its value varies in the range $[0.25, 0.5]$. Also, the expression A.5 allows one to conclude immediately that, since Y takes values in $[0,1]$, conditionally on $Y > 0.5$ it takes values in $[0.5, 1]$, and therefore the $C(\theta)$ must also take values in the same range.

Note 2 The shape of the density function $f_Y(y;\theta)$ is such that as $\theta \to 0$, it concentrates the probability mass more and more (and symmetrically) near zero and one. As a consequence, the conditional distribution of $Y \mid Y > 0.5$ becomes concentrated at one. $C(\theta)$ being the expected value of that random variable, it should indeed be expected to tend to one. Indeed, as $\theta \to 0$ the frailty random

variable is Gamma(θ, θ) with mean one and variance that tends to infinity, and the $[S_0(t)]^r$ survival functions corresponding to the widely moving values r will be very far indeed, and so will be the survival times that they produce. Conversely, as $\theta \to \infty$, $f_Y(y; \theta)$ becomes concentrated more and more around the mean, 0.5. Indeed, indexing θ in the natural numbers, as $\theta \to \infty$ one has that Y converges in probability to 0.5. This is the case of no frailty, since X becomes degenerate at one. In this case the fact that Harrell's index will tend to 0.5 is also, by a simple symmetry argument, to be expected.

Note 3 Given the closed form of $C(\theta)$, if the MLE $\widehat{\theta}$ of theta is available, then the approximate large sample distribution of $C(\widehat{\theta})$ can be obtained by a simple application of the delta method.

A.3 Data construction

There are several useful registries available, including:

1. “barn_clean” - A registry specifically focused on information about children.
2. “biomor_clean” - A registry specifically dedicated to biological mothers.
3. “cancer_clean” - A registry that focuses on cancer cases.
4. “death_clean” - A registry that records deaths.
5. “demografi_clean” - A general registry containing information about the subject’s birthday and sex.
6. “migrationer_clean” - A registry that tracks emigration and immigration events.
7. “syskon_clean” - A registry that documents sibling relationships.

Initially, our process involves cleaning these registries to extract or modify various variables, as also the survival couple, that are necessary for our analysis. We then proceed to merge all of these registries together.

A.3.1 Cleaning Registries

We will provide a description of the contents within each registry, outline our requirements, and explain our approach to managing each one effectively.

- BARN REGISTRY: this dataset contains information about the biological children of each subject. It includes the date of birth and sex of each child. By analyzing this dataset, we can derive valuable covariates such as the parity for each woman (whether she has at least one child or not), the number of biological children for each woman, and the age at which each child was born. Most importantly, we can identify the age at which each woman had her first child, which serves as an indicative risk factor for breast cancer.
- BIOMOR REGISTRY: the dataset provides information about the biological mother of each subject. There are no duplicate entries for the mothers, as each subject has a unique biological mother. Consequently, each row in the dataset corresponds to a different subject, ensuring that no ties or repetitions exist.
- CANCER REGISTRY: the dataset contains information on various tumor types, and our first step is to select cases related to breast cancer. Breast cancer cases are identified using different classifications, including International Classification of Diseases (ICD) codes. Specifically, "ICD7" identifies breast cancer through codes starting with 170 [3], "ICD9" with 174 [5], and both "ICDO10" and "ICDO3" with C050 [6, 4]. Since there are no missing entries in the ICD7

variable, we can use it to select breast cancer cases. Once breast cancer cases are selected, we examine whether the classification in other variables (ICDO3, ICD9, and ICDO10) aligns with the ICD7 classification. There are five special cases that require attention: one woman is identified as having a malignant neoplasm in genital organs (ICD9 = 1844), two with a non-specified neoplasm (ICD9 = 1991, ICDO10 = C809), one with a placenta neoplasm (ICDO3 = C589), and one with lymph nodes of axilla or arm neoplasm (ICDO3 = C773) [4, 5, 6]. Considering the nature of breast cancer, which is part of the genitalia, and the fact that a breast cancer can spread to lymph nodes, we retain four of these special cases. However, we remove the case of placenta neoplasm in ICDO3 because it does not coincide with a breast cancer case in ICD7. The dataset contains multiple rows for each subject, as breast cancer can develop over different visits. Since our focus is on the first occurrence of invasive cancer, we can identify the time of breast cancer by the visit when it was initially detected. Another issue to address is the presence of ties where multiple rows have the same ID number and visit date. This implies that a patient may have undergone multiple visits on the same day, all of which have been recorded. To handle this, we keep the first appearance of each visit in the dataset, as this information is randomly recorded. Thus, we retain the date of breast cancer onset, which remains consistent across visits on the same day. We lose the specific details of each visit, which may vary even within the same day of analysis. However, this level of detail is not relevant to our current analysis. In addition to the visit date, we are interested in the invasive nature of the cancer (`BEN`). Specifically, we consider ductal carcinoma in situ (DCIS) cases identified by `BEN=3` as censoring events. This approach allows us to focus solely on invasive breast cancer as the event of interest. It is important to notice that if a DCIS is detected followed by an invasive breast cancer, the time to breast cancer is censored at the DCIS diagnosis. This is because the treatment of DCIS can modify the natural progression of breast cancer and introduce bias. Hence, we exclude time-to-breast cancer records that occur after a DCIS diagnosis. After removing the ties, we are left with 20,216 cases of DCIS out of a total of 265,756, accounting for 7% of all subjects.

- DEMOGRAPHY REGISTRY: the data includes details about the subject's birthdate and date of death (`FODELSEMAN`, `DodDatum`). Each subject is uniquely identified by their ID number (`LOPNR`), and there are no duplicate entries. Consequently, each row corresponds to a distinct subject.

- MIGRATION REGISTRY: the dataset contains information about the emigration dates (`Unt`) and immigration dates (`Int`) of individuals, which can occur multiple times as they move in and out of Sweden. Each move is recorded in a separate row, indicating whether it is an emigration or immigration and when it occurred. This means that a subject's ID number may be repeated multiple times based on their movements. To eliminate this repetition, the data structure has been transformed from a multiple rows format to a multiple columns format. Several columns have been created to accommodate the maximum number of subject movements within the dataset. Each row's information has been transferred to the correspond-

ing column, identified by the movement time and nature (emigration or immigration). As a result, each subject now has one row and multiple columns indicating the dates of their movements. In cases where there have been no movements, the entry is filled with a missing value. For instance, the column `Unt1` represents the first (indicated by "1") emigration (indicated by "Unt"), while `Int2` represents the first immigration (indicated by "Int") back to Sweden after a recorded emigration, but it is the second movement in total for that subject (indicated by "2").

- SYSKON REGISTRY: the dataset includes information about sibling relationships on a one-to-one basis. However, there are ties within the data, as each subject is repeated based on the number of siblings they have. The dataset provides details about the nature of the sibling relationship, such as full siblings, half-siblings from the mother's side (with the same father or missing information on the father), and half-siblings from the father's side (with the same mother or missing information on the mother). For our analysis, we focus solely on females, as they are of interest to us. Among the female subjects, we specifically select full sisters for several reasons. Firstly, the genetic factor is considered to be stronger than the environmental factor in the development of breast cancer. This means that the same childhood environment does not necessarily cause familial aggregation of breast cancer [1]. Secondly, by selecting full sisters, we indirectly obtain genetic information from the father that can only be shared if the sisters have the same father. To address the issue of ties and ensure each subject has a unique row, we transform the data from a multiple rows structure to a multiple columns structure. We create a set of columns that correspond to the largest number of sisters within a family in the dataset, which is twelve (there is one family with thirteen daughters). These columns are labeled as "Sister 1" through "Sister 12", and each column contains the ID number of a sister if available. If a subject does not have a particular sister, a missing value is recorded in the corresponding column. Since we only consider subjects with sisters in this dataset, each individual has at least one sister, guaranteeing that the column for the first sister contains complete information regarding the ID numbers. However, as we move to subsequent sisters, the number of missing values in the columns naturally increases due to the varying number of sisters among the subjects.

