SURVIVAL MODELLING IN MATHEMATICAL AND MEDICAL STATISTICS

by

HAIRUI HUA

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Mathematics
College of Engineering and Physical Sciences
The University of Birmingham
December 2014

Abstract

An essential aspect of survival analysis is the estimation and prediction of survival probabilities for individuals. For this purpose, mathematical modelling of the hazard rate function is a fundamental issue. This thesis focuses on the novel estimation and application of hazard rate functions in mathematical and medical research. In mathematical research we focus on the development of kernel-based estimates of the hazard rate estimation, and in medical research we concentrate on the development and validation of survival models using individual participant data from multiple studies.

Our first proposal is a multiplicative semiparametric estimate of the hazard rate function. The semiparametric estimate starts from a crude guess of the true hazard rate function and then modifies it by a nonparametric correction factor. We utilize the shape parameter α to unify different types of multiplicative semiparametric estimates and then discuss how to estimate the data-driven version of the estimate in practice. The asymptotic analysis shows that the bias of our proposed estimate could be totally removed if our parametric guess of the underlying data is correct, and even if the assumed parametric model is wrong, the resulted estimate is still as good as the standard kernel hazard rate estimate.

We then investigate an approach to optimize a kernel hazard rate estimate by minimizing its L_1 error. Rather than using the traditional L_2 error criterion such as mean (integrated) squared error, we derive an optimal bandwidth to minimize the L_1 error (mean (integrated) absolute error) of the kernel hazard rate estimate and then develop a simple Newton algorithm to calculate the bandwidth. Theoretically, we demonstrate that the theoretic and adaptive versions of the bandwidth does minimize the L_1 error of the kernel estimate approximately.

We then consider application of more flexible survival methods to medical research, to examine whether the mortality risk of breast cancer patients is associated with their country of residence. The Royston-Parmar approach is used to flexibly model the baseline hazard using restricted cubic splines, and it reveals a significant association between country and survival probability, even adjusting for several confounding factors. The robustness of findings is also evaluated after multiple imputation is used to estimate missing values in the database.

This work is then extended to develop, implement and evaluate a prognostic model for individual mortality risk after breast cancer, using individual participant data (IPD) from multiple countries. We firstly utilize the Royston-Parmar approach to develop a prognostic model with a stratified intercept included to allow for a unique baseline of each included country. Then we utilize an internal-external cross validation method and meta-analysis to summarise the performance of the developed model in external data .

Finally, we consider how to fit survival models that predict individual response to treatment effectiveness, given IPD from multiple trials. We evaluate a range one-stage IPD meta analysis survival models, all of which can estimate the interaction between the treatment effect and patient-level factors. Using a large simulation study and a real case study in epilepsy, we show the necessity of separating within and across trial interaction effects to avoid potential ecological bias.

In conclusion, this thesis develops and apply novel statistical methods for survival analysis. Although many challenges still remain, the work provides an important contribution to both statistical and clinical research.

Acknowledgements

I would like to express my heartily appreciation and thanks to my supervisor, Professor Richard Riley. His advice and guidance help me go through all the time of research in medical statistics and writing of this thesis. Further, his advice and encouragement on both research as well as on my career motivates me to be a good medical statistician in the future. I would also express my sincere thanks to my co-supervisor, Dr Prakash Patil for guiding me in mathematics research in the past years. Although he is now working in America, he is always willing to correct my writing and answer my questions in mathematical proofs. I would also like to thank Dr Hui Li, who brought me to Birmingham and gave me great help both in my study and life.

A special thanks to my parents, grandmothers and all my family. Words cannot express how grateful I am for your supporting me go through all the good and hard times in the past three years. I am very sad that my paternal grandfather and maternal grandfather could not see my graduation, but your memories will be with me ever. At the end I would like express appreciation to all of my friends who have ever helped me in my PhD study.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

MSE	Mean squared error
MISE	Mean integrated squared error
IPD	Individual participant data
upa	Urokinase-type plasminogen activator
pai1	Plasminogen activator inhibitor
CI	Confidence interval
HR	Hazard ratio
MI	Multiple imputation
Ned	Netherland
Irl	Ireland
Swe	Sweden
Slo	Slovenia
Aut	Austria
Fra	France
Sui	Switzerland
Den	Denmark
PI	Prognostic Index
d.f.	Degree of freedom

CHAPTER 1

INTRODUCTION OF SURVIVAL MODELLING IN MATHEMATICAL AND MEDICAL STATISTICS

1.1 Overview of the thesis

Survival analysis is the study of data with the time-to-event outcome. In survival analysis, invariably one needs to answer questions such as how long is a subject likely to survive or which factors are predictors toward the mortality risk of the subject? To address these issues, we should establish appropriate statistical models to predict the occurrence of the event (failure) of interest and also to identify the prognostic factors associated with the event[83]. Such models are known as survival models and are seen in many fields of research such as reliability analysis in engineering and prognosis research in pharmacy[76]. However, in this thesis, we mainly focus on its theory and application in mathematical and medical research.

In mathematical research, in this thesis we are particularly interested in estimating a function from a noisy data set. For instance, while dealing with a data set contain values of a single variable Y from an unknown distribution, one hopes to estimate its true underlying probability distribution[68] or in regression analysis, one needs to estimate the relationship between the response variable Y and explanatory regressors $\underline{X}$[53]. In survival models, a key interest is estimating the hazard function which describes the failure rate of an item under risk. See Section 1.2 for the details of hazard rate function. One can classify the methods of estimating hazard functions into two categories, parametric and nonparametric methods. Particularly, many of the methods used for hazard rate estimation were originally utilized in density estimation, *i.e.* kernel methods. We introduce these approaches in the following subsections.

In medical research, in this thesis we discuss several survival modelling problems using time-to-event data arising from multiple clinical studies such as randomised trials. Time-to-event data provides the time of a patient from a well-defined origin point to the occurrence of an event of interest. The end-point of data could be non-fatal but more often, it is death. Time-to-event data from clinical trials owns two evident characteristics, firstly, the data is prone to be censored in practice which means that the patients in trials may quit the study (or the study ends) before the event occurs. Secondly, the data is generally asymmetrical distributed because survival time could not be negative. Therefore the common statistical model such as logistic regression is not amenable[24][77]. Instead, we introduce several useful approaches in this chapter such as parametric regression, Cox regression and flexible parametric regression to develop a survival model.

The goals of this thesis are to novelly apply and develop survival modes, and this can be summarized into 2 parts. Chapter 2 and 3 are devoted to the mathematical study on a new semiparametric kernel hazard estimate and a L_1 optimal kernel hazard rate estimate respectively. Chapter 4 and 5 discuss several survival modelling problems using individual participant data from multiple cohort studies in breast cancer, to compare mortality rates across countries and predict outcome risk for new individuals. Chapter 6 considers multiple trials, and uses survival models to evaluate treatment effects. We start with a broad introduction of existing methodologies which are important to our research in the rest of this chapter.

1.2 General functions in survival modelling

In survival modelling, there are several functions that are important in the analysis of event time data. For example, hazard rate function, cumulative hazard rate function and survival function are very useful in the reliability analysis concerning the lifetime of manufactured items in industries, in the duration modelling in economics research and in the survival analysis, one of the major areas in medical statistics.

A hazard rate function is defined as the probability of failure of an item in an interval $(x, x + dx)$ where dx converges to 0 given that it has survived until time x. The function $\lambda(x)$ referred to as a hazard rate function is then defined as

$$\lambda(x) = \lim_{dx \to 0^+} \frac{P(x \leq X < x + dx | x \leq X)}{dx}.$$

Using the definition of conditional probability and then finding the limit, one expresses it as,

$$\lambda(x) = \frac{f(x)}{1 - F(x)}, \text{ when } F(x) < 1$$

where $f(x)$ is the probability density function and $F(x)$ is its cumulative distribution function of x.

Another important function in survivor analysis is the cumulative hazard rate function which is defined as

$$\Lambda(x) = \int_0^x \lambda(t) dt.$$

This gives the total hazard accumulated by time x. Note that the cumulative hazard rate function can also be expressed as

$$\Lambda(x) = \int_0^x \frac{f(t)}{1 - F(t)} dt = -\log[1 - F(x)].$$

Further the survival function $S(x)$ usually describes the probability of an item that will survive beyond the specified time x. It is closely associated with the cumulative distribution which could be written as $1 - F(x)$.

It is well known that for the real event time data from statistics, engineering, economics, and medical research, censoring often occurs when the value of an observation is only partially known. In mathematical research, for the sake of simplicity, we will not discuss the data with censorship and always assume that the failure times of all the observations are known to us. However, our theoretical results can be easily extended to the censored cases. As for medical research, the methodologies and data we utilize always allow for the censored cases.

1.3 Parametric estimation

The classical approach to estimating an unknown function in statistics is first to assume a parametric model, *i.e.* a parameter function with unknown parameter values. For example, in the case of density estimation assume that the unknown density function $f(x)$ has a certain specified functional form $f(x;\underline{\theta})$ characterized by a parameter vector $\underline{\theta}$ where $\underline{\theta} \in \underline{\Theta}$ is unknown. In the case of regression, assume that the conditional mean $E[Y|\underline{X}]$ has some specified functional form $m(x;\underline{\theta})$, again characterized by $\underline{\theta}$. Then with these assumptions about the functional form, our problem reduces to the estimation of the parameter vector $\underline{\theta}$. Therefore the second step is the estimation of the parameters in the parametric model. This classical approach is referred to as parametric estimation.

To illustrate the important drawback of the parametric estimation, consider the data set of event time $(X_1, X_2, ..., X_{200})$ simulated from a Weibull distribution with the shape parameter α being 5 and the scale parameter β being 3.

Assume that the above data are modelled by Weibull distribution (α, β) and then the unknown parameters α and β are estimated by the maximum likelihood method. Thus the estimate of the Weibull (α, β) density $f(x;\alpha,\beta)$ is $f(x;\hat{\alpha},\hat{\beta})$ and it is plotted in the left panel of Fig 1.1. One finds that with a sample of size 200, the parametric estimate provides a reasonably accurate curve which is close to the true one.

Now assume (wrongly) that the same data were generated from the exponential distribution $f(x;\mu)$ with mean μ, and let $\hat{\mu}$ be the maximum likelihood estimator of μ and thus $f(x;\hat{\mu})$ is the estimate of $f(x;\mu)$. Again the estimated density function $f(x;\hat{\mu})$ and the true Weibull density are plotted in the right panel of Fig 1.1. One notices that because of the wrong model assumption, even with a maximum likelihood estimate of μ, as expected the estimate $f(x;\hat{\mu})$ fails to reveal the true shape of the underlying model.

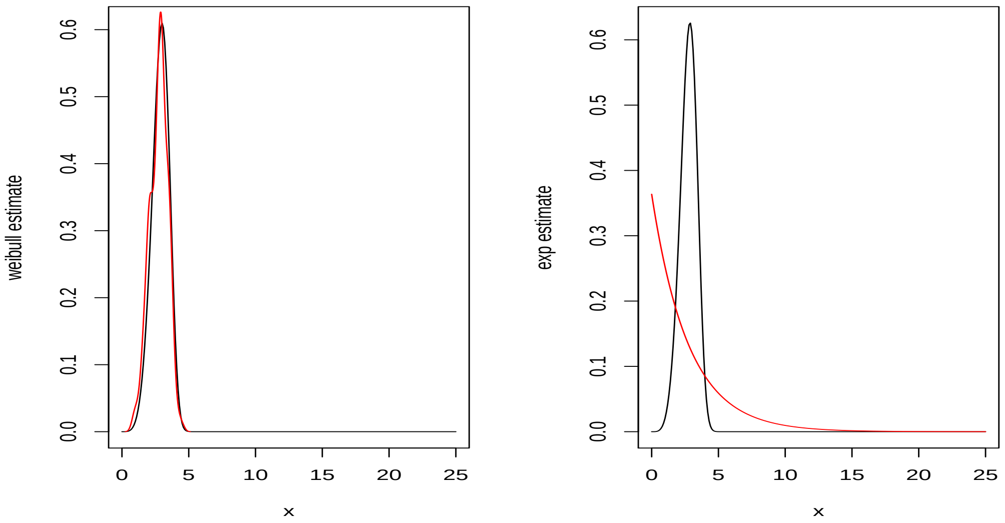

Figure 1.1: The left panel plots the estimator with the correct Weibull model and the right panel illustrates the estimator with the wrong assumption, exponential distribution. The curve in black denotes the true underlying curve and the curve in red denotes the estimated curve.

The above example illustrates the following important point about the parametric estimation. If the model assumption (*i.e.* functional form) is right, our estimate of the function is quite accurate. However, if the assumed parametric model is wrong, our estimate tends to be no longer reliable and diverges from the true model. This may lead to drawing meaningless inferences.

The theory of parametric inference for survival models is well studied and developed. There is a large body of literature on estimation of hazard rate function and related generalizations using parametric methods. For example, see Barlow and Proschan[9] and Lawless[82]. Here we mainly illustrate Weibull, log-normal, and log-logistic models[94] in modelling event-time data as they would be utilized in the thesis:

Weibull model

If one assumes the underlying event follows a Weibull distribution, then its density function, hazard rate function are characterized as

$$\begin{aligned} f(t) &= \frac{p}{\mu}\left(\frac{t}{\mu}\right)^{p-1} e^{-(t/\mu)^p} \\ h(t) &= \frac{p}{\mu}\left(\frac{t}{\mu}\right)^{p-1} \end{aligned}$$

and the survival function is

$$S(t) = e^{-(t/\mu)^p} \tag{1.3.1}$$

where $p > 0$ is the shape parameter and $\mu > 0$ is the scale parameter determining whether the hazard is increasing, decreasing, or constant over time.

Log-logistic model

Log-logistic hazard rate model is obtained if the natural logarithm of survival time t has a logistic density with location parameter μ and scale parameter σ. The log-logistic density and hazard rate functions are

$$\begin{aligned} f(t) &= \frac{2\exp\left(\frac{\ln t - \ln \mu}{\sigma}\right)}{t\sigma\{1+\exp\left(\frac{\ln t - \ln \mu}{\sigma}\right)\}^2} \\ h(t) &= \frac{2\exp\left(\frac{\ln t - \ln \mu}{\sigma}\right)}{t\sigma\{1+\exp\left(\frac{\ln t - \ln \mu}{\sigma}\right)\}} \end{aligned}$$

and its survival function is

$$S(t) = \left\{1 + \exp\left(\frac{\ln t - \ln \mu}{\sigma}\right)\right\}^{-1}. \tag{1.3.2}$$

Log-normal model

For a log-normal model, the natural logarithm of time follows a normal distribution, say, its density and hazard rate functions are given by

$$\begin{aligned} f(t) &= \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[\frac{1}{2\sigma^2}\{\log(t) - \mu\}^2\right], \\ h(t) &= \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[\frac{1}{2\sigma^2}\{\log(t) - \mu\}^2\right] \Big/ \left\{1 - \Phi\left(\frac{\log(t) - \mu}{\sigma}\right)\right\}. \end{aligned}$$

The survival function is,

$$S(t) = 1 - \Phi\left(\frac{\log(t) - \mu}{\sigma}\right) \tag{1.3.3}$$

where $\Phi(z)$ is the standard normal cumulative distribution function.

Having defined $f(t)$ and $S(t)$ for each of three parametric models (Weibull, log-logistic, log-normal) respectively, the parameters in the model could be easily estimated by the maximum likelihood method[26][40] when presented with a set of survival times. An observation with the occurrence of the event at time t contributes the likelihood function the value of the hazard rate $f(t)$ while a censored observation, known to leave the study at time t contributes $S(t)$, which is the probability that it survives till time t[21].

To make use of this well developed theory in survival modeling and analysis, it is essential that one has an appropriate parametric model for the collected data. But invariably we come across a situation where either it is not possible to propose an appropriate parametric model or the assumed model does not provide satisfactory fit to the data. In such cases, the parametric methods, as argued above, are not going to be helpful. So as to address this problem, nonparametric approaches have been suggested in the literature[139]. In the next section we give a brief introduction to nonparametric functional estimation for general statistics and survival modelling particularly.

1.4 Nonparametric estimation

The main idea of nonparametric procedures is to let the data speak for itself without forcing any specified structure or model on the data. For example, in the case of density estimation, it means removing the constraint imposed by the assumed functional form and allowing data to search an appropriate curve that best describes the original data.

Over the last few decades, many different nonparametric approaches have been proposed such as kernel estimation[166], spline smoothing[165] and wavelet approach[164]. Particulary for survival modelling, Kaplan-Meier estimate and Nelson-Aalen estimate[21] were proposed to account for possible censoring cases in the time-to-event data.

1.4.1 Kernel density estimation

Kernel-based estimation is one of the most commonly used methods in nonparametric curve estimation. We first illustrate its use in density setting as it provides very important guidelines to any kernel-based method in hazard rate estimation. For that suppose that we have a sample, $(X_1, X_2, ..., X_n)$, from a continuous probability density function f and let F be its cumulative distribution function.

Note that

$$f(x) = \frac{dF(x)}{dx} = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}.$$

It means when h is sufficiently small, one has

$$f(x) \simeq \frac{F(x-h) - F(x+h)}{2h}. \tag{1.4.1}$$

Now one can estimate $f(x)$ by plugging-in an estimate of $F(x)$ in (1.4.1). For example, let empirical cumulative distribution function

$$F_n(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n}, \tag{1.4.2}$$

estimate $F(x)$. Then it leads to the density estimate $\hat{f}(x)$ given by

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} I(X_i \in (x-h, x+h))}{2nh} = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$$

where $K(u) = \frac{I[-1<u<1]}{2}$ is the uniform probability density function. The idea behind the above estimate is that the density estimate at a point x is simply the proportion of observations out of n falling into the interval $(x-h, x+h)$ and then this proportion is adjusted for the length of the interval. Note that because of K being the uniform distribution, each of the observations located in the neighborhood of x of length h, $(x-h, x+h)$, is weighted by $1/n$ while others are weighted 0 and the estimate at x is then the sum of the weights of these observations. In general, it is not necessary for us to weight all the observations in $(x-h, x+h)$ equally, instead one can weight the observations with a different scheme. For example, if $K(u)$ is chosen to be the normal density function, then an observation closer to the point x is counted with a larger weight in estimation. That is to say, we do not need to restrict the kernel function to be uniform and many other choices such as triangular, biweight, triweight, Epanechnikov and normal densities are possible. In general, kernel functions are taken to be symmetric, since there is usually no reason to weight any two symmetric observations about the point x unequally. As discussed in literature (*i.e.* Wand and Jones[166]) in kernel estimation, the choice of kernel functions is not a crucial step in estimation. In fact the kernel density estimate does not vary significantly with different kernel functions.

In contrast, the length of the support of the kernel function exhibits a strong influence on the resulting estimate. This length is generally referred to as the bandwidth. With the very large bandwidth h, more observations are included into estimation for the density at the target x. This leads the estimate to be more stable. With a small h, only very few observations are used to construct the estimate. That leads the estimate to be more unstable. But from equation (1.4.1), though unstable, small h leads to less bias. Thus clearly the bandwidth h affects the bias and the variance of the estimate in the opposite direction.

To quantify the effect of bias and variance or to assess the goodness of a density estimator, one needs to define an appropriate error criterion. For that let

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K_h(x - X_i)$$

where $K_h(\cdot) = h^{-1}k(./h)$ is a symmetric kernel function and h the bandwidth. Typically, the goodness of the kernel estimator $\hat{f}(x)$ is measured by a suitable error criterion. To measure the accuracy of the estimator at a certain point, usually the squared error $(\hat{f}(x) - f(x))^2$ or the absolute deviation, $|\hat{f}(x) - f(x)|$ are used. But if the purpose is to assess the global performance of the estimator, commonly the L_1, L_p or L_∞ distance are used where

$$L_1(f, \hat{f}) = \int |\hat{f} - f|,$$

$$L_p(f, \hat{f}) = \int |\hat{f} - f|^p, 2 \leq p < \infty,$$

$$L_\infty(f, \hat{f}) = \sup |\hat{f} - f|.$$

In this section at first we restrict our attention only to the L_p error criteria with $p = 2$ namely, the squared errors, (SE) and the integrated squared errors (ISE) for its technical tractability. In Section 1.4.3 we provide a short discussion on L_1 criterion.

The MSE measures the expected squared difference between the estimate and true function, $(\hat{f}(x) - f(x))^2$, at the point x, *i.e.*

$$MSE(\hat{f}(x)) = E(f(x) - \hat{f}(x))^2 = (Bias(\hat{f}(x)))^2 + Var(\hat{f}(x))$$

where $Bias(\hat{f}(x)) = E\hat{f}(x) - f(x)$.

Naturally, a good nonparametric estimate is the one with a smaller MSE at a point x. That is, the MSE reflects the local performance of $\hat{f}(x)$. As discussed above, to evaluate the accuracy of the estimator globally, we use the expected value of integrated squared errors, generally referred to as MISE and defined as

$$\text{MISE}(\hat{f}(x)) = \int E(f(x) - \hat{f}(x))^2 w(x) dx.$$

The weight function $w(x)$ enables us to give more 'weight' to some interval of x than others. For example, in hazard rate estimation, we may set the weight of rightmost x to be 0 as its estimate tends to be very unstable.

The asymptotic properties of the kernel-based estimator have been discussed in great details in the literature, for example, see Wand and Jones[166] or Simonoff[139]. We summarized one of the important results in the following theorem which gives the bias and variance of a kernel density estimator.

Theorem 1.1. *Let X_1, X_2,...,X_n be i.i.d. random variables with common distribution $F(x)$ and density function $f(x)$, which is twice continuously differentiable. Let $K(\cdot)$ be a kernel function s.t $\int K(x)dx = 1$, $\int K(x)xdx = 0$ and $\int K(x)x^p dx < \infty$ for $p = 2$*

Then as $n \to +\infty$, $h \to 0$ and $nh \to +\infty$, for the kernel estimate $\hat{f}(x)$, we have

$$\begin{aligned} Bias(\hat{f}(x)) &= \frac{h^2 \mu_{2,K}}{2} f''(x) + o(h^2), \\ Var(\hat{f}(x)) &= \frac{R(K)}{nh} f(x) + o\left(\frac{1}{nh}\right). \end{aligned}$$

$$MSE(\hat{f}(x)) = \frac{h^4 \mu_{2,K}^2}{4} f''(x)^2 + \frac{R(K)}{nh} f(x) + o(h^4) + o\left(\frac{1}{nh}\right)$$

where $\mu_{2,K} = \int u^2 K(u) du$ and $R(K) = \int K(u)^2 du$.

Proof. Refer to Wand and Jones[166].

Remark 1.1. Theorem 1.1 makes two important contributions towards the understanding of a kernel estimator. First, it confirms our intuitive understanding that as the bandwidth h increases, the bias of a kernel estimator increases and the variance decreases, *i.e.* $Bias(\hat{f}(x)) = O(h^2)$ and $Var(\hat{f}(x)) = O(1/(nh))$.

Further the quantification of the bias and variance in terms of h provided by the above theorem helps one to determine the optimal bandwidth h which minimizes MSE (or MISE)[139].

1.4.2 Kernel hazard rate estimation

Kernel-based approaches are also frequently used to estimate hazard rate function in survival modelling. In our thesis, we mainly introduce kernel hazard rate estimate for the event time data without censorship. However, the estimate could be easily generalized to any censored data. We suggest the readers to see Watson and Leadbetter[167],[168], Singpurwalla and Wong[140], Tanner and Wong[153] and Muller and Wang[96] for example.

Now suppose $(X_1, X_2, ..., X_n)$ are failure times of n identical items and one of the interests is to estimate the hazard rate of the component under study. Since $\lambda(x)$ is a ratio of density function to the survival function, one approach is to estimate the density function and distribution function separately and then plug the estimates of these functions into the hazard rate function directly, to define

$$\hat{\lambda}_1(x) = \frac{\hat{f}(x)}{1 - F_n(x)}$$

where $\hat{f}(x)$ is the kernel density estimate and $F_n(x)$ is the empirical estimate of the distribution function $F(x)$. Notice that with the conventional definition of $F_n(x)$, defined in equation (1.4.2), $\hat{\lambda}_1(x)$ is not well defined since $(1 - F_n(x))$ equals 0 for $x > \max_i(X_i)$. Then the modified definition of $F_n(x)$ is considered and it is defined as

$$F_n(x) = \frac{\#(X_i \leq x) - 1}{n} \tag{1.4.3}$$

where $\#(X_i \leq x)$ denote the number of observations that are smaller than x.

Note that when n is very large, the modified and original empirical cumulative distribution functions are equivalent. Thus it gives us the first kernel hazard estimator as

$$\hat{\lambda}_1(x) = \frac{\hat{f}(x)}{1 - F_n(x)}.$$

As opposed to the ratio of two functions, the other approach treats the hazard rate as a function by itself. In this approach, first by Taylor expansion, we notice that, as $h \to 0$

$$\int \frac{1}{h} K\left(\frac{x-t}{h}\right) \lambda(t) dt = \int K(u) \lambda(x - uh) du = \lambda(x) + O(h^2).$$

Therefore, we could approximate $\lambda(x)$ as

$$\lambda(x) \approx \int \frac{1}{h} K\left(\frac{x-t}{h}\right) \lambda(t) dt = \lambda^*(x). \tag{1.4.4}$$

Let $\Lambda(x) = \int_0^x \lambda(t)dt$ be the cumulative hazard rate, then

$$\begin{aligned} \lambda(x) \simeq \lambda^*(x) &= \int K_h(x-t)\lambda(t)dt = \int K_h(x-t) d\Lambda(t) \\ &= \int K_h(x-t) d\Lambda_n(t) + \int K_h(x-t) d(\Lambda(t) - \Lambda_n(t)) \end{aligned}$$

where $\Lambda_n(t)$ is an empirical version of $\Lambda(t)$.

The first term provides us with an estimator of λ, *i.e.*

$$\hat{\lambda}_2(x) = \int K_h(x-t) d\Lambda_n(t) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{K_h(x - X_i)}{1 - F_n(X_i)} \right) = \sum_{i=1}^{n} \left(\frac{K_h(x - X_{(i)})}{n - i + 1} \right)$$

where $F_n(x)$ is of the modified version defined in (1.4.3) and $X_{(i)}$ is the ith order statistics. Note that $\lambda(x) - \lambda^*(x)$ is the bias, whereas the noise in the estimators comes from the term $\int K_h(x-t)d(\Lambda(t) - \Lambda_n(t))$. Like in the case of the density estimation setting, the bandwidth h plays the similar role in hazard rate estimation and controls the amount of smoothing applied to kernel estimates. That is, for large h, kernel hazard rate estimates are always more smooth with lower variance. On the other hand, from equation (1.4.4), it is clear that when the bandwidth h is very small, the approximation is closer to the true function with small bias. Thus, for both kernel hazard rate estimators $\lambda_1(x)$ and $\lambda_2(x)$, the bandwidth h influences their biases and variances in opposite directions.

Similar to the case of density estimation we use the MSE to evaluate the local performance of the hazard rate estimators. To evaluate its performance as an estimator of the whole function we use MISE, *i.e.*

$$\mathrm{MSE}(\hat{\lambda}(x)) = E(\hat{\lambda}(x) - \lambda(x))^2,$$

$$\mathrm{MISE}(\hat{\lambda}(x)) = E\int (\hat{\lambda}(x) - \lambda(x))^2 w(x)dx$$

where $w(x)$ is a nonnegative weight function.

The asymptotic properties of the two hazard rate estimators $\hat{\lambda}_1(x)$ and $\hat{\lambda}_2(x)$ have been fully studied by several authors such as Watson and Leadbetter[167], [168] from which we summarize the biases and variances of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ in the following theorem,

Theorem 1.2. *Let X_1, X_2,...,X_n be i.i.d. random non-negative variables with common distribution $F(x)$ and density $f(x)$, which is twice continuously differentiable. Let $K(\cdot)$ be a second-order kernel function such that $\int K(x)dx = 1$, $\int K(x)xdx = 0$ and $\int K(x)x^p dx \leq \infty$ for $p = 2$. Also suppose that there exists small enough h such that $K_h(y-x)/(1-F(y))$ is uniformly bounded for $|y-x| > M$ for any $M > 0$.*

Then as $n \to +\infty$, $h \to 0$ and $nh \to +\infty$, for the kernel hazard rate estimates, we have

$$\begin{aligned} Bias(\hat{\lambda}_1(x)) &= \frac{h^2\mu_{2,K}}{2}\frac{f''(x)}{1-F(x)} + o(h^2), \\ Var(\hat{\lambda}_1(x)) &= \frac{R(K)}{nh}\frac{\lambda(x)}{1-F(x)} + o\left(\frac{1}{nh}\right). \end{aligned}$$

$$MSE(\hat{\lambda}_1(x)) = \frac{h^4\mu_{2,K}^2}{4}\left(\frac{f''(x)}{1-F(x)}\right)^2 + \frac{R(K)}{nh}\frac{\lambda(x)}{1-F(x)} + o(h^4) + o\left(\frac{1}{nh}\right).$$

$$\begin{aligned} Bias\hat{\lambda}_2(x) &= \frac{h^2}{2}\mu_{2,K}\lambda''(x) + o(h^2), \\ Var(\hat{\lambda}_2(x)) &= \frac{R(K)}{nh}\frac{\lambda(x)}{1-F(x)} + o\left(\frac{1}{nh}\right). \end{aligned}$$

$$MSE(\hat{\lambda}_2(x)) = \frac{h^4\mu_{2,K}^2}{4}(\lambda''(x))^2 + \frac{R(K)}{nh}\frac{\lambda(x)}{1-F(x)} + o(h^4) + o\left(\frac{1}{nh}\right).$$

Proof. Refer to Watson and Leadbetter[167], [168] for the proof.

Remark 1.2. Theorem 1.2 helps us to determine the value of the bandwidth h which optimizes MSE or MISE. Similar to the case of density setting, for each of the two hazard rate estimators, the squared bias term and the variance term change in the opposite directions whenever there is a change in bandwidth h, *i.e.* $Bias = O(h^2)$ and $Variance = O(1/(nh))$. Hence in kernel estimation, the bandwidth h plays the role of balancing the squared bias term against the variance term of the MSE or MISE and ideally we can determine the optimal bandwidth h by minimizing the MSE or MISE w.r.t h.

Remark 1.3. The other point is that due to the presence of the survival function $1-F(x)$ in the denominators of the variances for both estimators, for large x, the variances of the above kernel hazard rate estimators are very large. Further, the bias term of $\lambda_1(x)$ will also go up quickly for large x since the denominator of its bias term also consists of $1-F(x)$.

In this thesis, we discuss hazard rate estimation problems only based on the second kernel hazard rate estimate $\hat{\lambda}_2(x)$.

1.4.3 L_1 error criterion

For its technical tractability and easy understanding, the L_2 error criterion is widely utilized to evaluate the performance of a nonparametric estimator. However, it was not the only method which could be utilized. Consider an alternative L_1 error criterion,

$$L_1(f(x), \hat{f}(x)) = \int |\hat{f}(x) - f(x)| dx,$$

to assess the accuracy of a kernel density estimator. Devroye and Gy$\ddot{o}$rfi[36] has demonstrated that the L_1 error is always well-defined and invariant under monotone transformation of the coordinate axes. Because of the appealing properties of L_1 error criterion emphasised by Devroye and Györfi, in Chapter 3, we propose an optimal kernel hazard rate estimator in the sense of minimizing its L_1 error but without considering censorship.

1.4.4 Nonparametric estimates for time-to-event data

In this section, we specifically introduce two commonplace nonparametric estimates for time-to-event data allowing for censorship, which are Kaplan-Meier estimate for survival function $S(t)$[75] and Nelson-Aalen estimate for cumulative hazard rate function $\Lambda(t)$[17].

Suppose that a sample of n observations are obtained from a population with $t_{(1)} < t_{(2)} < ... < t_{(k)}$ being the ordered sequence of observed event times (*i.e.* death), $(t_1, t_2, ..., t_n)$. Here $k \leq n$, as some patients may be censored. Let n_i be the number of subjects at risk at the time prior to $t_{(i)}$ and u_i be the number of events occurs at time $t_{(i)}$, then the survival function $S(t)$ can be estimated by Kaplan-Meier formula:

$$\hat{S}(t) = \prod_{t_{(i)}<t} \frac{n_i - u_i}{n_i}.$$

There are several ways to approximate the variance of the Kaplan-Meier estimate of which the most common one is Greenwood's formula:

$$\hat{V}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_{(i)}<t} \frac{u_i}{n_i(n_i - u_i)}.$$

The Nelson-Aalen estimate is a nonparametric estimate of the cumulative hazard rate function, given by

$$\hat{\Lambda}(t) = \sum_{t_{(i)}<t} \frac{u_i}{n_i},$$

and its variance is estimated by

$$\hat{V}(\hat{\Lambda}(t)) = \sum_{t_{(i)}<t} \frac{u_i}{(n_i)^2}.$$

The Kaplan-Meier survival estimate and Nelson-Aalen cumulative hazard rate estimate can be transformed to each other in the following way:

$$\hat{S}(t) = e^{-\hat{\Lambda}(t)}.$$

1.5 Semiparametric estimation

When one has the exact information of the underlying model before we collect the data, the parametric model approach of Section 1.3 is an ideal choice to our data analysis. On the other hand the nonparametric approaches of 1.4 are preferred when a researcher does not have the knowledge about the true underlying model. However, in practice, most often one has partial information of the underlying model but is not certain about it. In such cases, neither parametric nor nonparametric approaches are suitable since one wants to use the information at hand but is not willing to rely on it totally.

Thus to make use of the partial knowledge about the true model in the nonparametric estimation, a new approach referred to as *semiparametric* estimation has been proposed in the literature. A semiparametric model, as its name suggests, is a hybrid of the parametric and nonparametric approaches. For example, in regression, the semiparametric model are often used in the situations where one hopes to use a parametric model to fit the data but the functional form with respect to a subset of the regressors is not known, that is,

$$Y = \underline{X}'\beta + g(\underline{Z})$$

where Y is a dependent variable, $\underline{X}$ is the vector of explanatory variables that assumed to be in the linear relationship with Y and β is the corresponding parametric vector. The $\underline{Z}$ is the vector of regressors of which the parametric relationship with Y is not known and it is estimated by nonparametric approaches such as kernel regression , spline smoothing and *etc.* .

In the density estimation setting, different semiparametric smoothing approaches are proposed and discussed in the literature. One of the methods is to utilize a nonparametric method to estimate the underlying model initially and then adjust the nonparametric estimate by adding a possible parametric model which is referred to as additive correction factor[78][99]. For example, it is suspected that the data follows the density function described by the functional form $f(x;\underline{\theta})$ which is characterized by a parametric vector $\underline{\theta}$. However since one is not sure of the correctness of the model, one may consider an estimator of the underlying density function as,

$$\hat{f}(x,\pi) = \pi f(x;\hat{\underline{\theta}}) + (1-\pi)\hat{f}(x)$$

where $f(x;\hat{\underline{\theta}})$ is the parametric estimator and $\hat{f}(x)$ is the nonparametric estimator. The parameter $\pi \in [0,1]$ is usually estimated by the maximum likelihood method. It can be showed that if the parametric model prevails, the estimate of π is expected to be close to unity and $\hat{f}(x,\hat{\pi})$ converges to the parametric density function $f(x;\underline{\theta})$ as $n \to \infty$. However if the parametric assumption is wrong, $\hat{\pi}$ is expected to be close to zero and $\hat{f}(x,\hat{\pi})$ approaches the true $f(x)$ as $n \to \infty$. This approach will not be pursued in this dissertation.

An alternative method is to fit the data with an assumed parametric model and if the the crude guess is not satisfied then utilize a nonparametric correction factor to modify the estimator as,

$$\hat{f}(x) = f(x; \underline{\hat{\theta}})\xi(x)$$

where $f(x; \underline{\hat{\theta}})$ is an estimator obtained by estimating the parametric vector $\underline{\theta}$ in the assumed probability model $f(x; \underline{\theta})$ and $\xi(x)$ is a nonparametric multiplicative correction factor. The unknown parameter vector $\underline{\theta}$ is usually estimated by the maximum likelihood method. Since one suspects the accuracy of the parametric density assumption, here the multiplicative correction factor $\xi(x)$ is used to correct the possible inaccuracy in the guessed functional form. To calculate the correction factor, several kernel-type nonparametric methods have been proposed in literatures such as Hjort and Glad[62] who determined the factor by minimizing the squared error distance or the Kullback Leibler distance between the estimate $f(x; \underline{\hat{\theta}})\xi(x)$ and the true density function $f(x)$, and Hjort and Jones[63] who estimated $\xi(x)$ by the kernel-type estimate of $f(x)/f(x; \underline{\hat{\theta}})$.

Naito[97] developed a more generalized approach which includes the former work as its special cases. He showed that the performance of the generalized semiparametric density estimator is asymptotically better than their fully nonparametric kernel counterparts in a broad family of functions around the true model. Motivated by this appealing idea, in Chapter 2, we extend the generalized method of Naito to the setting of hazard rate estimation using data without censorship.

1.5.1 Cox proportional hazards model

Cox proportional hazard model is an important semiparametric regression approach for time-to-event data with censoring time[30]. The basic Cox model assumes the linear relationship for the covariates included but relaxes the fitting by letting the baseline hazards be nonparametric. It can be expressed as

$$\ln \lambda(t) = \ln \lambda_0(t) + \underline{\beta}^T \underline{X}$$

where $\lambda(t)$ is the hazard rates at time t, $\underline{X}$ is the vector of associated prognostic factors and $\lambda_0(t)$ is the baseline hazards for patients with $\underline{X}$ being 0. β is the log hazard ratio between two individuals whose values differ by one unit in the corresponding x but same in other covariates. The critical assumption of Cox proportional hazard model is that hazard ratio should be constant overtime for each of the included covariates (prognostic factors). There are many methods to test the proportional assumption and they will be discussed in Chapter 4.

The Cox proportional hazard model is a semi-parametric model where the coefficients $\underline{\beta}$ could be easily estimated by solving the partial likelihood function. See, for example, Collett[24] for the details. However, one evident disadvantage of Cox regression is that the baseline hazards are not directly estimated. This is not helpful where $S(t)$ is to be predicted in new individuals, as then $\lambda_0(t)$ is needed.

1.6 Prognostic model and factors in survival modelling

In health care, a common procedure is to establish a prognostic model to predict survival time of patients with certain disease[147] and this survival probability may depend on individual characteristics or measurements such as age and sex, referred to as prognostic factors or predictors[129].

Consider a dataset of n patients from a clinical study, (t_i, d_i) for $i = 1, 2, ..n$ where t_i is the follow-up time of the ith patient and d_i denotes whether the observation is censored (1=event and 0=censored). The patient-level characteristic vector $\underline{X}$ is recorded for each patient. An appropriate prognostic model is required which is able to utilize potential prognostic factors in $\underline{X}$, and thereby predict $S(t)$ for new individuals based on their $\underline{X}$.

1.6.1 Flexible parametric model via Royston-Parmar scheme

Royston and Parmar[79][122] proposed a flexible parametric model to fit the survival data. As indicated from this name, it is easy to imagine that the family of models were firstly generalized from conventional parametric models such as Weibull, log-normal, and log-logistic distributions. But, compared to these models, it provides much more flexible estimates by fitting the baseline hazards with smoothing splines[42].

Splines are one of most commonly used mathematical functions to establish an unknown curve. They are usually defined by piecewise polynomials with some constraints to ensure the smoothness, and the points which join the polynomials are named as 'knots'[27]. Restricted natural cubic splines[38] for Royston and Parmar model with $K+2$ knots (or $d.f. = K+1$) is defined as

$$s(\ln t|\underline{\gamma}, K) = \gamma_0 + \gamma_1 \ln t + \gamma_2 z_1 + ...\gamma_{K+1} z_K$$

where the derived variables, z_j (also known as the basis functions) for $j = 1, ..., K$ are calculated as

$$z_j = (\ln t - k_j)^3_+ - \psi_j(\ln t - k_{min})^3_+ - (1-\psi_j)(\ln t - k_{max})^3_+ \tag{1.6.1}$$

where $\psi_j = (k_{max} - k_j)/(k_{max} - k_{min})$ and $(\ln t - k)_+ = \max(0, \ln t - k)$ for any k. We will show that by incorporating splines, three different kinds of Royston-Parmar models can be derived from Weibull, log-logistic and log-normal parametric models respectively:

Weibull generalized model

If the survival function $S(t)$ in (1.3.1) is transformed to the log cumulative hazard scale, then

$$\ln \Lambda(t) = \ln[-\ln S(t)] = p \ln t - p \ln \mu = \gamma_0 + \gamma_1 \ln t$$

where $\gamma_0 = -p \ln \mu$ and $\gamma_1 = p$. The basic idea of the flexible parametric approach is to relax the assumption of linearity of log time by using restricted cubic splines and then add the covariate effects, *i.e.* the log cumulative hazard scale can be expressed as

$$\ln \Lambda(t) = \ln \Lambda_0(t) + \underline{\beta}^T \underline{X}$$

where $\ln \Lambda_0(t) = \gamma_0 + \gamma_1 \ln t + \sum_{j=1}^{k_0} \gamma_{j+1} z_j$ is the baseline function, $k_0 + 2$ denotes the number of knots and z_j is the basis function of natural cubic splines defined in (1.6.1). βs again provide log hazard ratios of included covariates, assumed proportional over time.

Log-logistic generalized model

The log odds of log-logistic model, $\ln O(t)$ is given by

$$\ln O(t) = \ln \frac{1 - S(t)}{S(t)} = \frac{\ln t - \ln \mu}{\sigma} = \gamma_0 + \gamma_1 \ln t$$

where $\gamma_0 = -(\ln \mu)/\sigma$, $\gamma_1 = 1/\sigma$, is linearly related to $\ln t$. Then restricted natural cubic splines and covariate effects could be introduced into the model by

$$\ln O(t) = \ln O_0(t) + \underline{\beta}^T \underline{X}$$

where $\ln O_0(t) = \gamma_0 + \gamma_1 \ln t + \sum_{j=1}^{k_0} \gamma_{j+1} z_j$ is the baseline function and $\underline{X}$ is the vector of covariates included. βs now represent log odds ratios of included covariates which are assumed to be proportional to time.

Log-normal generalized distribution

The flexible parametric model of $\ln t$ with a probit link function $-\Phi^{-1}(\ln t)$ is generalized from log-normal parametric model over time t where $\Phi^{-1}(.)$ is the inverse standard Normal distribution. Using (1.3.3), it is defined by

$$-\Phi^{-1}(S(t)) = \frac{\ln(t) - \mu}{\sigma} = \gamma_0 + \gamma_1 \ln t$$

where $\gamma_0 = -\mu/\sigma$ and $\gamma_1 = 1/\sigma$. The extended class of spline models from log-normal distribution is

$$-\Phi^{-1}(S(t)) = \gamma_0 + \gamma_1 \ln t + \sum_{j=1}^{k_0} \gamma_{j+1} z_j + \underline{\beta}^T \underline{X}.$$

where $\underline{X}$ is the vector of covariates included. Unfortunately, β^T here could not be interpreted directly. We recommend the readers to refer to Long and Freese[86] for the way to interpret the coefficient in a probit model.

The baseline hazard function of a Royston-Parmar model depends on the number of knots fitted in restricted cubic splines. By default, the knot locations are decided by centiles of log-time. Royston and Lambert[120] recommended 3 knots are sufficient for normal size datasets and 5 or 6 knots for really big datasets.

The full maximum likelihood method can be easily applied to estimate Royston-Parmar model[120]. For example, the likelihood function of ith observation in Weibull generalized model (Proportional Hazard model) with $d.f. = 3$ can be written as

$$\begin{aligned}\ln L_i &= d_i(\ln[\gamma_1 + \gamma_2 z_1'(\ln t) + \gamma_3 z_2'(\ln t)] + \gamma_0 + \gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t) \\ &\quad + \underline{\beta}^T \underline{X}_i) - \exp(\gamma_0 + \gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t) + \underline{\beta}^T \underline{X}_i)\end{aligned}$$

where d_i is the indicator variable to denote whether the sample is censored.

Using either of the three models (with proportional hazards, proportional odds and probit scales) to model the event-time data, we could establish the baseline hazards of the dataset and meanwhile evaluate the coefficient $\underline{\beta}^T$ for all the prognostic factors. In this thesis, we mainly focus on the model with proportional hazard scale. For a proportional hazard model, using the fact that $\Lambda(t) = -\ln(S(t))$, we could estimate $S_0(t)$ by the obtained $\exp(-\hat{\Lambda}_0(t))$ and thereby predict the survival probabilities of new individuals using the following formula:

$$\ln \hat{S}(t) = \ln \hat{S}_0(t) \exp(\underline{\hat{\beta}}^T \underline{X})$$

where $\hat{S}_0(t)$ and $\underline{\hat{\beta}}^T$ are obtained from the developed model[120].

In Chapter 4 and 5, we will use Royston-Parmar modelling to identify prognostic factors and models for breast cancer outcome risk and meanwhile highlight the advantages of this method compared with traditional parametric models or Cox regression in survival modelling.

Prognostic index and risk groups

The last issue to address in this section is about the definition of prognostic index (PI) and risk groups in flexible parametric regression.

For practical application, one of the main products of flexible parametric regression is a prognostic index. The straightforward way to construct a prognostic index is to take the linear predictor $\underline{\beta}^T\underline{X}$ from flexible parametric model. In usual, the patient with high prognostic index is expected to experience more risks.

Instead of directly implementing the prognostic index to predict the risk probabilities of patients, alternatively, we can divide the patients into several risk groups according to the prognostic index of individuals. This is can be achieved by placing cutpoints on the prognostic index of the underlying dataset and typically, Royston et al[119] suggested that between two and five groups are appropriate to the problem.

1.7 Clinical randomised trials and treatment effects

Clinical randomized trials are very important experiments in medical research which also require highly knowledge of survival modelling especially in analysing the data with time-to-event outcome. They are often used to investigate the effect of different types of treatments, for example, clinicians usually compare the risk outcomes of patients from the new treatment group with those from the placebo group to see whether the survival probabilities of patients are improved after receiving the new treatment[88].

The golden rule of clinical trial study is randomization. That is, all the participant patients are randomized to different treatment groups without knowing which group they stay and then are followed exactly in the same way except for the different treatment care they may receive. One of the great advantages of this design is that it minimizes the bias in allocation and further it balances both known and unknown confounding factors in the baseline of patients in different groups. Many references can be found to inform how to practice a clinical randomized trial, for example, see Jadad et al.[72] and Rosenberger and

Lachin[114].

In clinical randomized trials, we may be only interested in investigating the treatment effect on patients and how the treatment effect is impacted by any patient-level factor[106][110]. By identifying whether the treatment effects differ for certain patients, clinicians could offer more personalized and precise medical services to individual patients based on their own health measurements and characteristics. Therefore in Chapter 6 we evaluate how to estimate interaction effects between treatment and patient-level covariate using time-to-event data from multiple clinical trials.

1.8 Aims and outlines of the thesis

The aims of this thesis could be summarized into two parts: mathematical research in Chapter 2 and 3 and medical research from Chapter 4 to 6.

In Chapter 2, we extend the generalized method of Naito[97] to the setting of hazard rate estimation. We demonstrate the advantage in precision accuracy of the semiparametric estimate when the true function is close to our prior assumed model and introduce the ways to choose the shape parameter in the sense of minimizing the mean integrated squared errors of the estimator. In Chapter 3, we propose an optimal kernel hazard rate estimator in the sense of minimizing its L_1 error. Then we discuss how to derive a data-driven bandwidth to minimize the L_1 error of the estimator in practice.

In Chapter 4, we develop a prognostic model to investigate primary breast cancer mortality rates across countries. To model the baseline hazards, flexible parametric regression is utilized and to account for the missing values, multiple imputation strategy is adopted. In Chapter 5, we develop and validate prognostic models for breast cancer mortality using individual participant data from multiple studies. An internal and external cross-validation scheme is also introduced to validate a model on multiple occasions, and its performance is summarised by meta-analysis. In Chapter 6, we develop statistical methods to identify interaction effects between treatment and a patient-level factor on the survival probabilities of patients, when using data from multiple trials and a multi-level survival modelling framework.

Finally, in the last chapter, we summarize the findings and recommendations of the thesis and give suggestions for future work.

CHAPTER 2

SEMIPARAMETRIC HAZARD ESTIMATION

2.1 Introduction

In a survival analysis, a group of failure times may be observed. Suppose that all the observations in the study are non-censored and share with the same probability distribution such as Weibull, we are then interested to explore the hazard rate function of the dataset. However, mostly we only have the partial information about the model underlying the given data. Thus in such situations neither the use of a parametric approach nor the use of purely nonparametric approach to estimation seems appropriate, and a different kind of approach is required which incorporates the partial knowledge about the possible parametric model in nonparametric estimation. Such approaches to estimation are generally referred to as semiparametric approaches to estimation.

Typically, semiparametric methods are divided into two groups. One is to rely on a nonparametric method to estimate the underlying model initially and then make use of the partial knowledge to add a possible parametric model to the nonparametric estimate referred to as additive correction factor. This approach has been described briefly in the regression and density estimation settings in Section 1.5. The other is to fit the data with an assumed parametric model and if the fit is not satisfactory then use a nonparametric correction factor to modify the model that one started with. In this chapter, we will only focus on the latter approach referred to as multiplicative correction factor and propose a semiparametric hazard rate estimate using this approach. Notice that in this chapter, the proposed estimate is to model the data without censorship, however, it could be easily generalized to the censored case.

The various multiplicative correction factor approaches in the settings of density estimation are illustrated in Section 2.2. There a detailed discussion of a generalized multiplicative correction factor approach proposed by Naito[97] is provided. By introducing a parameter α, here referred to as shape parameter, Naito[97] unifies different approaches of devising multiplicative correction factors. Naito's generalized approach to select a multiplicative correction factor is then extended to the setting of hazard rate estimation in Section 2.3. The hazard rate estimator obtained using the generalized approach of Naito[97] is referred to as generalized hazard rate estimator. In Section 2.4, we exhibit and discuss the role α plays in the generalized density and hazard rate estimators. There by carrying out example studies with different parametric assumptions and different underlying models we provide insight into the role of this parameter. In Section 2.5, the asymptotic properties of the generalized estimator are investigated. The methods to estimate the shape parameter and bandwidth are given in Section 2.6. Section 2.7 is devoted to the proof and the final section summarizes the key findings of this chapter.

2.2 Semiparametric estimation of density function

In this section, we first describe the multiplicative correction factor approach in the estimation of density which provides important guidelines to develop a generalized estimate in the hazard rate setting. Let $(X_1, X_2, ..., X_n)$ be a random sample from a probability density function $f(x)$, where $f(x)$ is unknown. It is suspected that the unknown density function $f(x)$ can be approximated by the density function $f(x;\underline{\theta})$. Here although the parameter which characterizes the function $f(x;\underline{\theta})$ is unknown, the functional form of $f(x;\underline{\theta})$ is known. Now let $\hat{\underline{\theta}}$ denotes say, an estimator of $\underline{\theta}$. Then define $\hat{f}$ an estimator of f as a product,

$$\hat{f}(x) = f(x;\hat{\underline{\theta}})\xi(x)$$

where $\xi(x)$ is the nonparametric multiplicative correction factor. The parametric vector $\underline{\theta}$ may be estimated by the maximum likelihood method. The role of $\xi(x)$ is to provide a correction to the initial parametric guess if the parametric estimate $f(x;\hat{\underline{\theta}})$ does not provide a satisfactory fit to the data. Different criterion to select the correction factor $\xi(x)$ leads to different estimates which are described in the rest of this section.

A couple of estimators which are based on the above idea are proposed by Hjort and Jones[63]. To define them, assume $\hat{f}_i(x) = f(x;\hat{\underline{\theta}})\xi_i(x)$, $i = 0, 1$. That is $f(x;\hat{\underline{\theta}})$ is the parametric estimator and $\xi_i(x)$s are the correction factors. Then the correction factor $\xi_0(x)$ of the first of the two estimators is now the minimizer of the local squared distance between $f(x;\hat{\underline{\theta}})\xi(x)$ and the true density function,

$$q(x,\xi(x)) = \int K_h(t-x)\{f(t) - f(t;\hat{\underline{\theta}})\xi(t)\}^2 dt.$$

To minimize $q(x,\xi(x))$, we let the differential of $q(x,\xi(x))$ w.r.t $\xi(x)$ be 0 and solve it to obtain

$$\xi_0(x) = \frac{\int K_h(t-x)f(t)f(t;\hat{\underline{\theta}})dt}{\int K_h(t-x)f(t;\hat{\underline{\theta}})^2 dt}.$$

Then taking sample analogue of $\xi_0(x)$ we get the correction factor,

$$\hat{\xi}_0(x) = \frac{n^{-1}\sum_{i=1}^n K_h(X_i-x)f(X_i;\hat{\underline{\theta}})}{\int K_h(t-x)f(t;\hat{\underline{\theta}})^2 dt}.$$

This leads to the first of the two estimators proposed by Hjort and Jones[63],

$$\hat{f}_0(x) = f(x;\hat{\underline{\theta}})\frac{n^{-1}\sum_{i=1}^n K_h(X_i-x)f(X_i;\hat{\underline{\theta}})}{\int K_h(t-x)f(t;\hat{\underline{\theta}})^2 dt}. \quad (2.2.1)$$

The second correction factor $\xi_1(x)$ of the two estimators proposed by Hjort and

Jones[63] is determined by minimizing the local Kullback-Leibler distance w.r.t $\xi(x)$, *i.e.*

$$l(x, \xi(x)) = \int K_h(t-x) \left\{ f(t) \log \left(\frac{f(t)}{f(t; \hat{\underline{\theta}}) \xi(t)} \right) - (f(t) - f(t; \hat{\underline{\theta}}) \xi(t)) \right\} dt.$$

If the semiparametric estimate is close to the true function $f(x)$, then the distance $l(x, \xi(x))$ should be close to 0. This fact shows that minimizing the local Kullback-Leibler distance between the semiparametric estimate and true function is a reasonable way to determine the correction factor $\xi(x)$. To minimize $l(x, \xi(x))$, again we set the differential of $l(x, \xi(x))$ w.r.t $\xi(x)$ be 0 and solve it to obtain

$$\xi_1(x) = \frac{\int K_h(t-x) f(t) dt}{\int K_h(t-x) f(t; \hat{\underline{\theta}}) dt}.$$

Again taking the sample analogue of $\xi_1(x)$ we get the correction factor which is given by,

$$\hat{\xi}_1(x) = \frac{n^{-1} \sum_{i=1}^n K_h(X_i - x)}{\int K_h(t-x) f(t; \hat{\underline{\theta}}) dt}.$$

This leads to the second of the two estimators proposed by Hjort and Jones[63],

$$\hat{f}_1(x) = f(x; \hat{\underline{\theta}}) \frac{n^{-1} \sum_{i=1}^n K_h(X_i - x)}{\int K_h(t-x) f(t; \hat{\underline{\theta}}) dt}. \tag{2.2.2}$$

Hjort and Glad[62] proposed a density estimator that consists of a parametric start and a nonparametric correction factor. To explain their proposal, let the model which provides approximate description of the data be a parametric density $f(x; \underline{\theta})$ where $\underline{\theta}$ is estimated by the maximum likelihood method. Then multiply the guess $f(x; \hat{\underline{\theta}})$ by a correction factor $\xi_2(x) = \frac{f(x)}{f(x; \underline{\theta})}$ to modify the possible misspecfication of the parametric assumption where $f(x)$ is the true unknown density. Notice that in estimation of correction factor $\xi_2(x)$, if one estimates its numerator, $f(x)$, by the kernel method, and the dominator, $f(x; \underline{\theta})$, by the parametric method, then the multiplicative estimator will be reduced to a standard kernel estimator. However, Hjort and Glad[62] proposed a kernel-type estimator of $\xi_2(x)$ treating the correction factor as a function in itself and define $\hat{\xi}_2(x) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i)/f(X_i; \hat{\underline{\theta}})$. This results into the third density estimator which uses idea of multiplicative correction described at the beginning of this section and is given by

$$\hat{f}_2(x) = f(x; \hat{\underline{\theta}})\hat{\xi}_2(x) = \frac{f(x; \hat{\underline{\theta}})}{n} \sum_{i=1}^{n} \frac{K_h(x - X_i)}{f(X_i; \hat{\underline{\theta}})}. \qquad (2.2.3)$$

Naito[97] observed that the local L_2 fitting criteria that he proposed provides a more general method of deriving correction factors. According to this criterion, the correction factor $\xi_\alpha(x)$ is obtained by minimizing,

$$Q(x, \xi_\alpha(x)) = \int K_h(t - x)\frac{(f(t) - f(t; \hat{\underline{\theta}})\xi_\alpha(x))^2}{f(t; \hat{\underline{\theta}})^\alpha} dt$$

for a fixed point x when $\alpha \geq 0$. This is the local squared distance between the semiparametric estimate and the true function scaled by $f(t; \hat{\underline{\theta}})^\alpha$. The minimizer of $Q(x, \xi_\alpha(x))$ with respect to $\xi_\alpha(x)$ is attained at

$$\xi_\alpha(x) = \frac{\int K_h(t - x) f(t) f(t; \hat{\underline{\theta}})^{1-\alpha} dt}{\int K_h(t - x) f(t; \hat{\underline{\theta}})^{2-\alpha} dt}. \qquad (2.2.4)$$

Thus using the sample analogue of $\xi_\alpha(x)$, the kernel-type estimate of $\xi_\alpha(x)$ is

$$\hat{\xi}_\alpha(x) = \frac{n^{-1}\sum_{i=1}^{n} K_h(X_i - x) f(X_i; \hat{\underline{\theta}})^{1-\alpha}}{\int K_h(t - x) f(t; \hat{\underline{\theta}})^{2-\alpha} dt}.$$

The generalized estimator of the true but unknown density function is

$$\hat{f}_\alpha(x) = f(x; \hat{\underline{\theta}})\frac{n^{-1}\sum_{i=1}^{n} K_h(X_i - x) f(X_i; \hat{\underline{\theta}})^{1-\alpha}}{\int K_h(t - x) f(t; \hat{\underline{\theta}})^{2-\alpha} dt}.$$

That is, every $\alpha \geq 0$ leads to a distinct estimator. In fact, if one takes $\alpha = 0$, then the generalized estimator can be simplified to the one given by (2.2.1). Similarly, if α is set to equal to 1 or 2, the generalized estimator will be reduced to the estimators defined in

(2.2.2) or (2.2.3) respectively. Hence it can be seen that in the density estimation setting, the parameter α, unifies the different multiplicative methods discussed in the literature. It also provides more flexibilities in selecting the correction factor.

In terms of the precision of the proposed methodology for the density estimation, Naito[97] shows that asymptotically the semiparametric approach performs better than their fully nonparametric kernel counterparts in a broad family of functions around the true model. To appreciate the usefulness or the good performance of the proposed methodology, it will be useful to gain an intuitive insight into the role α plays. One of the ways this could be achieved is by graphical illustration. That is, first exhibiting the closeness of target function and its crude guess multiplied by a correction factor (as a function of α) and then illustrating the closeness as α varies. In fact, this investigation has been carried out in Section 2.4 and its admirable performance has been one of the motivating factors for extending the Naito's methodology to hazard rate estimation. However, before we consider such pictorial illustration, we will first consider the extension of the methodology to semiparametric hazard rate estimation.

2.3 Semiparametric estimation of hazard rate function

The multiplicative semiparametric methodology described in the settings of density estimation is considered for hazard rate estimation in this section. For instance, Hjort et al.[64] discussed a semiparametric estimator of hazard rate by minimizing the local Kullback-Leibler distance between the estimate and the true hazard rate function which is an extension of the density estimate by Hjort and Jones[63]. Anderson[5] proposed a semiparametric approach to hazard rate estimation with a parametric start modified by a nonparametric correction factor. However, there is no theoretical investigation of the individual approaches mentioned above in the settings of hazard rate estimation, let alone investigation into unifying different approaches. Thus here we first propose a generalized semiparametric hazard rate estimator and show that the approaches in Hjort et al.[64]

and Anderson[5] are special cases of the generalized estimator.

Assume that $(X_1, X_2, ..., X_n)$ is a random sample of event times from a probability density $f(x)$ with the distribution function $F(x)$ which are all known. The standard kernel hazard rate estimator as defined in Section 1.4.2 is then given by

$$\hat{\lambda}(x) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{K_h(x - X_i)}{1 - F_n(X_i)}\right) = \sum_{i=1}^{n}\left(\frac{K_h(x - X_{(i)})}{n - i + 1}\right).$$

Now let us assume that the true unknown hazard rate $\lambda(x)$ can be approximated by function $g(x; \underline{\theta})$, where functional form of g is known but $\underline{\theta}$ is unknown. Let $\hat{\underline{\theta}}$ be the maximum likelihood estimator of $\underline{\theta}$, then the semiparametric estimator of $\lambda(x)$ will have the functional form,

$$\hat{\lambda}(x) = g(x; \hat{\underline{\theta}})\xi(x)$$

where $\xi(x)$ is a nonparametric correction to modify the crude parametric guess.

Following Naito[97], the optimal $\xi_\alpha(x)$ is determined by minimizing the local L_2 errors between true $\lambda(x)$ and $g(x; \hat{\underline{\theta}})\xi_\alpha(x)$ scaled by $g(t; \hat{\underline{\theta}})^\alpha$ where α is real number called the index. That is, by minimizing,

$$Q(x, \xi_\alpha(x)) = \int K_h(t - x)\frac{\{\lambda(t) - g(t; \hat{\underline{\theta}})\xi_\alpha(x)\}^2}{g(t; \hat{\underline{\theta}})^\alpha}dt,$$

this yields the correction factor $\xi_\alpha(x)$ where

$$\xi_\alpha(x) = \frac{\int K_h(t - x)\lambda(t)g(t; \hat{\underline{\theta}})^{1-\alpha}dt}{\int K_h(t - x)g(t; \hat{\underline{\theta}})^{2-\alpha}dt}. \tag{2.3.1}$$

The sample counterpart of the correction factor is

$$\hat{\xi}_\alpha(x) = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{K_h(X_i - x)}{1 - F_n(X_i)}g(X_i; \hat{\underline{\theta}})^{1-\alpha}}{\int K_h(t - x)g(x; \hat{\underline{\theta}})^{2-\alpha}(t)dt}.$$

Thus the generalized estimator of the true but unknown hazard rate $\lambda(x)$ is

$$\hat{\lambda}_\alpha(x) = g(x;\underline{\hat{\theta}})\frac{\frac{1}{n}\sum_{i=1}^{n}\frac{K_h(X_i-x)}{1-F_n(X_i)}g(X_i;\underline{\hat{\theta}})^{1-\alpha}}{\int K_h(t-x)g(x;\underline{\hat{\theta}})^{2-\alpha}(t)dt}.$$

This estimator unifies the different multiplicative semiparametric estimators of hazard rate.

For the case $\alpha = 0$, the correction factor of a generalized estimator is given by

$$\xi_0(x) = \frac{\int K_h(t-x)\lambda(t)g(t;\underline{\hat{\theta}})dt}{\int K_h(t-x)g(t;\underline{\hat{\theta}})^2 dt}.$$

$\xi_0(x)$ is a reasonable correction factor since it minimizes the local squared distance $q(x,\xi)$ between the semiparametric estimate $g(x;\underline{\hat{\theta}})\xi(x)$ and the true hazard rate function $\lambda(x)$ where

$$q(x,\xi(x)) = \int K_h(t-x)\{\lambda(t) - g(t;\underline{\hat{\theta}})\xi(t)\}^2 dt.$$

The semiparametric hazard rate estimator corresponding to this correction factor is that

$$\hat{\lambda}_0(x) = g(x;\underline{\hat{\theta}})\frac{\frac{1}{n}\sum_{i=1}^{n}\frac{K_h(X_i-x)}{1-F_n(X_i)}g(X_i;\underline{\hat{\theta}})}{\int K_h(t-x)g(t;\underline{\hat{\theta}})^2 dt}.$$

As discussed in (2.2.1), a similar estimator has been discussed by Hjort and Jones[63] in the density setting.

For the case $\alpha = 1$, the correction factor of the generalized estimate is

$$\xi_1(x) = \frac{\int K_h(t-x)\lambda(t)dt}{\int K_h(t-x)g(t;\underline{\hat{\theta}})dt}.$$

This is a modified version of the correction factor first proposed by Hjort et al.[64]. The initial correction factor proposed by Hjort et al. is

$$\xi_1^*(x) = \frac{\int K_h(t-x)f(t)dt}{\int K_h(t-x)S(t)g(t;\underline{\hat{\theta}})dt}.$$

As discussed by Hjort et al.[64], the initial $\xi_1^*(x)$ is determined by minimizing the local Kullback-Leibler distance $l(x, \xi(x))$ from the semiparametric estimator $g(t; \underline{\hat{\theta}})\xi(x)$ to the true $\lambda(x)$, *i.e.*

$$l(x, \xi(x)) = \int K_h(t-x)[f(t)(\log \lambda(t) - \log g(t; \underline{\hat{\theta}})\xi(t)) - S(t)(\lambda(t) - g(t; \underline{\hat{\theta}})\xi(t))]dt.$$

However, note that asymptotically, both $\xi_1(x)$ and $\xi_1^*(x)$ converge to the same quantity, $\lambda(x)/g(x; \underline{\hat{\theta}})$, as $n \to \infty$ and $h \to 0$.

Now since the sample counterpart of $\xi_1(x)$ equals to

$$\hat{\xi}_1(x) = \frac{1}{n} \frac{\sum_{i=1}^{n} \frac{K_h(X_i - x)}{1 - F_n(X_i)}}{\int K_h(t-x) g(t; \underline{\hat{\theta}}) dt},$$

we have a kernel-type estimator of $\lambda(x)$ given by

$$\hat{\lambda}_1(x) = g(x; \underline{\hat{\theta}}) \frac{1}{n} \frac{\sum_{i=1}^{n} \frac{K_h(X_i - x)}{1 - F_n(X_i)}}{\int K_h(t-x) g(t; \underline{\hat{\theta}}) dt}.$$

For the case $\alpha = 2$, the estimator of the correction factor is given by

$$\hat{\xi}_2(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{K_h(X_i - x)}{(1 - F_n(X_i)) g(X_i; \underline{\hat{\theta}})}.$$

Notice that $\hat{\xi}_2(x)$ is a kernel-type estimator of the correction factor $\xi(x) = \lambda(x)/g(x; \underline{\theta})$ that treats $\xi(x)$ as a function itself. It was first proposed by Hjort and Glad[62] in the density estimation setting given by (2.2.3). Anderson[5] developed this correction factor $\hat{\xi}_2(x)$ in hazard rate estimation and proposed a multiplicative estimate with $\hat{\xi}_2$ as,

$$\hat{\lambda}_2(x) = g(x; \underline{\hat{\theta}}) \frac{1}{n} \sum_{i=1}^{n} \frac{K_h(X_i - x)}{(1 - F_n(X_i)) g(X_i; \underline{\hat{\theta}})}.$$

Before studying mathematical properties of the generalized hazard rate estimator, we now carry out the example studies to illustrate and to gain an insight into the role of α

in the generalized semiparametric estimators.

2.4 The role of α

In the last section, we have seen that the introduction of parameter α helps to unify different approaches of selecting a multiplicative correction factor in semiparametric hazard rate estimation just the way it does so in the semiparametric density estimation. Through mathematical analysis, Naito[97] has illustrated the positive influence of α on the mean squared analysis of the generalized semiparametric density estimator. This appealing positive influence of α on the mean squared analysis is one of main motivations of extending Naito's generalized semiparametric density estimation methodology to hazard rate estimation. However, there is no graphical illustration of the very appealing role α plays in approximating the true function (either density or hazard) by a corrective mechanism considered in the last section. Thus the main focus of this section is to carry out example studies which will help to exhibit the role of α pictorially, and to gain more understanding about the parameter α.

In each example study below, we utilize the true correction factor ξ_α instead of its kernel-based estimate $\hat{\xi}_\alpha$ to fit the underlying curve. The advantage of this setting is that, making use of true correction factors, the example studies of density and hazard rate functions can be held at the same time since the correction factors in density estimation and hazard rate estimation have the same functional form. Also, the vector $\underline{\theta}$ of the parameters is assumed to be known in studies for simplicity.

Detailed steps of each study are given as follows,

1. Decide the true underlying function $f(x)$ or $\lambda(x)$ and the parametric models $f(x;\underline{\theta})$ or $g(x;\underline{\theta})$ that will be used instead of the true models.

2. Fix the bandwidth h for each study.

From (2.2.4) and (2.3.1), it is clear that with a very small bandwidth h, regardless of the choice of α, the corresponding approximation, $f_\alpha(x)$ or $\lambda_\alpha(x)$ is expected to be almost equal to the true function. On the other hand, if h is extremely large, all the approximations will be far off from the true model. Hence so as to output the satisfactory results that the divergence between each approximations with different α can be clearly viewed on graphs, an appropriate h is chosen which compromises the two extreme cases.

For example studies in density estimation, we use

$$f_\alpha(x) = f(t;\underline{\theta})\xi_\alpha(x)$$

where

$$\xi_\alpha(x) = \frac{\int K_h(t-x)f(t)f(t;\underline{\theta})^{1-\alpha}dt}{\int K_h(t-x)f(t;\underline{\theta})^{2-\alpha}dt}.$$

In hazard rate estimation, we use

$$\lambda_\alpha(x) = g(t;\underline{\theta})\xi_\alpha(x)$$

where

$$\xi_\alpha(x) = \frac{\int K_h(t-x)\lambda(t)g(t;\underline{\theta})^{1-\alpha}dt}{\int K_h(t-x)g(t;\underline{\theta})^{2-\alpha}dt}.$$

4. Generate two panels in one figure for each example study. In the first panel, the curves of the true function and the parametric model that will be used to model the true function are plotted in order to provide an intuitive description of the difference between the two functions. The obtained four approximations of the true model by the product of assumed parametric model and correction factor with different values of α are plotted against the true function in the second panel.

2.4.1 Example of a normal data with a cauchy start

The first example is a simple application in the density setting. Let the true density function $f(x)$ be the standard Normal distribution (0,1) where 0 is the location parameter and 1 is the scale parameter.

Assume that the above curve is modeled by a density function $f(x;0,1)$ of a Cauchy distribution (0,1),

$$f(x;0,1) = \frac{1}{\pi(1+x^2)}$$

where 0 is the location parameter and 1 is the scale parameter.

The results are plotted in Fig 2.1. In the top panel, the density function of Normal distribution and Cauchy distribution are plotted and they are similar to one another except that Cauchy density function has thicker tails on both sides. Now note that the true model is $N(0,1)$ but we will model it as Cauchy(0,1) and multiply it by a correction factor. The bottom panel of the figure illustrates the approximations to the true density function, *i.e.* approximation of $N(0,1)$ by the product of Cauchy(0,1) and the correction factors. The bandwidth h is 0.5. It can be seen that even with such a large bandwidth, the four approximations with different α succeed to detect the shape of the underlying function (*i.e.* $N(0,1)$) except for little upwards bias around the original point 0.

This fact indicates that with a reasonable parametric model, the semiparametric approximations perform well. Different choices of α do not change the resulting approximation significantly. Here a reasonable choice of parametric model is the one that has the similar shape and location to the true function.

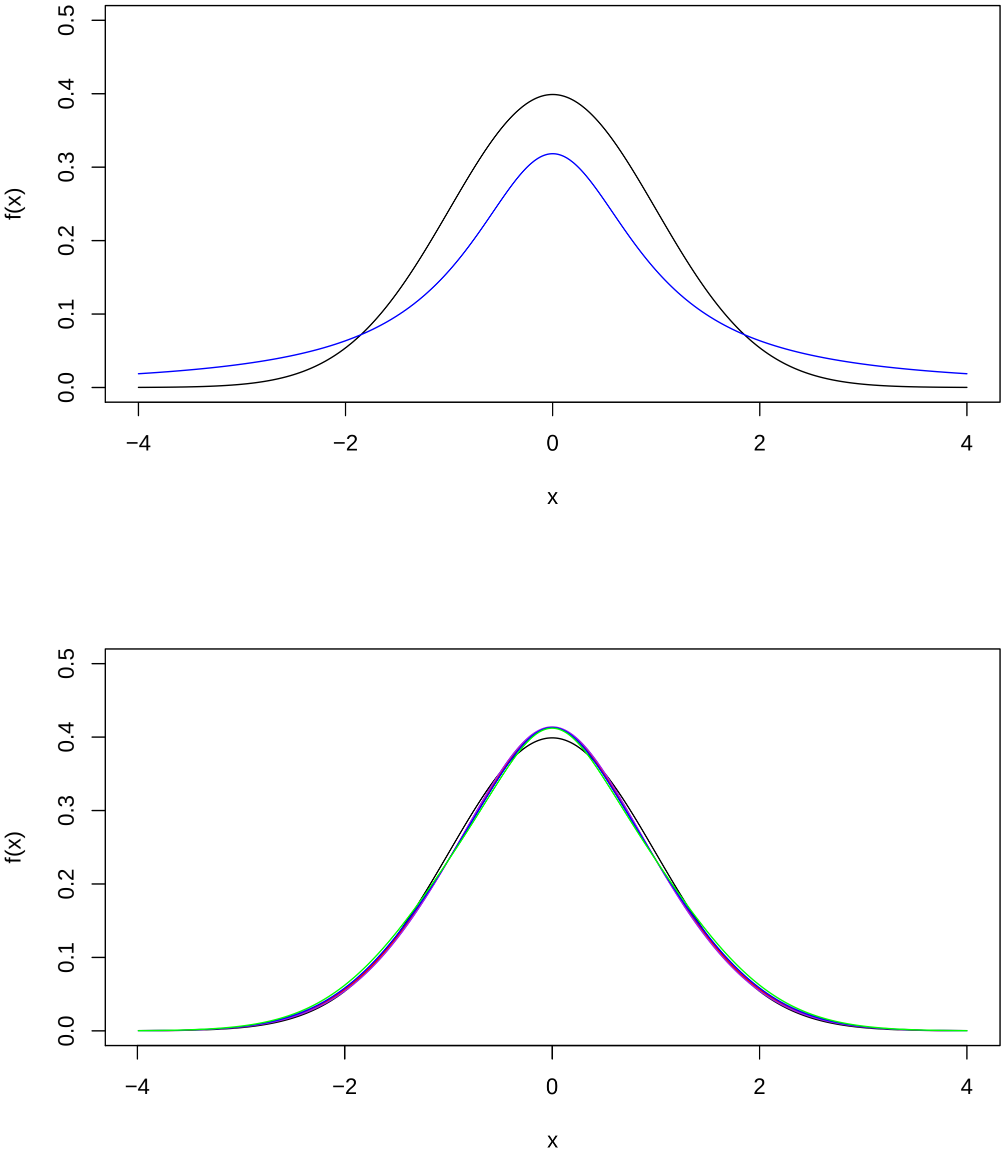

Figure 2.1: The true function is the standard normal density function $N(0, 1)$, and the parametric assumption is the cauchy distribution (0,1). In the top panel, the true curve is in black and the parametric function is in blue. The bandwidth h equals 0.5. In the bottom panel, four approximations are plotted against the true curve (in black) where the purple one is with $\alpha = 2$, the red one is with $\alpha = 1.5$, the blue one is with $\alpha = 1$ and the green one is with $\alpha = 0$.

2.4.2 Example of a normal data with a two-mixed normal start

The example in the last section was essentially to show that the generalized approximation performs well when the assumed parametric model is close to the true function in terms of its shape. Now we consider the opposite case, that is, the assumed parametric model is not an accurate guess of the true underlying model. The true curve here is chosen to be a density function of a standard normal distribution (0,1) and the assumed density function is a mixed distribution, $0.5\text{Normal}(3, 1) + 0.5\text{Normal}(-3, 1)$. The bandwidth h is fixed as 0.25.

The results of this study are depicted in Fig 2.2. From the top panel, it is clear that the assumed parametric function is far off from the true underlying model and thus one needs to modify this initial guess with correction factors. The modified results are plotted in the bottom panel and one finds that the behaviors of approximations are pretty decent even with wrong parametric model as a start. Among them, the approximation with $\alpha = 1.5$ (in red) is the most competitive in the sense that its curve is only different from the true one at the central peak.

Furthermore, from this figure, there is strong evidence to show that the choice of α is related to the shape of the assumed model but not to the smoothness of the assumed model.

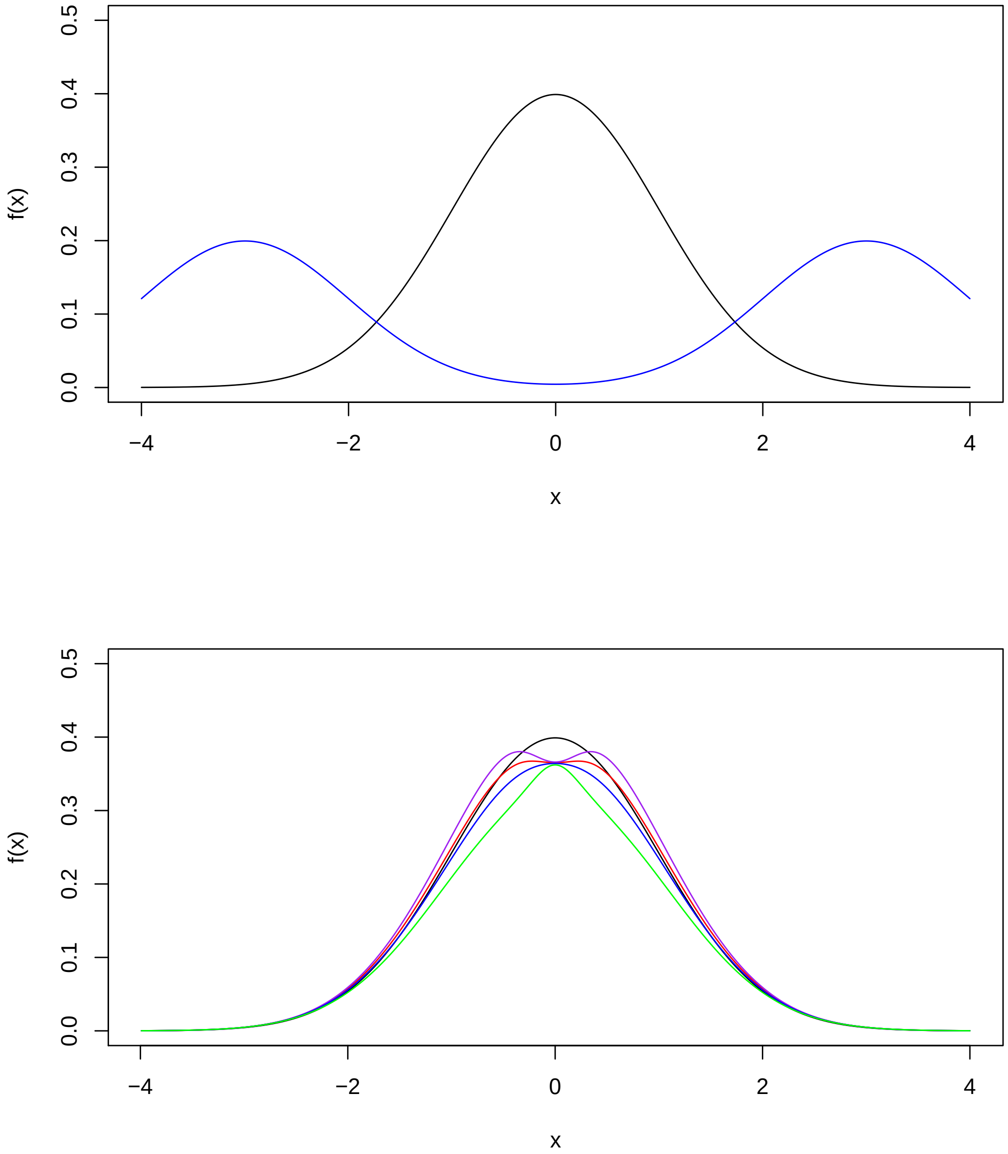

Figure 2.2: The parametric assumption is a mixed density distribution $0.5\text{Normal}(3, 1) + 0.5\text{Normal}(-3, 1)$ and the true density function is from a standard normal distribution (0,1). In the top panel, the true curve is in black and the parametric function is in blue. In the bottom panel, four approximations are plotted against the true curve (in black) where the purple one is with $\alpha = 2$, the red one is with $\alpha = 1.5$, the blue one is with $\alpha = 1$ and the green one is with $\alpha = 0$.

2.4.3 Example of a three-mixed normal data with a cauchy start

We consider another example where the assumed parametric model is not an appropriate guess of the underlying true model. In this study, the underlying curve is a density function of a three-mixed normal distributions, $0.2\text{Normal}(-3, 1) + 0.3\text{Normal}(0, 1) + 0.5\text{Normal}(3, 1)$ and the assumption is that it follows a Cauchy (0,1) density distribution. The bandwidth h is fixed as 0.8.

The approximations of the study are plotted in Fig 2.3. Although the parametric assumption is not correct, as seen in the top panel, the generalized approximations plotted in the bottom panel are still acceptable. One finds that the approximation with $\alpha = 1$ detects all the modes of the underlying function and appears to be the best of the four candidates. This result is different from the example discussed in Section 2.4.2 where the correction factor with $\alpha = 1.5$ is optimal of the four. It verifies our former viewpoint that the optimal value of α varies and relies on the parametric assumption and the true curve of the underlying model.

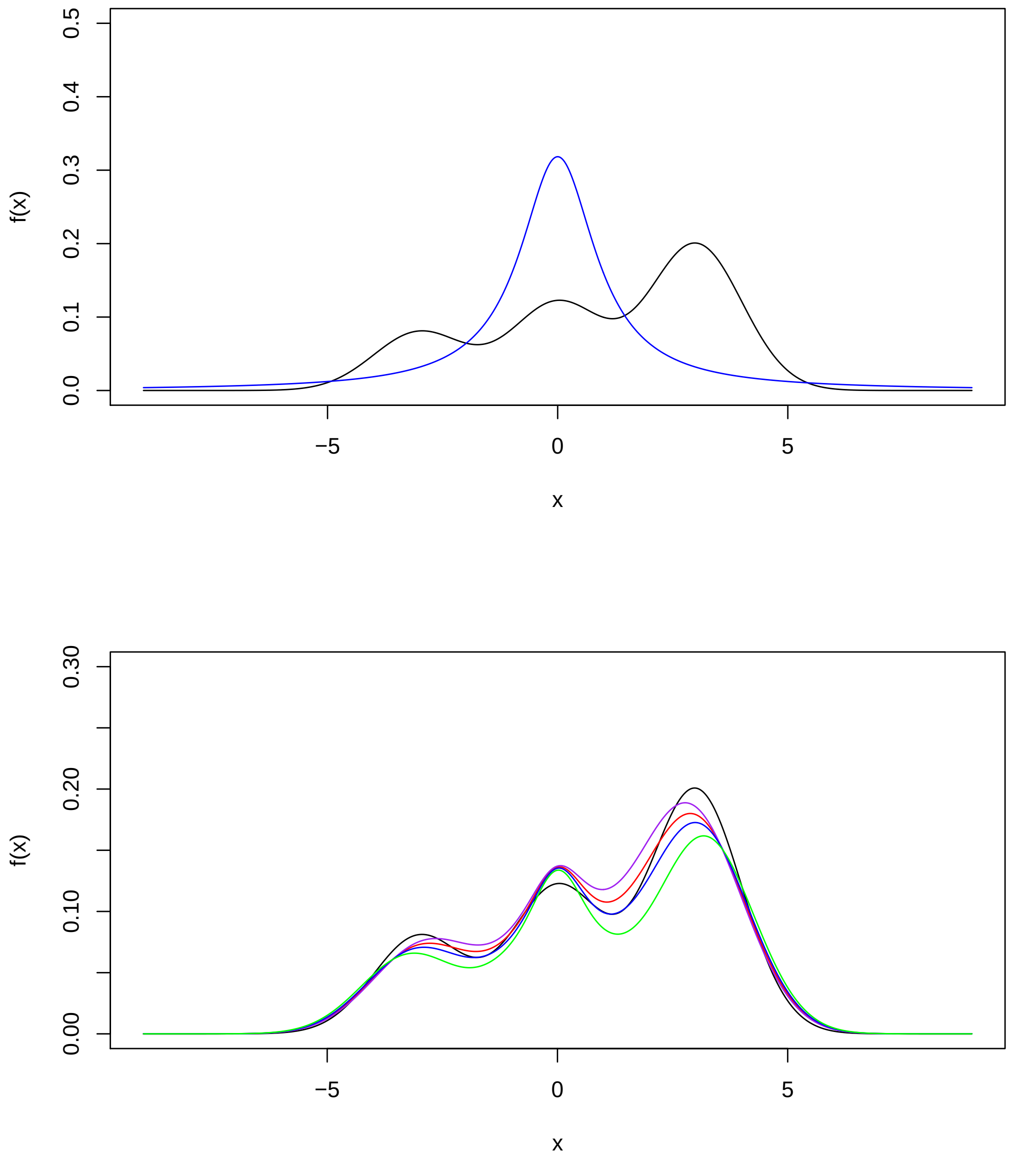

Figure 2.3: The true function is the density function from a three-mixed normal distribution $0.2\text{Normal}(-3, 1) + 0.3\text{Normal}(0, 1) + 0.5\text{Normal}(3, 1)$ and one assumes that the curve is from a Cauchy distribution (0,1). In the top panel, the true curve is in black and the parametric function is in blue where h is 0.8. In the bottom panel, four approximations are plotted against the true curve (in black) where the purple one is with $\alpha = 2$, the red one is with $\alpha = 1.5$, the blue one is with $\alpha = 1$ and the green one is with $\alpha = 0$.

2.4.4 Example of a log-normal data with a Weibull start

Here we consider the example in the settings of hazard rate estimation. Let the true curve be the hazard rate function associated with a log-normal distribution (0,1), that is,

$$\lambda(x) = \frac{\phi(\ln x)}{x\Phi(-\ln x)} \text{ for } x \geq 0$$

where ϕ is the probability density function of the standard normal distribution and Φ is the cumulative distribution function of the standard normal distribution.

We assume that the above curve is a hazard rate function of a Weibull distribution (3,3), *i.e.*

$$\lambda(x) = \left(\frac{x}{3}\right)^2 \text{ for } x \geq 0$$

where the shape parameter of the Weibull model equals to 3 and the scale parameter equals to 3.

The bandwidth h in this study is fixed as 0.3 and then the approximations of the true curve by the product of assumed curve and the correction factor are plotted in Fig 2.4. From the top panel, again it can be seen that the assumed parametric model is away from the true model. In the bottom panel, one finds that the approximation with $\alpha = 1.5$ is closest to the true function but somewhat higher than the true function before $x = 1$. Similar to the former two cases, the parameter α also determines the shape of an approximation in the example.