A.3.2 Building Survival Variables

- OBSERVED DATE: the observed date is the first event among diagnosis time, emigration time, death, DCIS diagnosis, or the end of the follow-up is considered. For all subjects, the last visit for invasive breast cancer is taken as the end of follow-up, which is set as December 30, 2016. Information on death and emigration is available until December 31, 2018. However, there will be censored observations with a probability of one, meaning that no information about breast cancer diagnosis is provided, so the follow-up ends two years prior to the date of December 31, 2018. In some cases, both information on diagnosis and emigration, or

diagnosis and death, are available. However, generally, there is not both information on emigration and death, as one event completely excludes the possibility of observing and recording the other event. Several subjects do not have information on any of these three events, and the date of the end of the study is considered as their observed date. The format of the date variable is set as "YYYY-MM-DD" (where Y stands for year, M stands for month, and D stands for day). Most observations in the raw dataset already have this format, but in some cases, manual modifications were made to retain information and ensure proper handling. There are nine special cases that required adjustments:

  - three cases were originally on February 29, but due to an error in the software R, they have been modified to February 28.
  - One case had "0" as the date, and it has been replaced with the end of the follow-up date.
  - Four dates were in the format "YYYY-00-00", so they have been adjusted by adding the middle day, 15th, of the middle month of the year, which is June. The computational process involves adding +615 to the format "YYYY0000" and then transforming it into "YYYY-MM-DD".
  - Similarly, for the only case without a day in the format "YYYY-MM-00", the day has been adjusted to coincide with the 15th of that month. Another possible approach could be to randomly sample a day of the month between the 1st and the 28th (or 30th, or 31st) based on the respective month.

- OBSERVED TIME: the follow-up length refers to the time duration between entering the risk group (typically at birth) and the dropout date. In this context, the follow-up begins at birth and ends on the observed date, as described in the previous bullet point. The observed time represents the number of days between birth and the first event that occurs, which can be either the onset of breast cancer or a censoring event. The observed time can be easily adjusted and converted from days to months or years, depending on the desired time unit for analysis. We transform it from days to years, dividing it for 365.25, adjusting for the leap year.

- DELTA: this variable serves as an indicator for observing the onset of invasive breast cancer. Initially, a vector is created in the cancer registry, assigning a value of one to cases of invasive breast cancer. During the merging process between the Cancer Registry and the Demographic Registry, the variable "delta" will contain missing values for subjects not included in the Cancer Registry. In order to handle these missing values, they are replaced with zeros. This is because subjects without a recorded breast cancer onset have not experienced the event yet, and a value of zero indicates the absence of the event.

- FH: the family history indicator of breast cancer among first-degree female relatives (mother and sisters) is determined based on their breast cancer cases occurring before the subject's

observed date. Main subjects are the women included in the dataset as observations (rows), while other subjects represent relatives (columns). If all family members are missing, the family history is also missing. Otherwise, with at least one relative, the family history is assigned a value of zero or one. If only censoring cases exist before the subject's observed time, the family history is negative; otherwise, it is positive. Ties in observed times lead to a negative family history, requiring breast cancer cases to occur at least one day before the subject's analysis time to be considered.

A.4 Reliability of the Cure-Rate assumption

We investigate the reliability of the tail of the Kaplan-Meier curve. Although the Swedish meticulous data collection process should alleviate any doubts, we decide to conduct an additional verification using the Swedish life tables, which are accessible online [2]. Specifically, we examine the ultracentenary women from 2017 and 2018, which correspond to the years immediately after the end of the follow-up into the dataset. Furthermore, in those years we have information about death and emigration. In the Swedish life tables, we focus on the column for hazard function and the column for survival function.

Age	Hazard 2017	Survival 2017	Hazard 2018	Survival 2018
100	0.36682	0.50	0.36518	0.50
101	0.39208	0.50	0.39067	0.50
102	0.41667	0.50	0.41549	0.50
103	0.44033	0.50	0.43938	0.50
104	0.46284	0.50	0.46212	0.50
105	0.48404	0.50	0.48353	0.50
106	0.50382	0.50	0.50349	0.50
107	0.52209	0.50	0.52192	0.50
108	0.53883	0.50	0.53880	0.50
109	0.55405	0.50	0.55414	0.50

The life table must be replicated by our dataset. If the tail is reliable, this means that ultracentenary subjects are truly still alive by the end of the study and they have their censoring event (either death or emigration) just after the end of the follow-up. If the recorded data are correct we expect to find a similar proportion of how many died in those years. Trivially, we only consider censoring until the 2018 because we do not have any information further.

The hazard of dying after being centenary is in mean 0.5315. We compute the proportion of deaths in the biennal out of the total of alive people. The result is 0.5037, that is very similar to the life table one. From these results we can claim that the cure-rate assumption holds, because of the presence of old alive women that do not experience the event breast cancer eventually, no matter how long they will live. For completeness of results, we report the table of frequency of

ultracentenary women in the dataset resulting in the higher number of women which are 101 years old (32.7%).

Age	Frequency
100	0.227
101	0.327
102	0.150
103	0.103
104	0.050
105	0.030
106	0.040
107	0.018
108	0.013
109	0.013
110	0.008
111	0.005
112	0.008
113	0.005
114	0.003

Appendix B

B.1 Two-latent-classes Cure-Rate model

Families are divided into two risk groups characterized by different hazard functions. We make the first assumption on the proportionality of the hazard functions, i.e. $\lambda_1(t) = \lambda_0(t)\alpha$. The baseline hazard $\lambda_0(t)$ needs to be fixed (e.g. Exponential, Weibull, Gamma).

We make a second assumption on the cure-rate. There is a proportion of all population that is cured i.e. that will never experience the event of interest no matter how long they live. This phenomenon is observed in both risk groups, but with different magnitudes. Indeed, in the high-risk group, the cure-rate is lower than in the low-risk group. We can appreciate this last assumption in representing the hazard function $\lambda(t) = f(t)/S(t)$ as composed of a mixture of two distributions in picture B.1: the closest distribution to zero is representing the observed events, while the other, around e.g. $T = 1000$, is the area of the events we never observe.

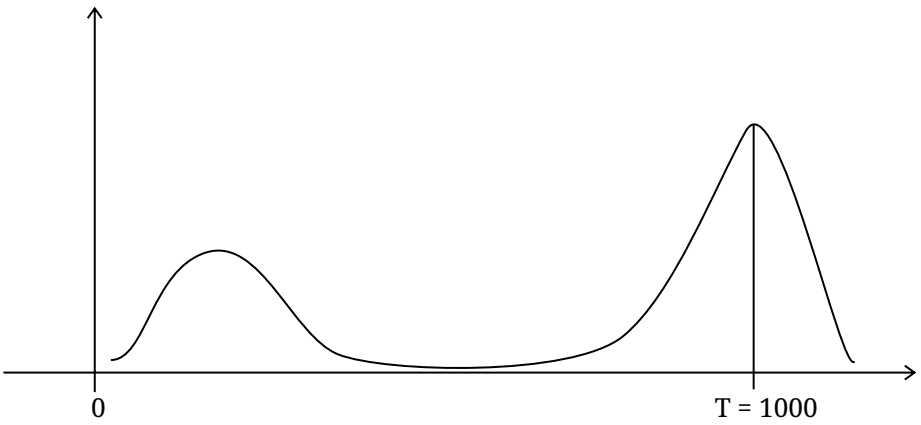

Figure B.1: cure-rate mixture density function.

The density function can be written as a mixture of two components, i.e.: $f_T(t) = pf_1(t) + (1 - p)f_2(t)$ where $(1 - p)$ is the cure-rate (then, p is the proportion of subjects experiencing the event). Similarly, the survival function is $S_T(t) = P(T \geq t) = pP(T_1 \geq t) + (1 - p)P(T_2 \geq t)$. Hence, the hazard functions for the low and high-risk groups are respectively:

$$\lambda_0(t) = \frac{pf_1(t) + (1-p)f_2(t)}{pS_1(t) + (1-p)S_2(t)} \qquad \lambda_1(t) = \alpha \left[\frac{pf_1(t) + (1-p)f_2(t)}{pS_1(t) + (1-p)S_2(t)} \right]$$

We can simplify the formulae by just giving a few considerations on the density and survival functions. We divide the time axis into two intervals, i.e.: I_1 and I_2 (see e.g. Figure B.2).