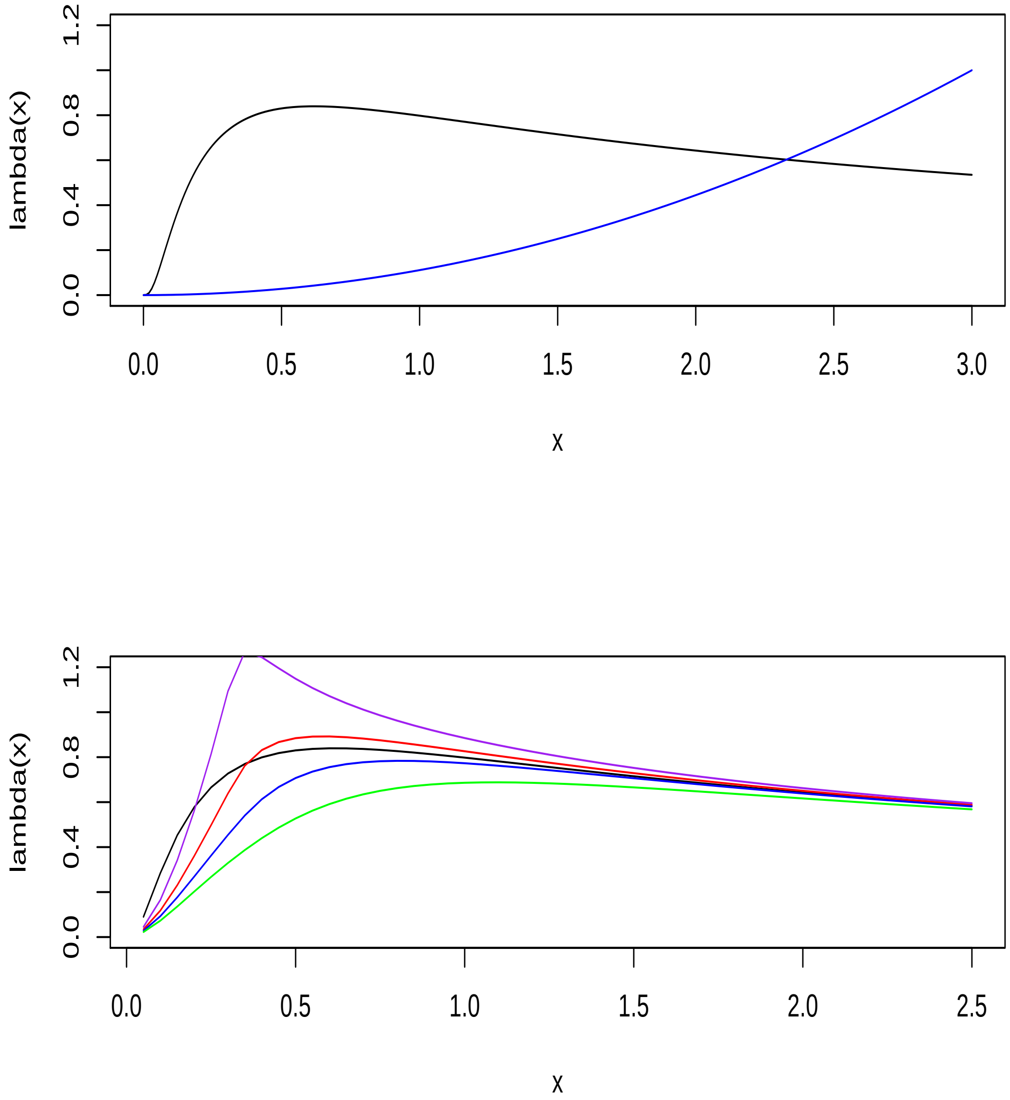

Figure 2.4: The true function is a hazard rate function from a log-normal distribution and the parametric assumption is a Weibull distribution. In the top panel, the true curve is in black and the parametric function is in blue where h is 0.3. In the bottom panel, four approximations are plotted against the true curve (in black) where the purple one is with $\alpha = 2$, the red one is with $\alpha = 1.5$, the blue one is with $\alpha = 1$ and the green one is with $\alpha = 0$.

2.4.5 Summary of example studies

From the studies illustrated in this section, we obtain an intuitive understanding of the relationship between the successive approximations to the true function and index α. That is, as α changes, the shape of the curve which approximates the true curve changes, with some values of α giving better approximations than the others. Making use of this important property, one may simply adjust the value of α so that the approximation is closer to the true function. For example, from the last example given in Fig 2.4, one may notice that the approximation with $\alpha = 1.5$ is over the true function before the point $x = 1$ while the one with $\alpha = 1$ is below. Thus one may choose a value of α between the interval $(1, 1.5)$ such as 1.3 to modify the approximation. This modified result is given in Fig 2.5. Obviously the new approximation with $\alpha = 1.3$ has a better accuracy than the four approximations given in Fig 2.4 and the bias in the approximation after $x = 0.5$ is also eliminated.

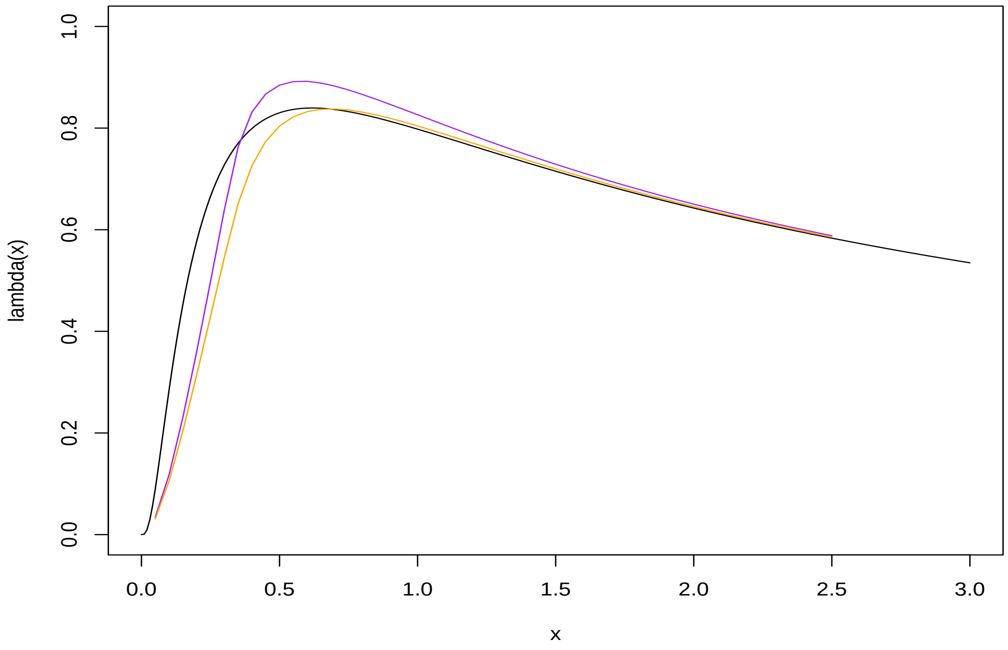

Figure 2.5: This figure is an extension of Fig 2.4, the true function is the hazard rate function from a log-normal distribution and the parametric assumption is a Weibull distribution. Two approximations are plotted against the true curve (in black) of which the purple one is with $\alpha = 1.5$ and the yellow one is with $\alpha = 1.3$.

Of course, in practice, one will not have the knowledge of the true function to decide the correct value of α and hence of the correction factor, just like we did in the example studies of this section. But these example studies are still important because they provided us with an intuitive way to understand the role of α. Now it is known that by changing α, one may adjust the position of the estimate to minimize its distance from the true function where the distance may be measured by L_2 error criterion. Hence in the next section, we will discuss analytical methods to measure this distance and propose the optimal values of α to minimize it in the hazard rate estimation.

2.5 Inference on α in semiparametric hazard rate estimation

In this section, the main problem we are concerned with is to determine the optimal value α of a generalized estimator which minimizes the mean integrated squared error.

2.5.1 Mean integrated squared errors analysis

The mean integrated squared error, which provides a reasonable quantification of the accuracy of $\hat{\lambda}_\alpha(x)$ as an estimator of the whole function $\lambda(x)$ is given by

$$\text{MISE}(\hat{\lambda}(x)) = \int \{Var(\hat{\lambda}(x)) + (Bias(\hat{\lambda}(x)))^2\} w(x) dx,$$

where

$$w(x) = \begin{cases} 1, & \text{if } 0 < x < T \\ 0, & \text{otherwise} \end{cases}$$

and $T = \inf\{x : 1 - F(x) < \epsilon\}$, $\epsilon > 0$.

In this section, we will determine the MISE of the generalized estimator, however, from the characteristic of $\hat{\lambda}_\alpha(x)$, it is trivial that its behavior still depends on that of $\hat{\underline{\theta}}$ included in the initial parametric start $g(x; \hat{\underline{\theta}})$. To address this issue, we need to define $\underline{\theta}_0$ to be the least false value of the best parametric approximation $g_0(x) = g(x; \underline{\theta}_0)$ to the true $\lambda(x)$ which minimizes the Kullback-Leibler distance measure

$$\int \lambda(x) \log\{\lambda(x)/g(x; \hat{\underline{\theta}})\} dx.$$

Then the important result which quantifies the bias and variance of the generalized hazard rate estimator is given in the next theorem and in the proof of the theorem, it will show that the difference between the maximum likelihood estimate $g(x;\hat{\underline{\theta}})$ and $g(x;\underline{\theta_0})$ could be ignored in semiparametric estimation as $n \to \infty$. Recall that $\mu_{2,K} = \int u^2 K(u)du$ and $R(K) = \int K(u)^2 du$. For the formal derivation in this chapter, we need the following assumption:

A. Hazard rate function of interest is sufficient smooths and bounded over $[0, T]$

Theorem 2.1. *Let $g(x;\hat{\underline{\theta}})$ be the maximum likelihood estimator of $g(x;\underline{\theta})$ and $K(\cdot)$ be a second-order kernel function s.t. $\int K(x)dx = 1$, $\int K(x)xdx = 0$ and $\int K(x)x^p dx < \infty$ for $p > 1$. As $n \to \infty$, $h \to 0$ and $nh \to \infty$, one has*

$$\begin{aligned}
Bias(\hat{\lambda}_\alpha(x)) &= \frac{h^2}{2}\mu_{2,K}\left[\frac{(\lambda(x)g_0^{1-\alpha}(x))''}{g_0(x)^{1-\alpha}} - \frac{\lambda(x)(g_0^{2-\alpha}(x))''}{g_0(x)^{2-\alpha}}\right] \\
&\quad + O\left(\frac{1}{n^2} + \frac{h^2}{n} + h^4\right), \\
Var(\hat{\lambda}_\alpha(x)) &= \frac{R(K)}{nh}\frac{\lambda(x)}{1-F(x)} + o\left(\frac{1}{nh}\right) + O\left(\frac{1}{n^2}\right).
\end{aligned}$$

Proof. The proof is given in Section 2.7.

Remark 2.1. The leading term of the variance of the new estimator $\hat{\lambda}_\alpha$ is same as the standard kernel estimator and is independent with α. On the other hand, the bias term of the generalized estimator is different from its purely nonparametric counterpart and is dependent on α. Notice that if the assumed model $g(x;\underline{\theta})$ is exactly same as the true $\lambda(x)$, then the bias term of the proposed estimate vanishes. This shows that a semiparametric estimator is expected to perform better than nonparametric competitors under the case when the parametric assumption is reasonable and close to the true model.

2.5.2 Choice of α with L_2 error criterion

From the example study of Section 2.4, α clearly plays vital role in the performance of the estimator $\hat{\lambda}_\alpha$. Thus in this subsection first we find an α which minimizes asymptotic

MISE, *i.e.* we find α which minimizes AMISE. Making use of Theorem 2.1, the AMISE of the proposed estimator is given by

$$AMISE(\hat{\lambda}_\alpha) = \frac{h^4}{4}\mu_{2,K}^2 R(M_\alpha) + \frac{R(K)}{nh}\int \frac{\lambda(x)}{1-F(x)} w(x)dx,$$

where $R(M_\alpha) = \int \left[\frac{(\lambda(x)g_0^{1-\alpha}(x))''}{g_0(x)^{1-\alpha}} - \frac{\lambda(x)(g_0^{2-\alpha}(x))''}{g_0(x)^{2-\alpha}}\right]^2 w(x)dx$.

The AMISE optimal choice of α can be obtained directly by differentiating the expression of AMISE w.r.t α and solving for the case when the differential equals to zero. Notice that the variance terms is independent with α, and thus the problem is reduced to searching for the value of α that minimizes $R(M_\alpha)$. For that first set,

$$b_1(x) = \lambda''(x) - \lambda(x)g_0''(x)/g_0(x),$$

$$b_2(x) = 2\left\{\frac{g_0'(x)\lambda'(x)}{g_0(x)} - \lambda(x)\left(\frac{g_0'(x)}{g_0(x)}\right)^2\right\}.$$

Then it can be seen that

$$\frac{(\lambda(x)g_0^{1-\alpha}(x))''}{g_0(x)^{1-\alpha}} - \frac{\lambda(x)(g_0^{2-\alpha}(x))''}{g_0(x)^{2-\alpha}} = b_1(x) + b_2(x) - \alpha b_2(x).$$

Also set

$$c_1 = \int b_2^2(x)w(x)dx,$$

$$c_2 = \int b_2(x)(b_1(x) + b_2(x))w(x)dx,$$

$$c_3 = \int (b_1(x) + b_2(x))^2 w(x)dx.$$

It can be showed that $R(M_\alpha) = c_1\alpha^2 - 2c_2\alpha + c_3$ and it is minimized over α at $\alpha_0 = c_2/c_1$.

This result is slightly disappointing since the AMISE optimal value of α_0 involves both c_1 and c_2 which depend on the unknown functions $\lambda^{(p)}$, $p = 0, 1, 2$. That is, to choose the optimal value of α one needs to know $\lambda^{(p)}$s which are unknown. This issue is addressed in the next section where we propose a plug-in idea to seek the data dependent value of α_0.

2.6 Plug-in estimate of the shape parameter and bandwidth

In this section, we discuss the methods to estimate the shape parameter α and the bandwidth h of the semiparametric estimator.

Two data-based methods are proposed to estimate α which is dependent on unknown quantities c_1 and c_2. The first method is a simple "plug-in" approach where the unknown functions are estimated by kernel-based methods and it is quite easy to implement. The second method provides us with a well-defined and systematic way to estimate the unknown quantities c_1 and c_2.

Plug-in method is a widely used tool in the selection of bandwidth in smoothing problems. For example, if one is intended to determine the AMISE optimal bandwidth h of a kernel density estimate $\hat{f}(x)$ which involves unknown quantities depending on the true $f^{(p)}$ for $p = 0, 1, 2$, then one replaces the unknown terms $f^{(p)}$ with their preliminary estimates $\hat{f}^{(p)}$ directly to estimate h. Similarly here we propose to estimate the terms c_1 and c_2 which involve unknown $\lambda^{(p)}$s for $p = 0, 1, 2$ on the basis of preliminary estimates of $\lambda^{(p)}$s for $p = 0, 1, 2$.

As for the choice of bandwidth, it is not the main issue of our work that the most common methods like plug-in or cross-validatory ones can be used. It will be simply discussed in the end of this section.

2.6.1 Simple plug-in method of α

Recall that the optimal value of α is $\alpha_0 = c_2/c_1$ where

$$c_1 = \int b_2^2(x)w(x)dx,$$

$$c_2 = \int b_2(x)(b_1(x) + b_2(x))w(x)dx$$

and

$$b_1(x) = \lambda''(x) - \lambda(x)g_0''(x)/g_0(x),$$

$$b_2(x) = 2\left\{\frac{g_0'(x)\lambda'(x)}{g_0(x)} - \lambda(x)\left(\frac{g_0'(x)}{g_0(x)}\right)^2\right\}.$$

To estimate α_0, we estimate the unknown $\lambda^{(p)}(x)$ by kernel-based methods, *i.e.*

$$\hat{\lambda}^{(p)}(x) = \frac{1}{nh_p^{p+1}}\sum_{i=1}^{n}\frac{K^{(p)}\left(\frac{x-X_i}{h_p}\right)}{1-F_n(X_i)},$$

where $K^{(p)}$ is the pth derivative of the kernel function K and $F_n(x)$ is the empirical distribution function. The bandwidth pilot h_p used in the kernel derivative estimate of hazard rate may be different from the bandwidth h implemented in our semiparametric estimator[166].

Then the plug-in estimators of c_1 and c_2 are given by

$$\bar{c}_1 = \int_0^T \bar{b}_2(x)^2 dx,$$

$$\bar{c}_2 = \int_0^T \bar{b}_2(x)(\bar{b}_1(x) + \bar{b}_2(x))dx$$

where

$$\bar{b}_1(x) = \hat{\lambda}''(x) - \hat{\lambda}(x)g''(x;\underline{\hat{\theta}})/g(x;\underline{\hat{\theta}}),$$

$$\bar{b}_2(x) = 2\left\{\frac{g'(x;\underline{\hat{\theta}})\hat{\lambda}'(x)}{g(x;\underline{\hat{\theta}})} - \hat{\lambda}(x)\left(\frac{g'(x;\underline{\hat{\theta}})}{g(x;\underline{\hat{\theta}})}\right)^2\right\},$$

and $T = \inf\{x : 1 - F(x) < \epsilon\}, \epsilon > 0$.

Thus,

$$\bar{\alpha}_0 = \bar{c}_2/\bar{c}_1.$$

2.6.2 Alternative method to select α

In this section, the other plug-in approach to estimate α_0 is introduced. For that write c_1 and c_2 as the sums of a family of integrals involving functional of hazard rate function $\lambda(x)$. That is

$$\begin{aligned}
c_1 &= 4\int \lambda'(x)^2 q_1(x)^2 w(x)dx + 4\int \lambda(x)^2 q_1(x)^4 w(x)dx \\
&\quad -8\int \lambda(x)\lambda'(x)q_1(x)^3 w(x)dx, \\
c_2 &= c_1 + 2\int \lambda'(x)\lambda''(x)q_1(x)w(x)dx - 2\int \lambda(x)\lambda'(x)q_1(x)q_2(x)w(x)dx \\
&\quad -2\int \lambda(x)\lambda''(x)q_1(x)^2 w(x)dx + 2\int \lambda(x)^2 q_1(x)^2 q_2(x)w(x)dx
\end{aligned}$$

where

$$q_1(x) = \frac{g_0'(x)}{g_0(x)},\ q_2(x) = \frac{g_0''(x)}{g_0(x)} = q_1'(x) + q_1(x)^2.$$

Notice that

$$\begin{aligned}
&\int \lambda'(x)^2 q_1(x)^2 w(x)dx \\
= &-2\int \lambda(x)\lambda'(x)q_1(x)q_2(x)w(x)dx + 2\int \lambda(x)\lambda'(x)q_1(x)^3 w(x)dx \\
&-\int \lambda(x)\lambda''(x)q_1(x)^2 w(x)dx + \int (\lambda(x)q_1(x)^2\lambda'(x))' w(x)dx.
\end{aligned}$$

Under Assumption A for $\lambda(x)$ over $(0,T)$,

$$\int (\lambda(x)q_1(x)^2\lambda'(x))' w(x)dx = \lim_{x\to T}\lambda(x)q_1(x)^2\lambda'(x) - \lim_{x\to 0}\lambda(x)q_1(x)^2\lambda'(x) = 0.$$

Similarly,

$$\begin{aligned}
&\int \lambda'(x)\lambda''(x)q_1(x)w(x)dx \\
= &-\int \lambda(x)\lambda'''(x)q_1(x)w(x)dx - \int \lambda(x)\lambda''(x)q_2(x)w(x)dx \\
&+\int \lambda(x)\lambda''(x)q_1(x)^2w(x)dx + \int (\lambda(x)\lambda''(x)q_1(x))'w(x)dx.
\end{aligned}$$

Under Assumption A for $\lambda(x)$ over $(0, T)$,

$$\int (\lambda(x)\lambda''(x)q_1(x))'w(x)dx = 0.$$

Therefore the above expressions for c_1 and c_2 can be written as

$$c_1 = 4\{\psi(0|4,0) - \psi(2|2,0) - 2\psi(1|1,1)\},$$

$$c_2 = c_1 + 2\{\psi(0|2,1) - \psi(3|1,0) - \psi(2|0,1) - \psi(1|1,1)\}$$

where $\psi(p|r,s) = \int \lambda(x)\lambda^{(p)}(x)q_1^r(x)q_2^s(x)w(x)dx$. Then the optimal α_0 is

$$\alpha_0 = \frac{c_2}{c_1} = 1 + \frac{N}{2D}$$

where

$$N = \psi(0|2,1) - \psi(3|1,0) - \psi(2|0,1) - \psi(1|1,1)$$

and

$$D = \psi(0|4,0) - \psi(2|2,0) - 2\psi(1|1,1).$$

It is clear that the terms N and D are dependent on the sum of $\psi(p|r,s)$s. Hence a systematic way to estimate $\hat{\alpha}_0$ is to estimate $\psi(p|r,s)$. That is,

$$\hat{\alpha}_0 = 1 + \frac{\hat{N}}{2\hat{D}}$$

where

$$\hat{N} = \hat{\psi}(0|2,1) - \hat{\psi}(3|1,0) - \hat{\psi}(2|0,1) - \hat{\psi}(1|1,1),$$

$$\hat{D} = \hat{\psi}(0|4,0) - \hat{\psi}(2|2,0) - 2\hat{\psi}(1|1,1).$$

Now to estimate $\psi(p|r,s)$, we use a kernel estimator given by

$$\begin{aligned}\hat{\psi}(p|r,s) &= \frac{1}{n^2}\sum\sum\nolimits_{i\neq j}\frac{L_g^{(p)}(X_i - X_j)}{(1-F_n(X_i))(1-F_n(X_j))}v(X_i) \\ &= \sum\sum\nolimits_{i\neq j}\frac{L_g^{(p)}(X_{(i)} - X_{(j)})}{(n-j+1)(n-i+1)}v(X_{(i)})\end{aligned}$$

where $v(X_i) = q_1^r(X_i)q_2^s(X_i)$, $L_g(x) = \frac{1}{g}L(\frac{x}{g})$ is a kernel with the bandwidth g and $L_g^{(p)}$ is the pth derivative of L_g. Kernel L is not necessarily the same as K. The inference of kernel-based estimates of the integral $\int \lambda(x)\lambda^{(p)}(x)v(x)dx$ for even p is discussed in Chapter 2 of Bagkavos[8]. Here we also consider the case for odd p. The disadvantage of this kernel type estimate is that it depends on the bandwidth g of the kernel L, so this method will not be fully automatic.

Hence to search for a proper g, again we choose the AMSE criterion to determine the optimal bandwidths $\hat{g}_N$ and $\hat{g}_D$ for the kernel-type estimates $\hat{N}$ and $\hat{D}$ respectively. For instance, to estimate $\hat{N}$, we are interested in an optimal value of g_N which will be taken as the bandwidth for all the four terms $\hat{\psi}(0|2,1)$, $\hat{\psi}(3|1,0)$, $\hat{\psi}(2|0,1)$ and $\hat{\psi}(1|1,1)$ of the functional form of N, and this optimal value is determined by minimizing the AMSE of $\hat{N}$. Further $v(x) = q_1^r(x)q_2^s(x)$ depends on the parametric model $g_0(x) = g(x;\underline{\theta}_0)$ with the unknown parameter vector $\underline{\theta}_0$. Here θ_0 is replaced by its maximum likelihood estimate. Note that by argument similar to the one used in Theorem 2.1, it can be easily shown that replacing $\underline{\theta}_0$ by $\hat{\underline{\theta}}_0$, won't change the leading terms of the asymptotic bias and variance of a kernel estimate $\hat{\psi}(p|r,s)$. The MSE of $\hat{N}$ and $\hat{D}$ can be expanded as

$$\begin{aligned}
&MSE[\hat{N}]\\
&= MSE[\hat{\psi}(0|2,1)] + MSE[\hat{\psi}(3|1,0)] + MSE[\hat{\psi}(1|1,1)] + MSE[\hat{\psi}(2|0,1)]\\
&\quad - 2E[\hat{\mu}(0|2,1)\hat{\mu}(3|1,0)] - 2E[\hat{\mu}(0|2,1)\hat{\mu}(2|0,1)] - 2E[\hat{\mu}(0|2,1)\hat{\mu}(1|1,1)]\\
&\quad + 2E[\hat{\mu}(3|1,0)\hat{\mu}(2|0,1)] + 2E[\hat{\mu}(3|1,0)\hat{\mu}(1|1,1)] + 2E[\hat{\mu}(2|0,1)\hat{\mu}(1|1,1)],
\end{aligned}$$

$$\begin{aligned}
&MSE[\hat{D}]\\
&= MSE[\hat{\psi}(0|4,0)] + MSE[\hat{\psi}(2|2,0)] + 4MSE[\hat{\psi}(1|1,1)]\\
&\quad - 2E[\hat{\mu}(0|4,0)\hat{\mu}(2|2,0)] - 4E[\hat{\mu}(0|4,0)\hat{\mu}(1|1,1)] + 4E[\hat{\mu}(2|2,0)\hat{\mu}(1|1,1)]
\end{aligned}$$

where $\hat{\mu}(p|r,s) = \hat{\psi}(p|r,s) - \psi(p|r,s)$. From the above expansions to derive formulas of $MSE(\hat{N})$ and $MSE(\hat{D})$, one needs to know the MSE of $\hat{\psi}(p|r,s)$ and the covariance between the paired $\hat{\psi}(p|r,s)$s. Therefore in the next two theorems, we summarize the important results of the MSE and covariance for $\hat{\psi}(p|r,s)$ respectively.

Theorem 2.2. *Let $q_1(x)$ and $q_2(x)$ be estimated by $g(x;\underline{\hat{\theta}})$ and $v(x) = q_1^r(x)q_2^s(x)$. Then as $n \to \infty$ and $g \to 0$,*

$$Bias(\hat{\psi}(p|r,s)) = \frac{g^2}{2}\mu_{2,L(u)}\psi(p+2|r,s) + o(g^2). \tag{2.6.1}$$

For the case when p is odd,

$$\begin{aligned}
Var(\hat{\psi}(p|r,s)) = \frac{\mu_{2,(L^{(p)})^2}}{2n^2g^{2p-1}} \int \frac{\lambda(x)}{1-F(x)} \Bigg\{ \left(\frac{v^2(x)\lambda(x)}{1-F(x)}\right)''\\
- v(x)\left(\frac{v(x)\lambda(x)}{1-F(x)}\right)'' \Bigg\} w(x)dx + O\left(\frac{1}{n}\right) + o\left(\frac{1}{n^2g^{2p-1}}\right).
\end{aligned} \tag{2.6.2}$$

For the case when p is even,

$$\begin{aligned} Var(\hat{\psi}(p|r,s)) &= \frac{2}{n^2 g^{2p+1}} R(L^{(p)}) \int v^2(x) \left(\frac{\lambda(x)}{1-F(x)} \right)^2 w(x)dx \\ &+ O\left(\frac{1}{n}\right) + o\left(\frac{1}{n^2 g^{2p+1}}\right). \end{aligned} \tag{2.6.3}$$

Proof. The proof is given in Section 2.7.

In the next theorem the asymptotic covariance of $\hat{\psi}(p_1|r_1, s_1)$ and $\hat{\psi}(p_2|r_2, s_2)$ is derived. To simplify the writing we denote $\hat{\psi}(p_1|r_1, s_1)$ and $\hat{\psi}(p_2|r_2, s_2)$ simply as $\hat{\psi}_1$ and $\hat{\psi}_2$ respectively.

Theorem 2.3. *Let $q_1(x)$ and $q_2(x)$ be calculated by $g(x;\hat{\underline{\theta}})$ and $v_i(x) = q_1^{r_i}(x) q_2^{s_i}(x)$. Then as $n \to \infty$ and $g \to 0$,*

$$E[(\hat{\psi}_1 - \psi_1)(\hat{\psi}_2 - \psi_2)] = Bias(\hat{\psi}_1) \times Bias(\hat{\psi}_2) + A + O\left(\frac{1}{n}\right).$$

When p_1 and p_2 are both even,

$$A = \frac{2}{n^2 g^{p_1+p_2+1}} \int L^{(p_1)}(u) L^{(p_2)}(u) du \int v_1(x) v_2(x) \left\{ \frac{\lambda(x)}{1-F(x)} \right\}^2 w(x)dx + o\left(\frac{1}{n^2 g^{p_1+p_2+1}}\right).$$

When p_1 and p_2 are both odd,

$$\begin{aligned} A &= \frac{1}{2n^2 g^{p_1+p_2-1}} \int L^{(p_1)}(u) L^{(p_2)}(u) u^2 du \int \frac{\lambda(x)}{1-F(x)} \Bigg\{ \left[\frac{v_1(x) v_2(x) \lambda(x)}{1-F(x)} \right]'' \\ &\quad - v_2(x) \left[\frac{v_1(x)\lambda(x)}{1-F(x)} \right]'' \Bigg\} w(x)dx + o\left(\frac{1}{n^2 g^{p_1+p_2-1}}\right). \end{aligned}$$

When $p_1 + p_2$ is odd (assuming p_2 is odd and p_1 is even),

$$\begin{aligned} A &= \frac{1}{n^2 g^{p_1+p_2}} \int L^{(p_1)}(u) L^{(p_2)}(u) u du \int \frac{\lambda(x)}{1-F(x)} \Bigg\{ \left[\frac{v_1(x) v_2(x) \lambda(x)}{1-F(x)} \right]' \\ &\quad + (-1)^{p_2} v_2(x) \left[\frac{v_1(x)\lambda(x)}{1-F(x)} \right]' \Bigg\} w(x) dx + o\left(\frac{1}{n^2 g^{p_1+p_2}} \right). \end{aligned}$$

Proof. The proof is given in Section 2.7.

Remark 2.2. In comparison to the similar study given by Bagkavos[8], the introduction of odd values for p of $\hat{\psi}(p|r,s)$ leads us to deal with more possibilities with respect to p.

Thus by using Theorem 2.2 and 2.3, and also restricting the function over the support $[0,T]$, for even p one achieves that

$$\begin{aligned} & AMSE[\hat{N}[p]] \\ &= \frac{g^4}{4} \mu_{2,L}^2 N[p+2]^2 + \frac{1}{n^2 g^{2p+5}} \Bigg\{ \frac{1}{2} \mu_{2,(L^{(p+3)})^2} \int \frac{\lambda(x)}{1-F(x)} \Bigg\{ \left[\frac{q_1^2(x)\lambda(x)}{1-F(x)} \right]'' \\ & - q_1(x) \left[\frac{q_1(x)\lambda(x)}{1-F(x)} \right]'' \Bigg\} w(x) dx + 2R(L^{(p+2)}) \int q_2^2(x) \left[\frac{\lambda(x)}{1-F(x)} \right]^2 w(x) dx \\ & + 2 \int L^{(p+3)}(u) L^{(p+2)}(u) u du \int \left[\frac{\lambda(x)}{1-F(x)} \right]^2 q_1'(x) q_2(x) w(x) dx \Bigg\} \end{aligned} \tag{2.6.4}$$

where $N[p] = \psi(p|2,1) - \psi(p+3|1,0) - \psi(p+2|0,1) - \psi(p+1|1,1)$ and $N = N[0]$. Similarly

$$\begin{aligned} & AMSE[\hat{D}[p]] \\ &= \frac{g^4}{4} \mu_{2,L}^2 D[p+2]^2 + \frac{2}{n^2 g^{2p+5}} R(L^{(2)}) \int q_1(x)^4 \left(\frac{\lambda(x)}{1-F(x)} \right)^2 w(x) dx \end{aligned} \tag{2.6.5}$$

where $D[p] = \psi(p|4,0) - \psi(p+2|2,0) - 2\psi(p+1|1,1)$ and $D = D[0]$.

So as to improve the accuracy of the kernel-based estimate $\hat{N}$ and $\hat{D}$, we need to estimate the the AMSE optimal bandwidth g_N and g_D for each of them. The MSE of $\hat{N}$ and $\hat{D}$ are quantified in the following theorem.

Theorem 2.4. *Let $q_1(x)$ and $q_2(x)$ be estimated by $g(x;\hat{\underline{\theta}})$, then as $n \to \infty$, $g \to 0$,*

$$
\begin{aligned}
MSE[\hat{N}] &= \frac{g^4}{4}\mu_{2,L}^2 N[2]^2 + \frac{1}{2n^2g^5}\int\int m_{2|3}(u,x)^2 w(x)dudx \\
&\quad + O\left(\frac{1}{n}\right) + o\left(g^4 + \frac{1}{n^2g^5}\right),
\end{aligned}
$$

$$
\begin{aligned}
MSE[\hat{D}] &= \frac{g^4}{4}\mu_{2,L}^2 D[2]^2 + \frac{1}{2n^2g^5}\int\int n_2(u,x)^2 w(x)dudx \\
&\quad + O\left(\frac{1}{n}\right) + o\left(g^4 + \frac{1}{n^2g^5}\right)
\end{aligned}
$$

where

$$
m_{p_2|p_1}(u,x) = L^{(p_1)}(u)u\left(\frac{q_1'(x)\lambda(x)}{1-F(x)}\right) + 2(L^{(p_2)}(u))\left(\frac{q_2(x)\lambda(x)}{1-F(x)}\right)
$$

and

$$
n_{p_2}(u,x) = 2L^{(p_2)}(u)q_1^2(x)\left(\frac{\lambda(x)}{1-F(x)}\right).
$$

Proof. The proof is given in Section 2.7.

From Theorem 2.4, we can easily calculate the AMSE optimal bandwidths for $\hat{N}$ and $\hat{D}$ which are respectively,

$$
g_N = \left[\frac{5}{2}\frac{\int\int m_{2|3}(u,x)^2 w(x)dudx}{\mu_{2,L}^2 N[2]}\right]^{1/9} n^{-2/9},
$$

$$
g_D = \left[\frac{5}{2}\frac{\int\int n_2(u,x)^2 w(x)dudx}{\mu_{2,L}^2 D[2]}\right]^{1/9} n^{-2/9}.
$$

Unfortunately, we find that the bandwidths g_N and g_D depend on unknown quantities $N[2]$ and $D[2]$. Moreover, the estimates of $N[2]$ and $D[2]$ depend on unknown $N[4]$ and $D[4]$, *i.e.* we have to have a recursive estimation of $N[p]$ or $D[p]$, $p = 2, 4, 6, 8, ...$ without end. A common way to overcome this problem is to calculate the kernel estimates for $N[2]$

and $D[2]$ by estimating $N[4]$ and $D[4]$ with reference to some distribution. Once $\hat{N}[2]$ and $\hat{D}[2]$ are found, we can use the above formula to calculate g_N and g_D for estimating N and D. A point that needs to be answered is the number of iterations that we need to estimate the optimal bandwidth gs for N and D. Wand and Jones[166] suggested that two stages are enough in the density setting and Bagkavos[8] gave the same comment in hazard rate estimation.

Note that the estimation of $N[p]$ and $D[p]$ equals to estimating a linear sums of functions $\psi(p|r,s) = \int \lambda(x)\lambda^{(p)}(x)v(x)w(x)dx$ and we assume $\lambda(x)$ is corresponding to some commonly frequently-used parametric distribution. For example, the Weibull distribution is a sensible choice for the parametric assumption of hazard rate model.

Let the scale parameter of the Weibull distribution to be $1/c$ and the shape index be b, then the pth derivative of the hazard rate corresponding to Weibull distribution is $c^{p+1}b(b-1)(b-2)(b-p)(cx)^{b-p-1}$. Thus $\psi(p|r,s)$ can be estimated by

$$\begin{aligned}
\bar{\psi}(p|r,s) &= \int \lambda^{(p)}(x)\lambda(x)v(x)w(x)dx \\
&= \int_0^T c^{p+1}b(b-1)(b-2)...(b-p)(cx)^{b-p-1}cb(cx)^{b-1}v(x)dx \\
&= c^{p+2}b^2(b-1)(b-2)...(b-p)\int_0^T (cx)^{2b-p-2}v(x)dx \\
&= c^{2b}b^2(b-1)(b-2)...(b-p)\int_0^T x^{2b-p-2}v(x)dx.
\end{aligned}$$

Then $\bar{N}[4]$, $\bar{D}[4]$ can be calculated as the sum of a series of $\bar{\psi}(p|r,s)$s.

In addition, notice the unknown functions $\lambda(x)$ and $F(x)$ that appear above in the expressions of $m_{p_2|p_1}(u,x)$ and $n_{p_2}(u,x)$. There we replace $\lambda(x)$ with a simple kernel hazard rate estimate discussed in the last section,

$$\hat{\lambda}(x) = \frac{1}{nh_n}\sum_{i=1}^{n}\frac{K\left(\frac{x-X_i}{h_n}\right)}{1-F_n(X_i)}, \tag{2.6.6}$$

where h_n is pilot bandwidth and $F(x)$ is replaced by the empirical function $F_n(x)$. That is to say,

$$\hat{m}_{p_2|p_1}(u,x) = L^{(p_1)}(u)u\left(\frac{q_1'(x)\hat{\lambda}(x)}{1-F_n(x)}\right) + 2(L^{(p_2)}(u))\left[\frac{q_2(x)\hat{\lambda}(x)}{1-F_n(x)}\right]$$

and

$$\hat{n}_{p_2}(u,x) = 2L^{(p_2)}(u)q_1^2(x)\left\{\frac{\hat{\lambda}(x)}{1-F_n(x)}\right\}.$$

Taking advantage of the above simple estimates of $N[4]$, $D[4]$, $m_{p_2|p_1}(u,x)$ and $n_{p_2}(u,x)$, we give a detailed algorithm to estimate the index α_0 as follows,

1.Compute $\bar{N}[4]$ and $\bar{D}[4]$ with the reference to a parametric distribution.

2.Compute $\hat{m}_{4|5}$ and $\hat{n}_4$ and then compute

$$\hat{g}_{N1} = \left[\frac{9}{2}\frac{\int\int \hat{m}_{4|5}(u,x)^2 w(x)dudx}{\mu_{2,L}^2\bar{N}[4]}\right]^{1/13} n^{-2/13}$$

and

$$\hat{g}_{D1} = \left[\frac{9}{2}\frac{\int\int \hat{n}_4(u,x)^2 w(x)dudx}{\mu_{2,L}^2\bar{D}[4]}\right]^{1/13} n^{-2/13}.$$

3.Use $\hat{g}_{N1}$ and $\hat{g}_{D1}$ to estimate $N[2]$ and $D[2]$ respectively.

4.Compute $\hat{m}_{2|3}$ and $\hat{n}_2$ and then compute

$$\hat{g}_{N} = \left[\frac{5}{2}\frac{\int\int \hat{m}_{2|3}(u,x)^2 w(x)dudx}{\mu_{2,L}^2\hat{N}[2]}\right]^{1/9} n^{-2/9}$$

and

$$\hat{g}_{D} = \left[\frac{5}{2}\frac{\int\int \hat{n}_2(u,x)^2 w(x)dudx}{\mu_{2,L}^2\hat{D}[2]}\right]^{1/9} n^{-2/9}.$$

5.Use $\hat{g}_N$ and $\hat{g}_D$ to estimate N and D respectively.

6.Computer the value of α as

$$\hat{\alpha}_0 = 1 + \frac{\hat{N}}{2\hat{D}}.$$

Remark 2.3. The above algorithm is an extension of the algorithm Naito[97] had used in the similar situation while computing shape parameter for his semiparametric density estimator. Compared with the first simple plug-in method, this algorithm seems to be more stable and reasonable where the two iterations of the algorithm from pilot estimates of $N[4]$ and $D[4]$ to the ultimate estimators of N and D reduce the potential bias caused by prior assumption.

2.6.3 Choice of the bandwidth h

In this section, we introduce the plug-in estimate of the bandwidth with respect to the MISE criterion. From Theorem 2.1, we see that bandwidth h that minimizes the AMISE of $\hat{\lambda}_\alpha$ is

$$h(\alpha) = \left[\frac{R(K)}{nR(M_\alpha)\mu_{2,K}^2} \int_0^T \frac{\lambda(x)}{1 - F(x)} dx \right]^{1/5}.$$

Then the unknown terms in the expression of $h(\alpha)$ can be replaced by the plug-in estimates as follows: $\hat{\alpha}_0$ is given by either of the two methods introduced above, $\hat{\lambda}(x)$ is given by (2.6.6) and $1 - F(x)$ is estimated by its empirical version, $(1 - F_n(x))$. That is to say, h of our generalized estimate is given by

$$h(\alpha) = \left[\frac{R(K)}{nR(\hat{M}_{\hat{\alpha}_0})\mu_{2,K}^2} \int_0^T \frac{\hat{\lambda}(x)}{1 - F_n(x)} dx \right]^{1/5}.$$

2.7 Proofs

We finish by proving the theorems presented in this chapter.

Proof. of Theorem 2.1

Technically the proof can be separated into two steps of which the first is to derive the bias and variance of the generalized estimator with $g(x;\hat{\underline{\theta}})$ being replaced by $g_0(x)$, and the second is to show that the difference between the generalized estimator with $g_0(x)$ and $\hat{\lambda}_\alpha(x)$ with $g(x;\hat{\underline{\theta}})$ does not change the leading terms of variance and bias.

Without loss of generality, one defines the semiparametric estimator with the least false parametric approximation $g(x;\underline{\theta}_0)$ as

$$\hat{\lambda}^*_\alpha(x) = \frac{g_0(x)\sum_{i=1}^n \frac{K_h(X_{(i)}-x)}{n-i+1} g_0(X_{(i)})^{1-\alpha}}{\int K_h(t-x) g_0^{2-\alpha}(t)dt}.$$

Observe that

$$\begin{aligned}
E & \left\{g_0(x)\sum_{i=1}^n \left(\frac{K_h(X_{(i)}-x)}{n-i+1} g_0(X_{(i)})^{1-\alpha}\right)\right\} \\
= & \ g_0(x)\int K_h(y-x)\frac{f(y)g_0(y)^{1-\alpha}}{1-F(y)} \\
& \times \left\{\sum_{i=1}^n F(y)^{i-1}(1-F(y))^{n-i+1}\frac{n!}{(n-i+1)!(i-1)!}\right\} dy \\
= & \ g_0(x)\int K_h(y-x)\lambda(y)g_0(y)^{1-\alpha}(1-F^n(y))dy \\
= & \ g_0(x)\int K_h(y-x)\lambda(y)g_0(y)^{1-\alpha}dy \\
& - g_0(x)\int K_h(y-x)\lambda(y)g_0(y)^{1-\alpha}F^n(y)dy. \qquad (2.7.1)
\end{aligned}$$

As $n \to \infty$, $F^n(y) \to 0$, thus the second term of (2.7.1) goes to 0. Hence

$$\int K_h(t-x)g_0^{2-\alpha}(t)dt \times E[\hat{\lambda}^*_\alpha(x)] = g_0(x)\int K_h(y-x)\lambda(y)g_0(y)^{1-\alpha}dy.$$

By setting $(y-x)/h = u$, it gives

$$\begin{aligned}
& E\left\{g_0(x)\sum_{i=1}^n \left(\frac{K_h(X_i-x)}{n-i+1} g_0(X_i)^{1-\alpha}\right)\right\} \\
= & \ g_0(x)\int K(u)\lambda(x+uh)g_0(x+uh)^{1-\alpha}du \\
= & \ g_0^{2-\alpha}(x)\lambda(x) + \frac{h^2}{2}\mu_{2,K}(\lambda(x)g_0(x)^{1-\alpha})'' g_0(x) + o(h^2).
\end{aligned}$$

That is to say, $Bias(\hat{\lambda}^*_\alpha) = E(\hat{\lambda}^*_\alpha) - \lambda(x)$ can be written as

$$\begin{aligned}
Bias(\hat{\lambda}^*_\alpha(x)) &= \frac{g_0^{2-\alpha}(x)\lambda(x) + \frac{h^2}{2}\mu_{2,K}(\lambda(x)g_0^{1-\alpha}(x))''g_0(x)}{\int K_h(t-x)g_0^{2-\alpha}(t)dt} - \lambda(x) \\
&= \frac{g_0^{2-\alpha}(x)\lambda(x) + \frac{h^2}{2}\mu_{2,K}(\lambda(x)g_0^{1-\alpha}(x))''g_0(x)}{g_0^{2-\alpha}(x) + \frac{h^2\mu_{2,K}}{2}[g_0^{2-\alpha}(x)]'' + o(h^2)} - \lambda(x) \\
&= \frac{\lambda(x) + \frac{h^2}{2}\mu_{2,K}(\lambda(x)g_0^{1-\alpha}(x))''g_0^{\alpha-1}(x)}{1 + \frac{h^2\mu_{2,K}}{2}[g_0^{2-\alpha}(x)]''g_0^{\alpha-2}(x) + o(h^2)} - \lambda(x) \\
&= \frac{\frac{h^2}{2}\mu_{2,K}(\lambda(x)g_0^{1-\alpha}(x))''g_0^{\alpha-1}(x) - \frac{h^2\mu_{2,K}}{2}[g_0^{2-\alpha}(x)]''g_0^{\alpha-2}(x)\lambda(x)}{1 + \frac{h^2\mu_{2,K}}{2}[g_0^{2-\alpha}(x)]''g_0^{\alpha-2}(x) + o(h^2)}.
\end{aligned}$$

Now note that $g_0^{2-\alpha}(x)$ and $[g_0^{2-\alpha}(x)]''$ are bounded, for h sufficiently small we could prove that $Bias(\hat{\lambda}^*_\alpha)$ follows

$$\begin{aligned}
Bias(\hat{\lambda}^*_\alpha(x)) &= \frac{h^2}{2}\mu_{2,K}\left[\frac{(\lambda(x)g_0^{1-\alpha}(x))''}{g_0(x)^{1-\alpha}} - \frac{\lambda(x)(g_0^{2-\alpha}(x))''}{g_0(x)^{2-\alpha}}\right] + O(h^4) \\
&= \frac{h^2}{2}\mu_{2,K}M_\alpha(x) + O(h^4)
\end{aligned}$$

where $M_\alpha(x) = \frac{(\lambda(x)g_0^{1-\alpha}(x))''}{g_0(x)^{1-\alpha}} - \frac{\lambda(x)(g_0^{2-\alpha}(x))''}{g_0(x)^{2-\alpha}}$.

The variance of $\hat{\lambda}^*_\alpha(x)$ can be expressed as,

$$Var(\hat{\lambda}^*_\alpha(x)) = E(\hat{\lambda}^*_\alpha)^2 - (E\hat{\lambda}^*_\alpha)^2.$$

The term $(g_0(x))^{-2}E(\hat{\lambda}^*_\alpha(x))^2(\int K_h(t-x)g_0^{2-\alpha}(t)dt)^2$ can be expressed as

$$
\begin{aligned}
& (g_0(x))^{-2}E(\hat{\lambda}^*_\alpha(x))^2\Big(\int K_h(t-x)g_0^{2-\alpha}(t)dt)^2 \\
= \; & E\left[\sum_{i=1}^n \frac{K_h(x-X_{(i)})}{n-i+1} g_0(X_{(i)})^{1-\alpha}\right]^2 \\
= \; & \sum_{i=1}^n E\left(\frac{K_h^2(x-X_{(i)})}{(n-i+1)^2} g_0(X_{(i)})^{2-2\alpha}\right) \\
& +2\sum\sum\nolimits_{i<j} E\left(\frac{K_h(x-X_{(i)})K_h(x-X_{(j)})}{(n-i+1)(n-j+1)} g_0(X_{(i)})^{1-\alpha} g_0(X_{(j)})^{1-\alpha}\right) \\
= \; & \sum_{i=1}^n \int \Bigg(\frac{K_h^2(x-y)}{(n-i+1)^2} f(y) \\
& \quad \times \frac{n!}{(i-1)!(n-i)!} F(y)^{i-1}(1-F(y))^{n-i} g_0(y)^{2-2\alpha}\Bigg) dy \\
& +2\int\int_{y\le z}\Bigg(K_h(x-y)K_h(x-z)(g_0(y)g_0(z))^{1-\alpha} f(y)f(z) \\
& \times \sum_{i=1}^{n-1} \frac{F(y)^{i-1}n!}{(n-i+1)!(i-1)!} \\
& \quad \times \sum_{j=i+1}^n \frac{(n-i)!(F(z)-F(y))^{j-i-1}(1-F(z))^{n-j}}{(j-i-1)!(n-j+1)!}\Bigg) dydz \\
= \; & \sum_{i=1}^n \int \Bigg(\frac{K_h^2(x-y)}{(n-i+1)^2} f(y) \frac{n!}{(i-1)!(n-i)!} \\
& \quad \times F(y)^{i-1}(1-F(y))^{n-i} g_0(y)^{2-2\alpha}\Bigg) dy \\
& +2\int\int_{y\le z}\Bigg(K_h(x-y)K_h(x-z)(g_0(y)g_0(z))^{1-\alpha} f(y)\lambda(z) \\
& \times \sum_{i=1}^{n-1} \frac{F(y)^{i-1}n!}{(n-i+1)!(i-1)!} \\
& \quad \times \sum_{j=i+1}^n \frac{(n-i)!(F(z)-F(y))^{j-i-1}(1-F(z))^{n-j+1}}{(j-i-1)!(n-j+1)!}\Bigg) dydz
\end{aligned}
$$

$$
\begin{aligned}
& (g_0(x))^{-2}E(\hat{\lambda}^*_\alpha(x))^2(\int K_h(t-x)g_0^{2-\alpha}(t)dt)^2 \\
= & \int I_n(F_n(y))\lambda(y)K_h^2(x-y)g_0(y)^{2-2\alpha}dy \\
& +2\int\int_{y\le z} K_h(x-y)K_h(x-z)(g_0(y)g_0(z))^{1-\alpha}\lambda(y)\lambda(z) \\
& \times\Bigg\{\sum_{i=1}^{n-1}\frac{n!F(y)^{i-1}(1-F(y))^{n-i+1}}{(n-i+1)!(i-1)!} \\
& -\frac{1-F(y)}{F(z)-F(y)}\sum_{i=1}^{n-1}\frac{n!F(y)^{i-1}(F(z)-F(y))^{n-i+1}}{(n-i+1)!(i-1)!}\Bigg\}dydz \\
= & \int I_n(F_n(y))\lambda(y)K_h^2(x-y)g_0(y)^{2-2\alpha}dy \\
& +2\int\int_{y\le z} K_h(x-y)K_h(x-z)\lambda(y)\lambda(z)(g_0(y)g_0(z))^{1-\alpha} \\
& \times\left\{(1-F(y)^n)-(1-F(y))\frac{F(z)^n-F(y)^n}{F(z)-F(y)}dydz\right\}
\end{aligned}
$$

where $I_n(F(y))=\sum_{i=1}^n\frac{1}{n-i+1}\binom{n}{i-1}F(y)^{i-1}(1-F(y))^{n-i+1}$.

Further it follows from (2.7.1),

$$
\begin{aligned}
& (g_0(x))^{-2}[E\hat{\lambda}^*_\alpha(x)]^2(\int K_h(t-x)g_0^{2-\alpha}(t)dt)^2 \\
= & \int\int K_h(x-y)K_h(x-z)\lambda(y)\lambda(z)(g_0(y)g_0(z))^{1-\alpha} \\
& (1-F(y)^n)(1-F(z)^n)dydz \\
= & 2\int\int_{y\le z} K_h(x-y)K_h(x-z)\lambda(y)\lambda(z)(g_0(y)g_0(z))^{1-\alpha} \\
& (1-F(y)^n)(1-F(z)^n)dydz.
\end{aligned}
$$

Also, notice that $\int K_h(t-x)g_0^{2-\alpha}(t)dt = g_0(x)^{2-\alpha}+O(h^2)$, we could write $Var(\hat{\lambda}^*_\alpha(x))$ as

$$
\begin{aligned}
Var(\hat{\lambda}^*_\alpha(x)) = & g_0(x)^{2\alpha-2}\int I_n(F_n(y))\lambda(y)K_h^2(x-y)g_0(y)^{2-2\alpha}dy+2g_0(x)^{2\alpha-2} \\
& \times\int\int_{y\le z} K_h(x-y)K_h(x-z)(g_0(y)g_0(z))^{1-\alpha}\lambda(y)\lambda(z)A_n(y,z)dydz
\end{aligned}
$$

where

$$A_n(y,z) = (1 - F(y)^n)F(z)^n - (1 - F(y))\frac{F(z)^n - F(y)^n}{F(z) - F(y)}.$$

By the Taylor Series Expansion and a result in Waston and Leadbetter[168], as $n \to \infty$, $nI_n(F(u))$ converges to $1/(1 - F(u))$ and A_n is negligible. Therefore the variance can be easily shown to be

$$Var(\hat{\lambda}^*_\alpha(x)) = \frac{R(K)}{nh}\frac{\lambda(x)}{1 - F(x)} + o\left(\frac{1}{nh}\right).$$

This completes the first step of the proof which verifies the bias and variance of $\hat{\lambda}^*_\alpha(x)$ for large n.

To consider a more general case where $\hat{\lambda}_\alpha(x)$ allows for a maximum likelihood estimate $\hat{\underline{\theta}}$ in the functional form of the parametric assumption $g(x, \underline{\theta})$, we need to introduce more asymptotic properties of the maximum likelihood estimator $\hat{\underline{\theta}}$ given by Hjort and Glad[62].

Let F be the true distribution function, f be the density function and F_n be the empirical distribution function. We consider the functional estimators of $\underline{\theta}$ of the form $\hat{\underline{\theta}} = T(F_n)$ for some smooth T having the influence function

$$I(T) = \lim_{\epsilon \to 0}[T((1 - \epsilon)F + \epsilon\delta_x) - T(F)]/\epsilon$$

where δ_x is the unit point mass at x and $\Sigma_I = E_f[I(X_i)I(X_i)^T]$ is finite. The least false value $\underline{\theta}_0$ of $g_0(x) = g(x; \underline{\theta}_0)$ is determined by $\underline{\theta}_0 = T(F)$. It is well known for the case of the maximum likelihood estimator that $T(F)$ is defined as the solution to $\int(\partial/\partial\underline{\theta})\log g(x; \underline{\theta})dF(x) = 0$, and so $I(x) = J^{-1}(\partial/\partial\underline{\theta})\log g(x; \underline{\theta}_0)dF(x)$, where $J = -E_f[(\partial^2/\partial\underline{\theta}\partial\underline{\theta}^T)\log g(X_i; \underline{\theta}_0)]$. With reference to Chapter 6, Serfling[135], Hjort and Glad found that, under regularity conditions given by [67][136],

$$\hat{\underline{\theta}} - \underline{\theta}_0 = \frac{1}{n}\sum_{i=1}^{n} I(X_i) + \frac{d}{n} + \epsilon_n \tag{2.7.2}$$

where $\epsilon_n = O_p(n^{-1})$ with mean $O(n^{-2})$ and d/n is essentially the bias of $\hat{\underline{\theta}}$.

In our case, set

$$\eta_0(x) = \int K_h(t-x)g_0(t)^{2-\alpha}dt,$$

$$\eta_1(x) = \int K_h(t-x)u_0(t)g_0(t)^{2-\alpha}dt,$$

$$\eta_2(x) = \int K_h(t-x)\{U_0(t) + (2-\alpha)u_0(t)u_0(t)^T\}g_0(t)^{2-\alpha}dt,$$

where

$$u_0(x) = \frac{\partial}{\partial \underline{\theta}} \log g(x;\underline{\theta})|_{\underline{\theta}=\underline{\theta}_0},$$

$$U_0(x) = \frac{\partial^2}{\partial \underline{\theta} \partial \underline{\theta}^T} \log g(x;\underline{\theta})|_{\underline{\theta}=\underline{\theta}_0}.$$

Expanding $g(x;\hat{\underline{\theta}})$ with respect to $g(x;\underline{\theta}_0)$ and excluding smaller order terms gives

$$\hat{\lambda}_\alpha(x) = \hat{\lambda}^*_\alpha(x) + (\hat{\underline{\theta}} - \underline{\theta}_0)^T \bar{B}_n(x) + \frac{1}{2}(\hat{\underline{\theta}} - \underline{\theta}_0)^T \bar{C}_n(\hat{\underline{\theta}} - \underline{\theta}_0)$$

where $\bar{B}_n(x) = \frac{1}{n}\sum_{i=1}^n B_i(x)$ and $\bar{C}_n(x) = \frac{1}{n}\sum_{i=1}^n C_i(x)$, and

$$B_i(x) = g_0(X_i)^{1-\alpha}\frac{g_0(x)K_h(X_i-x)}{\eta_0(x)(1-F_n(X_i))} \times \left[(1-\alpha)u_0(X_i) - \frac{2-\alpha}{\eta_0(x)}\eta_1(x) + u_0(x)\right],$$

$$\begin{aligned} C_i(x) &= g_0(X_i)^{1-\alpha}\frac{g_0(x)K_h(X_i-x)}{\eta_0(x)(1-F_n(X_i))} \times \Bigg[-\frac{2(1-\alpha)(2-\alpha)}{\eta_0(x)}\eta_1(x)u_0(X_i)^T \\ &+ 2(1-\alpha)u_0(x)u_0(X_i)^T + (1-\alpha)\{U_0(X_i) + (1-\alpha)u_0(X_i)u_0(X_i)^T\} \\ &- \frac{2(2-\alpha)}{\eta_0(x)}u_0(x)\eta_1(x)^T + \{U_0(x) + u_0(x)u_0(x)^T\} \\ &+ \frac{2(2-\alpha)}{\eta_0(x)^2}\left\{(2-\alpha)\eta_1(x)\eta_1(x)^T - \frac{1}{2}\eta_0(x)\eta_2(x)^T\right\}\Bigg]. \end{aligned}$$

Using the fat that $I_i = I(X_i)$ has the mean 0, we have

$$E[(\hat{\underline{\theta}} - \underline{\theta}_0)^T \bar{B}_n(x)] = n^{-1}E[B_i^T(x)I_i] + n^{-1}(E[B_i(x)])^T d + O(n^{-2}), \tag{2.7.3}$$

$$E[(\underline{\hat{\theta}} - \underline{\theta}_0)^T \bar{C}_n(x)(\underline{\hat{\theta}} - \underline{\theta}_0)] = n^{-1}\mathrm{Tr}(E[C_i(x)]E[I_iI_i^T]) + O(n^{-2}). \qquad (2.7.4)$$

Then let $y = vx + h$ to get

$$\begin{aligned}
& E[B_i(x)^T I_i] \\
= \; & E\left\{ g_0(X_i)^{1-\alpha} \frac{g_0(x)\frac{K_h(X_i-x)}{1-F_n(X_i)}}{\eta_0(x)} \left[(1-\alpha)u_0(X_i) - \frac{2-\alpha}{\eta_0(x)}\eta_1(x) + u_0(x)\right]^T I_i \right\} \\
= \; & E\left\{(1-\alpha)g_0(X_i)^{1-\alpha}\frac{K_h(X_i-x)}{1-F_n(X_i)}\frac{g_0(x)}{\eta_0(x)} \times (u_0(X_i) - u_0(x) + O(h^2))^T I_i\right\} \\
= \; & \frac{(1-\alpha)g_0(x)}{\eta_0(x)} \int \frac{K_h(y-x)}{1-F_n(y)} f(y) g_0(y)^{1-\alpha}(u_0(y) - u_0(x) + O(h^2))^T I(y)dy \\
= \; & \frac{(1-\alpha)g_0(x)}{\eta_0(x)} \int \frac{K(v)}{1-F_n(vh+x)} f(vh+x) g_0(vh+x)^{1-\alpha} \\
& (u_0(vh+x) - u_0(x) + O(h^2))^T I(vh+x)dv \\
= \; & \frac{(1-\alpha)g_0(x)}{\eta_0(x)} \int \frac{K(v)}{1-F_n(vh+x)} f(vh+x) g_0(vh+x)^{1-\alpha} \\
& (v^2h^2u_0''(x)/2)^T I(vh+x)dv + O(h^2).
\end{aligned}$$

Thus $E[B_i(x)^T I_i]$ is of order $O(h^2)$. Similar calculations show that $E[C_i(x)]$ and $E[B_i(x)]$ are of the size equal $O(h^2)$, however we omit the details.

From equations (2.7.3) and (2.7.4), $E[(\underline{\hat{\theta}} - \underline{\theta}_0)^T \bar{B}_n(x)]$ and $E[(\underline{\hat{\theta}} - \underline{\theta}_0)^T \bar{C}_n(\underline{\hat{\theta}} - \underline{\theta}_0)]$ are both of orders $O(\frac{1}{n^2} + \frac{h^2}{n})$, thus

$$E[\hat{\lambda}_\alpha] = E[\lambda_\alpha^*] + O\left(\frac{1}{n^2} + \frac{h^2}{n}\right).$$

In order to calculate the variance of $\hat{\lambda}_\alpha$, first we determine the orders of magnitude of its components, $Var[(\underline{\hat{\theta}} - \underline{\theta}_0)^T \bar{B}_n(x)]$, $Var[(\underline{\hat{\theta}} - \underline{\theta}_0)^T \bar{C}_n(x)(\underline{\hat{\theta}} - \underline{\theta}_0)]$ and $Cov[\lambda_\alpha^*(t), (\underline{\hat{\theta}} -$

$\underline{\theta}_0)^T\bar{B}_n(x)]$ respectively. For example,

$$\begin{aligned} & Var[(\hat{\underline{\theta}} - \underline{\theta}_0)^T\bar{B}_n(x)] \\ = \ & Var[n^{-1}\sum_i(B_i^T(x)I_i) + \bar{B}_n^T\frac{d}{n} + \bar{B}_n^T(x)\epsilon_n] \\ = \ & n^{-1}Var((B_n^T(x)I_n)) + O(n^{-2}) \\ = \ & n^{-1}(E[B_i(x)])^T\Sigma_I(E[B_i(x)]) - n^{-1}(E[B_i^T(x)I_i])^T(E[B_i^T(x)I_i]) + O(n^{-2}). \end{aligned}$$

Since $(E[B_i(x)])$ and $E(B_i^T(x)I_i)$ are of order $O(h^2)$, and $\Sigma_I = E_f[I(X_i)I(X_i)^T]$ is finite, then $Var((\hat{\underline{\theta}} - \underline{\theta}_0)^T\bar{B}_n(x))$ is of order $O(\frac{h^4}{n} + n^{-2})$. Similar calculations can be used to show that $Var((\hat{\underline{\theta}} - \underline{\theta}_0)^T\bar{C}_n(x)(\hat{\underline{\theta}} - \underline{\theta}_0))$ is of order $(\frac{h^4}{n^2})$, but we omit the details.

Furthermore

$$\begin{aligned} & Cov[\lambda_\alpha^*(t), (\hat{\underline{\theta}} - \underline{\theta}_0)^T\bar{B}_n(x)] \\ & = n^{-1}(E[B_i(x)])^TE\left[\frac{K_h(X_i - x)}{1 - F_n(X_i)}\frac{g_0(X_i)^{1-\alpha}}{g_0(x)^{1-\alpha}}I_i\right] + O(n^{-2}) \\ & = O\left(\frac{h^2}{n} + \frac{1}{n^2}\right). \end{aligned}$$

Thus,

$$Var(\hat{\lambda}_\alpha) = \frac{R(K)}{nh}\frac{\lambda(x)}{1 - F(x)} + o\left(\frac{1}{nh}\right) + O\left(\frac{1}{n^2}\right),$$

which complete the proof.

Proof. of Theorem 2.2

First consider $Bias(\hat{\psi}(p|r,s))$.

To prove (2.6.1) note that for $i < j$, one has

$$E\Big[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{L_g^{(p)}(X_{(i)}-X_{(j)})}{(n-j+1)(n-i+1)}v(X_{(i)})\Big]$$

$$=\sum_{i=1}^{n-1}\int\int_{x<y}\sum_{j=i+1}^{n}\frac{n!}{(i-1)!(j-i-1)!(n-j+1)!(n-i+1)}L_g^{(p)}(x-y)$$

$$\times F(x)^{i-1}(F(y)-F(x))^{j-i-1}(1-F(y))^{n-j}f(x)f(y)v(x)w(x)dxdy$$

$$=\sum_{i=1}^{n-1}\int\int_{x<y}\binom{n}{i-1}L_g^{(p)}(x-y)\times\sum_{j=i+1}^{n}\binom{n-i}{j-i-1}$$

$$\left(\frac{F(y)-F(x)}{1-F(y)}\right)^{j-i-1}\left(\frac{F(x)}{1-F(y)}\right)^{i-1}(1-F(y))^{n-2}f(x)f(y)v(x)w(x)dxdy$$

$$=\sum_{i=1}^{n-1}\int\int_{x<y}\binom{n}{i-1}L_g^{(p)}(x-y)[(1-F(x))^{n-i}-(F(y)-F(x))^{n-i}]F(x)^{i-1}$$

$$\lambda(y)f(x)v(x)w(x)dxdy$$

$$=\int\int_{x<y}L_g^{(p)}(x-y)f(x)\lambda(y)v(x)$$

$$\sum_{i=1}^{n-1}\binom{n}{i-1}[(1-F(x))^{n-i}-(F(y)-F(x))^{n-i}]F(x)^{i-1}w(x)dxdy$$

$$=\int\int_{x<y}L_g^{(p)}(x-y)f(x)\lambda(y)v(x)\Big\{\frac{1}{1-F(x)}(1-n(1-F(x))F(x)^{n-1}-F(x)^n)$$

$$-\frac{F(y)^n-F(x)^n}{F(y)-F(x)}\Big\}w(x)dxdy$$

$$=\int\int_{x<y}L_g^{(p)}(x-y)\lambda(x)\lambda(y)v(x)w(x)dxdy$$

The same conclusion can be obtained for the case $i > j$, that is

$$E\Big[\sum_{i=j+1}^{n}\sum_{j=1}^{n-1}\frac{L_g^{(p)}(X_{(i)}-X_{(j)})}{(n-j+1)(n-i+1)}v(X_{(i)})\Big]$$

$$=\int\int_{x>y}L_g^{(p)}(x-y)\lambda(x)\lambda(y)v(x)w(x)dxdy.$$

This completes the proof of (2.6.1) that

$$
\begin{aligned}
& E[\hat{\psi}(p|r,s)] \\
&= \int\int L_g^{(p)}(x-y)\lambda(x)\lambda(y)v(x)w(x)dxdy \\
&= \int\int L_g(x-y)\lambda(x)\lambda^{(p)}(y)v(x)w(x)dxdy \\
&= \int\int v(x)L(u)\lambda(x)(\lambda^{(p)}(x)+1/2(ug)^2\lambda^{(p+2)}(x)+o(g^2))w(x)dxdu \\
&= \int \lambda(x)\lambda^{(p)}(x)v(x)w(x)dx + \frac{g^2}{2}\mu_{2,L(u)}\psi(p+2|r,s)+o(g^2).
\end{aligned}
$$

Notice that by partial integration and Assumption A, $\int\lambda(x)v(x)w(x)\int L_g^{(p)}(x-y)\lambda(y)dydx$ equals to

$$
\int \lambda(x)v(x)w(x)\int L_g^{(p-1)}(x-y)\lambda^{(1)}(y)dydx,
$$

thus the last third equality can be finally proved by the p times partial integration in this way.

To establish (2.6.2) and (2.6.3) note that $|F_n(x)-F(x)|=O_p(n^{-1/2})$. Therefore the leading term of the variance expression of $\hat{\psi}(p|r,s)$ will not change if one replaces $1-F_n(x)$ appearing in $\hat{\psi}(p|r,s)$ by $1-F(x)$. Thus,

$$
Var\{\hat{\psi}(p|r,s)\} = \mathrm{I}+\mathrm{II}+\mathrm{III}+\mathrm{IV}+\mathrm{V}+\mathrm{VI}
$$

where

$$
\mathrm{I} = \frac{n(n-1)}{n^4}Var\left\{\frac{L_g^{(p)}(X_i-X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))}\right\},
$$

$$
\mathrm{II} = \frac{n(n-1)}{n^4}Cov\left\{\frac{L_g^{(p)}(X_i-X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))},\frac{L_g^{(p)}(X_j-X_i)v(X_j)}{(1-F(X_i))(1-F(X_j))}\right\},
$$

$$
\begin{aligned}
\mathrm{III} &= \frac{n(n-1)(n-2)}{n^4} \\
&\quad \times Cov\left\{\frac{L_g^{(p)}(X_i-X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))},\frac{L_g^{(p)}(X_i-X_k)v(X_i)}{(1-F(X_i))(1-F(X_k))}\right\},
\end{aligned}
$$

$$
\begin{aligned}
\text{IV} &= \frac{2n(n-1)(n-2)}{n^4} \\
&\quad \times Cov\left\{\frac{L_g^{(p)}(X_i - X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))}, \frac{L_g^{(p)}(X_k - X_i)v(X_k)}{(1-F(X_i))(1-F(X_k))}\right\},
\end{aligned}
$$

$$
\begin{aligned}
\text{V} &= \frac{n(n-1)(n-2)}{n^4} \\
&\quad \times Cov\left\{\frac{L_g^{(p)}(X_j - X_i)v(X_j)}{(1-F(X_i))(1-F(X_j))}, \frac{L_g^{(p)}(X_k - X_i)v(X_k)}{(1-F(X_i))(1-F(X_k))}\right\}
\end{aligned}
$$

and

$$
\begin{aligned}
\text{VI} &= \frac{n(n-1)(n-2)(n-3)}{n^4} \\
&\quad \times Cov\left\{\frac{L_g^{(p)}(X_j - X_i)v(X_j)}{(1-F(X_i))(1-F(X_j))}, \frac{L_g^{(p)}(X_k - X_l)v(X_k)}{(1-F(X_l))(1-F(X_k))}\right\}.
\end{aligned}
$$

Now we analyze I-VI separately. Before that, the two critical points needed to be highlighted and those are

1. $L_g^{(p)}(x-y) = \frac{1}{g^{p+1}} L^{(p)}\left(\frac{x-y}{g}\right)$ and $L(\cdot)$ is a even function.

2. The derivative of an even function is an odd function and vice versa.

For term I, as $n \to \infty$,

$$
\begin{aligned}
&\frac{n-1}{n^3} Var\left\{\frac{L_g^{(p)}(X_i - X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))}\right\} \\
&= \frac{1}{n^2} E\left(\frac{L_g^{(p)}(X_i - X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))}\right)^2 - \frac{1}{n^2}\left\{E\left(\frac{L_g^{(p)}(X_i - X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))}\right)\right\}^2.
\end{aligned}
$$

By setting $x = y + ug$, the first term in the last line can be written as

$$
\begin{aligned}
&\frac{1}{n^2} E\left(\frac{L_g^{(p)}(X_i - X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))}\right)^2 \\
&= \frac{1}{n^2}\int\int [L_g^{(p)}(x-y)]^2 v^2(x) \frac{\lambda(x)\lambda(y)}{(1-F(x))(1-F(y))} w(x)w(y)dxdy \\
&= \frac{1}{n^2 g^{2p+1}} \int L^{(p)}(u)^2 v^2(y+ug) \frac{\lambda(y+ug)\lambda(y)}{(1-F(y+ug))(1-F(y))} w(y)dydu.
\end{aligned}
$$

When p is even, by the standard Taylor series expansion, the first term of I can be expressed as

$$\frac{1}{n^2 g^{2p+1}} R(L^{(p)}) \int v^2(y) \left(\frac{\lambda(y)}{1-F(y)} \right)^2 w(y) dy + o\left(\frac{1}{n^2 g^{2p+1}} \right).$$

When p is odd, one expands the equation to its 2nd order to get

$$\begin{aligned} &\frac{1}{n^2 g^{2p+1}} R(L^{(p)}) \int v^2(y) \left\{ \frac{\lambda(y)}{1-F(y)} \right\}^2 w(y) dy \\ &+ \frac{\mu_{2,L^{(p)2}}}{2n^2 g^{2p-1}} \int \frac{\lambda(y)}{1-F(y)} \left\{ \frac{v^2(y)\lambda(y)}{1-F(y)} \right\}'' w(y) dy + o\left(\frac{1}{n^2 g^{2p-1}} \right). \end{aligned}$$

For term II, as $n \to \infty$, it can be written as

$$\begin{aligned} &\frac{1}{n^2} E\left\{ \frac{L_g^{(p)}(X_i - X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))} \frac{L_g^{(p)}(X_j - X_i)v(X_j)}{(1-F(X_i))(1-F(X_j))} \right\} \\ &- \frac{1}{n^2} E\left\{ \frac{L_g^{(p)}(X_i - X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))} \right\} E\left\{ \frac{L_g^{(p)}(X_j - X_i)v(X_j)}{(1-F(X_i))(1-F(X_j))} \right\}. \end{aligned}$$

The second term in the last line equals to $-\frac{1}{n^2}\left\{ E\left(\frac{L_g^{(p)}(X_i - X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))} \right) \right\}^2$ and the first term can be written as

$$\begin{aligned} &\frac{1}{n^2} E\left\{ \frac{L_g^{(p)}(X_i - X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))} \frac{L_g^{(p)}(X_j - X_i)v(X_j)}{(1-F(X_i))(1-F(X_j))} \right\} \\ &= \frac{1}{n^2 g^{2p+2}} \int\int L^{(p)}\left(\frac{x-y}{g} \right) L^{(p)}\left(\frac{y-x}{g} \right) \\ &\frac{v(x)v(y)\lambda(x)\lambda(y)}{(1-F(x))(1-F(y))} w(x)w(y) dx dy. \end{aligned}$$

When p is even, one has $L^{(p)}\left(\frac{x-y}{g}\right) = L^{(p)}\left(\frac{y-x}{g}\right)$, and

$$\begin{aligned} &\frac{1}{n^2} E\left\{ \frac{L_g^{(p)}(X_i - X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))} \frac{L_g^{(p)}(X_j - X_i)v(X_j)}{(1-F(X_i))(1-F(X_j))} \right\} \\ &= \frac{1}{n^2 g^{2p+1}} R(L^{(p)}) \int v^2(y) \left(\frac{\lambda(y)}{1-F(y)} \right)^2 w(y) dy + o\left(\frac{1}{n^2 g^{2p+1}} \right). \end{aligned}$$

When p is odd, one has $L^{(p)}\left(\frac{x-y}{g}\right) = -L^{(p)}\left(\frac{y-x}{g}\right)$, and by setting $x = y + ug$ we obtain,

$$\begin{aligned}
&\frac{1}{n^2}E\left\{\frac{L_g^{(p)}(x-y)v(x)}{(1-F(x))(1-F(y))}\frac{L_g^{(p)}(y-x)v(y)}{(1-F(x))(1-F(y))}\right\} \\
&= -\frac{1}{n^2g^{2p+1}}\int L^{(p)}(u)^2\frac{v(y+ug)v(y)\lambda(y+ug)\lambda(y)}{(1-F(y+ug))(1-F(y))}w(y)dydu \\
&= -\frac{1}{n^2g^{2p+1}}R(L^{(p)})\int v^2(y)\left\{\frac{\lambda(y)}{(1-F(y))}\right\}^2 w(y)dy \\
&-\frac{\mu_{2,L^{(p)2}}}{2n^2g^{2p-1}}\int\frac{v(y)\lambda(y)}{1-F(y)}\left\{\frac{v(y)\lambda(y)}{1-F(y)}\right\}'' w(y)dy + o\left(\frac{1}{n^2g^{2p-1}}\right).
\end{aligned}$$

For Term III, as $n \to \infty$, it approximates to

$$\begin{aligned}
&\frac{1}{n}E\left\{\frac{L_g^{(p)}(X_i-X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))}\frac{L_g^{(p)}(X_i-X_k)v(X_i)}{(1-F(X_i))(1-F(X_k))}\right\} \\
&-\frac{1}{n}E\left\{\frac{L_g^{(p)}(X_i-X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))}\right\}E\left\{\frac{L_g^{(p)}(X_i-X_k)v(X_i)}{(1-F(X_i))(1-F(X_k))}\right\}.
\end{aligned}$$

The second term of its expression equals to $-\frac{1}{n}\left\{E\left(\frac{L_g^{(p)}(X_i-X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))}\right)\right\}^2$. The first term can be written as

$$\frac{1}{n}\int\int\int L_g^{(p)}(x-y)L_g^{(p)}(x-z)\frac{v^2(x)\lambda(x)\lambda(y)\lambda(z)}{1-F(x)}w(x)dxdydz.$$

By setting $y = x - gu$ and $z = x - gv$, here irrespective of p is even or odd, the first term of the above equation equals,

$$\begin{aligned}
&\frac{1}{n}\int\int\int L_g^{(p)}(x-y)L_g^{(p)}(x-z)\frac{v^2(x)\lambda(x)\lambda(y)\lambda(z)}{1-F(x)}w(x)dxdydz \\
&= \frac{1}{n}\int\int\int L(u)L(v)\frac{v^2(x)\lambda(x)}{(1-F(x))}\lambda^{(p)}(x-gu)\lambda^{(p)}(x-gv)w(x)dxdudv \\
&= \frac{1}{n}\int\frac{v^2(x)\lambda(x)}{(1-F(x))}(\lambda^{(p)}(x))^2w(x)dx + o\left(\frac{1}{n}\right).
\end{aligned}$$

Using the similar arguments one can show that, the terms IV and V reduce to,

$$\text{IV} = \begin{cases} \frac{2}{n}\int (v(x)\lambda(x))^{(p)}\lambda^{(p)}(x)\frac{v(x)\lambda(x)}{(1-F(x))}w(x)dx \\ \qquad -\frac{2}{n}\left\{E\left\{\frac{L_g^{(p)}(X_i-X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))}\right\}\right\}^2 + o(\frac{1}{n}), \quad \text{when } p \text{ is even} \\ -\frac{2}{n}\int (v(x)\lambda(x))^{(p)}\lambda^{(p)}(x)\frac{v(x)\lambda(x)}{(1-F(x))}w(x)dx \\ \qquad -\frac{2}{n}\left\{E\left\{\frac{L_g^{(p)}(X_i-X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))}\right\}\right\}^2 + o(\frac{1}{n}), \quad \text{when } p \text{ is odd} \end{cases}$$

and

$$\text{V} = \frac{1}{n}\int \{(v(x)\lambda(x))^{(p)}\}^2\frac{\lambda(x)}{(1-F(x))}w(x)dx \\ -\frac{1}{n}\left\{E\left\{\frac{L_g^{(p)}(X_i-X_j)v(X_i)}{(1-F(X_i))(1-F(X_j))}\right\}\right\}^2 + o\left(\frac{1}{n}\right),$$

irrespective of whether p is odd or even.

Further it is straightforward to see that term VI is 0 since $\frac{L_g^{(p)}(X_j-X_i)v(X_j)}{(1-F(X_i))(1-F(X_j))}$ and $\frac{L_g^{(p)}(X_l-X_k)v(X_l)}{(1-F(X_k))(1-F(X_l))}$ are independent.

Thus when p is even, we get

$$Var(\hat{\psi}(p|r,s)) \\ = \frac{2}{n^2g^{2p+1}}R(L^{(p)})\int v^2(x)\left\{\frac{\lambda(x)}{1-F(x)}\right\}^2 w(x)dx \\ +\frac{1}{n}\int \frac{\lambda(x)}{1-F(x)}\{(v(x)\lambda(x))^{(p)} + v(x)\lambda^{(p)}(x)\}^2 w(x)dx - \frac{4}{n}\{E(\hat{\psi}(p|r,s))\}^2 \\ +o\left(\frac{1}{n}+\frac{1}{n^2g^{2p+1}}\right),$$

and when p is odd,

$$\begin{aligned}
&Var(\hat{\psi}(p|r,s)) \\
&= \frac{\mu_{2,L^{(p)2}}}{2n^2g^{2p-1}} \int \frac{\lambda(x)}{1-F(x)} \left\{ \left\{ \frac{v^2(x)\lambda(x)}{1-F(x)} \right\}'' - v(x) \left\{ \frac{v(x)\lambda(x)}{1-F(x)} \right\}'' \right\} w(x)dx \\
&+ \frac{1}{n} \int \frac{\lambda(x)}{1-F(x)} \{(v(x)\lambda(x))^{(p)} - v(x)\lambda^{(p)}(x)\}^2 w(x)dx - \frac{4}{n}\{E(\hat{\psi}(p|r,s))\}^2 \\
&+ o\left(\frac{1}{n} + \frac{1}{n^2g^{2p-1}}\right).
\end{aligned}$$

Hence the result follows.

Proof. of Theorem 2.3

Clearly, $E[(\hat{\psi}_1 - \psi_1)(\hat{\psi}_2 - \psi_2)] = E(\hat{\psi}_1\hat{\psi}_2) - E\hat{\psi}_1 E\hat{\psi}_2 + Bias(\hat{\psi}_1)Bias(\hat{\psi}_2)$.

Since $|F_n(x) - F(x)| = O_p(n^{-1/2})$, we could replace $1 - F_n(x)$ appearing in $\hat{\psi}(p|r,s)$ by $1 - F(x)$. Thus,

$$\begin{aligned}
E[\hat{\psi}_1\hat{\psi}_2] \quad &= E\Bigg\{ \frac{1}{n^4} \sum\sum\nolimits_{i_1 \neq j_1} \frac{L_g^{(p_1)}(X_{i_1} - X_{j_1})v_1(X_{i_1})}{(1-F_n(X_{i_1}))(1-F_n(X_{j_1}))} \\
&\quad \times \sum\sum\nolimits_{i_2 \neq j_2} \frac{L_g^{(p_2)}(X_{i_2} - X_{j_2})v_2(X_{i_2})}{(1-F_n(X_{i_2}))(1-F_n(X_{j_2}))} \Bigg\} \\
&= \text{I+II+III+IV+V+VI+VII}
\end{aligned}$$

where

$$\text{I} = \frac{n(n-1)}{n^4} E\left\{ \frac{L_g^{(p_1)}(X_i - X_j)L_g^{(p_2)}(X_i - X_j)v_1(X_i)v_2(X_i)}{(1-F(X_i))^2(1-F(X_j))^2} \right\},$$

$$\text{II} = \frac{n(n-1)}{n^4} E\left\{ \frac{L_g^{(p_1)}(X_i - X_j)L_g^{(p_2)}(X_j - X_i)v_1(X_i)v_2(X_j)}{(1-F(X_i))^2(1-F(X_j))^2} \right\},$$

$$\text{III} = \frac{n(n-1)(n-2)}{n^4} E\left\{ \frac{L_g^{(p_1)}(X_i - X_j)L_g^{(p_2)}(X_k - X_i)v_1(X_i)v_2(X_k)}{(1-F(X_i))^2(1-F(X_j))(1-F(X_k))} \right\},$$

$$\text{IV} = \frac{n(n-1)(n-2)}{n^4} E\left\{ \frac{L_g^{(p_1)}(X_i - X_j)L_g^{(p_2)}(X_i - X_k)v_1(X_i)v_2(X_i)}{(1-F(X_i))^2(1-F(X_j))(1-F(X_k))} \right\},$$

$$\text{V} = \frac{n(n-1)(n-2)}{n^4} E\left\{ \frac{L_g^{(p_1)}(X_j - X_i)L_g^{(p_2)}(X_i - X_k)v_1(X_j)v_2(X_i)}{(1-F(X_i))^2(1-F(X_j))(1-F(X_k))} \right\},$$

$$\text{VI} = \frac{n(n-1)(n-2)}{n^4} E\left\{ \frac{L_g^{(p_1)}(X_j - X_i) L_g^{(p_2)}(X_k - X_i) v_1(X_j) v_2(X_k)}{(1-F(X_i))^2(1-F(X_j))(1-F(X_k))} \right\}$$

and

$$\begin{aligned} \text{VII} = & \frac{n(n-1)(n-2)(n-3)}{n^4} \\ & E\left\{ \frac{L_g^{(p_1)}(X_j - X_i) v_1(X_j)}{(1-F(X_j))(1-F(X_i))} \right\} E\left\{ \frac{L_g^{(p_2)}(X_l - X_k) v_1(X_l)}{(1-F(X_l))(1-F(X_k))} \right\}. \end{aligned}$$

Since the derivations of the terms III - VI are identical, we only analyze term III to illustrate.

By setting $y = x - gu, z = x + gv$, term III approximates to

$$\begin{aligned} & \frac{1}{n} E\left\{ \frac{L_g^{(p_1)}(X_i - X_j) L_g^{(p_2)}(X_k - X_i) v_1(X_i) v_2(X_k)}{(1-F(X_i))^2(1-F(X_j))(1-F(X_k))} \right\} \\ = & \frac{1}{n} \int\int\int \frac{L_g^{(p_1)}(x-y) L_g^{(p_2)}(z-x) v_1(x) v_2(z)}{1-F(x)} \lambda(x)\lambda(y)\lambda(z) w(x) dx dy dz \\ = & \frac{1}{n} \int\int\int \frac{(-1)^{p_2} L_g(x-y) L_g(z-x) v_1(x)}{1-F(x)} \\ & \qquad \times \lambda(x)\lambda^{(p_1)}(y)[v_2(z)\lambda(z)]^{(p_2)} w(x) dx dy dz \\ = & \frac{(-1)^{p_2}}{n} \int \frac{v_1(x)}{1-F(x)} \lambda(x)\lambda^{(p_1)}(x)[v_2(x)\lambda(x)]^{(p_2)} w(x) dx + o(n^{-1}). \end{aligned} \tag{2.7.5}$$

Using the similar argument get

$$\text{IV} = \frac{1}{n} \int \frac{v_1(x) v_2(x) \lambda(x)}{1-F(x)} \lambda^{(p_1)}(x) \lambda^{(p_2)}(x) dx + o(n^{-1}), \tag{2.7.6}$$

$$\text{V} = \frac{(-1)^{p_1}}{n} \int \frac{v_2(x)\lambda(x)}{1-F(x)} \lambda^{(p_2)}(x) [v_1(x)\lambda(x)]^{(p_1)} dx + o(n^{-1}), \tag{2.7.7}$$

and

$$\text{VI} = \frac{(-1)^{p_1+p_2}}{n} \int \frac{\lambda(x)}{1-F(x)} [v_1(x)\lambda(x)]^{(p_1)} [v_2(x)\lambda(x)]^{(p_2)} dx + o(n^{-1}). \tag{2.7.8}$$

The expectations involved in terms I and II are more tricky for being dependent on odd or evenness of p_1 and p_2. Set $x = gu + y$ and then

$$\begin{aligned}
\mathrm{I} &= \frac{1}{n^2} E\left\{ \frac{L_g^{(p_1)}(X_i - X_j) L_g^{(p_2)}(X_i - X_j) v_1(X_i) v_2(X_i)}{(1 - F(X_i))^2 (1 - F(X_j))^2} \right\} \\
&= \frac{1}{n^2 g^{p_1+p_2+2}} \int\int L^{(p_1)}\left(\frac{x-y}{g}\right) L^{(p_2)}\left(\frac{x-y}{g}\right) \\
&\qquad \times \frac{v_1(x) v_2(x) \lambda(x) \lambda(y)}{(1-F(x))(1-F(y))} w(x) w(y) dx dy \\
&= \frac{1}{n^2 g^{p_1+p_2+1}} \int L^{(p_1)}(u) L^{(p_2)}(u) du \int v_1(y) v_2(y) \left\{\frac{\lambda(y)}{1-F(y)}\right\}^2 w(y) dy \\
&+ \frac{1}{n^2 g^{p_1+p_2}} \int L^{(p_1)}(u) L^{(p_2)}(u) u du \int \frac{\lambda(y)}{1-F(y)} \left\{\frac{v_1(y) v_2(y) \lambda(y)}{1-F(y)}\right\}' w(y) dy \\
&+ \frac{1}{2n^2 g^{p_1+p_2-1}} \int L^{(p_1)}(u) L^{(p_2)}(u) u^2 du \int \frac{\lambda(y)}{1-F(y)} \left\{\frac{v_1(y) v_2(y) \lambda(y)}{1-F(y)}\right\}'' w(y) dy.
\end{aligned} \tag{2.7.9}$$

Similarly,

$$\begin{aligned}
\mathrm{II} &= \frac{1}{n^2} E\left\{ \frac{L_g^{(p_1)}(X_i - X_j) L_g^{(p_2)}(X_j - X_i) v_1(X_i) v_2(X_j)}{(1 - F(X_i))^2 (1 - F(X_j))^2} \right\} \\
&= \frac{1}{n^2 g^{p_1+p_2+2}} \int\int L^{(p_1)}\left(\frac{x-y}{g}\right) \\
&\qquad \times L^{(p_2)}\left(\frac{y-x}{g}\right) \frac{v_1(x) v_2(y) \lambda(x) \lambda(y)}{(1-F(x))(1-F(y))} w(x) w(y) dx dy \\
&= \frac{(-1)^{p_2}}{n^2 g^{p_1+p_2+1}} \int L^{(p_1)}(u) L^{(p_2)}(u) du \int v_1(y) v_2(y) \left\{\frac{\lambda(y)}{1-F(y)}\right\}^2 w(y) dy \\
&+ \frac{(-1)^{p_2}}{n^2 g^{p_1+p_2}} \int L^{(p_1)}(u) L^{(p_2)}(u) u du \int \frac{v_2(y)\lambda(y)}{1-F(y)} \left\{\frac{v_1(y) \lambda(y)}{1-F(y)}\right\}' w(y) dy \\
&+ \frac{(-1)^{p_2}}{2n^2 g^{p_1+p_2-1}} \int L^{(p_1)}(u) L^{(p_2)}(u) u^2 du \int \frac{\lambda(y) v_2(y)}{1-F(y)} \left\{\frac{v_1(y) \lambda(y)}{1-F(y)}\right\}'' w(y) dy.
\end{aligned} \tag{2.7.10}$$

Now consider three different cases namely, when both p_1 and p_2 are even, when both p_1 and p_2 are odd and when $p_1 + p_2$ is odd.