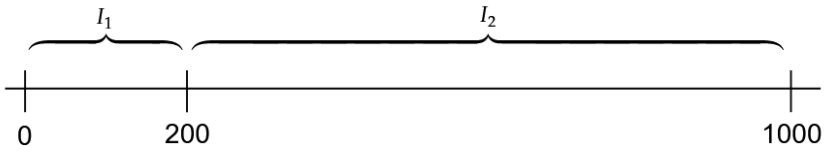

Figure B.2: two time intervals of the total follow-up.

In interval I_1 the density function and the survival function of the right tail population assume values $f_2(t) = 0, S_2(t) = 1$, then the overall quantities are $f_T(t) = pf_1(t), S_T(t) = pS_1(t) + (1 - p)$. While in interval I_2 the density and survival functions from population one are null: $f_1(t) = 0$, $S_1(t) = 0$. Then, the overall quantities are $f_T(t) = (1 - p)f_2(t), S_T(t) = (1 - p)S_2(t)$. Hence, the hazard functions take different values according to the interval:

$$\lambda_L(t) = \begin{cases} \frac{pf_1(t)}{pS_1(t)+(1-p)} & I_1 : [0 - 200]; \\ \lambda_2(t) = \frac{f_2(t)}{S_2(t)} & I_2 : [200 - 1000]. \end{cases} \qquad \lambda_H(t) = \alpha\lambda_L(t) = \begin{cases} \alpha \cdot \frac{pf_1(t)}{pS_1(t)+(1-p)} & I_1; \\ \alpha \cdot \lambda_2(t) = \alpha\frac{f_2(t)}{S_2(t)} & I_2, \end{cases}$$

where the subscript “L” and “H” stay for low-risk and high-risk group hazard function.

Notice that the cure-rate structure does not hold for the high-risk group hazard function. Similarly, where the cure-rate structure is introduced the PH assumption does not hold, i.e.:

$$\alpha \frac{pf_1(t)}{pS_1(t) + (1-p)} \neq \frac{\widetilde{p}f_1(t)}{\widetilde{p}S_1(t) + (1-\widetilde{p})}$$

where at left only the PH assumption holds, and at right only the cure-rate structure holds. We can handle this issue by finding a relation between α and $\widetilde{p}$. We start from the rate of the two

quantities:

$$\alpha\frac{\frac{pf_1(t)}{pS_1(t)+(1-p)}}{\frac{\widetilde{p}f_1(t)}{\widetilde{p}S_1(t)+(1-\widetilde{p})}}=0 \iff \alpha p f_1(t)(\widetilde{p}S_1(t)+(1-\widetilde{p}))=\widetilde{p}f_1(t)(pS_1(t)+(1-p))$$

$$\iff \widetilde{p}(S_1(t)-1)\alpha p f_1(t)+\alpha p f_1(t)-\widetilde{p}f_1(t)(pS_1(t)+(1-p))=0$$

$$\iff \widetilde{p}=0\frac{-\alpha p f_1(t)}{(S_1(t)-1)\alpha p f_1(t)-f_1(t)(pS_1(t)+(1-p))}$$

$$\iff \frac{1}{\widetilde{p}}=0-\frac{(S_1(t)-1)\alpha p f_1(t)}{\alpha p f_1(t)}+\frac{f_1(t)(pS_1(t)+(1-p))}{\alpha p f_1(t)}=-(S_1(t)-1)+\frac{pS_1(t)+(1-p)}{\alpha p}$$

$$=F_1(t)+\frac{S_1(t)}{\alpha}+\frac{1-p}{p}\frac{1}{\alpha}=F_1(t)+\frac{1}{\alpha}\left[S_1(t)+\frac{1-p}{p}\right] \iff \frac{1}{\widetilde{p}}=1-S_1(t)\left[\frac{1}{\alpha}-1\right]+\frac{1}{\alpha}\left(\frac{1-p}{p}\right)$$

Hence, say we use $S_1(t)=1/2$ because we approximate $\int S_1(t)\mathrm{d}t=\mathbb{E}(T_1)$. Then, the computation brings to

$$\frac{1}{\widetilde{p}}=1-\frac{1}{2}\left(\frac{1}{\alpha}-1\right)+\frac{1}{\alpha}\left(\frac{1-p}{p}\right)=1+\frac{1}{2}-\frac{1}{2}\frac{1}{\alpha}+\frac{1}{\alpha}\left(\frac{1}{p}-1\right)=\frac{3}{2}+\frac{1}{\alpha}\left(\frac{1}{p}-\frac{3}{2}\right)$$

$$\iff \widetilde{p}=\left[\frac{3}{2}+\frac{1}{\alpha}\left(\frac{1}{p}-\frac{3}{2}\right)\right]^{-1}.$$

Approximately, the difference between the values of $1/\widetilde{p}$ at $S_1(t)=0$ and $S_1(t)=1$ is $|1/\alpha-1|$. We prove this through the following computation:

$$\frac{1}{\widetilde{p}}=\begin{cases}1+\frac{1}{\alpha}\left(\frac{1-p}{p}\right) & S_1(t)=0\\ 1-\left(\frac{1}{\alpha}-1\right)+\frac{1}{\alpha}\left(\frac{1-p}{p}\right) & S_1(t)=1\end{cases}$$

$$1+\frac{1}{\alpha}\left(\frac{1-p}{p}\right)-1+\left(\frac{1}{\alpha}-1\right)-\frac{1}{\alpha}\left(\frac{1-p}{p}\right)=\frac{1}{\alpha}-1\Rightarrow\frac{1}{\widetilde{p}(0)}-\frac{1}{\widetilde{p}(1)}\leq|\frac{1}{\alpha}-1|$$

We want to explore the link between $\widehat{\alpha}$ and $\widetilde{p}=\left[\frac{3}{2}+\frac{1}{\alpha}\left(\frac{1}{p}-\frac{3}{2}\right)\right]^{-1}$. We would like to use a PH estimate when the PH assumption does not hold. We apply another approach to have a formula of the $\widetilde{p}$ starting from the definition with cure-rate structure, i.e.:

$$p_L=P(T\leq 150|L)=1-S_T(150|L)=1-\mathrm{e}^{-\int_0^{150}\lambda_L(u)\mathrm{d}u}$$

$$p_H=P(T\leq 150|H)=1-S_T(150|H)=1-\mathrm{e}^{-\alpha\int_0^{150}\lambda_L(u)\mathrm{d}u}=1-[S_T(150|L)]^{\alpha}$$

$$p_H=1-(1-p_L)^{\alpha}$$

where 150 is just an arbitrary end of the study. An interesting question is about the difference between

$$1-(1-p_L)^{\alpha}\text{ vs. }\left[\frac{3}{2}+\frac{1}{\alpha}\left(\frac{1}{p_L}-\frac{3}{2}\right)\right]^{-1}.$$

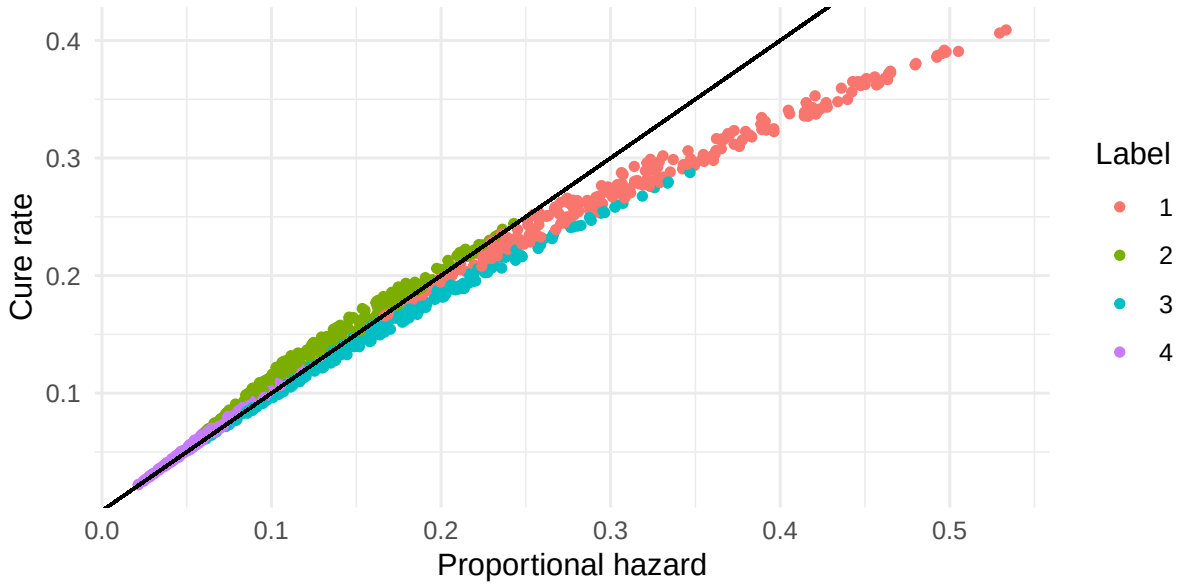

Figure B.3: proportional hazard vs cure-rate

where at left the PH holds, and at right the cure-rate holds. In Figure B.3 we study the relation between the aforementioned quantities. We can notice that they are similar in each of the four considered cases (in black the 45 degrees bisector), given different thresholds for p_L and α, respectively at 0.15 and 0 and the labels are 1) $\alpha > 0$ and $p_L > 0.15$, 2) $\alpha < 0$ and $p_L > 0.15$, 3) $\alpha > 0$ and $p_L < 0.15$, 4) $\alpha < 0$ and $p_L < 0.15$.

B.2 The observed data likelihood for the Lehmann cure-rate model

Recall the usual notation $X = \min(T, U)$ and $\Delta = \mathbb{I}(T \leq U)$ for the bivariate observed random variable arising from survival data T independently right censored by the random variable U. It is easy to check that when (T, U) has joint density function $f_{(T,U)}(t, u) = f_T(t) f_U(u)$, the distribution of (X, Δ) is proportional to $[f_T(x)]^{\delta} [S_T(x)]^{1-\delta}$. When the pairs (T_i, U_i) are *i.i.d.* for $i = 1, \ldots, n$, the product of such terms represents the observed data likelihood that can be maximized to learn about the distribution $F_T(t)$ ($F_U(u)$ is typically not of interest). In the following, U is still assumed to be independent of T.

Now, consider the cure-rate model $S(t) = p + (1 - p)\widetilde{S}(t)$, with $\widetilde{f}(t)$ the (proper) conditional density function of the time-to-event random variable for the "cases," i.e. for those subjects who will eventually experience the event of interest. Notice that T has a positive probability p of being equal to $+\infty$ (or to an extremely large number, as this model is sometimes also described). For ease of notation, below we write "∞" for "$+\infty$."

Proposition A For the cure-rate model $S(t) = p + (1-p)\widetilde{S}(t)$, the contribution to the observed data

likelihood by one observation (X, Δ) is proportional to the quantity $\left[(1-p)\widetilde{f}(x)\right]^{\delta}\left[p+(1-p)\widetilde{S}(x)\right]^{1-\delta}$.

Proof. Consider the probability $P(X \in [x, x+\Delta x), \Delta = 0)$ for a non-negative, finite x. Define the set $A_T(x) = \{(t,u) \in \mathbb{R}^+ \times \mathbb{R}^+ : u \in [x, x+\Delta x), t \geq u\}$. We have

$$\begin{aligned} &P(X \in [x, x+\Delta x), \Delta = 0) = P((T,U) \in A_T(x)) \\ &= P((T,U) \in A_T(x) \mid T < \infty)P(T < \infty) + P((T,U) \in A_T(x) \mid T = \infty)P(T = \infty). \end{aligned}$$

It is easy to check that conditionally on $T < \infty$, T and U remain independent, with joint density function $f_{(T,U)|T<\infty}(t,u) = \widetilde{f}(t)f_U(u)$ on $\mathbb{R}^+ \times \mathbb{R}^+$. Therefore,

$$\begin{aligned} &P(X \in [x, x+\Delta x), \Delta = 0) = (1-p)\int_x^{x+\Delta x}\int_u^{\infty} \widetilde{f}(t)f_U(u)\mathrm{d}t\,\mathrm{d}u + p\int_x^{x+\Delta x} f_U(u)\mathrm{d}u \\ &= (1-p)\int_x^{x+\Delta x} f_U(u)\widetilde{S}(u)\mathrm{d}u + p\int_x^{x+\Delta x} f_U(u)\mathrm{d}u \approx (1-p)(\Delta x)f_U(x)\widetilde{S}(x) + p\,(\Delta x)f_U(x) \\ &= (\Delta x)\left[f_U(x)\left(p+(1-p)\widetilde{S}(x)\right)\right]. \end{aligned}$$

Now, define the set $A_U(x) = \{(t,u) \in \mathbb{R}^+ \times \mathbb{R}^+ : t \in [x, x+\Delta x), u \geq t\}$. For $\Delta = 1$, slightly different steps yield

$$\begin{aligned} &P(X \in [x, x+\Delta x), \Delta = 1) = P((T,U) \in A_U(x)) \\ &= P((T,U) \in A_U(x) \mid T < \infty)P(T < \infty) + P((T,U) \in A_U(x) \mid T = \infty)P(T = \infty) \\ &= (1-p)\int_x^{x+\Delta x}\int_t^{\infty} f_U(u)\widetilde{f}(t)\mathrm{d}u, \mathrm{d}t + 0 = (1-p)\int_x^{x+\Delta x} \widetilde{f}(t)S_U(t)\mathrm{d}t \approx (\Delta x)(1-p)\widetilde{f}(x)S_U(x). \end{aligned}$$

Dividing by Δx, letting $\Delta x \to 0$, and writing the two terms in compact form produces the contribution

$$\begin{aligned} &\left[f_U(x)\left(p+(1-p)\widetilde{S}(x)\right)\right]^{\delta}\left[(1-p)\widetilde{f}(x)S_U(x)\right]^{1-\delta} \\ &= \left(p+(1-p)\widetilde{S}(x)\right)^{\delta}\left[(1-p)\widetilde{f}(x)\right]^{1-\delta}[f_U(x)]^{\delta}[S_U(x)]^{1-\delta} \propto \left(p+(1-p)\widetilde{S}(x)\right)^{\delta}\left[(1-p)\widetilde{f}(x)\right]^{1-\delta}. \end{aligned}$$

□

Let us now turn to the Lehmann cure-rate model structure.

Proposition B If $S_r(t) = S(t \mid R = r) = \left[p+(1-p)\widetilde{S}(t)\right]^r$ $(r > 0)$, the contribution to the observed data likelihood provided by one observation (X, Δ) is proportional to the quantity

$$\left[\frac{(1-p)\widetilde{f}(x)}{p+(1-p)\widetilde{S}(x)}\right]^{\delta} S_r(x)\, r^{\delta}.$$

Proof. From earlier results, we can write $S_r(t) = \left[p+(1-p)\widetilde{S}(t)\right]^r = p^r + (1-p^r)\widetilde{S}_r(t)$, for

$$\widetilde{S}_r(t) = \frac{\left[p+(1-p)\widetilde{S}(t)\right]^r - p^r}{1-p^r},$$

and

$$\widetilde{f}_r(t) = \frac{1-p}{1-p^r} r \left(p + (1-p)\widetilde{S}(t)\right)^{r-1} \widetilde{f}(t).$$

One can then use Proposition A for this new cure-rate model, replacing p by p^r, $S(t)$ by $S_r(t)$, and $\widetilde{f}(t)$ by $\widetilde{f}_r(t)$. Simple algebra then yields the result. □

B.3 Agreement probabilities

We compute the probabilities of agreement $P(FH = R \mid R = 1)$, $P(FH = R \mid R = 0)$ and the more interesting probabilities of correct classification and misclassification:

$$\begin{aligned} P(FH = 1 \mid R = 1) &= p_{11} \\ P(FH = 0 \mid R = 0) &= p_{00} \\ P(FH = 1 \mid R = 0) &= p_{10} \\ P(FH = 0 \mid R = 1) &= 1 - p_{11} \end{aligned}$$

First, we compute the probability that $FH(t) = 0$, with hypothetically the three closest family member, the grandmother, the mother and the first sister, so that we have a first measure of how well the indicator represents the true latent risk group membership:

$$P(FH(b+t) = 0) \stackrel{\perp}{=} P(t_g \geq t + 60)P(t_m \geq t + 30)P(t_s \geq t) = S_{T_g}(t + 60)S_{T_m}(t + 30)S_{T_s}(t),$$

assuming that there are, easily, constant 30 years gap of age difference between each generation [1].