When p_1 and p_2 are both even, from (2.7.9) and (2.7.10) the sum of terms I and II is

$$\frac{2}{n^2 g^{p_1+p_2+1}} \int L^{(p_1)}(u) L^{(p_2)}(u) du \int v_1(y) v_2(y) \left\{ \frac{\lambda(y)}{1-F(y)} \right\}^2 w(y) dy$$
$$+ o\left(\frac{1}{n^2 g^{p_1+p_2+1}} \right).$$

When p_1 and p_2 are both odd, we have that $\int L^{(p_1)}(u) L^{(p_2)}(u) u du = 0$. Hence from (2.7.9) and (2.7.10) the sum of terms I and II reduces to

$$\frac{1}{2n^2 g^{p_1+p_2-1}} \int L^{(p_1)}(u) L^{(p_2)}(u) u^2 du \int \frac{\lambda(y)}{1-F(y)} \Bigg\{ \left[\frac{v_1(y) v_2(y) \lambda(y)}{1-F(y)} \right]''$$
$$- v_2(y) \left[\frac{v_1(y)\lambda(y)}{1-F(y)} \right]'' \Bigg\} w(y) dy + o\left(\frac{1}{n^2 g^{p_1+p_2-1}} \right).$$

Finally when $p_1 + p_2$ is odd, without losing generalities, by assuming that p_2 is odd and p_1 is even, the sum of terms I and II is

$$\frac{1}{n^2 g^{p_1+p_2}} \int L^{(p_1)}(u) L^{(p_2)}(u) u du \int \frac{\lambda(y)}{1-F(y)} \Bigg\{ \left[\frac{v_1(y) v_2(y) \lambda(y)}{1-F(y)} \right]'$$
$$+ (-1)^{p_2} v_2(y) \left[\frac{v_1(y)\lambda(y)}{1-F(y)} \right]' \Bigg\} w(y) dy + o\left(\frac{1}{n^2 g^{p_1+p_2}} \right).$$

Further, it is easy to show that as $n \to \infty$, VII is asymptotically equivalent to $E\hat{\psi}_1 E\hat{\psi}_2$.

Then the conclusion of Theorem 2.3 follows once we add up the asymptotic expressions of the terms I to VI given in (2.7.5)-(2.7.10).

Proof. of Theorem 2.4

Notice that

$$\begin{aligned}
&\int\int m_{2|3}(u,x)^2 w(x)dudx \\
&= \int\int \left\{ L^{(3)}(u)u\left(\frac{q_1'(x)\lambda(x)}{1-F(x)}\right) + 2(L^{(2)}(u))\left(\frac{q_2(x)\lambda(x)}{1-F(x)}\right) \right\}^2 w(x)dxdu \\
&= \mu_{2,(L^{(3)})^2}\int\left(\frac{q_1'(x)\lambda(x)}{1-F(x)}\right)^2 w(x)dx + 4R(L^{(2)})\int\left(\frac{q_2(x)\lambda(x)}{1-F(x)}\right)^2 w(x)dx \\
&+4\int L^{(3)}(u)L^{(2)}(u)udu\int\left(\frac{\lambda(x)}{1-F(x)}\right)^2 q_1'(x)q_2(x)w(x)dx.
\end{aligned}$$

Since $\int\left[q_1(x)q_1'(x)\left(\frac{\lambda(x)}{1-F(x)}\right)^2\right]' w(x)dx = 0$ as $\lim_{x\to T} q_1(x)q_1'(x)\left(\frac{\lambda(x)}{1-F(x)}\right)^2 = 0$ and $\lim_{x\to 0} q_1(x)q_1'(x)\left(\frac{\lambda(x)}{1-F(x)}\right)^2 = 0$, so

$$\begin{aligned}
&\int\int m_{2|3}(u,x)^2 w(x)dudx \\
&= \mu_{2,(L^{(3)})^2}\int\left\{(q_1'(x))^2\left(\frac{\lambda(x)}{1-F(x)}\right)^2 + \left(q_1(x)q_1'(x)\left(\frac{\lambda(x)}{1-F(x)}\right)^2\right)'\right\}w(x)dx \\
&+4R(L^{(2)})\int q_2^2(x)\left[\frac{\lambda(x)}{1-F(x)}\right]^2 w(x)dx \\
&+4\int L^{(3)}(u)L^{(2)}(u)udu\int\left[\frac{\lambda(x)}{1-F(x)}\right]^2 q_1'(x)q_2(x)w(x)dx \\
&= \mu_{2,(L^{(3)})^2}\int\frac{\lambda(x)}{1-F(x)}\left\{2(q_1'(x))^2\frac{\lambda(x)}{1-F(x)} + q_1(x)q_1''(x)\frac{\lambda(x)}{1-F(x)}\right. \\
&\left.+2q_1(x)q_1'(x)\left(\frac{\lambda(x)}{1-F(x)}\right)'\right\}w(x)dx + 4R(L^{(2)})\int q_2^2(x)\left[\frac{\lambda(x)}{1-F(x)}\right]^2 w(x)dx \\
&+4\int L^{(3)}(u)L^{(2)}(u)udu\int\left[\frac{\lambda(x)}{1-F(x)}\right]^2 q_1'(x)q_2(x)w(x)dx.
\end{aligned}$$

With reference to (2.6.4), one can easily show the first equation of Theorem 2.4 holds for $MSE(\hat{N})$. The second equation for $\hat{D}$ is straightforward,l so it is not proved here.

2.8 Discussions

In this chapter, we began with the introduction of several multiplicative semiparametric estimates in density estimation and illustrated a generalized estimate defined by Naito[97] which includes the previous ones as its special cases.

Then we propose a multiplicative semiparametric estimate in the hazard rate setting which could be seen as the generalization of Naito's work. The semiparametric estimate starts from a crude guess of the true hazard rate function but modified by a nonparametric correction factor. From our example studies, we show that the shape parameter α plays an important role in determining the accuracy of the estimate.

In the following section, we also investigated the asymptotic properties of the generalized estimate and found that the proposed estimator performs better than the traditional nonparametric kernel estimate if the parametric guess of the data locates very close to the true model. Even if the assumed parametric model is not correct, the resulted estimate will converge to the true function with the same rate as the nonparametric counterpart. In the end of this chapter, we proposed two adaptive estimates of the shape parameter α and the bandwidth h using the 'plug-in' method.

In conclusion, the main findings of this chapter are summarized in Fig 2.1. In the next chapter, we will discuss the standard kernel hazard rate estimator in terms of L_1 error criterion.

What is already known on this topic:

- In density estimation, one common semiparametric method is to fit the data with an assumed parametric model and then use a nonparametric correction factor to modify the model that one started with.

- There are various multiplicative correction factor approaches in the settings of density estimation and Naito[97] unifies different approaches of devising multiplicative correction factors.

What this study adds

- We generalize Naito's approach to the setting of hazard rate estimation using the survival data without censorship.

- We use the example studies to exhibit and discuss the role of the shape parameter α that plays in the multiplicative semiparametric estimators.

- The asymptotic analysis proves that the proposed estimator performs better than its nonparametric counterpart when our prior assumption is close to the true model.

- We introduce the 'plug-in' approaches to estimate the shape parameter α and the bandwidth h of the generalized estimator.

Table 2.1: Summary of the main issues and key findings in Chapter 2.

CHAPTER 3

L_1 ERRORS ANALYSIS IN KERNEL ESTIMATION

3.1 Introduction

We have mentioned general error criterion to measure the accuracy of a kernel estimate in Section 1.4. Usually, in kernel estimation, for its technical tractability and easy understanding, L_2 error is most commonly used. For the same reasons, in the last chapter, we utilize the L_2 error criterion, namely the MISE, $E\int|\hat{\lambda}(x)-\lambda(x)|^2 w(x)dx$ to judge the performance of the semiparametric estimator $\hat{\lambda}(x)$ of the true hazard rate function $\lambda(x)$. The bandwidth and shape parameters are then obtained so as to optimize L_2 error.

However see Devroye and Gy$\ddot{o}$rfi[36] for the shortcomings of L_2 error and a case for L_1 error,

$$L_1(f,\hat{f}) = \int |\hat{f}(x) - f(x)|dx,$$

to assess the accuracy of a kernel density estimator. In particular, Devroye and Gy$\ddot{o}$rfi[36] show that L_1 error is, well-defined and invariant under monotone transformation of the coordinate axes. To make it precise, consider two d-dimensional random vectors X and Y with densities f and g respectively. Now suppose a transformation T follows $\{T^{-1}B|B \in \mathcal{B}\} \in \mathcal{B}$ where $\mathcal{B}$ is the class of all Borel sets of R^d and the densities of $T(X)$ and $T(Y)$ are f^* and g^*, then regardless of T, one has

$$\int |f-g| = \int |f^* - g^*|.$$

Specially, for $d = 1$, the L_1 distance is invariant under continuous strictly monotone transformation.

Of the recent developments of the L_1 analysis of a kernel density estimator, it is important to mention the work of Hall and Wand[51]. They have considered a simple, rapidly converging, iterative algorithm allowing for the minimization of L_1 distance w.r.t bandwidth h in the setting of density estimation. Based on that, they also developed an adaptive (data-driven) bandwidth choice which minimizes L_1 errors asymptotically.

The reasons for which Devroye and Györfi have advocated L_1 error criteria in density estimation settings apply to hazard rate estimation as well. Therefore in this chapter, we extend the concept of the L_1 optimal kernel density estimation to kernel-based hazard rate estimation. For that in Section 3.2 we propose a general asymptotic expression for the L_1 distance between the true hazard rate function $\lambda(x)$ and a kernel estimator $\hat{\lambda}(x)$ and then derive the theoretical asymptotic L_1 optimal bandwidth. We then utilize the Newton method to develop an iterative algorithm to calculate the asymptotic L_1 optimal bandwidth in Section 3.3. Unfortunately, due to the fact that the derived algorithm depends on the unknown terms such as the derivatives of the true hazard rate function it cannot be implemented in practice. We, therefore, propose a data-driven version of the L_1 optimal bandwidth in Section 3.4. The detailed proofs of the main theorems are given in Section 3.5 and in the last section, the main findings of this chapter are concluded.

3.2 Hazard rate estimation and its L_1 properties

In this chapter, we focus on the hazard rate function over a bounded support of x, that is,

$$\lambda(x) = \frac{f(x)}{1 - F(x)} \text{ for } 0 < x < T$$

where $T = \inf\{x|1 - F(x) < \epsilon\}$ for $\epsilon > 0$, can be large enough. The reasons to consider the truncated hazard rate function here are that the true hazard rate function $\lambda(x)$ is not necessarily a bounded function at the right end and that a kernel estimator is expected to be unreliable for large x. The unreliability mentioned in the latter reason stems from the fact that the variance of a kernel estimator is very high for large x, see for example Theorem 1.2.

Recall that

$$w(x) = \begin{cases} 1, & \text{if } 0 < x < T \\ 0, & \text{otherwise.} \end{cases}$$

Suppose that

$$\hat{\lambda}(x) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{K_h(x-X_i)}{1-F_n(X_i)}\right) = \sum_{i=1}^{n}\left(\frac{K_h(x-X_{(i)})}{n-i+1}\right).$$

This estimator will be referred to as the standard kernel hazard rate estimator. For $\hat{\lambda}(x)$, the bias and standard deviation are of the same order of magnitude. Let δ denote the common order of magnitude. Lo et al.[85] showed that $\hat{\lambda}(x)$ is asymptotically normally distributed, hence it can be represented as

$$\hat{\lambda}(x) - \lambda(x) = \delta(b(x) - \sigma(x)Z(x)) \tag{3.2.1}$$

where $\delta b(x)$ is asymptotic to the bias of $\hat{\lambda}$, $\delta\sigma(x)$ is asymptotic to the standard deviation of $\hat{\lambda}$, and $Z = Z(x)$ is the density function of Normal distribution (0,1). From (3.2.1), we may write the L_1 error of $\hat{\lambda}(x)$ as

$$\int E|\hat{\lambda}(x) - \lambda(x)|w(x)dx = \delta\eta + o(\delta)$$

where $\eta = \int w(x)dx \int_{-\infty}^{\infty} |b(x) - \sigma(x)z|\phi(z)dz$.

From Theorem 1.2, by setting the bandwidth $h_u = n^{-1/5}(u)^2$, asymptotically one can represent that δ by $n^{-2/5}$, $b(x)$ by $(u^2)^2 b_0(x)$, and $\sigma(x)$ by $(u^2)^{-1/2}\sigma_0(x)$ where $u > 0$ is a parameter not depending on x, $b_0(x) = \frac{\mu_{2,K}}{2}\lambda''(x)$ and $\sigma_0(x) = \frac{\sqrt{R(K)\lambda(x)}}{\sqrt{(1-F(x))}}$ which are independent with u. Hence η can be seen as a function of u and we have

$$\eta(u) = \int w(x)dx \int_{-\infty}^{\infty} |u^4 b_0(x) - u^{-1}\sigma_0(x)z|\phi(z)dz.$$

For a formal derivation to show that η could be represented by $\eta(u)$, we need the following assumptions.

A1. $K(\cdot)$ is a compactly supported, symmetric function on $\mathcal{R}$, with Hölder continuous derivatives, and satisfies

$$\int K(u)du = 1, \int u^2 K(u)du \neq 0.$$

.

A2. The density function $f(x)$ and hazard rate function $\lambda(x)$ are twice differentiable, and $f''(x)$ and $\lambda''(x)$ are uniformly continuous.

A3. $f(x)$ and $\lambda(x)$ are bounded and continuous over a bounded support.

Then we have the following theorem which gives precise asymptotic formula for the intergraded mean absolute error for $\hat{\lambda}(x)$.

Theorem 3.1. *Under the assumptions, A1, A2 and A3, we have*

$$n^{2/5} \int E|\hat{\lambda}(x|h_u) - \lambda(x)|w(x)dx = \eta(u) + o(1)$$

uniformly in $u \in [C^{-1}, C]$, for each $C > 1$.

Proof. The proof is given in Section 3.5.

Recall that in the standard kernel hazard rate estimation, by balancing the bias against standard deviation of the estimator one obtains the bandwidth which when used gives the 'optimal' estimator. That is the estimator which minimizes the L_2 errors, MISE. On the similar lines here we propose to find the bandwidth which minimizes integrated absolute error and thus the optimal kernel estimator in L_1-sense. For that let u^* be the value of u which minimizes $\eta(u)$. Then we show that the corresponding bandwidth $h_{u^*} = n^{-1/5}(u^*)^2$ leads to the optimal kernel estimator in sense of L_1 errors.

Theorem 3.2. *Under the assumptions A1, A2 and A3, by setting u^* to be the value of u which minimizes $\eta(u)$, we have*

$$\inf_{h>0} \int E|\hat{\lambda}(x|h) - \lambda(x)|w(x)dx \sim n^{-2/5}\eta(u^*). \tag{3.2.2}$$

Proof. The proof is given in Section 3.5.

Remark 3.1. The expected L_1 error of a standard kernel hazard rate estimator can be minimized asymptotically by balancing the order of magnitude of its bias and standard deviation. From the above theorem clearly, as in L_1 density estimation, the optimal value of the bandwidth in L_1 hazard rate estimator is also of the functional form $n^{-1/5}u^2$ where u may vary.

3.3 Theoretical L_1 optimal bandwidth

From Theorem 3.1 and 3.2, we showed that if there is a point u^* which minimizes $\eta(u)$, then the bandwidth u^* leads to the optimal kernel hazard rate estimator in sense of L_1 error. In this section , we show the existence and uniqueness of u^* and then develop a simple and rapid converging, iterative way to calculate u^*.

Recall that by balancing the order of magnitude of the bias and standard deviation of a kernel hazard rate estimator, we have the L_1 error as

$$\int E|\hat{\lambda}(x) - \lambda(x)|w(x)dx = \delta\eta + o(\delta)$$

where $\eta = \int w(x)dx \int_{-\infty}^{\infty} |b(x) - \sigma(x)z|\phi(z)dz$.

We simply it as

$$\begin{aligned}
\delta\eta &= \delta\int w(x)dx\int_{-\infty}^{b(x)/\sigma(x)}(b(x)-\sigma(x)z)\phi(z)dz \\
&\quad -\delta\int w(x)dx\int_{b(x)/\sigma(x)}^{\infty}(b(x)-\sigma(x)z)\phi(z)dz \\
&= 2\delta\int w(x)dx\int_{-\infty}^{b(x)/\sigma(x)}(b(x)-\sigma(x)z)\phi(z)dz \\
&= 2\delta\int w(x)dx\int_{-\infty}^{b(x)/\sigma(x)}(b(x)-\sigma(x)z)d\Phi(z) \\
&= 2\delta\int w(x)dx\Bigg[\bigg(\lim_{z\to b(x)/\sigma(x)}((b(x)-\sigma(x)z)\Phi(z))-\lim_{z\to-\infty}((b(x)-\sigma(x)z)\Phi(z))\bigg) \\
&\quad +\int_{-\infty}^{b(x)/\sigma(x)}\Phi(z)\sigma(x)dz\Bigg] \\
&= 2\delta\int \sigma(x)w(x)dx\int_{-\infty}^{b(x)/\sigma(x)}\Phi(z)dz
\end{aligned}$$

where Φ is the c.d.f of a standard normal distribution. From Theorem 3.1, we have proved that by setting the bandwidth $h_u = n^{-1/5}u^2$, approximately one can represent $b(x)$ by $u^4 b_0(x)$ and $\sigma(x)$ by $u^{-1}\sigma_0(x)$ in the functional form of $\delta\eta$ and thus have

$$\begin{aligned}
\delta\eta(u) &= 2\delta\int u^{-1}\sigma_0(x)w(x)dx\int_{-\infty}^{u^5 b_0(x)/\sigma_0(x)}\Phi(z)dz \\
&= 2\delta\int \sigma_0(x)w(x)dx\int_{-\infty}^{u^4 b_0(x)/\sigma_0(x)}\Phi(uz)dz.
\end{aligned}$$

Therefore to minimize the integrated mean absolute error of $\hat{\lambda}(x)$, under the case that its bias and standard deviation are of the same order, is to minimize $\eta(u)$ w.r.t u. This step can be implemented by letting the differential of $\eta(u)$ w.r.t u to be equivalent to 0. Noting that $\int_{z<y} z\phi(z)dz = -\phi(y)$, it gives

$$\begin{aligned}
&\frac{1}{2}\eta'(u) \\
&= \int\Bigg[\int_{-\infty}^{u^4 b_0(x)/\sigma_0(x)} z\phi(uz)dz + 4u^3\frac{b_0(x)}{\sigma_0(x)}\times\Phi(u^5 b_0(x)/\sigma_0(x))\Bigg]\sigma_0(x)w(x)dx \\
&= u^{-2}\Lambda(u^5)
\end{aligned}$$

where

$$\Lambda(v) = \int [4vb_0(x)\Phi(vb_0(x)/\sigma_0(x)) - \sigma_0(x)\phi(vb_0(x)/\sigma_0(x))]w(x)dx. \tag{3.3.1}$$

Let u^* be a solution to $\Lambda(u^5) = 0$. From the expression of $\Lambda(v)$, it is very difficult to find any explicit solution of u to $\Lambda(u^5) = 0$ directly. Hence to solve this equation, we employ the Newton's method. The next theorem demonstrates that there exists a unique and positive solution u^* to $\Lambda(u^5) = 0$.

Theorem 3.3. *For a standard kernel hazard rate estimator $\hat{\lambda}$, if the orders of its bias and standard error are balanced against each other, then there always exists a unique and positive solution u^* which minimizes the leading term of its integrated mean absolute error term, $\delta\eta(u)$.*

Proof. The proof is given in Section 3.5.

Remark 3.2. For a kernel hazard rate estimator with $h = n^{-1/5}u^2$,

$$\eta(u) = 2 \int u^{-1}\sigma_0(x)w(x)dx \int_{-\infty}^{u^5 b_0(x)/\sigma_0(x)} \Phi(z)dz$$

where $b_0(x) = \frac{\mu_{2,K}}{2}\lambda''(x)$ and $\sigma_0(x) = \frac{\sqrt{R(K)\lambda(x)}}{\sqrt{(1-F(x))}}$, let u^* denote the value of u which minimizes $\eta(u)$ for a particular second order kernel K. Also let $\eta_K(u)$ denote the function $\eta(u)$ when kernel K is used. Let $c_1 = \mu_{2,K}$ and $c_2 = (R(K))^{1/2}$ for kernel K while $c_{0,1}$ and $c_{0,2}$ being the versions of c_1 and c_2, respectively, for another kernel function K_0. Then for any u^* which minimizes $\eta_K(u)$,

$$u_0^* = u^*\{(c_{0,2}c_1)/(c_2 c_{0,1})\}^{1/5}$$

minimizes the quantity of $\eta_{K_0}(u_0)$. That is to say that once the value of u^* for a particular kernel is known, the optimal integrated mean absolute error bandwidth of other kernel functions can be derived easily.

Remark 3.3. From the last remark, it is clear that by changing 2nd-order kernel from say K to K_0, the integrated mean absolute error of a kernel estimator alters only by the quantities involving c_1 and c_2. That is to say, when we change the kernel function from K to K_0 that $c_{0,1} = a_1 c_1$ and $c_{0,2} = a_2 c_2$ for constants a_1 and a_2, then

$$\eta_{K_0}(u_0) = (a_1 a_2^4)^{1/5} \eta_K(u).$$

It implies that $\inf_u \eta_{K_0}(u) = (a_1 a_2^4)^{1/5} \inf_u \eta_K(u)$, and the optimal kernel is the one which minimizes $a_1 a_2^4$.

With the setting of $L(v) = v^{-1}\Lambda(v)$, let

$$\begin{aligned}
&H(v) = L(v)/L'(v) \\
&= \left[\int (4b_0(x)\Phi(vb_0(x)/\sigma_0(x)) - v^{-1}\sigma_0(x)\phi(vb_0(x)/\sigma_0(x)))w(x)dx\right] \\
&\times \left[\int (5b_0(x)^2\sigma_0(x)^{-1} + \sigma_0(x)v^{-2})\phi(vb_0(x)/\sigma_0(x))w(x)dx\right]^{-1}.
\end{aligned}$$

If v' is an approximation to the solution of $L(v) = 0$ then $v'' = v' - H(v')$ is a better approximation, and approximations converges to one point v^*quickly on iteration. Hence the optimal value u^* which minimizes $\eta(u)$ is given by $u^* = (v^*)^{-1/5}$.

Based on the above derivations, Newton method gives the the way to determine the optimal value of u^* which minimizes $\eta(u)$. $h = n^{-1/5}(u^*)^2$ is the optimal integrated mean absolute bandwidth of a kernel hazard rate estimator and u^* can be obtained by minimizing $\eta(u)$. However the algorithm could not be implemented directly since the derivation procedure of u^* using the Newton method above still depends on the unknown terms in the functional forms of $b_0(x)$ and $\sigma_0(x)$. To further address this issue, in the next section, we propose a data-driven bandwidth that can be used in practice and which does minimize the integrated mean absolute error in an asymptotic sense.

3.4 Adaptive L_1 optimal bandwidth selection

Still represent the expected L_1 distance of a kernel hazard rate estimator as $J(h) = \int E|\hat{\lambda}(x|h) - \lambda(x)|w(x)dx$. As showed in Theorem 3.1, when $h_u = n^{-1/5}u^2$, then

$$J(h_u) \sim n^{-2/5}\eta(u)$$

uniformly in $u \in [C^{-1}, C]$ for $C > 1$ and further by setting u^* be the unique value to minimize $\eta(u)$, $\inf_{h>0} J(h) \sim n^{-2/5}\eta(u^*)$.

Unfortunately, both $b_0(x) = \frac{\mu_{2,K}}{2}\lambda''(x)$ and $\sigma_0(x) = \sqrt{\frac{R(k)\lambda(x)}{1-F(x)}}$ are not known because of the presence of the unknown terms $\lambda''(x)$, $(\lambda(x))^{1/2}$ and $F(x)$. So any attempt at estimating u^* should involve explicit estimation of the unknown terms. In this section, we utilize the simple plug-in method to estimate the unknown terms. Specifically, $F(x)$ is replaced by its empirical cdf $F_n(x)$ while $\lambda^{(p)}(x)$ for positive integer p and $(\lambda(x))^{1/2}$ are estimated by

$$\hat{\lambda}_1^{(p)}(x) = \frac{1}{nh_1^{p+1}} \sum_{i=1}^{n} \frac{K_1^{(p)}\{(x - X_i)/h_1\}}{1 - F_n(X_i)}$$

and

$$(\hat{\lambda}_2(x))^{1/2} = \left[\frac{1}{nh_2} \sum_{i=1}^{n} \frac{K_2\{(x - X_i)/h_2\}}{1 - F_n(X_i)}\right]^{1/2}$$

respectively. Note that one could simply take the pth derivative of $\hat{\lambda}$, $\hat{\lambda}^{(p)}(x)$ and the square-root of $\hat{\lambda}$, $(\hat{\lambda}(x))^{1/2}$ as estimators for $\lambda^{(p)}(x)$ and $(\lambda(x))^{1/2}$. However, to allow the possibilities of using a different kernel than the one used to define $\hat{\lambda}(x)$, we have $\hat{\lambda}_1^{(p)}(x)$ as the derivative of the standard kernel estimator with kernel K_1 and $(\hat{\lambda}_2(x))^{1/2}$ as the square-root of the standard kernel estimator with kernel K_2.

The strong consistency properties of the above estimators of $\lambda^{(p)}(x)$ and $(\lambda(x))^{1/2}$ are demonstrated in the following theorem.

Theorem 3.4. *Assume that $K_1^{(p)}$ is well-defined and bounded and $E(|X_1|^\gamma) < \infty$ for some $\gamma > 1$. Then under the assumptions A1, A2, A3, as $h_1, h_2 \to 0$, $nh_1^{2p+1}/\log n \to \infty$*

$$\int |\hat{\lambda}_1^{(p)}(x) - \lambda^{(p)}(x)|w(x) \to 0 \ a.s. \tag{3.4.1}$$

and

$$\int |\hat{\lambda}_2^{1/2}(x) - \lambda^{1/2}(x)|w(x) \to 0 \ a.s.. \tag{3.4.2}$$

Proof. The proof is given in Section 3.5.

Remark 3.4. The condition $E(|X|^\gamma)$ where $\gamma > 1$ imposed in this theorem is to exclude the probability models for which $E(|X|) = \infty$.

Remark 3.5. In estimation of both $(\lambda(x))^{(p)}$ and $(\lambda(x))^{1/2}$, we need to use the initial (pilot) bandwidths, h_1 and h_2. See Section 3.2.1 of [166] for the details.

From Theorem 3.4, it is easy to verify that the plug-in estimators of $\hat{b}_0(x)$ and $\hat{\sigma}_0(x)$ are consistent in the sense that,

$$\int (|\hat{b}_0 - b_0| + |\hat{\sigma}_0 - \sigma_0|)w(x) \to 0 \ a.s. \tag{3.4.3}$$

as $n \to \infty$.

We could define $\hat{u}^*$ to be the value of u minimizing $\hat{\eta}(u)$ with $\hat{b}_0$ and $\hat{\sigma}_0$, and $\hat{h}^*$ be the corresponding optimal data-driven integrated mean absolute bandwidth. In order to identify whether $\hat{u}^*$ is also optimal in the sense of integrated mean absolute errors as $n \to \infty$, we summarize the important asymptotic properties of $\hat{u}^*$ in the next theorem.

Theorem 3.5. *Under the assumptions, A1, A2 and A3, we have*

$$J(\hat{h}^*)/\inf_{h>0} J(h) \to 1 \ a.s.$$

as $n \to \infty$.

Proof. The proof is given in Section 3.5.

From Theorem 3.5, it is clear that the data-driven bandwidth $\hat{h}^*$ proposed in this section is asymptotically optimal w.r.t integrated mean absolute error in the setting of hazard rate estimation.

However, it is important to keep in mind that $\hat{h}^*$ is a random variable that depends on the underlying data sample $\mathcal{X}$. That is to say, in a particular data sample, we may be also interested to see whether the proposed optimal bandwidth also minimizes the integrated absolute error,

$$\hat{J}(h) = \int |\hat{\lambda}(x|h) - \lambda(x)|w(x)dx.$$

between the estimator and the true hazard rate function. So it is more natural for us to ask a question that whether $\hat{h}^*$ derived from one dataset is asymptotically as good as the true optimal bandwidth in minimizing the integrated absolute error, $\hat{J}(h)$? In others words, that is whether

$$\hat{J}(\hat{h}^*)/\inf_{h>0} J(h) \to 1 \ a.s. \tag{3.4.4}$$

holds as $n \to \infty$. If that is the case, it will provide us another evidence to believe that $\hat{h}^*$ should be an optimal L_1 error bandwidth for a hazard rate estimator. So as to address this issue, we show that (3.4.4) does hold under appropriate conditions in the next theorem,

Theorem 3.6. *Assume A1, A2 and A3, then*

$$\{\inf_{h>0} \hat{J}(h)\}/\{\inf_{h>0} J(h)\} \to 1 \text{ and } \hat{J}(\hat{h}^*)/J(\hat{h}^*) \to 1 \ a.s. \tag{3.4.5}$$

as $n \to \infty$.

Proof. The proof is given in Section 3.5.

In summary, from Theorem 3.5 and Theorem 3.6, we have demonstrated that with the consistent estimators of $b_0(x)$ and $\sigma_0(x)$, the obtained bandwidth $\hat{h}^*$ minimizes the asymptotic integrated mean absolute error and integrated absolute error as $n \to \infty$. That is to say, the adaptive bandwidth choice, $\hat{h}^*$ can be seen as the optimal bandwidth in minimizing the L_1 error in hazard rate kernel estimation.

To illustrate the computation of $\hat{u}^*$ and the bandwidth $\hat{h}^*$ using the Newton method, define,

$$\hat{\eta}(u) = \int w(x)dx \int_{-\infty}^{\infty} |u^4 \hat{b}_0(x) - u^{-1}\hat{\sigma}_0(x)z|\phi(z)dz,$$

where $\hat{b}_0$ and $\hat{\sigma}_0$ are estimated by the plug in approaches discussed above. To minimize $\hat{\eta}(u)$ w.r.t u using the Newton method, we start with an initial value v_0 and then iteratively estimate v_{j+1} by v_j as

$$v_{j+1} = v_j - \hat{H}(v_j), \ j \geq 0$$

where

$$\begin{aligned}\hat{H}(v) = & \left[\int w(x)(r\hat{b}_0\Phi(v\hat{b}_0/\hat{\sigma}_0) - v^{-1}\hat{\sigma}_0 - \phi(v\hat{b}_0/\hat{\sigma}_0))dx\right] \\ & \times \left[\int w(x)((r+1)\hat{b}_0^2\hat{\sigma}_0^{-1} + \hat{\sigma}_0 v^{-2})\phi(v\hat{b}_0/\hat{\sigma}_0)dx\right]^{-1}.\end{aligned}$$

After finite number of iterations, the sequence $\{v_j\}$ converges to a fixed value $\hat{v}^*$ and the algorithm terminates. Then the adaptive optimal L_1 bandwidth $\hat{h}^*$ equals to $n^{-1/5}(\hat{u}^*)^2$ where $\hat{u}^* = (\hat{v}^*)^{1/5}$.

3.5 Proofs

We finish with the proofs of the theorems in this chapter. First we state and prove some important lemmas which are required to prove the main theorems in this chapter. Throughout the proofs given below, the symbols $C, C_1, C_2, ...$ denote positive generic constants.

Lemma 3.1. Let $(0, T)$ be a bounded interval. Then, for all $h > 0$ and 2nd order kernel K, the standard deviation $V(x)$ of a standard hazard rate kernel estimator satisfies that

$\int_0^T \left| V(x) - \sqrt{\frac{R(K)\lambda(x)}{nh(1-F(x))}} \right| dx = o(\frac{1}{\sqrt{nh}})$as $h \to 0$.

Proof. With reference to Watson and Leadbetter[168], it is known that the variance of standard kernel hazard rate estimator can be written as

$$\begin{aligned} V^2(x) &= \int I_n(F_n(u))\lambda(u)K_h^2(x-u)du \\ + & \ 2\int\int_{u\le z} A_n(u,z)K_h(x-u)K_h(x-z)\lambda(u)\lambda(z)dudz \end{aligned}$$

where $I_n(F_n(u)) = \sum_{i=1}^n \frac{1}{n-i+1}\binom{n}{i-1}F(u)^{i-1}(1-F(u))^{n-i+1}$ and $A_n(u,z) = (1-F^n(u))F^n(z) - (1-F(u))\frac{F^n(z)-F^n(u)}{F(z)-F(u)}$.

It has been demonstrated in [168] that $I_n(F(u))$ converges to $\frac{1}{n(1-F(u))}$, then

$$\begin{aligned} & V^2(x) \\ = & \ \frac{1}{nh^2}\int K^2\left(\frac{x-u}{h}\right)\frac{\lambda(u)}{1-F(u)}du \\ & \qquad +2\int\int_{u\le z} A_n(u,z)K_h(x-u)K_h(x-z)\lambda(u)\lambda(z)dudz \\ = & \ \frac{1}{nh}\int \frac{1}{h}\frac{K^2(\frac{x-u}{h})}{R(K)}\frac{\lambda(u)}{1-F(u)}R(K)du \\ & \qquad +2\int\int_{u\le z} A_n(u,z)K_h(x-u)K_h(x-z)\lambda(u)\lambda(z)dudz \\ = & \ \frac{1}{nh}\int K_h^*(x-u)\frac{\lambda(u)}{1-F(u)}R(K)du \\ & \qquad +2\int\int_{u\le z} A_n(u,z)K_h(x-u)K_h(x-z)\lambda(u)\lambda(z)dudz \\ = & \ \frac{R(K)\lambda(x)}{nh(1-F(x))} + \left(\frac{\lambda}{1-F} * K_h^*(x) - \frac{\lambda(x)}{1-F(x)}\right)\frac{R(K)}{nh} \\ & \qquad +2\int\int_{u\le z} A_n(u,z)K_h(x-u)K_h(x-z)\lambda(u)\lambda(z)dudz \end{aligned}$$

where $R(K) = \int K^2$, $K^* = K^2/\int K^2$ and the notation $*$ is the convolution operator that for example, $\lambda * K(x)$ can be written as $\int \lambda(y)K(x-y)dy$ or $\int K(y)\lambda(x-y)dy$. Denote

$$D(x) = 2\int\int_{u\le z} A_n(u,z)K_h(x-u)K_h(x-z)\lambda(u)\lambda(z)dudz,$$

then we can define

$$\alpha(x) = \frac{R(K)\lambda(x)}{nh(1-F(x))},$$

$$\beta(x) = \left(\frac{\lambda}{1-F} * K_h^*(x)\right)\frac{R(K)}{nh} - \left(\frac{\lambda(x)}{1-F(x)}\right)\frac{R(K)}{nh} + D(x)$$

and thus achieve that $V^2(x) = \alpha(x) + \beta(x)$.

Clearly, $\int_0^T \sqrt{\lambda(x)}dx < +\infty$ and $\alpha(x) \geq 0$. Thus, it gives $\sqrt{\alpha(x)+\beta(x)} \leq \sqrt{\alpha(x)} + \sqrt{|\beta(x)|}$ and $\sqrt{\alpha(x)+\beta(x)} \geq \sqrt{\alpha(x)} - \sqrt{|\beta(x)|}$. Integrating over $(0,T)$ gives

$$\int_0^T V(x)dx \leq \sqrt{\frac{R(K)}{nh}}\Bigg(\int_0^T \sqrt{\frac{\lambda(x)}{1-F(x)}}dx \\ + \int_0^T \sqrt{\left|\frac{\lambda}{1-F} * K_h^*(x) - \frac{\lambda(x)}{1-F(x)} + \frac{nh}{R(K)}D(x)\right|}dx\Bigg)$$

and also,

$$\int_0^T V(x)dx \geq \sqrt{\frac{R(K)}{nh}}\Bigg(\int_0^T \sqrt{\frac{\lambda(x)}{1-F(x)}}dx \\ - \int_0^T \sqrt{\left|\frac{\lambda}{1-F} * K_h^*(x) - \frac{\lambda(x)}{1-F(x)} + \frac{nh}{R(K)}D(x)\right|}dx\Bigg).$$

As $F(x)$ is between 0 and 1, it is easy to see that $A_n(u,z)$ is of order $o(\frac{1}{nh})$ and thus $D(x)$ is of order $o(\frac{1}{nh})$. Note that by applying the Cauchy-Schwarz inequality, it gives

$$\begin{aligned}
& \int_0^T \sqrt{\left|\frac{\lambda}{1-F} * K_h^*(x) - \frac{\lambda(x)}{1-F(x)} + \frac{nh}{R(K)}D(x)\right|}dx \\
= & \int_0^T \sqrt{\left|\frac{\lambda}{1-F} * K_h^*(x) - \frac{\lambda(x)}{1-F(x)} + o(1)\right|} \times 1dx \\
\leq & \left(\int_0^T \left|\frac{\lambda}{1-F} * K_h^*(x) - \frac{\lambda(x)}{1-F(x)} + o(1)\right|dx \int_0^T 1dx\right)^{1/2} \\
= & \sqrt{T\int_0^T \left|\frac{\lambda}{1-F} * K_h^*(x) - \frac{\lambda(x)}{1-F(x)} + o(1)\right|dx}.
\end{aligned}$$

Then we have that (excluding $o(1)$ term)

$$\begin{aligned}\int_0^T V(x)dx &\leq \sqrt{\frac{R(K)}{nh}}\Bigg(\int_0^T \sqrt{\frac{\lambda(x)}{1-F(x)}}dx \\ &+ \sqrt{T\int_0^T \left|\frac{\lambda}{1-F} * K_h^*(x) - \frac{\lambda(x)}{1-F(x)}\right|dx}\Bigg)\end{aligned}$$

and

$$\begin{aligned}\int_0^T V(x)dx &\geq \sqrt{\frac{R(K)}{nh}}\Bigg(\int_0^T \sqrt{\frac{\lambda(x)}{1-F(x)}}dx \\ &- \sqrt{T\int_0^T \left|\frac{\lambda}{1-F} * K_h^*(x) - \frac{\lambda(x)}{1-F(x)}\right|dx}\Bigg).\end{aligned}$$

From Theorem 1 on p.6 of [36], $\int_0^T \left|\frac{\lambda}{1-F} * K_h^*(x) - \frac{\lambda(x)}{1-F(x)}\right|dx$ converges to 0 as $h \to 0$, so it can be verified that $\int_0^T \left|\frac{\lambda}{1-F} * K_h^*(x) - \frac{\lambda(x)}{1-F(x)}\right|dx \leq o(1)$.

It thus gives

$$\int_0^T V(x)dx \leq \sqrt{\frac{R(K)}{nh}}\left(\int_0^T \sqrt{\frac{\lambda(x)}{1-F(x)}}dx + o(1)\right)$$

and

$$\int_0^T V(x)dx \geq \sqrt{\frac{R(K)}{nh}}\left(\int_0^T \sqrt{\frac{\lambda(x)}{1-F(x)}}dx - o(1)\right).$$

Therefore $\int_0^T V(x)dx = \sqrt{\frac{R(K)}{nh}}\left(\int_0^T \sqrt{\frac{\lambda(x)}{1-F(x)}}dx + o(1)\right)$.

Lemma 3.2. For any K satisfying A1 and A2, it is known that the bias term $B(x)$ of a standard kernel hazard rate estimator is asymptotic to $K_h * \lambda(x) - \lambda(x)$, then it satisfies $\int_0^T \left||B(x)| - \frac{h^2\mu_{2,K}}{2}|\lambda''(x)|\right| dx = o(h^2)$ as $h \to 0$.

Proof. Making use of Taylor series expansion, we can write

$$\lambda(y) = \lambda(x) + (y-x)\lambda'(x) + \int_x^y (y-z)\lambda''(z)dz.$$

The bias term $B(x)$ could be written as

$$\begin{aligned}
& K_h * \lambda(x) - \lambda(x) \\
= & \int \frac{1}{h} K\left(\frac{x-y}{h}\right) \int_x^y (y-z)\lambda''(z) dz dy \\
= & \int \frac{1}{h} K\left(\frac{x-y}{h}\right) \times \\
& \int [I_{y \geq x} I_{x \leq z \leq y} (y-z)\lambda''(z) dz + I_{y \leq x} I_{x \geq z \geq y} (z-y)\lambda''(z) dz] dy \\
= & \int \frac{1}{h} K\left(\frac{x-y}{h}\right) \int [I_{y \geq x} I_{x \leq z \leq y} + I_{y \leq x} I_{x \geq z \geq y}] |y-z| \lambda''(z) dz dy \\
= & \int K(u) \int [I_{u \leq 0} I_{x \leq z \leq x-hu} + I_{u \geq 0} I_{x \geq z \geq x-hu}] |x-hu-z| \lambda''(z) dz du \\
= & h^2 \int \frac{1}{h} K(u) \int [I_{u \leq 0} I_{0 \leq \frac{z-x}{h} \leq -u} + I_{u \geq 0} I_{0 \geq \frac{z-x}{h} \geq -u}] \left| u + \frac{z-x}{h} \right| \lambda''(z) dz du \\
= & h^2 \int \frac{1}{h} \tilde{K}\left(\frac{x-z}{h}\right) \lambda''(z) dz = h^2 \tilde{K}_h * \lambda''(z)
\end{aligned}$$

where $\tilde{K}(y) = \int K(u)[I_{u \leq 0} I_{u \leq y \leq 0} + I_{u \geq 0} I_{0 \leq y \leq u}] \, |u-y| \, du$. It is straightforward to see that $\tilde{K}$ is bounded and symmetric when K is. Note that

$$\begin{aligned}
\int \tilde{K}(y) dy &= \int \int K(u)[I_{u \leq 0} I_{u \leq y \leq 0} + I_{u \geq 0} I_{0 \leq y \leq u}] \, |u-y| \, du dy \\
&= \int K(u)(I_{u \leq 0} \int_u^0 |u-y| dy + I_{u \geq 0} \int_0^u |u-y| dy) du \\
&= \frac{1}{2} \int u^2 K(u) du = \frac{\mu_{2,K}}{2}.
\end{aligned}$$

Then we have

$$\begin{aligned}
& \int_0^T \left| |B(x)| - \frac{h^2 \mu_{2,K}}{2} |\lambda''(x)| \right| dx \\
= & \int_0^T \left| |h^2 \tilde{K}_h * \lambda''(x)| - \frac{h^2 \mu_{2,K}}{2} |\lambda''(x)| \right| dx \\
\leq & \int_0^T \left| h^2 \tilde{K}_h * \lambda''(x) - \frac{h^2 \mu_{2,K}}{2} \lambda''(x) \right| dx.
\end{aligned}$$

From Theorem 1 on p.6 of [36], it is known that

$$\int_0^T \left| h^2 \tilde{K}_h * \lambda''(x) - \frac{h^2 \mu_{2,K}}{2} \lambda''(x) \right| dx$$

is of order $o(h^2)$ as $h \to 0$, hence the result follows.

Lemma 3.3. With a bounded, continuous, 2nd order kernel K, if the density function $f(x)$ and the hazard rate function $\lambda(x)$ are bounded and continuous, then the standard kernel hazard rate estimator based on K satisfies

$$\int E|\hat{\lambda}(x|h) - E\hat{\lambda}(x|h)|w(x)dx > C \min\{(nh)^{-1/2}, 1\}.$$

Proof. Note that

$$\begin{aligned}
& \int E|\hat{\lambda}(x|h) - E\hat{\lambda}(x|h)|w(x)dx \\
= & \int E|\hat{\lambda}(x|h) - \lambda^*(x|h) + \lambda^*(x|h) - E\hat{\lambda}(x|h)|w(x)dx \\
\geq & \int E[|\lambda^*(x|h) - E\hat{\lambda}(x|h)| - |\hat{\lambda}(x|h) - \lambda^*(x|h)|]w(x)dx \\
\geq & \int E|\lambda^*(x|h) - E\hat{\lambda}(x|h)|w(x)dx - \int E|\hat{\lambda}(x|h) - \lambda^*(x|h)|w(x)dx \qquad (3.5.1)
\end{aligned}$$

where $\lambda^*(x) = \frac{1}{n} \sum \frac{K_h(x-X_i)}{1-F(X_i)}$.

We begin with proving that the first term of (3.5.1) has

$$\int E|\lambda^*(x|h) - E\hat{\lambda}(x|h)|w(x)dx > C \min\{(nh)^{-1/2}, 1\}. \qquad (3.5.2)$$

It is straightforward to see that $E\hat{\lambda}(x|h) = E\lambda^*(x|h)$, therefore we set that $Y_i = \frac{K_h(x-X_i)}{1-F(X_i)} - E\left(\frac{K_h(x-X_i)}{1-F(X_i)}\right)$ and obtain that $\sum Y_i/n = \lambda^*(x) - E\hat{\lambda}(x)$. By defining M to be any term of order $o((nh)^{-1})$ and using Lemma 8 on p.90 of [36], one has

$$\begin{aligned}
\sup\left|E\left|\frac{\sqrt{n}}{\sigma_Y n}\sum_{i=1}^{n} Y_i - \frac{M}{V}\right| - \psi\left(\left|\frac{M}{V}\right|\right)\right| &\leq \frac{C\rho_Y\sigma_Y^{-3}}{\sqrt{n}} \\
\sup\left|E\left|\frac{\lambda^* - E\hat{\lambda}}{V} - \frac{M}{V}\right| - \psi\left(\left|\frac{M}{V}\right|\right)\right| &\leq \frac{C\rho_Y\sigma_Y^{-3}}{\sqrt{n}} \\
\sup\left|E\left|\lambda^* - E\hat{\lambda} - o\left(\frac{1}{nh}\right)\right| - V\psi\left(\left|\frac{M}{V}\right|\right)\right| \leq \frac{C\rho_Y\sigma_Y^{-2}}{n} &\leq \frac{CK_{max}}{nh} \\
\sup\left(E\left|\lambda^* - E\hat{\lambda}\right| - V\psi\left(\left|\frac{M}{V}\right|\right)\right) &\leq \frac{CK_{max}}{nh}
\end{aligned}$$

where C is a positive constant, V^2 is the variance of $\lambda^*(x)$ (which can be shown to be asymptotically same as that of $\hat{\lambda}(x)$), $\sigma_Y^2 = E(Y_i^2) = nV^2$ for $i = 1, 2, ..., n$, $\rho_Y = E(|Y_i|^3) < \infty$, K_{max} is the upper bound for kernel function K and

$$\psi(u) = \sqrt{2}\pi\left(u\int_0^u e^{-x^2/2}dx + e^{-u^2/2}\right),\ u \geq 0.$$

Then with the inequality given on p.94 of [36], one has

$$V\psi\left(\left|\frac{M}{V}\right|\right) \leq |M| + \sqrt{\frac{2V^2}{\pi}}.$$

It gives

$$\sup(E|\lambda^*(x) - E\hat{\lambda}(x)|) - V\sqrt{\frac{2}{\pi}} \quad \leq \quad \frac{CK_{max}}{nh} \tag{3.5.3}$$

where C does not depend on x, n and h.

Recall that $nh \geq 1$, simple calculations can prove that

$$\int V(x)w(x)dx = \int (E|\lambda^*(x) - E\hat{\lambda}(x)|^2)^{1/2}w(x)dx \geq C_2(nh)^{-1/2}. \tag{3.5.4}$$

Therefore, if $nh > C_3$ and C_3 is sufficiently large, in views of (3.5.3) and (3.5.4), one has that

$$\int E|\lambda^*(x) - E\hat{\lambda}(x)|w(x)dx \geq C_2(nh)^{-1/2}. \tag{3.5.5}$$

If $nh \leq C_3$, suppose K vanishes outside $[-s, s]$, centered at the origin, then $\lambda^*(x|h) = 0$ if $|x - X_j| > sh$ for each j, $1 \leq j \leq n$. Therefore with reference to the proof of Lemma 5.1 in [51], for the case that $nh < C_3$ and n is sufficiently large, the chance that $\lambda^*(x|h)$ equals to 0 exceeds

$$\begin{aligned} p(x,n) &= (P(|x - X| > sh))^n \geq (1 - Uvsh)^n \\ &\geq C_4 \exp(-nUvsh) \geq C_5 > 0 \end{aligned}$$

where U is an upper bound to f, v equals to the content of the unit radius, and C_5 does not depend on x, n or h. Therefore,

$$\begin{aligned} & \int E|\lambda^*(x) - E\hat{\lambda}(x)|w(x)dx \\ \geq & \int p(x,n)|E(\hat{\lambda}(x|h))|w(x)dx \\ \geq & C_5|\int E(\hat{\lambda}(x|h))w(x)dx| = C_6. \end{aligned} \tag{3.5.6}$$

Equation (3.5.2) follows from (3.5.5) and (3.5.6).

Next we prove that the second term of (3.5.1) is of order $o(1)$. Consider that

$$\begin{aligned} & \int E|\hat{\lambda}(x|h) - \lambda^*(x|h)|w(x)dx \\ = & \int w(x)dx \int \left| \frac{K_h(x-y)}{1 - F_n(y)} - \frac{K_h(x-y)}{1-F(y)} \right| f(y)w(y)dy \\ = & \int w(x)dx \int \frac{|F_n(y) - F(y)|}{1 - F_n(y)} K_h(x-y)\lambda(y)w(y)dy. \\ \leq & \sup_{y\in[0,T]} |F_n(y) - F(y)| \int w(x)dx \int \frac{K_h(x-y)}{1-F_n(y)}\lambda(y)w(y)dy. \end{aligned}$$

From Glivenko-Cantelli theorem, it is clear that $\sup_{x\in[0,T]} |F_n(x) - F(x)| = o(1)$ as $n \to \infty$.

Notice that $\lambda(x)$ is bounded on $[0, T]$, so the second term of (3.5.1) is of order $o(1)$. That is to say, the order of the leading term of (3.5.1) won't be changed by its second term at all. Hence, it is ready to show that

$$\int E|\hat{\lambda}(x|h) - E\hat{\lambda}(x|h)|w(x)dx > C \min\{(nh)^{-1/2}, 1\}.$$

Lemma 3.4. Consider an increasing sequence $\Sigma_0, ..., \Sigma_n$ of sub-σ-fields of a basic probability space, where Σ_0 is trivial. A sequence of random variables Z_i, $1 \leq i \leq n$, is called a martingale difference sequence if each Z_i is Σ_i-measurable, and $E(Z_i|\Sigma_{i-1}) = 0$ for each i. It gives

$$P\left(\left|\sum_{i=1}^{n} Z_i\right| > \varepsilon\right) \leq 2\exp\left(-\frac{\varepsilon^2}{2\int_1^n ||Z_i||_\infty^2}\right) \tag{3.5.7}$$

where $\varepsilon > 0$ and $||Z_i||_\infty$ is the essential supremum norm of Z_i.

Proof. Please refer to the proof of Lemma 2 of Devroye[35].

Lemma 3.5. Let X be any random variable with finite mean, and let a be an arbitrary real number. Then

$$||X - a| - E(|X - a|)| \leq |X - E(X)| + E(|X - E(X)|).$$

Proof. Please refer to the proof of Lemma 1 of Devroye[35].

Proof. of Theorem 3.1

The first step is to show that for any $C \geq 1$ and $h \leq 1$,

$$I(n, h, C) = \int_{|x|>C} E|(\hat{\lambda}(x|h) - \lambda(x)|w(x)dx \leq g(C)((nh)^{-1/2} + h^2) \tag{3.5.8}$$

where $g(C)$ does not depend on n or h and converges to zero as $C \to T$.

By Taylor Expansion with the Lagrange remainder, we have

$$\lambda(x-hz)=\lambda(x)-hz\lambda'(x)+\int_x^{x-hz}\lambda''(u)(x-hz-u)du.$$

Then by setting set $u=x-thz$, one obtains

$$\lambda(x-hz)=\lambda(x)-hz\lambda'(x)+\int_0^1(hz)^2\lambda''(x-thz)(1-t)dt.$$

Therefore, $|E\hat{\lambda}(x|h)-\lambda(x)|$ could be written as

$$\begin{aligned}
&|E\hat{\lambda}(x|h)-\lambda(x)|\\
&=\left|\int\lambda(x-hz)K(z)dz-\lambda(x)\right|\\
&=\left|\int K(z)\left(\lambda(x)-hz\lambda'(x)+\int_0^1 h^2z^2(\lambda(x-thz))''(1-t)dt\right)dz\right.\\
&\left.-\lambda(x)\right|\\
&=h^2\left|\int K(z)z^2dz\int_0^1(\lambda(x-thz))''(1-t)dt\right|. \qquad (3.5.9)
\end{aligned}$$

Notice that if $|hz|\le C$ and $0<t<1$, then $\{x:x>2C\}\subseteq\{x:x-thz>C\}$. Thus for the case that $h\le 1$, one has

$$\begin{aligned}
I_1(n,h,2C) &= \int_{|x|>2C}|E\hat{\lambda}(x)-\lambda(x)|w(x)dx\\
&\le h^2\Big\{\int K(z)z^2dz\int_{|x|>C}|\lambda''(x)|w(x)dx\\
&+\int_{|hz|>C}K(z)z^2dz\int|\lambda''(x)|w(x)dx\Big\}\\
&\le g_1(2C)h^2
\end{aligned}$$

where

$$g_1(2C)=\int K(z)z^2dz\int_{|x|>C}|\lambda''(x)|w(x)dx+\int_{|hz|>C}K(z)z^2dz\int|\lambda''(x)|w(x)dx.$$

$g_1(2C)$ is bounded since both $\int_{|x|>C} |\lambda''(x)|w(x)dx$ and $\int |\lambda''(x)|w(x)dx$ are finite and converges to 0 as $2C \to T$.

On the other hand, from the variance expression given in Lemma 3.1, we could easily show that as $n \to \infty$ and C_0 is sufficiently large,

$$Var(\hat{\lambda}(x)) \leq \frac{C_0}{nh} \int \frac{K^2(z)\lambda(x-hz)}{1-F(x-hz)} dz.$$

Assume that $\alpha > 1$ and set

$$g_2(C) = \left\{ \int_{|x|>C} (1+|x|^\alpha)^{-1} w(x)dx \right\}^{1/2}$$

where $1+|x|^\alpha \leq 2^\alpha(1+|x-hz|^\alpha+|hz|^\alpha)$. Then using $h \leq 1$ we obtain

$$\begin{aligned}
&I_2(n,h,C) \\
&= \int_{|x|>C} [Var(\hat{\lambda}(x|h))]^{1/2} w(x)dx \\
&\leq g_2(C) \left[\int Var(\hat{\lambda}(x|h))(1+|x|^\alpha)w(x)dx \right]^{1/2} \\
&\leq g_2(C) \left[\frac{C_0}{nh} \int (1+|x|^\alpha)w(x) \int \frac{K^2(z)\lambda(x-hz)}{1-F(x-hz)} dzdx \right]^{1/2} \\
&\leq g_2(C) \left[\frac{C_0}{nh} \int 2^\alpha (1+|x-hz|^\alpha+|hz|^\alpha) w(x) \int \frac{K^2(z)f(x-hz)}{(1-F(x-hz))^2} dzdx \right]^{1/2} \\
&\leq g_2(C) \left[\frac{C_0}{nh} \int_0^T 2^\alpha (1+|x|^\alpha+z^\alpha) \int \frac{K^2(z)}{(1-F(T))^2} dzdx \right]^{1/2} \\
&\leq g_3(C)(nh)^{-1/2}
\end{aligned}$$

where the first inequity is given by Hölder inequality and

$$g_3(C) = g_2(C)(2^\alpha C_0)^{1/2} \left[\int K^2(z)dz \int_0^T \frac{(1+|x|^\alpha)}{(1-F(T))^2} dx + \int T \frac{|z|^\alpha K^2(z)dz}{(1-F(T))^2} \right]^{1/2}.$$

This quantity is finite if α is sufficiently close to unity. Notice $g_2(C)$ converges to 0 as $C \to T$, so $g_3(C)$ converges 0 simultaneously. Therefore we show that $I_1(n,h,C) \leq g_1(C)h^2$

and $I_2(n,h,C) \leq g_3(C)(nh)^{-1/2}$.

Notice that the mean squared error of estimator, $E(|\hat{\lambda}(x) - \lambda(x)|^2)$ equals to the sum of the square of the bias, $|E\hat{\lambda}(x) - \lambda(x)|^2$ and the variance, $Var(\hat{\lambda}(x))$, therefore $\sqrt{E(|\hat{\lambda}(x) - \lambda(x)|^2)} \leq |E\hat{\lambda}(x) - \lambda(x)| + \sqrt{Var(\hat{\lambda}(x))}$. Cauchy-Schwarz inequality gives $E|\hat{\lambda}(x) - \lambda(x)| \leq \sqrt{E(|\hat{\lambda}(x) - \lambda(x)|^2)}$ and thus $E|\hat{\lambda}(x) - \lambda(x)| \leq |E\hat{\lambda}(x) - \lambda(x)| + \sqrt{Var(\hat{\lambda}(x))}$. Now it is ready to verify that $I(n,h,C) \leq I_1(n,h,C) + I_2(n,h,C)$, thereby selecting $g(C) = g_1(C) + g_3(C)$, we prove (3.5.8).

Next, we set that

$$\eta(u, C_2) = \int_{|x|\leq C_2} dx \int_{-\infty}^{\infty} |u^4 b_0(x) - u^{-1}\sigma_0(x)z|\phi(z)dz,$$

then it is clear that

$$\lim_{C_2 \to T} \sup_{u\in[C_1^{-1}, C_1]} |\eta(u, C_2) - \eta(u)| = 0, \tag{3.5.10}$$

for any $C_1 > 1$, where

$$\eta(u) = \int w(x)dx \int_{-\infty}^{\infty} |u^4 b_0(x) - u^{-1}\sigma_0(x)z|\phi(z)dz.$$

Observe that

$$\begin{aligned}
&\sup_{u\in[C_1^{-1}, C_1]} \left| \int_{|x|\leq C_2} E|\hat{\lambda}(x|h_u) - \lambda(x)|w(x)dx - \delta\eta(u, C_2) \right| \\
= &\sup_{u\in[C_1^{-1}, C_1]} \delta \left| \int_{|x|\leq C_2} dx \int (|b(x) - \sigma(x)z| - |u^4 b_0(x) - u^{-1}\sigma_0(x)z|)\phi(z)dz \right| \\
\leq &\sup_{u\in[C_1^{-1}, C_1]} \delta \left| \int_{|x|\leq C_2} dx \int [(b(x) - u^4 b_0(x))\phi(z) - (\sigma(x) - u^{-1}\sigma_0(x))z\phi(z)]dz \right|.
\end{aligned}$$

Recall that when the bias term and standard deviation term of the kernel estimator are balanced against each other, the common magnitude δ should be $n^{-2/5}$ and it is achieved by setting $h_u = n^{-1/5}u^2$, thus using Lemma 3.1, we have $\delta \int_{|x|\leq C_2} |\sigma(x) - u^{-1}\sigma_0(x)|dx =$

$o(\delta)$ and in the same way, using Lemma 3.2, we have $\delta \int_{|x|\le C_2} |b(x) - u^4 b_0(x)|dx = o(\delta)$. Therefore it is ready to show that

$$\sup_{u\in[C_1^{-1},C_1]} \left| \int_{|x|\le C_2} E|\hat{\lambda}(x|h_u) - \lambda(x)|w(x)dx - \delta\eta(u, C_2) \right| \le o(\delta). \tag{3.5.11}$$

Finally using uniform convergence, Theorem 3.1 can been verified if we combing the conclusions from (3.5.8), (3.5.10) and (3.5.11) respectively.

Proof. of Theorem 3.2

We need to prove that for some $C > 0$,

$$\int E|\hat{\lambda}(x) - \lambda(x)|w(x)dx \ge C[\min((nh)^{-1/2}, 1) + \min(h^2, 1)]$$

whenever $n \ge 1$ and $h > 0$.

Since $3E|\hat{\lambda}(x) - \lambda(x)| \ge E|\hat{\lambda}(x) - E\hat{\lambda}(x) + E\hat{\lambda}(x) - \lambda(x)|$, we have

$$3J > J_1 + J_2$$

where $J = \int E|\hat{\lambda}(x) - \lambda(x)|w(x)dx$, $J_1 = \int E|\hat{\lambda}(x) - E\hat{\lambda}(x)|w(x)dx$ and $J_2 = \int |E\hat{\lambda}(x) - \lambda(x)|w(x)dx$. From (3.5.9), it is straightforward to see that for some C_1 and C_2, $J_2(h) \ge C_1 h^2$ whenever $0 \le h \le C_2$. Further, if $h > C_2$, then $J_2(h) \ge C_3$. Hence $J_2(h) \ge C_4 \min(h^2, 1)$, for all $h > 0$. On the other hand, from Lemma 3.3 it follows that $J_1 > C_5 \min((nh)^{-1/2}, 1)$. Thus (3.2.2) of the theorem can now be verified easily.

Proof. of Theorem 3.3

Clearly, we only need to show that there is a unique solution to $\Lambda(v) = 0$ and this solution is always positive where $\Lambda(v)$ is given by equation (3.3.1).

Let $L(v) = v^{-1}\Lambda(v)$, that is

$$L(v) = \int [4b_0(x)\Phi(vb_0(x)/\sigma_0(x)) - v^{-1}\sigma_0(x)\phi(vb_0(x)/\sigma_0(x))]w(x)dx.$$

Now since

$$\begin{aligned}
L'(v) &= \int \{4b_0^2\sigma_0^{-1}\phi(vb_0/\sigma_0) + \sigma_0 v^{-2}\phi(vb_0/\sigma_0) - v^{-1}\sigma_0(\partial\phi(vb_0/\sigma_0)/\partial v)w(x)dx \\
&= \int \{4b_0^2\sigma_0^{-1}\phi(vb_0/\sigma_0) + \sigma_0 v^{-2}\phi(vb_0/\sigma_0) + b_0^2\sigma_0^{-1}\phi(vb_0/\sigma_0)\}w(x)dx \\
&= \int \{5b_0^2\sigma_0^{-1} + \sigma_0 v^{-2}\}\phi(vb_0/\sigma_0)w(x)dx > 0,
\end{aligned}$$

$L(v)$ is continuous and strictly increasing. Also, $L(0) = -\infty$, and as $v \to \infty$,

$$\begin{aligned}
L(v) &= \int [4b_0(\Phi(vb_0/\sigma_0)) - v^{-1}\sigma_0\phi(vb_0/\sigma_0)]w(x)dx \\
&\to \int 4b_0 I(b_0 > 0)w(x)dx > 0.
\end{aligned}$$

Therefore, the equation $L(v) = 0$ has a unique positive solution and so is $\Lambda(u^5) = 0$.

Proof. of Theorem 3.4

First of all, in estimating $\hat{\lambda}_1^{(p)}(x)$ and $(\hat{\lambda}_2)^{1/2}(x)$, the difference between $F(x)$ and $F_n(x)$ can be ignored as $n \to \infty$, hence we utilize $F(x)$ rather than $F_n(x)$ in the following proof.

First we will establish (3.4.1). But observe that to establish (3.4.1) it is enough to prove

$$\int |E\hat{\lambda}_1^{(p)}(x) - \lambda^{(p)}(x)|w(x)dx \to 0 \tag{3.5.12}$$

and

$$\int |\hat{\lambda}_1^{(p)}(x) - E\hat{\lambda}_1^{(p)}(x)|w(x)dx \to 0 \ a.s.. \tag{3.5.13}$$

Now we demonstrate these two separately.

(i) To prove (3.5.12) observe that

$$
\begin{aligned}
&E\hat{\lambda}_1^{(p)}(x) - \lambda^{(p)}(x) \\
&= \frac{1}{h_1^{p+1}} \int K_1^{(p)}\left(\frac{x-y}{h_1}\right) \lambda(y)dy - \lambda^{(p)}(x) \\
&= \frac{1}{h_1} \int K_1\left(\frac{x-y}{h_1}\right) \lambda^{(p)}(y)dy - \lambda^{(p)}(x) \\
&= \int K_1(z)\{\lambda^{(p)}(x-h_1z) - \lambda^{(p)}(x)\}dz
\end{aligned}
$$

and so by continuity of $\lambda^{(p)}$ and compact support of K_1,

$$
\sup_{x\in(0,T)} |E\hat{\lambda}_1^{(p)}(x) - \lambda^{(p)}(x)| \to 0.
$$

Thus (3.5.12) follows since K vanishes at both tails.

(ii) To prove (3.5.13), first we introduce Bernstein's inequality (p.17, Hoeffding[65]):

Bernstein's inequality: If $Y_1, ..., Y_n$ are independent and identically distributed with zero mean and variance σ^2, and if each $|Y_j| \le c$, then

$$
P\left(\left|\sum_{j=1}^{n} Y_j\right| > t\right) \le 2\exp\left\{-\frac{1}{2}t^2(n\sigma^2+ct)^{-1}\right\},
$$

all $t > 0$.

For any $\tau > 0$, the integral on the left-hand side of (3.5.13) is

$$
\begin{aligned}
&\int_{|x|\le\tau} |\hat{\lambda}_1^{(p)}(x) - E\hat{\lambda}_1^{(p)}(x)|w(x)dx \\
&\quad + \int_{|x|>\tau} |\hat{\lambda}_1^{(p)}(x) - \lambda^{(p)}(x) - (E\hat{\lambda}_1^{(p)}(x) - \lambda^{(p)}(x))|w(x)dx \\
&\le \int_{|x|\le\tau} |\hat{\lambda}_1^{(p)}(x) - E\hat{\lambda}_1^{(p)}(x)|w(x)dx + \int_{|x|>\tau} |\hat{\lambda}_1^{(p)}(x)|w(x)dx \\
&\quad + \int_{|x|>\tau} |\lambda^{(p)}(x)|w(x)dx + \int |E\hat{\lambda}_1^{(p)}(x) - \lambda^{(p)}(x)|w(x)dx.
\end{aligned}
$$

When τ goes to T, it is straightforward to see that the third term converges to 0 as

$\lambda^{(p)}(x)$ is truncated at $x = T$ and the fourth term converges to 0 as $\sup_{x\in(0,T)} |E\hat{\lambda}_1^{(p)}(x) - \lambda^{(p)}(x)| \to 0$. Therefore it suffices to show that for some sequence $\tau = \tau(n)$ diverging to $+\infty$,

$$\int_{|x|>\tau} |\hat{\lambda}_1^{(p)}(x)|w(x)dx \to 0 \ a.s. \tag{3.5.14}$$

and

$$\int_{|x|\le\tau} |\hat{\lambda}_1^{(p)}(x) - E\hat{\lambda}_1^{(p)}(x)|w(x)dx \to 0 \ a.s.. \tag{3.5.15}$$

Therefore proof of the theorem will be complete if we prove (3.5.14) and (3.5.15). Now to prove (3.5.14) assume that the support of K_1 is $[-s, s]$ and h_1 is so small and τ so large that $h_1 s \le \tau/2$, then the left-hand side of (3.5.14) is dominated by

$$\begin{aligned} & \frac{1}{nh_1^{p+1}} \sum_{j=1}^{n} \int_{|x|>\tau} \left| \frac{K_1^{(p)}((x - X_j)/h_1)}{1 - F_n(X_j)} \right| w(x)dx \\ \le \ & C_1(nh_1^p)^{-1} \sum_{j=1}^{n} \int_{|X_j+h_1s|>\tau, |h_1s|\le\tau/2} w(x)dx \\ \le \ & 2C_1T(nh_1^p)^{-1} \sum_{j=1}^{n} I(|X_j| > \tau/2), \end{aligned}$$

where $C_1 = \sup(|K_1^{(p)}|/(1 - F(T))$. Suppose $E(|X_1|^\gamma) < \infty$, where $\gamma > 1$ and define $\pi = P(|X_i| > \tau/2)$, by Markov's inequality, which is less than $C_2\tau^{-\gamma}$. If one takes $\tau = h_1^{-p/\beta}$, where $(2\gamma/(\gamma + 1)) < \beta < \gamma$, one has

$$\begin{aligned} E\{(nh_1^p)^{-1} \sum_{j=1}^{n} I(|X_j| > \tau/2)\} &= (nh_1^p)^{-1} \sum_{j=1}^{n} P(|X_j| > \tau/2) \\ &\le C_2 h_1^{p(\gamma-\beta)/\beta} \to 0. \end{aligned}$$

Furthermore, for each $\varepsilon > 0$, by Bernstein's inequality one has

$$\begin{aligned} q &= P\left[\left|\sum_{j=1}^{n}\left\{I(|X_j| > \frac{\mathcal{T}}{2}) - \pi\right\}\right| > \varepsilon n h_1^p\right] \\ &\leq 2\exp[-\frac{1}{2}(\varepsilon n h_1^p)^2\{n\pi(1-\pi) + \varepsilon n h_1^p\}^{-1}]. \end{aligned}$$

Observe that $\pi(1-\pi) \leq \pi \leq C_2 h_1^{p\gamma/\beta} \leq C_2 h_1^p$ and $nh_1^{2p+1} \to \infty$, then

$$q \leq 2\exp\{-C_3(\varepsilon) n h_1^p\} = o(n^{-k})$$

for all $k > 0$. Therefore, by the Borel-Cantelli lemma,

$$(nh_1^p)^{-1}\sum_{j=1}^{n} I(|X_j| > \tau/2) \to 0 \ a.s..$$

Then (3.5.14) follows.

To establish (3.5.15), set $\zeta^2(x) = \max\{\zeta_1^2(x), (1+|x|^{2\gamma})^{-1}\}$, where

$$\zeta_1^2(x) = \int K_1^{(p)}(z)^2 \frac{f(x - h_1 z)}{(1 - F(T))^2} dz,$$

and let $\mathcal{T}$ denote the set of values of $x \in (0, T)$ such that $(1 + |x|^\gamma)\zeta^2(x) > 2$. It is easy to see that the Lebesgue measure of $\mathcal{T}$, which we denote by $\mathcal{L}(\mathcal{T})$, is bounded as $(0, T)$ is bounded and $(1 + |x|^\gamma)\zeta^2(x) > 2$ if and only if $(1 + |x|^\gamma)\zeta_1^2(x) > 2$. Now we will show separately that

$$\int_{|x|\leq\tau, x\in\mathcal{T}} |\hat{\lambda}_1^{(p)} - E\hat{\lambda}_1^{(p)}| \to 0 \ a.s. \qquad (3.5.16)$$

and

$$\int_{|x|\leq\tau, x\notin\mathcal{T}} |\hat{\lambda}_1^{(p)} - E\hat{\lambda}_1^{(p)}| \to 0 \ a.s.. \qquad (3.5.17)$$

For each $\varepsilon > 0$, the left-hand side of (3.5.16) is dominated by

$$\varepsilon\mathcal{L}(\mathcal{T}) + h_1^{-(p+1)}(2\sup|K_1^{(p)}|/(1-F(T)))M_1,$$

where

$$M_1 = \int_{\mathcal{T}} I\Bigg[\Bigg|\sum_{j=1}^{n}\Bigg\{\frac{K_1^{(p)}((x-X_j)/h_1)}{1-F(X_j)} - E\Bigg(\frac{K_1^{(p)}((x-X_j)/h_1)}{1-F(X_j)}\Bigg)\Bigg\}\Bigg|$$
$$> \varepsilon nh_1^{p+1}\Bigg]dx.$$

Define $Y_j = \frac{K_1^{(p)}((x-X_j)/h_1)}{1-F(X_j)} - E\left(\frac{K_1^{(p)}((x-X_j)/h_1)}{1-F(X_j)}\right)$, $t = \varepsilon nh_1^{p+1}$ and $c = 2\sup|K_1^{(p)}|/(1-F(T))$. It is easy to know that $E(Y_j) = 0$ and

$$\begin{aligned}
\sigma_Y^2 &= E(Y_j^2) - (E(Y_j))^2 = E(Y_j^2) \\
&= E\left(\left(\frac{K_1^{(p)}((x-X_j)/h_1)}{1-F(X_j)}\right)^2\right) - \left(E\left(\frac{K_1^{(p)}((x-X_j)/h_1)}{1-F(X_j)}\right)\right)^2 \\
&\leq \int \frac{K_1^{(p)}((x-y)/h_1)^2}{1-F(y)} f(y)dy \\
&\leq h_1 \int \frac{K_1^{(p)}(z)^2}{1-F(x-hz)} f(x-hz)dz \\
&\leq h_1\zeta_1^2(x) \leq C_1h_1,
\end{aligned}$$

hence from Bernstein's inequality, one obtains that

$$\begin{aligned}
&P\left(\left|\frac{K_1^{(p)}((x-X_j)/h_1)}{1-F(X_j)} - E\left(\frac{K_1^{(p)}((x-X_j)/h_1)}{1-F(X_j)}\right)\right| > \varepsilon nh_1^{p+1}\right) \\
\leq\ & 2\exp\left\{-\frac{1}{2}\left(\varepsilon nh_1^{p+1}\right)^2\left(n\sigma_Y^2 + 2\frac{\sup|K_1^{(p)}|\varepsilon nh_1^{p+1}}{(1-F(T))}\right)^{-1}\right\} \\
\leq\ & 2\exp\{-C_2(\varepsilon)(nh_1^{p+1})^2(nh_1C_1 + C_3(\varepsilon)nh_1^{p+1})^{-1}\} \\
\leq\ & \exp\{-C_4(\varepsilon)nh_1^{2p+1}\}.
\end{aligned}$$

As $nh^{2p+1}/\log n \to \infty$,

$$\begin{aligned}E(M_1) &\leq \int_{\mathcal{T}} \exp\{-C_4(\varepsilon)nh_1^{2p+1}\}dx \\ &\leq \mathcal{L}(\mathcal{T})\exp\{-C_4(\varepsilon)nh_1^{2p+1}\} = o(n^{-k})\end{aligned}$$

for all $k > 0$. Then if one sets ε to be arbitrary small and then with the fact that $E(M_1) = o(n^{-k})$, we have $\int_{x\leq\tau, x\in\mathcal{T}} |\hat{\lambda}_1^{(p)} - E\hat{\lambda}_1^{(p)}| \to 0$ as $n \to \infty$.

To prove (3.5.17), observe that for each $\varepsilon > 0$, the left-hand side of (3.5.17) is dominated by

$$\varepsilon \int w(x)\zeta(x)^{\beta/\gamma}dx + h_1^{-(p+1)}(2\sup|K_1^{(p)}|/(1-F(T)))M_2$$

where

$$\begin{aligned}M_2 = \int_{x\leq\tau, x\notin\mathcal{T}} I\Bigg\{\Bigg|\sum_{j=1}^{n}\Bigg(&\frac{K_1^{(p)}((x-X_j)/h_1)}{1-F_n(X_j)} \\ -E&\left(\frac{K_1^{(p)}((x-X_j)/h_1)}{1-F_n(X_j)}\right)\Bigg)\Bigg| > \varepsilon nh_1^{p+1}\zeta(x)^{\beta/\gamma}\Bigg\}dx.\end{aligned}$$

M_2 is asymptotic to

$$\int_{x\leq\tau, x\notin\mathcal{T}} I\{|\sum_{j=1}^{n} Y_j| > \varepsilon nh_1^{p+1}\zeta(x)^{\beta/\gamma}\}dx$$

and Y_j is defined as in the previous paragraph. Now by Hölder's inequality,

$$\begin{aligned}&\int w(x)\zeta(x)^{\beta/\gamma}dx \leq \left\{\int w(x)\zeta(x)^2(1+|x|)^{\gamma}dx\right\}^{\beta/2\gamma} \\ &\times\left\{\int w(x)(1+|x|)^{-\gamma\beta/(2\gamma-\beta)}dx\right\}^{(2\gamma-\beta)/(2\gamma)} \\ &< \infty\end{aligned}$$

uniformly in $h_1 \leq 1$, using the fact that $E(|X_1|^{\gamma}) < \infty$ and $\gamma\beta > 2\gamma - \beta$. With the same definition of Y_j and c given in the previous paragraph, $t = \varepsilon nh_1^{p+1}\zeta(x)^{\beta/\gamma}$ and

$\sigma_Y^2 \leq h_1\zeta^2(x)$, it gives

$$\begin{aligned}
&P\left(\left|\frac{K_1^{(p)}((x-X_j)/h_1)}{1-F(X_j)} - E\left(\frac{K_1^{(p)}((x-X_j)/h_1)}{1-F(X_j)}\right)\right| > \varepsilon n h_1^{p+1}\zeta(x)^{\beta/\gamma}\right) \\
\leq\ & 2\exp\left\{-\frac{1}{2}\left(\varepsilon n h_1^{p+1}\zeta(x)^{\beta/\gamma}\right)^2\left(n\sigma_Y^2 + 2\frac{\sup|K_1^{(p)}|\varepsilon n h_1^{p+1}\zeta(x)^{\beta/\gamma}}{(1-F(T))}\right)^{-1}\right\} \\
\leq\ & 2\exp\{-C_2(\varepsilon)(nh_1^{p+1}\zeta(x)^{\beta/\gamma})^2(nh_1\zeta^2(x) + C_1(\varepsilon)nh_1^{p+1}\zeta(x)^{\beta/\gamma})^{-1}\}.
\end{aligned}$$

On one hand, suppose $h_1^p \leq \zeta(x)^{2-(\beta/\gamma)}$, then

$$\begin{aligned}
&P\left(\left|\frac{K_1^{(p)}((x-X_j)/h_1)}{1-F_n(X_j)} - E\left(\frac{K_1^{(p)}((x-X_j)/h_1)}{1-F_n(X_j)}\right)\right| > \varepsilon n h_1^{p+1}\zeta(x)^{\beta/\gamma}\right) \\
\leq\ & \exp\{-C_3(\varepsilon)nh_1^{2p+1}\zeta(x)^{2\beta/\gamma-2}\} \\
\leq\ & \exp\{-C_3(\varepsilon)nh_1^{2p+1}(1+|x|^\gamma)^{(\gamma-\beta)/\gamma}\}.
\end{aligned}$$

The second inequality holds because $x \notin \mathcal{T}$ where $\zeta(x)^{-2} > (1+|x|^\gamma)/2$. Further $(1+|x|^\gamma)^{(\gamma-\beta)/\gamma}$ is bounded.

On the other hand, suppose $h_1^p > \zeta(x)^{2-(\beta/\gamma)}$, then

$$\begin{aligned}
&P\left(\left|\frac{K_1^{(p)}((x-X_j)/h_1)}{1-F_n(X_j)} - E\left(\frac{K_1^{(p)}((x-X_j)/h_1)}{1-F_n(X_j)}\right)\right| > \varepsilon n h_1^{p+1}\zeta(x)^{\beta/\gamma}\right) \\
&\leq \exp\{-C_4(\varepsilon)nh_1^{p+1}\zeta(x)^{\beta/\gamma}\} \\
&\leq \exp\{-C_4(\varepsilon)nh_1^{p+1}(1+|x|^{2\gamma})^{-\beta/2\gamma}\} \\
&\leq \exp\{-C_5(\varepsilon)nh_1^{2p+1}\}
\end{aligned}$$

where the second inequality holds since $\zeta(x)^2 \geq (1+|x|^{2\gamma})^{-1}$ and the third holds since $0 \leq x \leq \tau = h_1^{-p/\beta}$ and h_1 converges to 0.

Hence combining these two different cases, one has that

$$E(M_2) = \int_{x\leq\tau, x\notin\mathcal{T}} \exp\{-C_6(\varepsilon)nh_1^{2p+1}\}dx = o(n^{-k})$$

for all $k > 0$.

This completes the proof of (3.5.17) and thus (3.5.15) holds.

To prove (3.4.2) of the theorem, observe that by Cauchy-Schwarz inequality

$$\begin{aligned}\int w(x)|\hat{\lambda}_2^{1/2}(x) - \lambda^{1/2}(x)|dx &\leq \int w(x)|\hat{\lambda}_2(x) - \lambda(x)|^{1/2}(x)dx \\ &\leq \left\{\int w(x)|\hat{\lambda}_2(x) - \lambda(x)|(1 + |x|^\gamma)dx\right\}^{1/2} \\ &\quad \times \left\{\int w(x)(1 + |x|^\gamma)^{-1}dx\right\}^{1/2}\end{aligned}$$

where $\gamma > 1$ is again chosen so that $E(|X|^\gamma) < \infty$. Now applying (3.4.1) with $p = 1$ we have $\int w(x)|\hat{\lambda}_2(x) - \lambda(x)|dx \to 0$ almost surely. This completes the proof of (3.4.2).

Proof. of Theorem 3.5

If (3.4.3) holds, it is straightforward to see that for any $C > 1$,

$$\sup_{C^{-1} \leq u \leq C} |\hat{\eta}(u) - \eta(u)| \to 0 \ a.s.,$$

$\hat{u}^* \to u^*$ almost surely and $\lambda(\hat{u}^*) \to \lambda(u^*)$ almost surely.

Recall that $J(\hat{h}^*) = \int \{E|\hat{\lambda}(x|h) - \lambda(x)|\}_{h=\hat{h}^*} w(x)dx$ and that is a random variable. By defining $h^* = n^{-1/5}(u^*)^2$ which is asymptotically optimal bandwidth of a standard kernel hazard rate estimate in the L_1 sense, we achieve that $\hat{h}^*/h^* \to 1$ almost surely and $J(\hat{h}^*)/J(h^*) \to 1$ almost surely, using the fact that $\hat{u}^* \to u^*$ and $\lambda(\hat{u}^*) \to \lambda(u^*)$. Note that $\inf_{h>0} J(h) \sim n^{-2/5}\eta(u^*)$, this then justifies the claim

$$J(\hat{h}^*)/\inf_{h>0} J(h) \to 1 \ a.s..$$

Proof. of Theorem 3.6

Define the σ field $\Sigma_i = \sigma(X_1, ..., X_i)$, $Y = \int |\hat{\lambda}(x) - \lambda(x)|w(x)dx - E(\int |\hat{\lambda}(x) - \lambda(x)|w(x)dx)$ and $Z_i = E(\int |\hat{\lambda}(x) - \lambda(x)|w(x)dx \parallel \Sigma_i) - E(\int |\hat{\lambda}(x) - \lambda(x)|w(x)dx \parallel \Sigma_{i-1})$. It is easy to see that Z_i is Σ_i-measurable and $Y = \sum_{i=1}^n Z_i$. Then to apply

(3.5.7) given in Lemma 3.4, we need to show that there is an upper bound on $|Z_i|$. Set $W_{i,k} = \frac{1}{n}\sum_{i\leq j\leq k}\left[\frac{K_h(x-X_j)}{1-F_n(X_i)} - \lambda\right]$ and $\tilde{W}_{i,k} = \frac{1}{n}\sum_{i\leq j\leq k}\left[\frac{K_h(x-X_j)}{1-F(X_i)} - \lambda\right]$. As $n \to \infty$, it gives

$$\begin{aligned} W_{i,k} - \tilde{W}_{i,k} &= \frac{1}{n}\sum_{i\leq j\leq k}\frac{K_h(x-X_j)(F_n(X_i)-F(X_i))}{(1-F_n(X_i))(1-F(X_i))} \\ &\leq \frac{\sup_{x\in[0,T]}|F_n(x)-F(x)|}{n}\sum_{i\leq j\leq k}\frac{K_h(x-X_j)}{(1-F_n(X_i))(1-F(X_i))} \end{aligned}$$

Using Glivenko-Cantelli theorem, we have $\sup_{x\in[0,T]}|F_n(x)-F(x)| = o(1)$, then it is easy to see that the difference between $W_{i,k}$ and $\tilde{W}_{i,k}$ is of order $o(1)$. Denote that $a = W_{1,i-1} + W_{i+1,n}$, then we have

$$\begin{aligned} |Z_i| &\leq \int |E(|\hat{\lambda}-\lambda|w(x) \parallel \Sigma_i) - E(|\hat{\lambda}-\lambda|w(x) \parallel \Sigma_{i-1})|dx \\ &= \int |E(|W_{1,i-1} + W_{i,i} + W_{i+1,n}|w(x) \parallel \Sigma_i) \\ &\qquad - E(|W_{1,i-1} + W_{i,i} + W_{i+1,n}|w(x) \parallel \Sigma_{i-1})|dx \\ &\leq \int \sup_a |E(|a + \tilde{W}_{i,i} + o(1)|w(x) \parallel \Sigma_i) \\ &\qquad - E(|a + \tilde{W}_{i,i} + o(1)|w(x) \parallel \Sigma_{i-1})|dx \\ &\leq \int \sup_a ||a + \tilde{W}_{i,i}| - E(|a + \tilde{W}_{i,i}|)|w(x)dx. \end{aligned}$$

The last inequality is due to the fact that $\tilde{W}_{i,i}$ is independent with Σ_{i-1}. With reference to Lemma 3.5 we get

$$\begin{aligned} |Z_i| &\leq \int |\tilde{W}_{i,i} - E(\tilde{W}_{i,i})|w(x)dx + \int E(|\tilde{W}_{i,i} - E(\tilde{W}_{i,i})|)w(x)dx \\ &\leq \frac{C\int|K|}{n} \end{aligned}$$

where C is some sufficient large constant. Therefore Z_i is bounded above and by Lemma

3.4, one obtains that

$$P\left(\left|\int|\hat{\lambda}(x)-\lambda(x)|w(x)dx-E\int|\hat{\lambda}(x)-\lambda(x)|w(x)dx\right|>\varepsilon\right)$$
$$\leq C_1exp(-C_2n\varepsilon^2) \qquad (3.5.18)$$

where C_1 and C_2 are large enough constants and $C_3n^{-1/2}\leq\varepsilon\leq 1$.

Let h_0 and $\hat{h}_0$ be the values of h which minimize J and $\hat{J}$ respectively. It is clear that for sufficiently large a, we have $n^{-a}\leq h\leq n^a$ for all large n and also

$$P\{n^{-a}\leq\hat{h}_0(n),\hat{h}^*\leq n^a,\ \text{all}\ n\geq\tilde{n}\}\to 1$$

as $\tilde{n}\to\infty$. Given $c>0$, define $\mathcal{H}=\mathcal{H}(a,c)=\{h_1,h_2,...\}$ to be a nonrandom sequence that is defined by $n^{-a}=h_1<h_2<...<h_{m-1}\leq n^a<h_m<...$ and $h_{i+1}-h_i=n^{-c}$, $i\geq 1$. For each $h\in\mathcal{K}=[n^{-a},n^a]$, define $H(h)$ to be a value in $\mathcal{H}$ which minimizes $|h-H(h)|$. Using Hölder continuity and compact support of K, we could let $c=c(a)$ be so large that for some $C>0$,

$$\sup_{\mathcal{K}}|\hat{J}(h)-\hat{J}(H(h))|\leq Cn^{-1}.$$

Similar, for any sample $\mathcal{X}=\{X_1,X_2,...,X_n\}$, it is easy to verify that $|J(h)-J(H(h))|\leq Cn^{-1}$ for all $h\in\mathcal{K}$. So with the setting that $\Delta=\hat{J}-J$, we achieve

$$\sup_{h\in\mathcal{K}}|\Delta(h)-\Delta(H(h))|\leq 2Cn^{-1} \qquad (3.5.19)$$

uniformly in samples $\mathcal{X}$.

Taking $\varepsilon=n^{-(1-\zeta)/2}$ where $0<\zeta<1$ in (3.5.18), it follows that, for large n,

$$\begin{aligned}P\{\sup_{1\leq j\leq m}|\Delta(h_j)|>n^{-(1-\zeta)/2}\} &\leq \sum_{j=1}^{m}P\{|\Delta(h_j)|>n^{-(1-\zeta)/2}\}\\ &\leq C_1m\exp(-C_2n^{\zeta}).\end{aligned}$$

Since $m = O(n^{a+c})$ and $\exp(-C_2 n^{\zeta}) = o(n^{-(a+c)})$ as $n \to \infty$, then

$$\sum_{n=1}^{\infty} P\{\sup_{1\leq j\leq m} |\Delta(h_j)| > n^{-(1-\zeta)/2}\} \to 0.$$

It implies (by Borel-Cantelli lemma) that

$$n^{(1-\zeta)/2} \sup_{1\leq j\leq m} |\Delta(h_j)| \to 0 \ a.s.$$

Then by applying (3.5.19),

$$n^{(1-\zeta)/2} \sup_{h\in\mathcal{K}} |\Delta(h)| \to 0 \ a.s..$$

To verify (3.4.5), we also need to show that for some constant $C_0 > 0$ and $0 < \zeta < 1$,

$$\inf_{h\in\mathcal{K}} J(h) \geq C_0 n^{-(1-\zeta)/2}. \tag{3.5.20}$$

As indicated earlier, $3J \geq J_1 + J_2$ where $J_1 = \int |\hat{\lambda}(x) - E\hat{\lambda}(x)|w(x)dx$ and $J_2 = \int |E\hat{\lambda}(x) - \lambda(x)|w(x)dx$. From Lemma 3.3, one knows that for $h \leq 1$, $J_1 \geq C\min((nh)^{-1/2}, 1)$. On the other hand, for K is bounded and $x \in (0, T)$,

$$\begin{aligned} \hat{\lambda}(x) &= \frac{1}{n}\sum_{i=1}^{n} \frac{K_h(x - X_i)}{1 - F_n(X_i)} \\ &\leq \frac{K_{max}}{(1 - F(T))h} \end{aligned}$$

where K_{max} is the upper bound of K. Hence for each $\epsilon > 0$, there exists $\tilde{h} > 0$ sufficiently large, such that $|\hat{\lambda}(x)| < \epsilon$ for all $h > \tilde{h}$, $n \geq 1$ and all sample $\mathcal{X}$. In particularly, with small enough ϵ and $h > \tilde{h}$,

$$J_2 = \int |\epsilon - \lambda(x)|w(x)dx > C_1$$

where C_1 is some constant. For $h < \tilde{h}$, then

$$|E\hat{\lambda}(x) - \lambda(x)| \le C = 2(\sup \lambda(x)) \int |K|/(1 - F(T))$$

so that

$$J_2 C \ge \int |E\hat{\lambda}(x) - \lambda(x)|^2 w(x) dx \ge C_2 h^{\tau}$$

for some constant C_2 and $\tau > 0$, where the second inequality can be easily derived from Lemma 1 of Stone[150] in density setting to hazard rate setting. Thus (3.5.20) is proved.