For simplicity, we start computations from the trivial Exponential distribution and we recall that there is not a generational Exponential survival change, so survival functions are $S_{T_g} = S_{T_m} = S_{T_s} = S_T$. The marginal probability of the indicator in the traditional survival case is:

$$P(FH(b+t) = 0) = \mathrm{e}^{-\lambda(3t+90)} = \mathrm{e}^{-3\lambda(t+30)}.$$

This probability in the cure-rate case is:

$$P(FH(b+t) = 0) = \left(p + (1-p)\mathrm{e}^{-\lambda(t+60)}\right)\left(p + (1-p)\mathrm{e}^{-\lambda(t+30)}\right) \cdot \left(p + (1-p)\mathrm{e}^{-\lambda t}\right).$$

The plot is in Figure B.4. The difference between the two indicators can assume the values: $(FH(t) - R) \in \{-1, 0, 1\}$. We would like to be as close as possible to the scenario with no difference between the indicators. Importantly, R is fixed while $FH(t)$ depends on t.

We now compute the probabilities of the agreement for low-risk group and high-risk group membership. For the survival case we recall that the hazard function in the low (high) risk group

[1] To incorporate the possibility of improved survival across generations, we introduce distinct survival functions for family members: S_{T_g}, S_{T_m}, and S_{T_s}. This allows us to capture any potential improvements in survival over time. In the generational survival improvement case we obtain the probability $P(FH(b+t) = 0) = [S_T(t + 60)]^{\beta_m^2}[S_T(t + 30)]^{\beta_m}[S_T(t)]$, where β_m represents a parameter that measures the generational difference in survival.

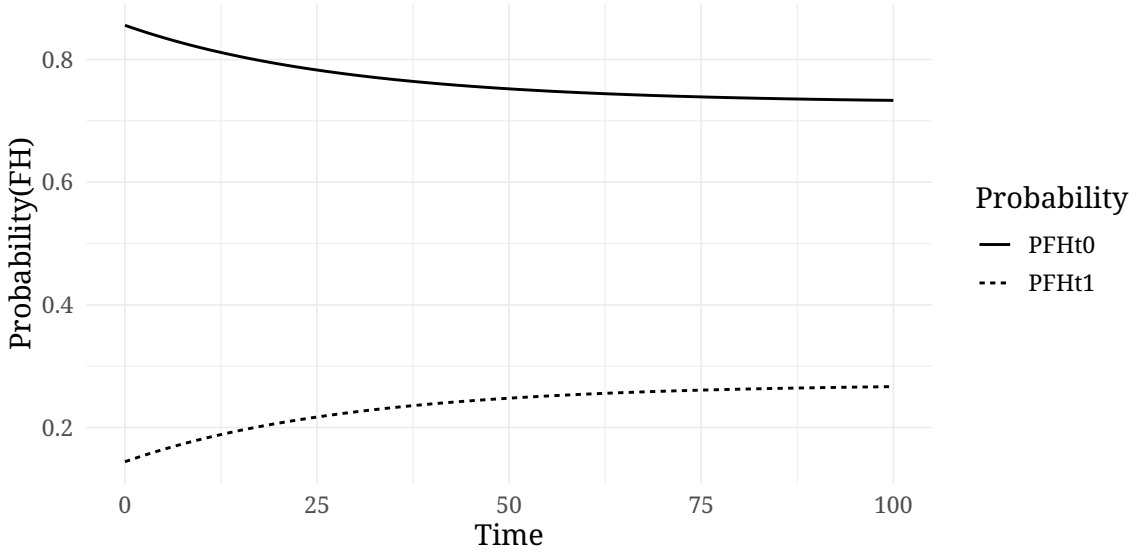

Figure B.4: probability of FH = 0 (PFHt0) vs Probability of FH = 1 (PFHt1).

is $\lambda_0(t)$ ($\lambda_1(t) = \alpha \cdot \lambda_0(t)$), keeping in mind that $\lambda(t \mid R = 0) = \lambda_0 = \lambda$ and $\lambda(t \mid R = 1) = \lambda_1(t) = \beta\lambda_0(t) = \beta\lambda$. But this does not hold in the disease development case. The probability of agreement in the low-risk group over the whole $\mathbb{R}^+$ time axis is:

$$P(FH(b+t) = R \mid R = 0) = \int_0^\infty P(FH(b+t) = 0) f_0(t) \mathrm{d}t$$
$$= \int_0^\infty \left(p + (1-p)\mathrm{e}^{-\lambda(t+60)}\right)\left(p + (1-p)\mathrm{e}^{-\lambda^*(t+30)}\right)\left(p + (1-p)\mathrm{e}^{-\lambda^* t}\right)(1-p)\lambda\mathrm{e}^{-\lambda t}\mathrm{d}t.$$

Similarly, we compute the probability of agreement for the high-risk group:

$$P(FH(b+t) = R \mid R = 1) = \int_0^\infty P(FH(b+t) = 1) f_1(t) \mathrm{d}t$$
$$= \int_0^\infty (1 - P(FH(b+t) = 0)) f_1(t, \lambda) \mathrm{d}t = \int_0^\infty f_1(t, \lambda) - P(FH(b+t) = 0) f_1(t, \lambda) \mathrm{d}t$$
$$= \int_0^\infty f_1(t, \lambda) \mathrm{d}t - \int_0^\infty P(FH(b+t) = 0) f_1(t, \lambda) \mathrm{d}t = 1 - \int_0^\infty P(FH(b+t) = 0) f_1(t, \lambda) \mathrm{d}t$$
$$= 1 - \int_0^\infty P(FH(b+t) = 0)(1 - \widetilde{p}) \left(\frac{\widetilde{f}(t)\alpha}{1 - \widetilde{p}}\right)(p + (1-p)\widetilde{S}(t))^{\alpha - 1} \mathrm{d}t.$$

One would like both probabilities to be large. For the cure-rate case, the conditional probability of agreement in the low-risk group over the whole $\mathbb{R}^+$ time axis is

$$P(FH(b+t) = R \mid R = 0) = \int_0^\infty P(FH(b+t) = 0) f_0(t) \mathrm{d}t.$$

Easily, we obtain the conditional probability of agreement for the high-risk group, such as

$$P(FH(b+t) = R \mid R = 1) = \int_0^\infty P(FH(b+t) = 1) f_1(t) \mathrm{d}t$$

We can also analyse the misclassification probabilities (notice that we implicitly assume that the proportion of high-risk families $P(R=1)=h$ is constant over time). For the cure-rate case the conditional correct classification probabilities are

$$
\begin{aligned}
&P(FH(t)=0 \mid R=0) = S_0(t+60)S_0(t+30)S_0(t) \\
&= (p+(1-p)\widetilde{S}(t+60))(p+(1-p)\widetilde{S}(t+30))(p+(1-p)\widetilde{S}(t)) \\
&= (p+(1-p)e^{-\lambda(t+60)})(p+(1-p)e^{-\lambda(t+30)})(p+(1-p)e^{-\lambda(t)}) \\
&P(FH(t)=0 \mid R=1) = S_1(t+60)S_1(t+30)S_1(t) = (S_0(t+60)S_0(t+30)S_0(t))^{\alpha} \\
&= ((p+(1-p)e^{-\lambda(t+60)})(p+(1-p)e^{-\lambda(t+30)})(p+(1-p)e^{-\lambda(t)}))^{\alpha}
\end{aligned}
$$

Clearly,

$$
\begin{aligned}
P(FH(t)=1 \mid R=0) &= 1 - S_0(t+60)S_0(t+30)S_0(t) \\
P(FH(t)=1 \mid R=1) &= 1 - S_1(t+60)S_1(t+30)S_1(t)
\end{aligned}
$$

where $S_0(t) = p+(1-p)e^{-\lambda(t)}$, and $S_1(t) = (p+(1-p)e^{-\lambda(t)})^{\alpha}$.