Consequently, $(\inf_{h>0} \hat{J}(h))/(\inf_{h>0} J(h)) \to 1$ and $\hat{J}(\hat{h}^*)/J(\hat{h}^*) \to 1$ a.s..

3.6 Discussion

In this chapter, we consider the approach to optimize a hazard rate kernel estimator by minimizing its L_1 error. In particular, we investigate the expression of the optimal L_1 error bandwidth for the kernel estimator and then develop a simple algorithm based on the Newton method to calculate an adaptive version of the bandwidth. This work can be seen as the extension of Devroye and Györfi[36] and Hall and Wand[51] in density estimation to the setting of hazard rate function. We show theoretically that both theoretical and adaptive forms of the bandwidth do minimize the L_1 distance from the true hazard rate function to the kernel estimator asymptotically. The key findings are summarized in Table 3.1. In the next chapter, we will discuss how to develop a survival model using the censored data from multiple studies.

What is already known on this topic:

- Devroye and Györfi[36] showed that in kernel estimation, L_1 error is well-defined and invariant under monotone transformation of the coordinate axes.

- Hall and Wand[51] proposed a simple, rapidly converging, iterative algorithm allowing for the minimization of L_1 distance w.r.t bandwidth h in the setting of density estimation.

What this study adds

- We propose a general asymptotic expression for the L_1 error distance between the true hazard rate function and a kernel estimator.

- We derive the theoretical asymptotic L_1 optimal bandwidth and utilize the Newton method to develop an iterative algorithm to calculate the bandwidth.

- We propose a data-driven version of the L_1 optimal bandwidth in practice and prove that the obtained estimate reaches the L_1 optimality asymptotically.

Table 3.1: Summary of the main issues and key findings in Chapter 3.

CHAPTER 4

EXAMINING THE OVERALL PROGNOSIS OF BREAST CANCER PATIENTS USING IPD FROM MULTIPLE COUNTRIES

4.1 Introduction

Chapter 2 and 3 focused on mathematical research for hazard rate estimation in survival models when event times are known for all observations. However, in practice censored event times will be an issue. We now concentrate on applications of survival models for breast cancer in this situation.

In both developing and developed countries, breast cancer is the leading cause of oncological death amongst women. Coleman[22] reported that the breast cancer patients account for 20% or more of all cancers in total and of every 12 women, one will develop the disease before 75 years old. Therefore, breast cancer is a major threat to women's health, and thus much time and money are devoted to investigating ways to improve outcomes in those with breast cancer.

The current research has showed the survival chances of cancer patients not only depend on their own health conditions but also are closely related to the environment they live[92][100]. According to current statistics by Ferlay et al.[43], it is known that the cancer patients observed in Europe accounts for 25% of total cases found in the world despite the population of Europe being only one-ninth of the global population. These findings motivate investigations into country-specific survival probabilities of patients with primary breast cancer within Europe[13][127]. In particular, which countries have a better breast cancer prognosis than others? This chapter investigates this issue by using a large, multi-country database collected by the European Organization for Research and Treatment of Cancer-Receptor and Biomarker Group. This was provided by Maxime Look (Josephine Nefkens Institute, Rotterdam) and the outcome of our interest is the time to the death of a patient following their diagnosis of breast cancer, and the aim here is to examine if the mortality rate in different European counties is the same overtime.

The dataset will be introduced in Section 4.2, and we will introduce key modelling concepts and objectives in Section 4.3. The necessary methodologies for the three objectives are illustrated in Section 4.4. Then the analysis and results on the breast cancer data are summarized in Section 4.5. Throughout this section, the advantages of Royston-Parmar models and multiple imputation techniques will be also highlighted. In the end, we conclude with a summary of the key findings in Section 4.6.

4.2 Data

The breast cancer incidence and mortality data in this study covers 15 laboratories and the quality-assurance programs are utilized in all these participating laboratories to measure the biologic variables in tumor tissue approved by the local reviews boards. It incorporates 7435 selected patients who have accepted the primary surgeries for breast cancer ranging from September 1978 to December 1995 and then being followed-up for 10 years. The corresponding inclusion and exclusion criteria of the patients were detailed introduced in Look et al.[87].

4.2.1 Available variables

We group 15 laboratories into 8 countries and define it as the categorical factor, 'country', containing: Netherland, Slovenia, Switzerland, France, Ireland, Austria, Sweden and Denmark. See the list of the laboratories of each country in Table 4.4. The laboratory in Rotterdam, Netherland, contributed the largest amount of data samples accounting for

Var	**Type**	**Description**
Age	Continuous	Age in years of patients
Upa	Continuous	Urokinase-type plasminogen activator antigen level determined by immunoassays
Pai1	Continuous	Plasminogen activator inhibitor antigen level determined by immunoassays
Rupa	Continuous	Fractional rank of upa within study (between 0 and 1)
Rpai1	Continuous	Fractional rank of pai1 within study (between 0 and 1)
Tumor type	Categorical	Tumor type of patient including idc, ilc, col, tubul, medull, papil, other, unknown
Tumor grade	Categorical	Tumour grade of patient including good, moderate, poor, unknown
Lymph nodes status	Categorical	Number of positive nodes involved: np= 0, np< 3, 3< np <10, np> 10
Menopasual status	Categorical	Menopasual status of patient including: premenopausal and postmenopausal age
Tumor size	Categorical	Tumor size: pT1 $\leq$ 2cm, 2cm< pT2 <5cm, pT34 > 5cm
Adjuvant treatment	Categorical	Indicator whether adjuvant systemic therapy is given
Hormone receptor	Categorical	Steroid hormone receptor status
Country	Categorical	Country of study
OSi	Categorical	Indicator whether the death is recorded or censored
OS	Continuous	Time from primary surgery to death or end of follow-up

Table 4.1: Variables in the breast cancer dataset with brief description. The cutoff points for the variable, lymph node status and tmuor size were chosen by Look et al.[87].

37% of the whole database, while the laboratory in Ljubljana, Slovenia provided the least patients ($n = 69$). Patients in each laboratory were followed up from the date of primary surgery until death (from any cause) and this outcome was defined as 'OS' for overall survival. Patients who survived till the end of their observation period were censored for the analysis of overall survival, and this was assumed to be right, non-informative censoring.

Table 4.1 lists all the variables recorded in the database with descriptions. It incorporates the following traditional confounding factors as age, tumor size, tumor type, lymph node status, hormone-receptor status, tumour grade and menopasual status where only age was available as a continuous variable. Tumor size was provided as an ordinate variable by two cutoff points, 2 centimeters and 5 centimeters. Lymph node status was the number of involved lymph nodes provided as 0, 1-3, 4-10, and more than 10. Low hormone-receptor status was defined as either estrogen receptor status or progesterone receptor was low; high hormone-receptor status was defined as at least one of estrogen receptor or progesterone receptor stratus was high. Tumour grade was treated as four categories, well differentiated (grade I), moderately differentiated (grade II), undifferentiated (grade III), and unknown[34][47]. It was also of concern that whether a patient received adjuvant systemic treatment and hence an indicator variable was set to denote whether the patient received this[50].

Further, as pointed by Look et al.[87], the use of above traditional confounding factors may not be sufficient to model the mortality rate of primary breast cancer. Thus, two additional body measurements, Urokinase-type plasmmogen activator (upa) and its inhibit (pai1) were also available where high level of upa or pai1 is closely related to poor prognosis in patients[6][45][60]. Due to the fact that the levels of upa and pai1 in different laboratories were not measured within the same scale, it was difficult to compare them directly across studies[152]. To address this issue, within each individual study, Look et al. provided rupa and rpai1 which denoted the fractional ranks of upa and pai1 respectively. That is, within one study, upa and pai1 were rescaled into fractional ranks (between 0 and 1) via dividing the two ranked variables by the number of patients of the study. In this way, the fractional ranks with the same scale could be comparable across different studies.

4.2.2 Descriptive statistics

The descriptive statistics of the data are now summarized. The median follow-up of patients alive ranged from 52 months to 120 months in different countries. The average age of Austria, France and Switzerland were more than 57 years old while Ireland was only 51.539. There were 40% of the patients who were premenopasual. The 45% of the patients had small tumors (pT1), while 48% had middle tumors (pT2) and others had large tumors. The tumour grade of 37% patients was unknown, 56% were lymph node negative and 43% received systemic adjuvant treatment. As for the outcome, 27% of patients died within 10 years.

Variable	No. of Obs	Mean	s.d.	median	Min	Max
Age						
Netherland	3242	56.777	13.366	56.000	22.000	90.000
Ireland	184	51.539	11.887	50.524	30.051	85.988
Sweden	914	54.365	11.748	51.717	23.000	89.000
Slovenia	69	55.275	10.923	55.485	31.797	76.172
Austria	88	57.638	13.140	56.070	30.229	89.588
France	1509	57.625	11.778	58.000	24.000	85.142
Switzerland	663	58.190	11.402	57.100	23.780	85.160
Denmark	766	54.851	10.857	55.000	28.000	80.000
Total	7435	56.447	12.461	56.000	22.000	90.000
Upa						
Netherland	3227	1.087	1.273	0.730	0.000	24.400
Ireland	184	0.564	0.920	0.355	0.000	10.200
Sweden	914	0.452	0.456	0.310	0.000	3.190
Slovenia	69	0.384	0.281	0.340	0.070	1.370
Austria	88	5.243	4.030	4.450	0.750	18.840
France	1470	0.714	0.789	0.440	0.000	7.220
Switzerland	653	1.004	0.848	0.730	0.000	5.040
Denmark	762	3.369	5.890	0.632	0.000	60.641
Total	7367	1.192	2.355	0.570	0.000	60.641
Pai1						
Netherland	3236	18.316	25.502	12.330	0.000	479.38
Ireland	184	2.270	3.841	0.581	0.000	20.900
Sweden	226	1.184	2.268	0.700	0.010	23.000
Slovenia	69	8.934	10.275	6.200	0.270	75.910
Austria	88	3.781	3.653	2.715	0.270	28.150
France	687	5.716	5.594	4.060	0.000	46.430
Switzerland	485	7.414	5.905	5.600	0.100	47.380
Denmark	766	1.970	4.089	1.027	0.000	68.925
Total	5741	12.182	20.706	6.250	0.000	479.380
Survival time						
Netherland	3242	75.178	32.850	78.193	1.051	120.000
Ireland	184	65.558	33.470	64.624	2.333	120.000
Sweden	914	93.958	34.790	120.000	4.172	120.000
Slovenia	69	51.253	11.243	54.735	8.903	63.671
Austria	88	55.923	26.814	52.073	1.478	120.000
France	1509	72.042	29.694	70.275	1.084	120.000
Switzerland	663	45.139	15.880	44.353	1.708	85.224
Denmark	766	64.07	28.035	63.622	1.281	120.000
Total	7435	72.339	32.878	69.651	1.051	120.000

Table 4.2: Summary of baseline characteristics of continuous variables by countries (s.d.=standard deviation).

Variable	**Ned**	**Irl**	**Swe**	**Slo**	**Aut**	**Fra**	**Sui**	**Den**
Survival status								
alive	2058	126	770	54	59	1235	587	503
deceased	1184	58	144	15	29	274	76	263
Tumour type								
idc	2136			61	60	1009	470	635
ilc	234			6	12	81	88	76
colloid	3			1		16		
tubul	21				1	3	21	4
medull	58				5	5	28	26
papil	13			1	3	2	2	
other	155				7	55	53	21
unknown	622	184	914			338	1	4
Tumour grade								
good	122		71	5	29	234	37	
moderate	654		81	22	41	776	183	
poor	1613		64	40	16	385	300	
unknown	853	184	698	2	2	114	143	766
Lymph nodes								
nodenegative	1656	87	491	25	36	1015	369	451
1-3	863	49	256	21	20	317	176	184
$> 3 - 10$	524	40	135	15	21	140	69	103
> 10	199	8	32	8	11	37	49	28
Menopausal status								
pre	1311	123	482	26	28	485	220	324
post	1931	61	432	43	60	1024	443	442
Tumour size								
pT1	1396	64	422	18	29	801	308	275
pT2	1533	91	472	42	51	676	307	424
pT3, pT4	313	29	20	9	8	32	48	67
Adjuvant treatment								
no or unknown	2353	44	411	10	22	787	137	442
yes	889	140	503	59	66	722	526	324
Hormone receptor								
0 or unknown	588	77	249	42	22	300	56	182
at least one	2654	107	665	27	66	1209	607	584
Total	3242	184	914	69	88	1509	663	766

Table 4.3: Summary of baseline characteristics of category variables by country.

Country	**Laboratory**	**Total**	**Missing upa**	**Missing pai1**
Netherland	Rotterdam	2722	12	1
	Nijmegen	321	3	5
	Utrecht	199		
Ireland	Ireland	184		
Sweden	Sweden1	688		688
	Sweden2	226		
Slovenia	Ljubljana	69		
France	Fr-lia	554		554
	Paris	188	39	
	StCloud	499		
	Lille	268		268
Austria	Oost	88		
Switzerland	Switzerland	663	10	178
Denmark	Denmark12	444	4	
	Denmark3	332		

Table 4.4: Summary of laboratories of each country and the corresponding missing values in upa and pai1.

Table 4.2 gives the basic characteristics of patients and their follow-up information (e.g. number censored or dead), plus descriptive statistics for continuous factors, and Table 4.3 summarizes other variables. Of the available variables, only upa and pai1 had missing values, which are summarized in Table 4.4 by laboratory and country.

4.3 Objectives

Three objectives are set up in this study.

4.3.1 Objective 1: Obtain a survival curve for each country and overall

The first objective of the analysis is to investigate the hazard rate of primary breast cancer within each country and then on the entire database respectively. Flexible parametric survival models via Royston-Parmar scheme are used, with all potential confounding factors ignored. Although no confounding factors are included in this step, the pure overall prognosis is still very useful for each country[58]. It will not only provide us an intuitive way to compare the precision accuracy of parametric survival models and Royston-Parmar models visually in each study but also to compare the shape of the hazard rates curves amongst the different countries.

4.3.2 Objective 2: Country-specific survival curve using prognostic model

The second objective is to formally compare the country-specific survival probabilities of patients. In other words, is country a prognostic factor? To quantify this, Royston-Parmar approaches are implemented where 'country' is included into the model to denote where patients come from. Two possible models are available in this case:

Unadjusted model

A straightforward idea is to incorporate 'country' as the only covariate in the Royston-Parmar model. However, the weakness of this unadjusted model is quite obvious: since mortality rates of breast cancer depend on confounding factors, comparisons between countries are confounded in the absence of confounding factors. Ignoring confounding can lead to a biased estimate of the true association between the outcome risks of breast cancer and 'country'.

Adjusted model

An alternative method is to let the Royston-Parmar model adjust for all the significant confounding variables in estimation and prediction. In the breast cancer dataset, the candidate confounding factors such as age, tumour type, tumour grade, number of lymph nodes, menopausal status, tumour size, adjuvant treatment and hormone receptor status listed in Table 4.1 could be added to the model directly as the covariates[20]. In comparison to common Cox models, adjusted average survival curves can be produced following a Royston-Parmar model (see Section 4.4.2).

4.3.3 Objective 3: Comparison of countries after including upa and pai1

One problem existing in the database is that subgroups of two confounding variables in our database (upa, pai1) were missing. Additionally, the measurement methods of upa and pari1 were not uniform across laboratories.

In Objective 2, the two variables, upa and pai1 were not considered as confounding factors. However, it was not a perfect solution since upa and pai1 play important roles in health measurements of patients with primary breast cancer[6][60][131]. Then the deletion of the two variables may lead to a waste of costly collected data and further, the omission of the key covariates in the regression may cause bigger unexplained variation in the final model[107]. As it has raised concerns to keep all information of the underlying data, in our study, an imputation approach by Buuren et al.[160] is proposed to estimate the missing values in the data sample.

4.4 Methods

To meet the objectives, a pre-specified statistical analysis plan was written and this is now summarised.

4.4.1 Methods for Objective 1

Three main statistical methods, parametric, nonparametric and Royston-Parmar models were used to fit mortality rates of patients on the entire database.

Initially, the estimators (hazard rate function $h(t)$, cumulative hazard rate function $H(t)$ and survival function $S(t)$) using three parametric models, Weibull, log-normal and log-logistic, were produced in three panels respectively. Royston-Parmar survival models were then fitted to the same data. To visualize the difference between the conventional parametric method and Royston-Parmar model in the same scale, we developed Royston-Parmar proportional hazards, proportional odds and probit models (seen in Section 1.6) and plotted the estimated curves against their parametric counterparts in each figure.

To consider country-specific hazards, the above procedure was repeated for each country separately. In this study, two questions may incur our interests, that is, whether Royston-Parmar approaches perform better than parametric counterparts and whether the hazard rate functions do vary across 8 European countries judging from the shapes of corresponding survival curves.

4.4.2 Methods for Objective 2

The second goal of this chapter is to compare country-specific mortality rate of breast cancer from 8 different European countries using Royston-Parmar regression models.

The hazard ratios, reflecting the difference in risks between two countries, were firstly estimated using an unadjusted Royston-Parmar proportional hazard model[2][144]. Netherland was selected to be the reference level which accounts for the biggest population of the total database. An alternative multivariable Royston-Parmar proportional hazards model is then provided to incorporate all the potential confounding variables such as age, tumour type, menopasual status, lymph node status, tumor size, adjuvant systemic treatment, tumor grade and hormone-receptor status.

Using the estimates from the multivariable model, adjusted survival curves could be predicted and plotted by calculating the population-averaged survival curve for each country[120]. This is done by predicting the survival curve for each individual in the dataset, using their own covariate values but assuming they were from a specified country. All these individual survival curves were then averaged to give the population-averaged survival curve for that country. For example, to estimate the mean survival curve in jth country, the log cumulative hazard scale for the ith sample can be expressed as

$$\ln(H_i(t)) = \ln(H_0(t)) + \hat{\beta}_{0j} + \underline{\hat{\beta}}^T \underline{X}_i$$

where $\hat{\beta}_{0j}$ is the coefficient estimate of jth country (0 for Netherland) and the covariate vector $\underline{X}$ represent all the confounding factors given above. Then the mean survival curve at time t for jth country could be written as

$$N^{-1}\sum_{i=1}^{N}\exp(-\exp(\ln(H_0(t))+\hat{\beta}_{0j}+\underline{\hat{\beta}}^T\underline{X}_i)).$$

Plotting over many t values gives the summary survival curve.

This adjusted model also allows to predict the mean absolute survival probabilities between countries at fixed time points, after adjusting for the other variables in the model[120].

Assessing the proportional hazards assumption

In Royston Parmar proportional hazards model, it usually assumes that the hazard ratio of a variable remains constant over time. If it is violated, it means that the hazard ratio of this variable changes over time and thus the time dependent effect should be taken into consideration[83].

There are a number of methods to test proportionality[59][74][132]. As guidance, the Cox models are fitted for the data from each country separately. If the proportional hazards assumption holds, then the graph of the survival function for each involved country versus the survival time should results in a graph with parallel curves, similarly the graph of the log(-log(survival)) versus log of survival time should result in parallel lines as well ('log-log' plot). To further adjust for any confounding variables, we could add the associated covariates to the Cox model w.r.t each country and then the estimated survival curves plotted in the graph are determined by fixing the value of each covariate being its average.

4.4.3 Methods for Objective 3

Multiple imputation technique was implemented to estimate the missing values for upa and pai1 (rupa and rpai1) in the database. Multiple imputation is a statistical technique for handling missing data, which is increasingly popular due to its generality and efficiency[117][118] (see Appendix A). The key idea of multiple imputation is to use the observed data to estimate the missing data and also take the uncertainly of estimation into consideration. In particular, m imputations of the data, each of which has the missing values properly estimated are produced and then treated equally to achieve m copies of parameter estimates. Afterwards, the point estimates, variance and confidence intervals are computed by averaging the m estimates and computing its standard errors using Rubin's Rule[125] (see Appendix A).

There is debate over using the original scale or ranked scale of upa and pai1 in clinical researches. As upa and pai1 were not measured by the same extraction and assay methods, the measurements were not comparable across the studies[87]. Arguably, the fractional ranks, rupa and rpai1 could be applied to rescale the measurements in each laboratory between 0 and 1. However, rupa and rpai1 forced the difference between any two neighbour entries of upa or pai1 to be equal within one study which was not a realistic setting. To be cautious, we proposed to run both multivariate models employing the original scale and ranked scale of upa and pai1 respectively and then looked into the difference between these two models with the different settings.

The multiple imputation procedures to reconstruct the two variables (in ranked scale or original scale) can be schemed as follows,

Multiple imputation for the model using rupa and rpai1

Within laboratory imputation: For laboratories in Rotterdam, Nijmegen, Paris and Switzerland, only a small proportion of upa and pai1 were not recorded. We suggested to estimate these missing values within each laboratory separately using multiple imputation technique. As a general rule, using all available information yields multiple imputations that have minimal bias and maximal certainty[160], hence we included all the confounding variables into this model such as age, tumour type, menopasual status, lymph node status, tumor size, adjuvant systemic treatment, tumor grade and steroid hormone-receptor status. Further, the survival or censoring time, OS and its censoring indicator, OSi were also added to the imputation procedure.

After each imputation, the observed and imputed upa and pai1 of each laboratory were pooled together and ranked again to produce the updated complete rupa and rpai1 respectively.

Between laboratory imputation: Unfortunately, for the laboratories, Swe1, Fr-lia and Lille where no pai1 was reported, the within laboratory imputation could not be applied directly since multiple imputation technique cannot estimate an variable without any known observations[117]. Then we could to make use of the observed dataset from Paris and StCloud to estimate the missing pai1 in Fr-lia and Lille because all the 4 laboratories are in France and should have similar baseline characteristics. In the same way, the values of rpai1 in Swe1 can be imputed in use of the observations from Swe2 since both of the two laboratories belong to Sweden.

In particular, for Fr-lia or Lille, firstly we combined their samples with Paris and StCloud together to form a big dataset and then estimated the missing values of rpai1 using known rupa and rpai1 as well as the other confounding variables. After each imputation, the estimated rpai1 in Fr-lia or Lille needed to be adjusted additionally to follow the standard setting of the fractional rank. By analogy, we pooled the data from Swe1 and Swe2 and then apply the within imputation method to estimate pai1 in Swe1 with pai1 from Swe2 and the other covaraites. Finally, rpai1 in Swe1 was computed based on the imputed pai1.

Multiple imputation for the model using upa and pai1

Within laboratory imputation: For laboratories in Rotterdam, Nijmegen, Paris and Switzerland, similar to the case of rank scale, the missing values of upa and pai1 within each laboratory were estimated by the known upa, pai1 and associated covariates separately using multiple imputation technique.

Between laboratory imputation: In analogy to estimating rpai1, we combined the samples from Fr-lia and Lille with those in Paris and StCloud to form a big dataset and then imputed the missing values of pai1 in Fr-lia and Lille using known upa and pai1. As for Swe1, missing values in pai1 were estimated using the combined data from Swe1 and Swe2.

Combining dataset for regression

No matter which scale we applied in multiple imputation to estimate upa and pai1, after each imputation across all the studies, we append the imputed datasets with the studies without missing values to form a complete imputed data set for the entire sample. Consequently, m copies of the whole database with imputed upa and pai1 or rupa and rpai1 could be obtained after m imputations.

Once the multiple imputed dataset was generated, each imputed data set was analyzed separately by using Royston-Parmar regression. For each model, country, age, tumour type, menopasual status, Lymph node status, tumor size, adjuvant systemic treatment, tumor grade, steriod hormone-receptor status, upa and pai1 (original or ranked scale) were included as the covariates. Finally the estimates of coefficients of the covariates were averaged across the m copies and the standard error of any estimate was computed based on the 'Rubin rule'[125].

4.5 Results

4.5.1 Results for Objective 1

We first determine the number of knots required for the Royston-Parmar model in our research. Then we show the parametric estimators (Weibull, log-normal, log-logistic) and their Royston-Parmar generalized estimators for the whole database and each individual country as follows.

Number of knots for Royston-Parmar model

When the Royston-Parmar approach is applied to fit the underlying data, we need to determine the knots required in estimating the baseline hazard function of the model. It is also recognized as the problem to decide the degrees of freedom (d.f.) for the Royston-Parmar model where the degrees of freedom equal to the number of knots minus 1.

Figure 4.1 illustrates the baseline hazard functions on the proportional hazards scale for the whole database using different degrees of freedom. We find that it makes little difference in estimation when 3 knot was used (d.f.=2) compared to as many as 7 knots (d.f.=6) for the outcomes of mortality[120]. Further with reference to AIC and BIC, we find that adding the degree of freedom to the model cannot result in any big reduction in AIC and BIC. On the contrary, AIC and BIC reached the minimum at d.f.=3 and d.f.=2 respectively. Therefore it was determined that 3 degrees of freedom (4 knots) would be adequate in this case. In the rest of this chapter, we do not discuss the choice of knots for the Royston-Parmar model further but based on our study, d.f.=2 or 3 is always sufficient to model the underlying breast cancer dataset.

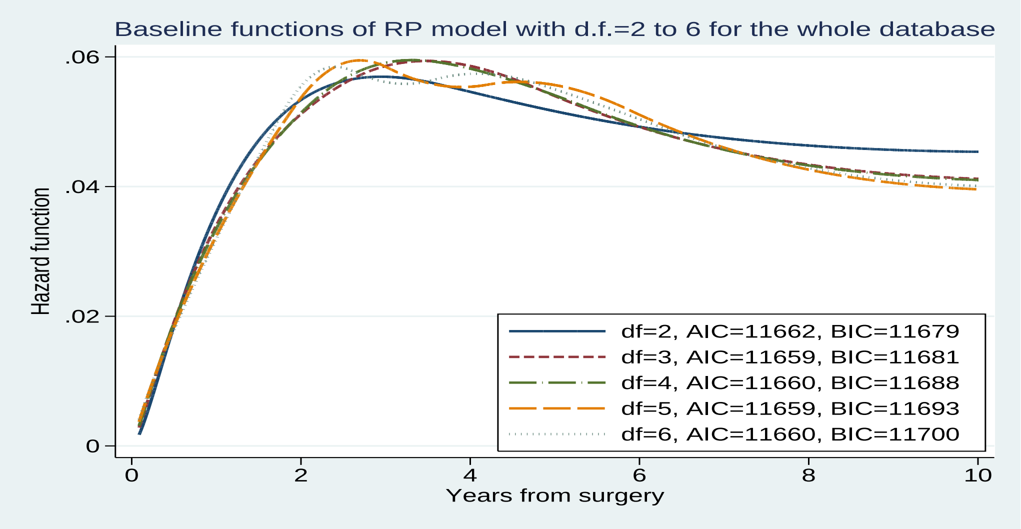

Figure 4.1: Baseline hazard functions after fitting the Royston-Parmar model on proportional hazards scale with d.f.=2 to 6 respectively.

Weibull distribution

The parametric estimator based on the Weibull parametric survival model and the Royston-Parmar proportional hazards model generalized from the Weibull distribution were plotted against each other in Figure 4.2. The parametric hazard rate estimator performed poorly that failed to capture the drop of the true curve around 4 years. Comparably, Royston-Parmar model performed very well except for a bit overestimating of the true risk for the patients who had survived between 2.2 and 5 years. This exemplifies why Royston-Parmar models are more flexible and needed.

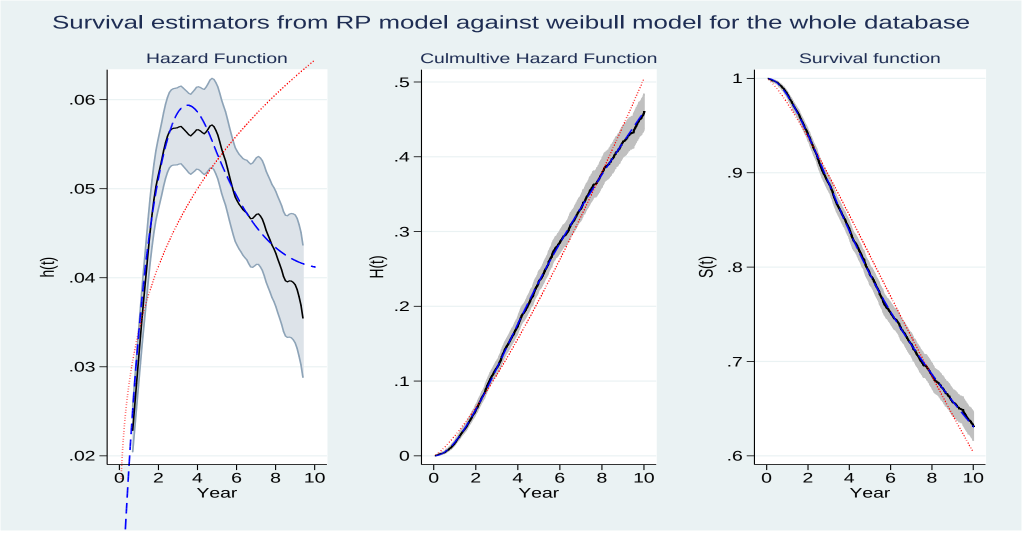

Figure 4.2: The entire data: Nonparametric estimator (solid) with pointwise confidence band in gray, parametric estimator (dot), Royston-Parmar estimator (dash). Left panel: Kernel estimator of hazard rate function compared with estimates from Royston-Parmar model and Weibull model. Middle panel: same comparison for the cumulative hazard rate function. Right panel: same comparison for the survival function.

Log-logistic distribution

The parametric log-logistic estimators for hazard rates were plotted in Figure 4.3. It is clear that the parametric estimator was again very poor in comparison to its nonparametric counterpart with 95% confidence band. Conversely, the difference between the Royston-Parmar proportional odds estimator and nonparametric estimator was much smaller implying the Royston-Parmar model is superior to its parametric counterpart.

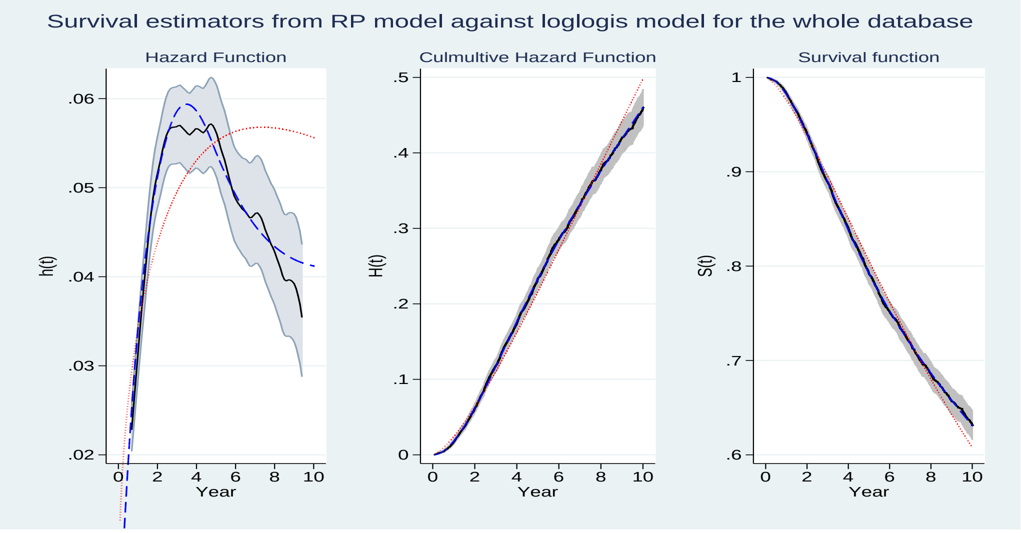

Figure 4.3: The entire data: Nonparametric estimator (solid) with pointwise confidence band in gray, parametric estimator (dot), Royston-Parmar estimator (dash). First panel: Kernel estimator of hazard rate function compared with estimates from Royston-Parmar model and log-logistic model. Middle panel: same comparison for the cumulative hazard rate function. Right panel: same comparison for the survival function.

Log-normal distribution

From Figure 4.4, it can be seen that the log-normal hazard rate estimator might capture the turning point (around $t = 4$) of the hazard curve of the underlying risk, however the Royston-Parmar probit estimator generalized from log-normal distribution still outperformed in estimation with the reference to the observed true curves (nonparametric estimators).

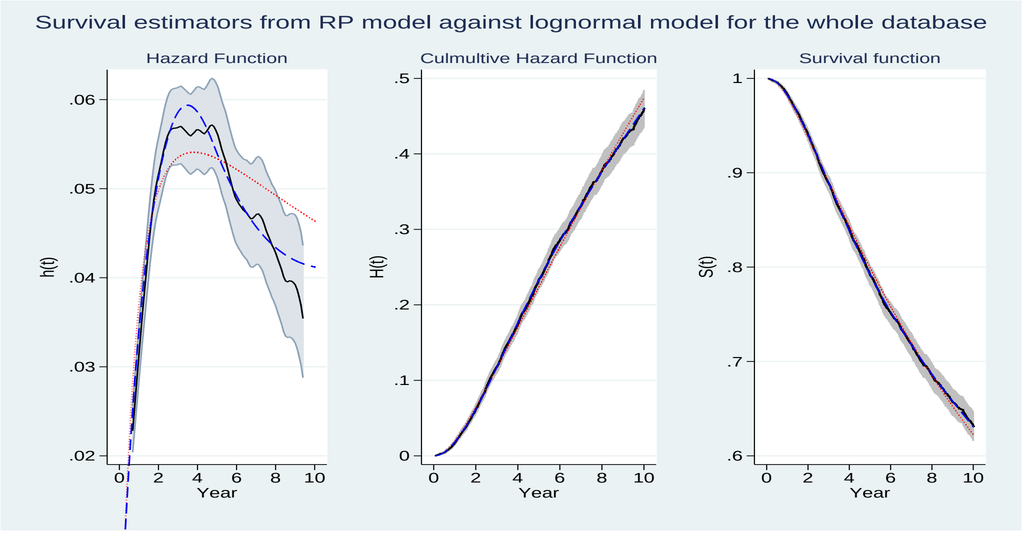

Figure 4.4: The entire data: Nonparametric estimator (solid) with pointwise confidence band in gray, parametric estimator (dot), Royston-Parmar estimator (dash). First panel: Kernel estimator of hazard rate function compared with estimates from Royston-Parmar model and log-normal model. Middle panel: same comparison for the cumulative hazard rate function. Right panel: same comparison for the survival function.

The same procedures were repeated for each European country separately. Figure 4.5 to 4.7 show the hazard rate estimators of the eight countries within the scale of Weibull, log-logistic and log-normal distributions respectively.

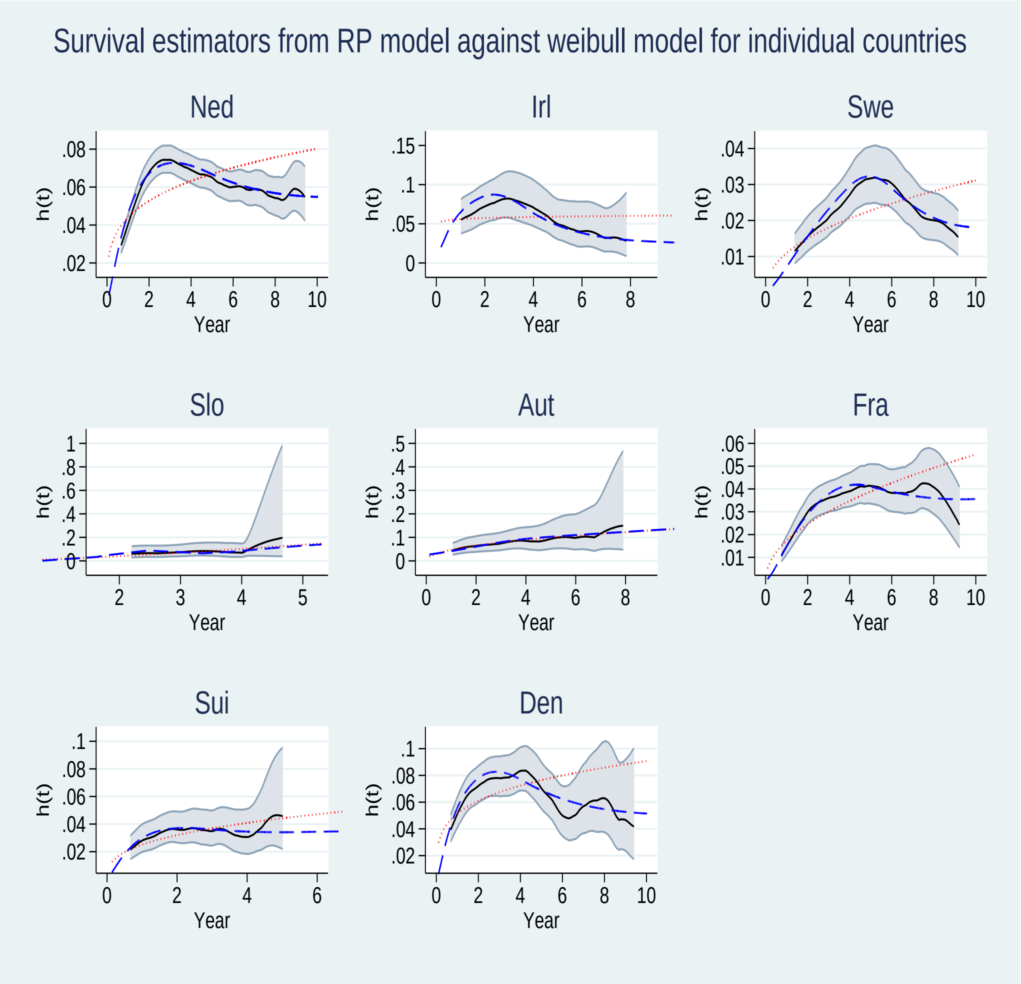

Figure 4.5: The kernel estimates of hazard rate function compared with the estimates from flexible parametric models with d.f.=3 and Weibull models for 8 countries, Netherland, Slovenia, Switzerland, France, Ireland, Austria, Sweden and Denmark respectively: nonparametric estimator (solid) with pointwise confidence band in gray, parametric estimator (dot), Royston-Parmar estimator (dash).

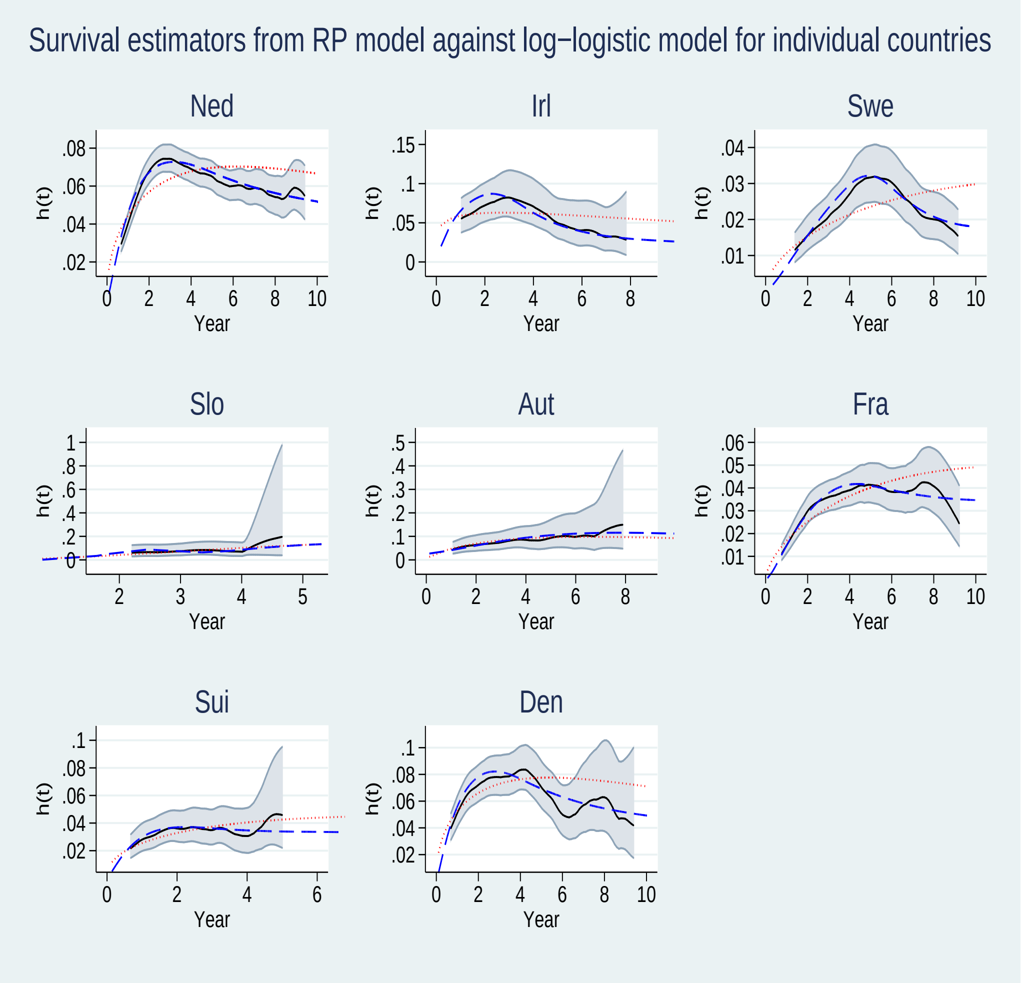

Figure 4.6: The kernel estimates of hazard rate function compared with the estimates from flexible parametric models with d.f.=3 and log-logistic models for 8 countries, Netherland, Slovenia, Switzerland, France, Ireland, Austria, Sweden and Denmark respectively: nonparametric estimator (solid) with pointwise confidence band in gray, parametric estimator (dot), Royston-Parmar estimator (dash).