A graphical visualization of these probabilities is illustrated in Figure B.5. We also compute the

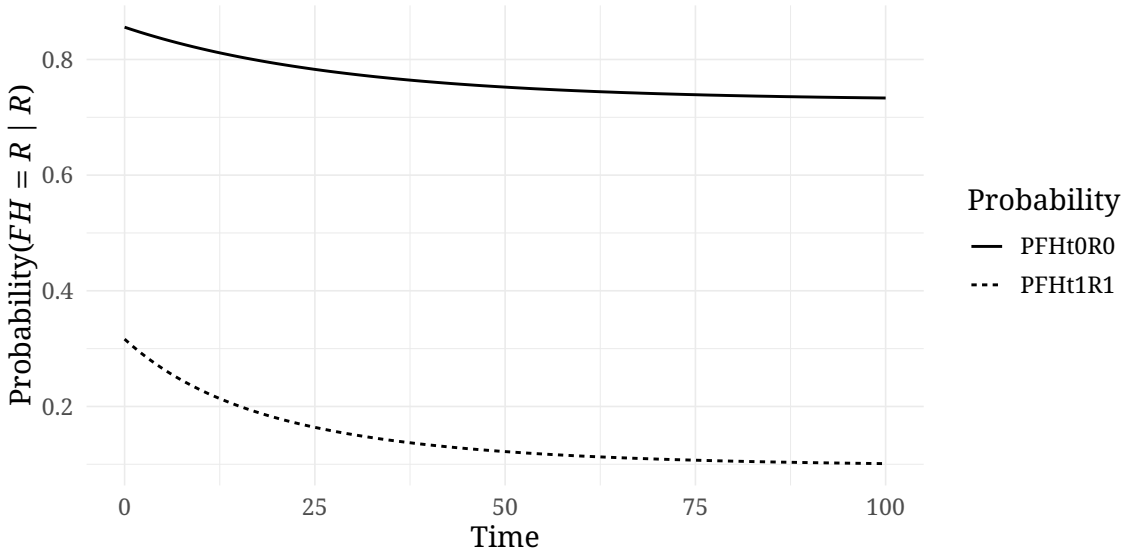

Figure B.5: probability of $FH=0$ conditional to $R=0$ (PFHt0R0) vs. Probability of ($FH=1$ conditional to $R=1$ (PFHt1R1).

inverse probabilities of correct classification only for the survival case, i.e.: (i) $P(R=0 \mid FH(t)=0)$ and (ii) $P(R=1 \mid FH(t)=1)$. These are:

$$
\begin{aligned}
&(i)\ P(R=0 \mid FH(t)=0) = \frac{P(FH(t)=0 \mid R=0)P(R=0)}{P(FH(t)=0)} \\
&= \frac{P(FH(t)=0 \mid R=0)(1-h)}{P(FH(t)=0 \mid R=0)(1-h) + P(FH(t)=0 \mid R=1)h} = \frac{f(t,p,\lambda)(1-h)}{f(t,p,\lambda)(1-h) + (f(t,p,\lambda)^{\alpha})\,h}
\end{aligned}
$$

and

$$
\begin{aligned}
(ii)\ P(R=1 \mid FH(t)=1) &= \frac{P(FH(t)=1 \mid R=1)P(R=1)}{P(FH(t)=1)} \\
&= \frac{P(FH(t)=1 \mid R=1)h}{P(FH(t)=1 \mid R=1)h + P(FH(t)=1 \mid R=0)(1-h)} \\
&= \frac{(1-f(t,p,\lambda)^{\alpha})h}{(1-f(t,p,\lambda)^{\alpha})h + (1-f(t,p,\lambda))\,(1-h)}
\end{aligned}
$$

with $f(t,p,\lambda) = (p+(1-p)e^{-\lambda(t+60)})(p+(1-p)e^{-\lambda(t+30)})(p+(1-p)e^{-\lambda(t)})$. With these probabilities, we describe the distribution of the measurement error when using the observed $FH(t)$ instead of R in the observed data model. A graphical representation of the trend of these probabilities for the fixed values $\lambda = 1/90$, $\alpha = 2$, $h = 0.7$ is illustrated in Figure B.6.

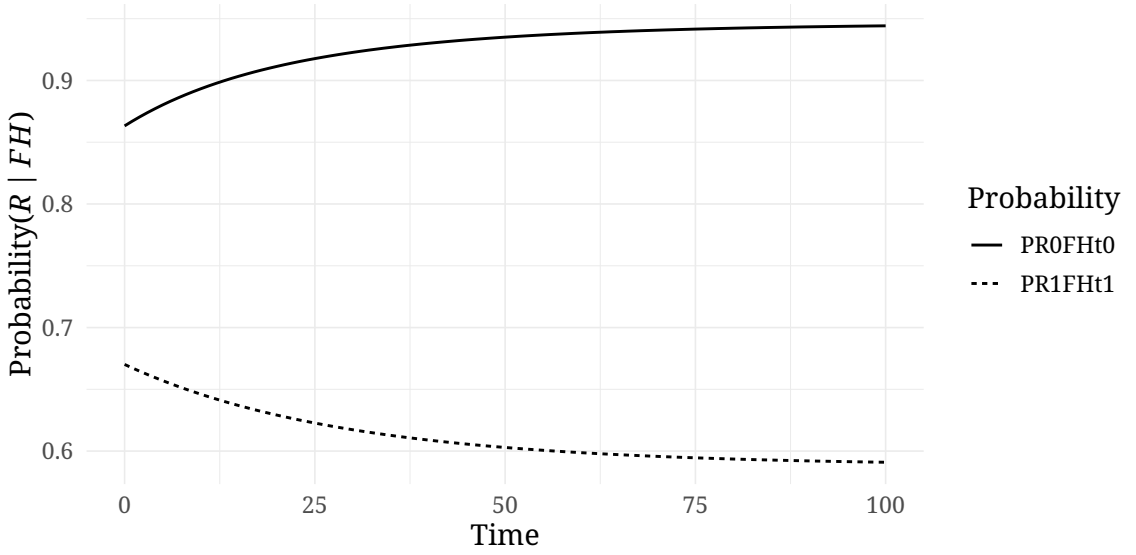

Figure B.6: probability of $R = 0$ conditional to $FH = 0$ (PR0FHt0) vs. Probability of $R = 1$ conditional to $FH = 1$ (PR1FHt1).

B.4 Development of a new indicator FH

Rather then considering $FH(t)$ as a binary indicator we may consider it as a function of the event indicators of the relatives at t. One may define:

$$FH^*(t) = w_g\mathbb{I}(T_g \leq t+60) + w_m\mathbb{I}(T_m \leq t+30) + w_s\mathbb{I}(T_s \leq t)$$

with $w_g \leq w_m \leq w_s$, and $w_g + w_m + w_s = 1$, so that the death of the grandmother has the smallest weight since it is more likely to occur. The weights could assume the form $w_g = (1 - P(\mathbb{I}(T_g \leq t+60)))$ so that as closer to one is the probability of dying, as closer to zero is the weight: $P(\mathbb{I}(T_g \leq t+60)) \approx 1 \implies w_g \approx 0$. We can rewrite the indicator as:

$$\begin{aligned} FH^*(t) &= (1 - P(\mathbb{I}(T_g \leq t+60)))\mathbb{I}(T_g \leq t+60) + \\ &+ (1 - P(\mathbb{I}(T_m \leq t+30)))\mathbb{I}(T_m \leq t+30) + (1 - P(\mathbb{I}(T_s \leq t)))\mathbb{I}(T_s \leq t). \end{aligned}$$

To fix ideas, the FH indicator is a linear combination of three indicators: $\mathbb{I}(T_g \leq t{+}60)$, $\mathbb{I}(T_m \leq t{+}30)$ and $\mathbb{I}(T_s \leq t)$. We wonder now how to optimally combine the three indicators in order to assign the families in the correct risk group with the higher probability. We intend to further develop this part.

B.5 Extension to subject-specific covariates

In our analysis, we previously assumed that breast cancer onset occurred equally across generations. However, considering the advancements in detection tools over the years, it is plausible to assume that breast cancer onset may differ among generations. To capture this generational difference, we introduce a subject-specific covariate indicating the calendar year of breast cancer detection. This covariate allows us to account for variations in breast cancer onset between daughters, mothers, and grandmothers.

To incorporate this covariate into our model, we assign a weight to the likelihood contribution of breast cancer that is inversely proportional to the time of onset. This means that the weight decreases as the onset of breast cancer occurs further in time.

Additionally, we simulate a birthday time window to group women of the same generation together, facilitating the analysis of generational differences.

It is worth noting that in our model, R represents the binary true genetic risk indicator, which is latent and remains unchanged from birth. To simplify the computational aspect of our analysis, we adopt an exponential multiplicative structure for the hazard function. This structure allows us to incorporate frailty risk through an exponential form. Specifically, when considering two risk groups, we express the hazard function as follows: $\lambda_1(t) = \mathrm{e}^{\alpha}\lambda_0(t)$. The hazard function and the survival function are given by

$$\lambda_R(t) = \lambda_0(t)\mathrm{e}^{\alpha R}\mathrm{e}^{\beta_1 z_1 + \cdots + \beta_k z_k} \qquad S_R(t) = [S(t)]^{\mathrm{e}^{\alpha R}\mathrm{e}^{\beta_1 z_1 + \cdots + \beta_k z_k}}$$

with $S(t) = p + (1-p)\widetilde{S}(t)$. So, clearly, when $R = 0$, the low-risk survival function is $S_0(t) = [S(t)]^{e^{\beta_1 z_1 + \cdots + \beta_k z_k}}$, while, with $R = 1$, the high-risk survival function is $S_1(t) = [S(t)]^{e^{\alpha+\beta_1 z_1 + \cdots + \beta_k z_k}}$. The improper density function is obtained:

$$f_R(t) = -\frac{\partial}{\partial t}[S(t)]^{e^{\alpha R+\beta' z}} = e^{\alpha R+\beta' z}[S(t)]^{e^{\alpha R+\beta' z}-1} f(t),$$

with $z = (z_1, \ldots, z_k)$ covariate vector, and β' the parameter collection. Clearly, the improper density function will be $f_0(t) = e^{\beta' z}[p + (1-p)\widetilde{S}(t)]^{e^{\beta' z}-1}(1-p)\widetilde{f}(t)$, and $f_1(t) = e^{\alpha+\beta' z}[p + (1-p)\widetilde{S}(t)]^{e^{\alpha+\beta' z}-1}(1-p)\widetilde{f}(t)$ respectively for $R = 0/1$.

For the low-risk group we have the cure-rate survival function as

$$\begin{aligned} S_0(t) &= [p + (1-p)\widetilde{S}(t)]^{e^{\beta' z}} = p^{e^{\beta' z}} + (1 - p^{e^{\beta' z}})\widetilde{S}_0(t) \\ \widetilde{S}_0(t) &= \frac{[p + (1-p)\widetilde{S}(t)]^{e^{\beta' z}} - p^{e^{\beta' z}}}{1 - p^{e^{\beta' z}}}. \end{aligned}$$

Similarly, for the high-risk group the survival function is

$$\begin{aligned} S_1(t) &= [p + (1-p)\widetilde{S}(t)]^{e^{\alpha+\beta' z}} = p^{e^{\alpha+\beta' z}} + (1 - p^{e^{\alpha+\beta' z}})\widetilde{S}_1(t) \\ \widetilde{S}_1(t) &= \frac{[p + (1-p)\widetilde{S}(t)]^{e^{\alpha+\beta' z}} - p^{e^{\alpha+\beta' z}}}{1 - p^{e^{\alpha+\beta' z}}}. \end{aligned}$$

Then, also the density can be rewritten so that the new form coincides with a proper density function multiplied by a constant that is at most equal to one.

$$\begin{aligned} f_0(t) &= (1 - p^{e^{\beta' z}})\widetilde{f}_0(t) = (1 - p^{e^{\beta' z}}) \left[\frac{e^{\beta' z}[p + (1-p)\widetilde{S}(t)]^{e^{\beta' z}-1}}{1 - p^{e^{\beta' z}}}(1-p)\widetilde{f}(t) \right] \\ \widetilde{f}_0(t) &= -\frac{\partial}{\partial t}\widetilde{S}_0(t) = -\frac{\partial}{\partial t}\frac{[p + (1-p)\widetilde{S}(t)]^{e^{\beta' z}} - p^{e^{\beta' z}}}{1 - p^{e^{\beta' z}}} \\ &= \frac{e^{\beta' z}[p + (1-p)\widetilde{S}(t)]^{e^{\beta' z}-1}}{1 - p^{e^{\beta' z}}}(1-p)\left(-\frac{\partial}{\partial t}\widetilde{S}(t)\right) = \frac{e^{\beta' z}[p + (1-p)\widetilde{S}(t)]^{e^{\beta' z}-1}}{1 - p^{e^{\beta' z}}}(1-p)\widetilde{f}(t) \end{aligned}$$

Similarly, for the high-risk group density function, the density function is

$$\begin{aligned} f_1(t) &= \frac{\partial}{\partial t}S_1(t) = \frac{\partial}{\partial t}\left(p^{e^{\alpha+\beta' z}} + (1 - p^{e^{\alpha+\beta' z}})\widetilde{S}_1(t)\right) = (1 - p^{e^{\alpha+\beta' z}})\left(-\frac{\partial}{\partial t}\widetilde{S}_1(t)\right) = (1 - p^{e^{\alpha+\beta' z}})\widetilde{f}_1(t) \\ &= (1 - p^{e^{\alpha+\beta' z}}) \left[\frac{e^{\alpha+\beta' z}[p + (1-p)\widetilde{S}(t)]^{e^{\alpha+\beta' z}-1}}{1 - p^{e^{\alpha+\beta' z}}}(1-p)\widetilde{f}(t) \right] \\ \widetilde{f}_1(t) &= -\frac{\partial}{\partial t}\widetilde{S}_1(t) = -\frac{\partial}{\partial t}\frac{[p + (1-p)\widetilde{S}(t)]^{e^{\alpha+\beta' z}} - p^{e^{\alpha+\beta' z}}}{1 - p^{e^{\alpha+\beta' z}}} \\ &= \frac{e^{\alpha+\beta' z}[p + (1-p)\widetilde{S}(t)]^{e^{\alpha+\beta' z}-1}}{1 - p^{e^{\alpha+\beta' z}}}(1-p)\left(-\frac{\partial}{\partial t}\widetilde{S}(t)\right) = \frac{e^{\alpha+\beta' z}[p + (1-p)\widetilde{S}(t)]^{e^{\alpha+\beta' z}-1}}{1 - p^{e^{\alpha+\beta' z}}}(1-p)\widetilde{f}(t) \end{aligned}$$

The closed formula for accurately obtaining the value of t depends on the specific survival baseline distribution $\widetilde{S}(t)$. However, when transitioning from the simple case of the Exponential

distribution, it becomes challenging to invert the survival distribution and obtain an exact closed-form solution. Therefore, we do not rely on this method and instead utilize an approximation using the survival function. Firstly, let us present the closed-form data generation method, and later we will discuss the approximated procedure.

The data generation with covariates for times-to-event in the low-risk group is given by the following procedure.

- The time-to-event takes value $T = +\infty$ with probability $p^{e^{\beta' z}}$ following a Bernoulli distribution.
- With probability $(1 - p^{e^{\beta' z}})$ the time-to-event is obtained from the inverse survival function, that is given by

$$\widetilde{S}_0(t) = \frac{(p + (1-p)\widetilde{S}(t))^{e^{\beta' z}} - p^{e^{\beta' z}}}{1 - p^{e^{\beta' z}}} = y \sim \mathrm{U}[0,1] \iff t = \widetilde{S}_0(y)^{-1}.$$

Similarly, for the high-risk group,

- the time-to-event takes value $T = +\infty$ with probability $p^{e^{\alpha+\beta' z}}$, following a Bernoulli distribution.
- With probability $(1-p^{e^{\alpha+\beta' z}})$ the time-to-event is obtained from the inverse survival function, that is given by

$$\widetilde{S}_1(t) = \frac{(p + (1-p)\widetilde{S}(t))^{e^{\alpha+\beta' z}} - p^{e^{\alpha+\beta' z}}}{1 - p^{e^{\alpha+\beta' z}}} = y \sim \mathrm{U}[0,1] \iff t = \widetilde{S}_1(y)^{-1}$$

The approximated data generation for the low-risk group is based on the generation and comparison of $u \sim U(0,1)$ to $p^{e^{\beta' z}}$:

- if $u < p^{e^{\beta' z}} \Rightarrow T = +\infty$,
- if else $u \geq p^{e^{\beta' z}} \Rightarrow u = [S(t)]^{e^{\beta' z}} \iff u^{1/e^{\beta' z}} = p + (1-p)\widetilde{S}(t) \iff t = \widetilde{S}^{-1}\left[\frac{u^{1/e^{\beta' z}} - p}{1-p}\right]$

Similarly, for the high-risk groups $u \sim U(0,1)$ is compared to $p^{e^{\alpha+\beta' z}}$:

- if $u < p^{e^{\alpha+\beta' z}} \Rightarrow T = +\infty$
- if else $u \geq p^{e^{\alpha+\beta' z}} \Rightarrow u = [S(t)]^{e^{\alpha+\beta' Z}} \iff u^{1/e^{\alpha+\beta' z}} = p + (1-p)\widetilde{S}(t) \iff t = \widetilde{S}^{-1}\left[\frac{u^{1/e^{\alpha+\beta' z}} - p}{1-p}\right]$

Appendix C

C.1 The Breslow estimator in the EM algorithm

We go deeply into the EM algorithm structure regarding the frailty context. The first step of the EM algorithm consists in estimating the frailty parameter $\widehat{\theta} = \widehat{\theta}(\underline{x}_1, \ldots, \underline{x}_n)$ in function of the data. In the `R` package `frailtyEM` we consider the frailty as distributed according to a Gamma($shape = \theta, rate = \theta$). The algorithm uses a general full likelihood estimation procedure. The baseline hazard is estimated through the Breslow estimator. We consider the expected number of events occurring at time $[\tau_j, \tau_{j+1}]$, given the lth subject "at risk" at time τ_j^- i.e. belonging to $R(\tau_j)$:

$$\sum_{l \in R(\tau_j)} (\tau_{j+1} - \tau_j)\lambda(\tau_j \mid \underline{z}_l) = \sum_{l \in R(\tau_j)} (\tau_{j+1} - \tau_j) e^{\underline{\beta}' \underline{z}_l} \lambda_0(\tau_j)$$

and setting this equal to the observed number of events d_j, it is then

$$d_j = \lambda_0(\tau_j)(\tau_{j+1} - \tau_j) \sum_{l \in R(\tau_j)} e^{\underline{\beta}' \underline{z}_l} \Rightarrow \lambda_0(\tau_j)(\tau_{j+1} - \tau_j) \approx \int_{t_j}^{t_{j+1}} \lambda_0(u)\mathrm{d}u = \frac{d_j}{\sum\limits_{l \in R(\tau_j)} e^{\widehat{\underline{\beta}}' \underline{z}_l}}$$

$$\Rightarrow \widehat{\Lambda}_0(t) = \sum_{\tau_j < t} \left[\frac{d_j}{\sum\limits_{l \in R(\tau_j)} e^{\widehat{\underline{\beta}}' \underline{z}_l}} \right]$$

where $\underline{z}_l$ is the covariate vector for lth subject, and $\widehat{\Lambda}_0(t)$ is the cumulative hazard Breslow estimator.

Appendix D

D.1 The non-negligibility of censored bservations

In what follows, we assume that the support of $f_T(t)$ is $\mathbb{R}^+$. We show that one cannot simply ignore the censored cases ($\Delta_i = \mathbb{I}(T_i \leq C_i) = 0$). We consider the empirical survival function estimator and we assess the case when it is consistent for the true survival function, for fixed t.

$$\begin{aligned}
&\widehat{S}^*(t) = \frac{\sum_{i=1}^n \Delta_i \cdot \mathbb{I}(X_i \geq t)}{\sum_{i=1}^n \Delta_i} \\
&\widehat{S}^*(t) \xrightarrow{p} \frac{\mathbb{E}(\Delta \cdot \mathbb{I}(X \geq t))}{\mathbb{E}(\Delta)} = \frac{\mathbb{E}(\mathbb{I}(T \leq C)\mathbb{I}(X \geq t))}{\mathbb{E}(\mathbb{I}(T \leq C))} \text{ as } n \to \infty \\
&= \frac{\mathbb{E}(\mathbb{I}(T \leq C)\mathbb{I}(X \geq t)\mathbb{I}(C \geq t))}{\mathbb{E}(\mathbb{I}(T \leq C))} \text{ since } X = \min(T, C) \\
&= \frac{P(T \leq C, X \geq t, C \geq t)}{P(T \leq C)},
\end{aligned}$$

to be compared to $S(t) = P(T \geq t)$. Suppose that the equality holds. Since $\{T \geq t\} \cap \{C \geq T\} \subseteq \{C \geq t\}$, the equality is equivalent to

$$\frac{P(T \leq C, T \geq t)}{P(T \geq t)} = P(T \leq C) = k \quad \forall t \geq 0.$$

Hence:

$$\begin{aligned}
&P(T \leq C, T \geq t) = k \cdot P(T \geq t) \\
&\frac{\mathrm{d}}{\mathrm{d}t}\left[\int_t^\infty \int_u^\infty f_{C|T}(c \mid u) f_T(u) \mathrm{d}c \mathrm{d}u\right] \overset{T \perp C}{=} \frac{\mathrm{d}}{\mathrm{d}t}\left[\int_t^\infty S_C(u) f_T(u) \mathrm{d}u\right] \overset{Leibnitz}{=} -S_C(t) f_T(t).
\end{aligned}$$

Since $\frac{\mathrm{d}}{\mathrm{d}t} S_T(t) = -f_T(t)$, we have $\frac{\mathrm{d}}{\mathrm{d}t} P(T \leq C, T \geq t) = -S_C(t) f_T(t) = -k f_T(t)$, i.e. $S_C(t) = k \; \forall t \geq 0$, which implies $S_C(t) = 1 \; \forall t \geq 0$ since $S_C(0) = 1$. Finally, $C = +\infty$ with probability one, and immediately $\Delta = 1$ with probability one. Notice that without the assumption that the censoring and the cases are independent, the censoring time does not need to be degenerate at $+\infty$.

References

[1] Czene, K., Lichtenstein, P., and Hemminki, K. (2002). Environmental and heritable causes of cancer among 9.6 million individuals in the swedish family-cancer database. *International journal of cancer*, 99(2):260–266.

[2] Human Mortality Database (2021). Sweden total population. https://www.mortality.org/File/GetDocument/hmd.v6/SWE/STATS/fltper_1x1.txt.

[3] IARC, World health organization (2010). Icd7 category. https://ci5.iarc.fr/CI5I-X/Pages/cancer.aspx.

[4] National cancer institute (2013). Icdo3 category. https://training.seer.cancer.gov/breast/abstract-code-stage/codes.html.

[5] National library of medicine (2013). Icd9 category. https://www.ncbi.nlm.nih.gov/books/NBK367629/table/sb201.t4/.

[6] World health organization (2019). Icdo10. https://icd.who.int/browse10/2019/en#/II.