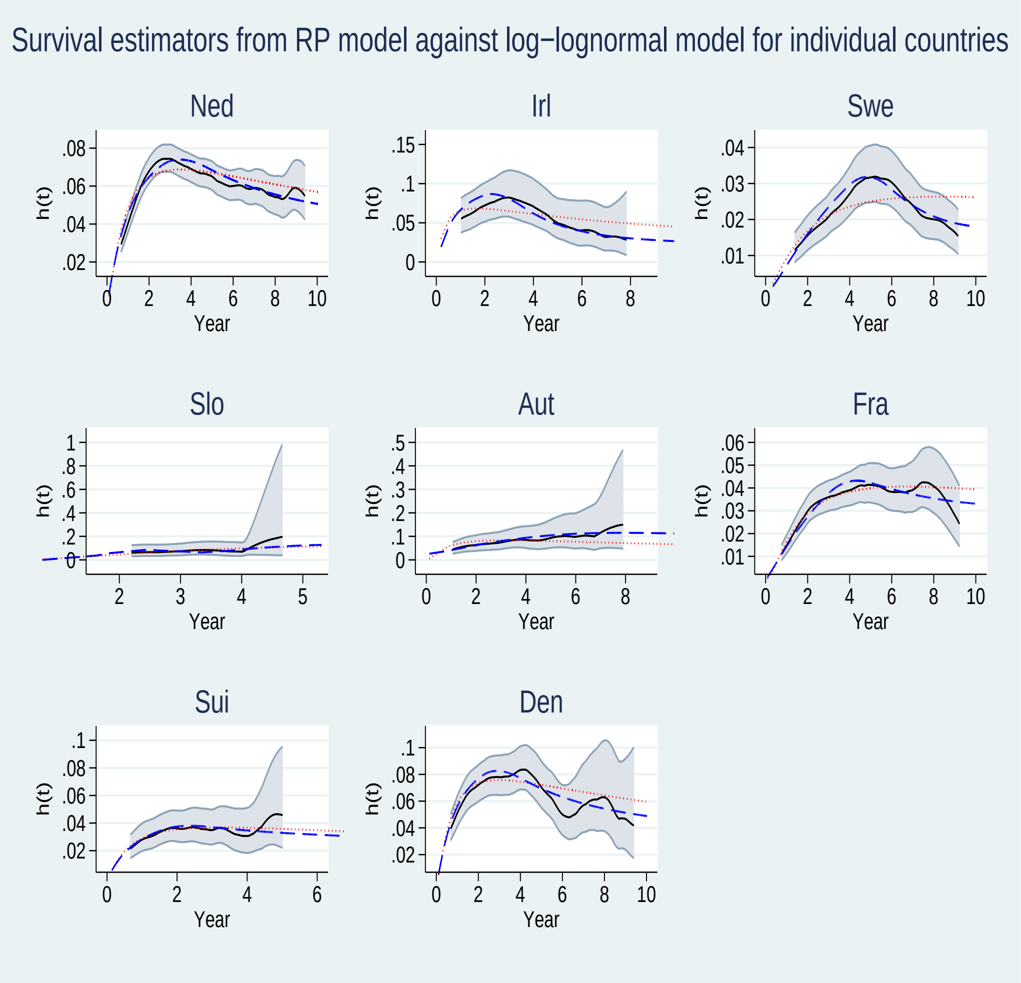

Figure 4.7: The kernel estimates of hazard rate function compared with the estimates from flexible parametric models with d.f.=3 knots and log-normal models for 8 countries, Netherland, Slovenia, Switzerland, France, Ireland, Austria, Sweden and Denmark respectively: nonparametric estimator (solid) with pointwise confidence band in gray, parametric estimator (dot), Royston-Parmar estimator (dash).

Two important conclusions can be summarized from these figures. Firstly is that, in terms of smooth kernel estimators with 95% confidence band, Royston-Parmar methods provide much better fits than their parametric counterparts. Regardless of the parametric start of Royston-Parmar estimators, no visual difference can be found amongst the three Royston-Parmar estimators, suggesting that the choice of the scale appear not crucial, at least in our study.

Secondly, the fitted curves across 8 countries appear noticeably different from each other. The shapes of hazard rate estimators were of the same type for Netherland, Ireland, Sweden, France and Denmark where the curves went up in the beginning and then dropped down, although there was some disagreement over the time at which the hazard function reached its highest point. However the mortality rates of the other three counties, Slovenia, Austria and Switzerland, had totally different trends, where the hazard rates gradually increased as time went. This fact implies that the hazard rate curves of individual risks across 8 European countries might not be same at all, and thus we should look into the potential country-level effects and whether differences were due to confounders.

4.5.2 Results for Objective 2

Unadjusted survival analysis

The unadjusted result of the Royston Parmar proportional hazards model is summarized in Table 4.5 where Netherland is set as the reference country and proportional hazards assumed. The patients in Sweden have the highest probabilities to survive amongst 8 countries where its hazard ratio versus Netherland is as low as 0.341 (CI=(0.286, 0.405)) and the *p*-values of log hazard ratio estimators for Ireland, Slovenia, Austria and Denmark are more than 0.05 which suggest that the unadjusted survival probabilities of patients in these four countries are not significantly different from those in Netherland. The hazard ratios of France and Switzerland are 0.520 (CI=(0.456, 0.593)) and 0.562 (CI=(0.445, 0.711)) indicating that unadjusted hazard rates of mortality are 48.0% and 43.8% lower than Netherland respectively.

Variable	**HR**	**s.e. of ln(HR)**	*P* **Value**	**95% CI for HR**
Netherland	.	.	.	.
Ireland	1.006	0.135	0.966	(0.773, 1.309)
Sweden	0.341	0.030	< 0.001	(0.286, 0.405)
Slovenia	0.908	0.236	0.710	(0.545, 1.512)
Austria	1.255	0.236	0.227	(0.868, 1.815)
France	0.520	0.035	< 0.001	(0.456, 0.593)
Switzerland	0.562	0.067	< 0.001	(0.445, 0.711)
Denmark	1.118	0.076	0.103	(0.978, 1.278)

Table 4.5: Unadjusted estimates for mortality from Royston-Parmar proportional hazards model.

The unadjusted survival curves for 8 countries are shown in Figure 4.8 which cluster into 3 groups. The patients in Netherland, Ireland, Slovenia, Austria and Denmark have highest mortality rates, the patients in France and Switzerland are second, and Sweden has the lowest risk over time.

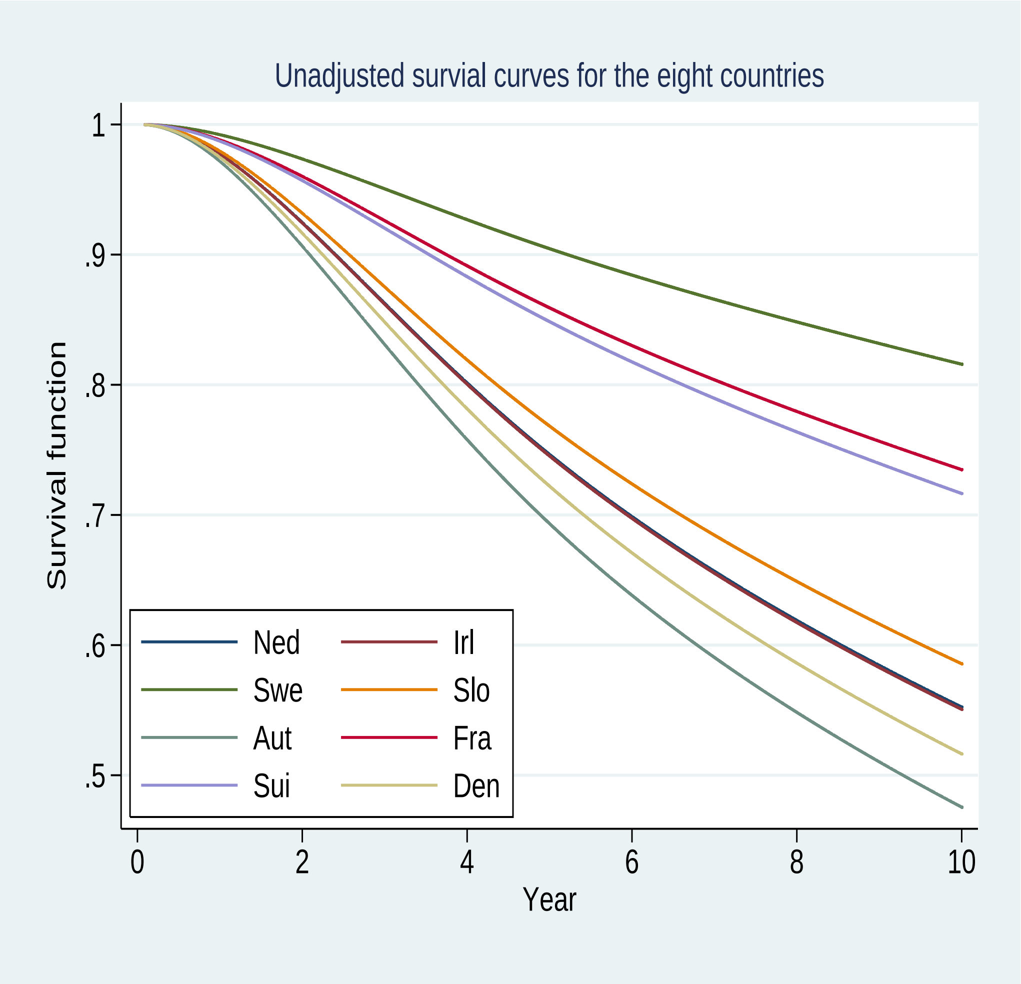

Figure 4.8: Unadjusted estimates of mortality for 8 included countries using the Royston-Parmar proportional hazards model.

Adjusted survival analysis

Do differences remain after we adjust for confounding factors? To adjust for baseline confounding, besides ‘country’, the variables age, tumour type, tumour grade, lymph nodes, menopausal status, tumour size, adjuvant treatment and hormone receptor status were also included into the model.

The estimates of the full model are summarized in Table 4.6. In comparison to the unadjusted case, the adjusted hazard ratios for France and Denmark substantially increased by approximately 20%. Sweden was still in the best position of all 8 countries with the hazard ratio equaling to 0.344 (CI=(0.272, 0.435)). The hazard ratio for Denmark increased dramatically from 1.118 (CI=(0.978, 1.278)) to 1.425 (CI=(1.203, 1.689)) which implied that the unadjusted model underestimated the hazard ratio of Denmark versus Netherland. The hazard ratio of Slovenia dropped from 0.908 (CI=(0.545, 1.512)) to 0.622 (CI=(0.371, 1.045)), but the other countries's risk stayed at a similar level as the unadjusted model.

Other estimates from this model suggest that a one-year increase in age increased the hazard ratio of death by 1%. Among all types of tumors, only 'medull' type was statistically significant (HR=0.462, CI=(0.299, 0.714)). For the tumour grade, the prognosis of patients in good status and moderate status were better than those in unknown and poor status. The hazard ratio for number of lymph nodes and tumor size gradually went up with increasing levels. The variable, menopasual status was not statistically significant in the model. The hazard rate of death for the patients with adjusted treatment were 10.9% lower than those with conventional treatment and the risk of death for the patients with at least one of estrogen receptor or progesterone receptor status being high was 43% lower than their counterparts.

The population-averaged survival curves for the 8 countries are plotted in Figure 4.9. It shows that Denmark dropped to the bottom of 8 countries while France and Switzerland are now in the 3rd and 4th position.

Variable	**HR**	**s.e. of ln(HR)**	*P* **Value**	**95% CI for HR**
Age	1.011	0.003	< 0.001	(1.006, 1.017)
Tumour type	Reference level: idc			
ilc	0.908	0.087	0.313	(0.753, 1.095)
col	0.745	0.530	0.679	(0.185, 3.001)
tubul	0.591	0.297	0.294	(0.221, 1.580)
medull	0.462	0.103	0.001	(0.299, 0.714)
papil	0.662	0.335	0.414	(0.245, 1.784)
other	0.929	0.121	0.573	(0.719, 1.200)
unknown	1.024	0.068	0.716	(0.900, 1.166)
Tumour grade	Reference level: good			
moderate	1.524	0.242	0.008	(1.116, 2.082)
poor	2.179	0.341	< 0.001	(1.604, 2.962)
unknown	1.832	0.299	< 0.001	(1.331, 2.521)
Tumour size	Reference level: np= 0			
np< 3	2.030	0.131	< 0.001	(1.788, 2.305)
$3 < np < 10$	3.490	0.233	< 0.001	(3.062, 3.978)
np> 10	5.494	0.459	< 0.001	(4.663, 6.472)
Menopasual status	0.944	0.071	0.442	(0.814, 1.094)
Tumour size	1.383	0.051	< 0.001	(1.287, 1.487)
Adjuvant treatment	0.891	0.026	< 0.001	(0.842, 0.943)
Hormone receptor	0.565	0.029	< 0.001	(0.511, 0.626)
Country	Reference level: Netherland			
Ireland	0.954	0.152	0.769	(0.698, 1.305)
Sweden	0.344	0.041	< 0.001	(0.272, 0.435)
Slovenia	0.622	0.165	0.073	(0.371, 1.045)
Austria	1.360	0.268	0.119	(0.924, 2.000)
France	0.744	0.057	< 0.001	(0.641, 0.864)
Switzerland	0.756	0.095	0.025	(0.591, 0.966)
Denmark	1.425	0.123	< 0.001	(1.203, 1.689)

Table 4.6: Adjusted estimates for mortality from Royston-Parmar proportional hazards regression.

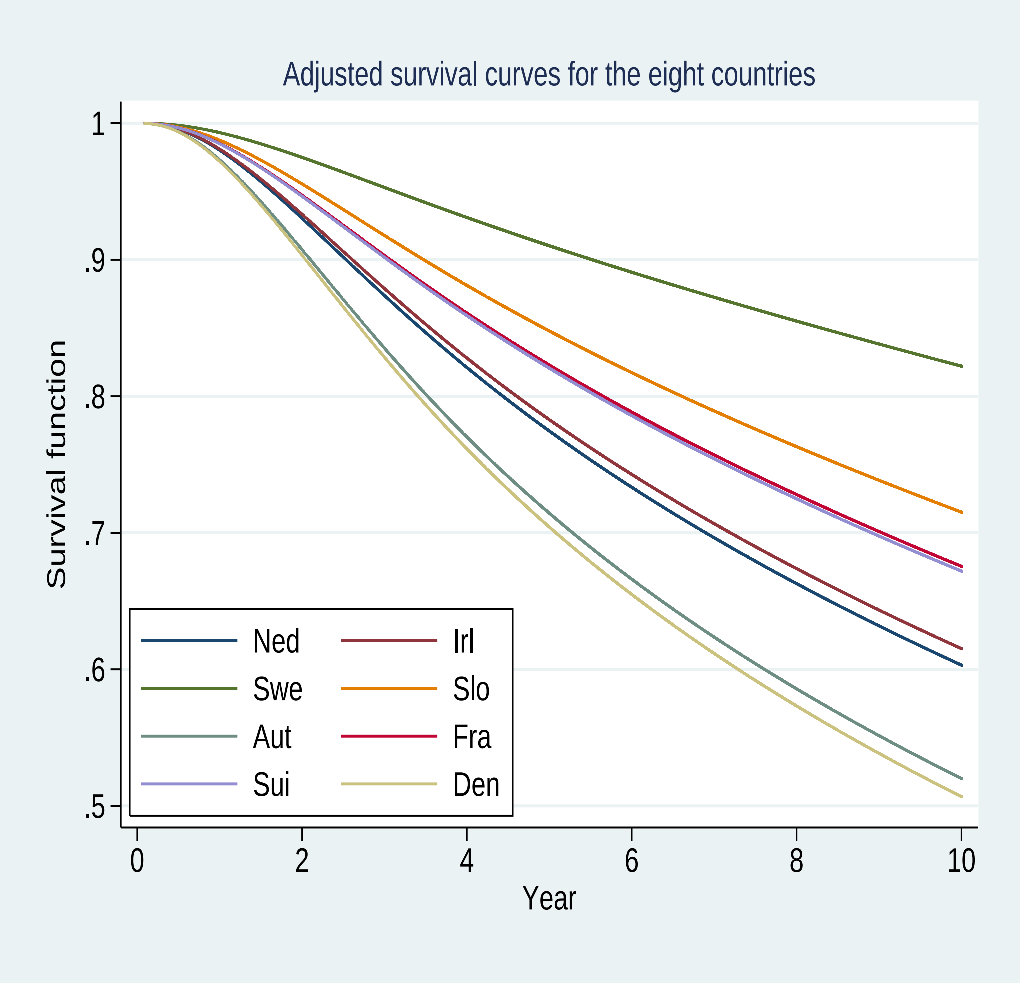

Figure 4.9: Population-averaged estimates of mortality for 8 included countries using the Royston-Parmar proportional hazards model adjusting for the potential confounding factors.

The absolute differences in predicted mean survival probabilities for each country at 1, 2, 3, 5 and 8 years following surgery are listed in Table 4.7. There is an increasing trend in the maximum difference in predicted mean survival probabilities which are 0.021 (CI=(0.021, 0.022)) for 1-year period, 0.072 (CI=(0.070,0.073)) for 2-year period, 0.125 (CI=(0.123 ,0,127)) for 3-year period, 0.206 (CI=(0.204, 0.208)) for 5-year period and 0.282 (CI=(0.280, 0.284)) for 8-year period.

Proportional hazards assumption

By visual inspection of 'log-log' plots, the proportional hazards assumption was assessed for all the possible outcomes of 'country' (Figure 4.10) accounting for confounders. Parallel curves suggest that the proportional hazards assumption appears reasonable in our

Period	Country	Mean	95% CI for HR
Year 1	Netherland	0.980	(0.980, 0.981)
	Ireland	0.981	(0.981, 0.981)
	Sweden	0.993	(0.993, 0.993)
	Slovenia	0.988	(0.987, 0.988)
	Austria	0.973	(0.973, 0.974)
	France	0.985	(0.985, 0.985)
	Switzerland	0.985	(0.985, 0.985)
	Denmark	0.972	(0.971, 0.972)
Year 2	Netherland	0.930	(0.929, 0.932)
	Ireland	0.933	(0.932, 0.935)
	Sweden	0.975	(0.974, 0.975)
	Slovenia	0.955	(0.955, 0.956)
	Austria	0.907	(0.906, 0.909)
	France	0.947	(0.946, 0.948)
	Switzerland	0.946	(0.945, 0.948)
	Denmark	0.903	(0.901, 0.905)
Year 3	Netherland	0.874	(0.871, 0.876)
	Ireland	0.879	(0.877, 0.881)
	Sweden	0.953	(0.952, 0.954)
	Slovenia	0.918	(0.916, 0.919)
	Austria	0.835	(0.832, 0.838)
	France	0.903	(0.901, 0.905)
	Switzerland	0.902	(0.900, 0.904)
	Denmark	0.828	(0.825, 0.831)
Year 5	Netherland	0.774	(0.771, 0.778)
	Ireland	0.783	(0.779, 0.786)
	Sweden	0.910	(0.908, 0.912)
	Slovenia	0.848	(0.845, 0.850)
	Austria	0.714	(0.710, 0.718)
	France	0.823	(0.820, 0.826)
	Switzerland	0.821	(0.818, 0.824)
	Denmark	0.704	(0.700, 0.708)
Year 8	Netherland	0.663	(0.658, 0.667)
	Ireland	0.674	(0.669, 0.678)
	Sweden	0.855	(0.852, 0.858)
	Slovenia	0.763	(0.759, 0.767)
	Austria	0.586	(0.581, 0.591)
	France	0.728	(0.724, 0.732)
	Switzerland	0.725	(0.721, 0.729)
	Denmark	0.573	(0.568, 0.578)

Table 4.7: Mean survival probabilities for eight countries at Year 1, 2, 3, 5 and 8 using the adjusted model.

research.

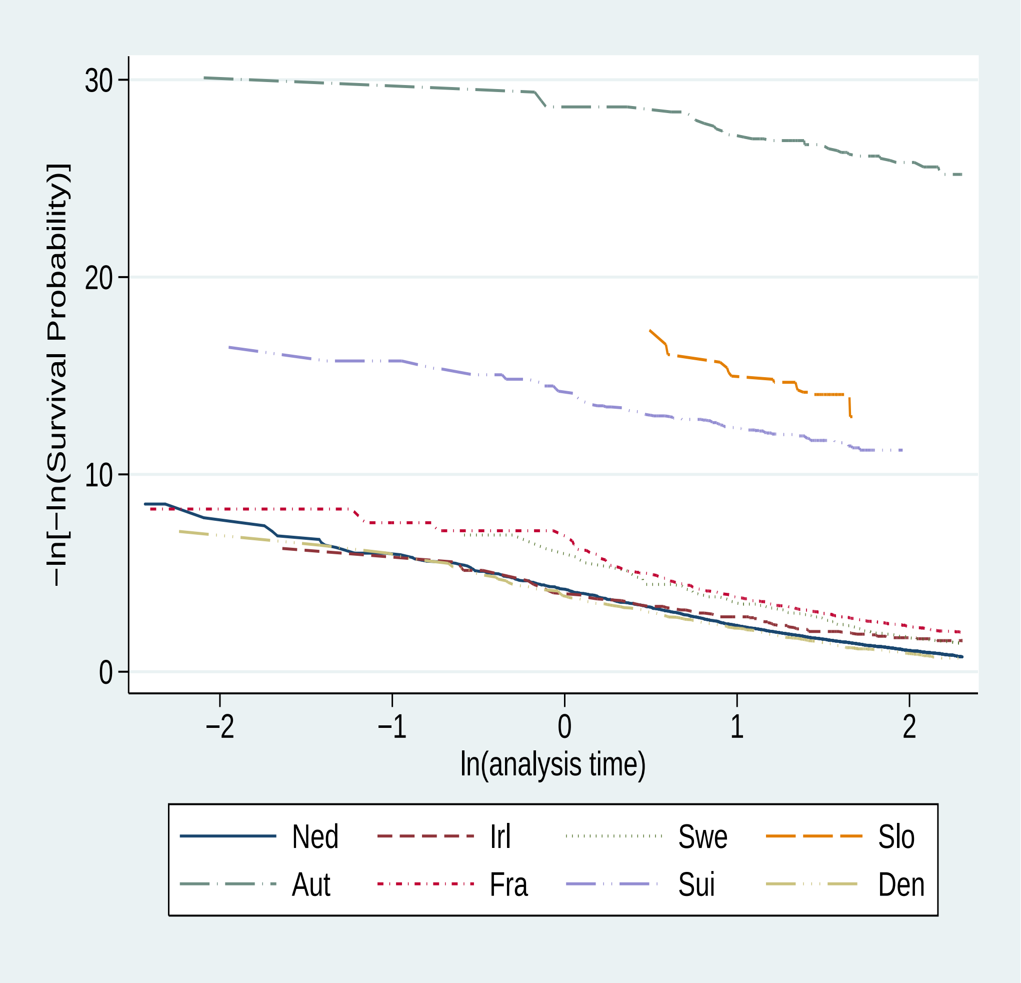

Figure 4.10: 'Log-log' plot for the outcome of mortality for 8 countries, Netherland, Slovenia, Switzerland, France, Ireland, Austria, Sweden and Denmark adjusted for all the related confounding factors.

4.5.3 Results for Objective 3

According to the choices of the scale for upa and pai1, two copies of the imputed datasets were constructed with the multiple imputation technique.

Estimation after including multiple imputation of rupa and rpai1

With the 10 imputations of rupa and rpai1, each imputed dataset was fitted separately where rupa and rpai1 entered the Royston-Parmar model as confounding factors. The final parameters estimates and standard errors of the model are summarized in Table 4.8.

Variable	**HR**	**s.e. of ln(HR)**	*P* **Value**	**95% CI for HR**
Age	1.011	0.003	< 0.001	(1.005, 1.017)
Rupa	1.644	0.101	< 0.001	(1.447, 1.841)
Rpai1	2.080	0.114	< 0.001	(1.856, 2.304)
Tumour type	Reference level: idc			
ilc	1.091	0.096	0.184	(0.902, 1.280)
col	0.691	0.711	0.302	(-0.703, 2.086)
tubul	0.616	0.502	0.167	(-0.368, 1.600)
medull	0.466	0.222	< 0.001	(0.031, 0.901)
papil	0.666	0.508	0.212	(-0.330, 1.662)
other	0.976	0.131	0.428	(0.720, 1.233)
unknown	1.026	0.066	0.346	(0.897, 1.156)
Tumour grade	Reference level: good			
moderate	1.387	0.160	0.020	(1.074, 1.701)
poor	1.974	0.157	< 0.001	(1.666, 2.282)
unknown	1.680	0.163	0.001	(1.360, 2.000)
Tumour size	Reference level: np= 0			
np< 3	1.990	0.065	< 0.001	(1.863, 2.117)
3 <np< 10	3.487	0.066	< 0.001	(3.357, 3.617)
np> 10	5.445	0.083	< 0.001	(5.283, 5.608)
Menopasual status	0.945	0.075	0.225	(0.797, 1.092)
Tumour size	1.389	0.037	< 0.001	(1.316, 1.463)
Adjuvant treatment	0.896	0.029	< 0.001	(0.839, 0.952)
Hormone receptor	0.625	0.053	< 0.001	(0.521, 0.728)
Country	Reference level: Netherland			
Ireland	0.978	0.160	0.443	(0.664, 1.291)
Sweden	0.343	0.120	< 0.001	(0.109, 0.578)
Slovenia	0.678	0.264	0.071	(0.160, 1.197)
Austria	1.307	0.197	0.087	(0.921, 1.694)
France	0.736	0.076	< 0.001	(0.588, 0.884)
Switzerland	0.724	0.125	0.005	(0.479, 0.970)
Denmark	1.417	0.087	< 0.001	(1.247, 1.587)

Table 4.8: Royston-Parmar model parameter estimates after multiple imputation of rupa and rpai1.

The population-averaged survival curves after this multiple imputation are shown in Figure 4.11 with the Royston-Parmar models now also including rupa and rpai1. The estimates of the hazard ratios of each country have not changed too much in comparison to the previous adjusted model, and thus the ordering of countries stays the same as before.

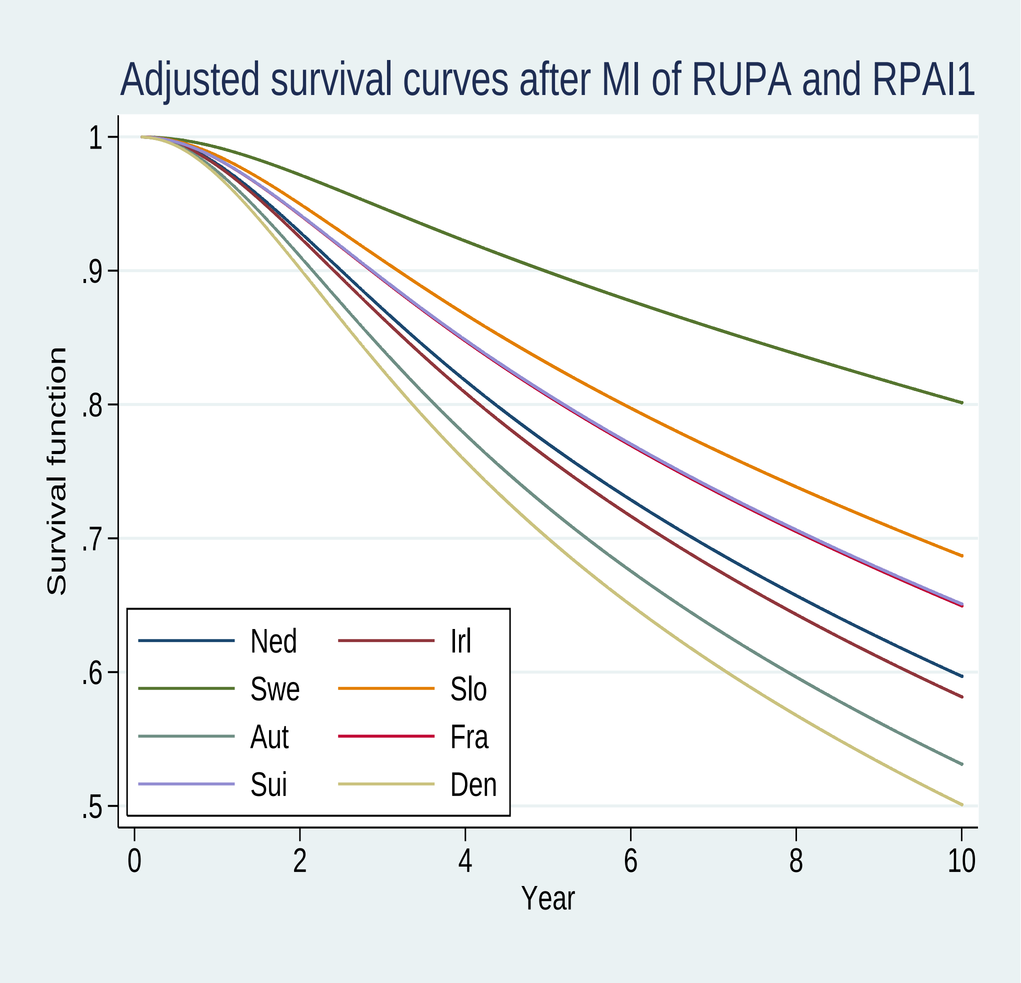

Figure 4.11: Population-averaged estimates of mortality for 8 included countries using the adjusted Royston-Parmar proportional hazards model after including the imputed variables, rupa and rpai1.

Estimation after including multiple imputation of upa and pai1

Multiple imputation in the previous section was repeated on the original values of up-a and pai1. The obtained hazard ratio estimates are summarized in Table 4.9 and the population-averaged survival curves of 8 countries are plotted in Figure 4.12 after additionally adjusting for upa and pai1. Again no dramatic changes in the hazard ratios of countries could be found compared with the adjusted model in the last section. Hence it implies that the inclusion or exclusion of upa and pai1 did not influence the association between the risks of patients and the place they lived.

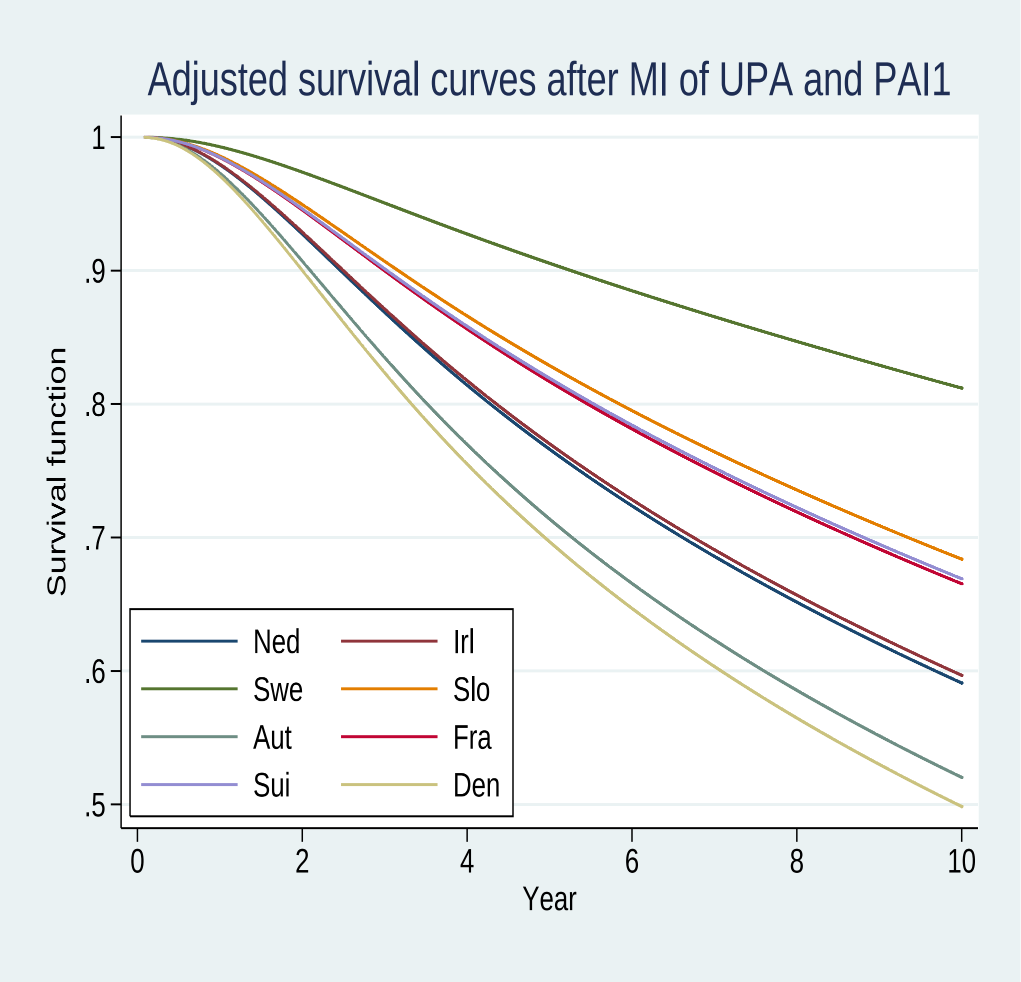

Figure 4.12: Population-averaged estimates of mortality for 8 included countries using the adjusted Royston-Parmar proportional hazards model after including the imputed variables, upa and pai1.

4.6 Discussion

This study of the EORTC-RBG data applied a newly proposed flexible parametric model approach (Royston-Parmar modelling) to estimate and compare the mortality rates of patients with primary breast cancer from 8 countries.

Variable	**HR**	**s.e. of ln(HR)**	P **Value**	**95% CI for HR**
Age	1.012	0.001	< 0.001	(1.010, 1.014)
Upa	1.030	0.002	< 0.001	(1.025, 1.034)
Pai1	1.005	< 0.001	< 0.001	(1.004, 1.005)
Tumour type	Reference level: idc			
ilc	0.946	0.029	0.072	(0.891, 1.005)
col	0.845	0.190	0.456	(0.544, 1.314)
tubul	0.612	0.097	0.002	(0.449, 0.836)
medull	0.479	0.034	< 0.001	(0.417, 0.549)
papil	0.694	0.111	0.023	(0.507, 0.951)
other	0.872	0.038	0.002	(0.800, 0.951)
unknown	1.050	0.022	0.020	(1.008, 1.094)
Tumour grade	Reference level: good			
moderate	1.474	0.077	< 0.001	(1.331, 1.632)
poor	2.077	0.106	< 0.001	(1.878, 2.296)
unknown	1.771	0.094	< 0.001	(1.595, 1.965)
Tumour size	Reference level: np= 0			
np< 3	2.120	0.044	< 0.001	(2.035, 2.209)
3 <np< 10	3.684	0.079	< 0.001	(3.532, 3.842)
np> 10	5.663	0.152	< 0.001	(5.374, 5.969)
Menopasual status	0.925	0.022	0.001	(0.883, 0.970)
Tumour size	1.373	0.016	< 0.001	(1.342, 1.406)
Adjuvant treatment	0.879	0.008	< 0.001	(0.863, 0.895)
Hormone receptor	0.570	0.010	< 0.001	(0.551, 0.589)
Country	Reference level: Netherland			
Ireland	1.040	0.053	0.436	(0.942, 1.149)
Sweden	0.372	0.014	< 0.001	(0.345, 0.401)
Slovenia	0.694	0.058	< 0.001	(0.589, 0.818)
Austria	1.288	0.082	< 0.001	(1.137, 1.459)
France	0.737	0.019	< 0.001	(0.700, 0.776)
Switzerland	0.823	0.033	< 0.001	(0.762, 0.890)
Denmark	1.446	0.042	< 0.001	(1.366, 1.530)

Table 4.9: Royston-Parmar model parameter estimates after multiple imputation of upa and pai1.

4.6.1 Key findings

In the first objective, the hazard rate function was estimated in each country separately to show the average (overall) prognosis in their populations[58]. The Royston-Parmar models fitted the observed data well regardless of the chosen scale whereas standard parametric models (such as Weibull) did not. The curves of Austria, Switzerland and Slovenia were monotone, others were unimodal.

In the second objective of the chapter, initially only the geographical factor, 'country' was included into Royston-Parmar regression (unadjusted model) as the explanatory variable. By setting Netherland to be the reference level, the hazard ratio of the other 7 countries ranged from 0.34 (Sweden) to 1.25 (Austria). After adjusting for age, tumour type, tumour grade, number of lymph nodes, menopausal status, tumour size, adjuvant treatment and hormone receptor status, the hazard ratios in France and Switzerland increased by approximately 0.2 and the hazard of death in Slovenia now dropped by 0.28. Sweden had the best mortality rates but Denmark had the worst.

The third goal of this chapter was to impute the missing values in upa and pai1 and then add them to the Royston-Parmar model. It was found that the inclusion or exclusion of upa and pai1 brought no significant influence to the estimates of hazard ratios of 'country'. This is because the association between mortality rates of breast cancer and 'country' did not depend on upa and pai1.

Beside the clinical achievements, the study also highlights the efficiency of multiple imputation technique in practice. The old-fashioned approach was to replace missing values with the mean or mode of the nonmissing values for that variable. However that approach is now thought to be insufficient since no randomness is considered in estimation[169]. Multiple imputation proposed by Burren[160] merged the proper degree of randomness into the imputed values and also considered the uncertainty when computing standard errors and confidence intervals for parameters of interests. Using this property, we are able to calculate the mean predicted survival probabilities of patients in the imputed models at any given time[117].

Statistically, the study also reflects the outstanding performance of Royston-Parmar model in fitting procedures, and more advantages are summarized as follows: Firstly, unlike Cox-regression, the baseline hazard functions can be estimated in this model[116]. Secondly, population-averaged ('adjusted') survival curves can be plotted to graphically show survival functions for groups of patients. Further from the population-averaged survival functions, the differences in absolute $S(t)$ can be calculated, after adjusting for covariates in the model[120].

4.6.2 Limitations

One of the major limitations in our work is that in Royston-Parmar regression, the effect of a covariate may wane with time, as opposed to being a constant multiplicative effect[83]. It was shown that the proportional assumption held for the effects of 'country', however, we might find that the time-dependent effects still exist in other predictors. This could form further research.

Except for the pitfalls of statistical modellings, our dataset also have clinical limitations. The ordering of countries according to their mortality rate should be interpreted with caution, as other unmeasured factors may still be at work. Firstly, our database could be affected by unknown confounding factors such as the inconsistent diagnosis time amongst different countreis[14]. For example, the survival duration from one country may be over measured when it could detect the disease of patients earlier than others (lead time bias); In other cases, one country may wrongly include the observations that would not progress to be overt or very slowly into the research and subsequently overestimate the overall survival time of patients (overdiagnosis bias)[15]. Secondly, it is noticed that we categorized the dataset by its registry nation. But for some participated countries (*i.e.* Slovenia: 1 study with 69 observations), the sample size and events were too small to represent the characteristics of the whole population[54][55]. Thirdly, our dataset is now old, as it was collected up to 1995. Finally, regardless of the reason of patients' death, they were all counted in the database, which may lead to a upwardly biased mortality rate estimator of the true risk of death due to breast cancer alone[87].

4.6.3 Link to other research

Recently, thanks to the new development of medical statistical methodologies and well constructed databases on cancer patients in Europe, the studies at regional or international levels have provided sufficient evidence to show the significant geographical effects on the survivals probabilities of breast cancer patients[92][100][102]. .

For example, Parkin et al.[100] summarized the incidence, mortality, and prevalence of dying from cancer in each national population and concluded that the risks of cancers are highest in Eastern Europe. From recently-published data in 2012, Ferlay et al.[43] estimated incidence and mortality estimates for the 40 countries in Europe and achieved the similar conclusion that Easter Europeans were mostly likely to have cancer problems. In particular, McPherson et al.[92] illustrated the geographical factor in the analysis of breast cancer and indicated that environmental factors are of greater importance than genetic factors of human beings. Further, their findings were echoed with our work that Danes has the lowest survival probabilities from cancers within Europe. A report from a more comprehensive project, namely as EUROCARE which collected cancer survival data from 45 population-based cancer registries in 17 European countries, was given by Sant et al.[126]. In this report, it revealed wide international differences in cancer survival where survival was generally highest in Northern Europe, followed by Western Europe, Denmark and the UK, and the Eastern European countries of Estonia and Poland.

The reasons behind the international differences are likely to be multiple. Some studies attributed them to artifact factors, for example, Sant et al.[128] concluded that some registries that only used linkage with death certificates to establish vital status of patients (died or alive) might overestimate survival because of linkage failures. Sant et al.[127] used multiple regression models to assess the influence of the stage of diagnosis on survivals of patients and stated that, longer survival could be simply due to early diagnosis without any advantage to the patient (lead-time bias) and the regional survival differences should therefore diminish if appropriate stage-adjusted comparisons are performed. The model proposed in our research has not adjusted for the stage of diagnosis, but it could be easily generalized if we know the diagnostic time for patients in different countries. Also, Vercelli[163] casted doubt on the large geographic variations in relative survival rates among European countries. They stated that the improvement of survivals from breast cancer were surprisingly big that might not be due to the real prognosis, but rather to a selection bias. For example, in elderly patients with a very bad prognosis, who are often suffering from other serious co-morbid conditions, cancer diagnoses could be under-notified and not reach at all the data sources commonly monitored by cancer registries.

On the contrary, Coleman et al.[23] opposed the above viewpoints. He listed several evidence suggesting that many of the observed differences in breast cancer survival between countries, regions and population subgroups are systematic, and can be largely attributed to differences in access to health services, including delay in presentation and diagnosis, and the overall quality of care. This argument was also supported by Engeland et al.[41] in the cancer survival in Denmark, the country with the lowest survivals rates in our study, because the cancer was usually found and treated too late in Denmark compared than other Nordic countries. In the discussion of geographical variations associated with breast cancer by Bray et al.[18], the differences in survival by stage at diagnosis were again marked as the important factor and the authors thought international comparisons of disease rates by area can provide important clues to the underlying causes of diseases, the effects of natural or planned interventions, and serve as indicators of the scope for preventive strategies. Further, they also believed that the studies of migrants provided the solid evidence that environmental (rather than genetic) determinants are responsible for most of the observed international and inter-ethnic differences in breast cancer incidence.

4.6.4 Conclusion

The key findings and what this chapter adds are shown in Table 4.10. The next chapter extends this work by developing a prognostic model using the same data.

What is already known on this topic:

- Breast cancer is a global threat to women's health which accounts for 20% or more of all cancers in the world.

- One conventional way to estimate the time-to-event dataset is Cox regression. But a more flexible parametric regression was also proposed via Royston-Paramar scheme.

- When we develop a multivariable model to investigate the association between the outcome of events and the covariate of our interests, it is important for us to adjust for all the confounding factors to avoid the potential bias.

- To deal with the missing values in the dataset, a multiple imputation method is proposed by Buuren et al.[160] to re-estimate the missing data using the observed data.

What this study adds

- We utilized Royston-Parmar proportional hazards, proportional odds and probit models to estimate the hazard rates of breast cancer patients respectively and showed the Royston-Parmar models fitted the observed data well regardless of the choice of the scale.

- Adjusting for the confounding factors, we explored the association between the incidence outcome of breast cancer and country using Royston-Parmar regression and showed that the geographical factor has significant influences on the survival probabilities of patients. Among the 8 countries included, Sweden had the best mortality rates but Denmark had the worst. However, unmeasured confounders, selection and lead-time bias may be the reason for this.

- By using multiple imputation method to estimate the missing values in upa and pai1, we demonstrated that the upa and pai1 did not impact the association between risks of breast cancer and the geographical factor country.

Table 4.10: Summary of the main issues and key findings in Chapter 4.

CHAPTER 5

INTERNAL-EXTERNAL VALIDATION OF A ROYSTON-PARMAR MODEL DEVELOPED USING IPD FROM MULTIPLE STUDIES

5.1 Introduction

It has become commonplace for clinical centres in Europe to cooperate in medical research, and share information about patients with the same disease to form a large dataset. The benefits of such communication are obvious, including, for example, sufficient data to support the clinical study of rare diseases and more stable estimators of extreme observations. One use of large data is to construct an appropriate prognostic model to predict the survival probabilities of the underlying disease in new patients, to inform treatment decision and patient support[111][121][147]. See Section 1.6 for the an introduction to prognostic models.

The first stage of prognostic model development is to establish an appropriate statistical equation for the underlying dataset[107]. In a single clinical study, we only need to determine the important patient level factors on individual risks, however within IPD meta analysis, we should additionally account for the heterogeneities in baseline risks across studies. Therefore, a flexible parametric model with study-specified baseline hazards are potentially very useful[120]. The second stage of prognostic model research is to identify whether the derived models have good generalizability in new individuals[95].

It requires us to validate the proposed model in an external population[4][16]. However, the additional validation dataset is often lacking or it is costly to collect[89]. To address this issue, an internal-external cross validation framework can be utilized to explore the prediction ability of the derived model within the database used to develop it given IPD from multiple studies[123].

The aim of this chapter is therefore to use Royston-Parmar modelling to develop a prognostic model for mortality risk of breast cancer patients, and to validate if using the Look et al.[87] data introduced in Chapter 4. The structure of the chapter is schemed as follows. In Section 5.2 we introduce the main steps to develop a prognostic model from an IPD meta analysis and optimally adjust its intercept to a new study population. Then we apply it to the breast cancer data in Section 5.3. We validate the generalizability of our model using meta analysis of validation studies as proposed by Snell et al.[142], and then derive a final model from the study. In the final section, we discuss the findings and limitations of our methodology and research.

5.2 Methods

In this section, we describe four main steps used to develop and validate a prognostic model in breast cancer: 1) constructing a Royston-Parmar flexible model in IPD meta analysis, 2) deriving an appropriate intercept for the new population, 3) validating model performance using internal-external cross validation scheme, 4) assessing the model within the internal-external cross validation framework to determine the final prognostic model.

5.2.1 Develop a Royston-Parmar prognostic model

Consider an IPD meta analysis of time-to-event data from K individual studies that the number of patients in each study may have not to be same. Let Y_{ij} denote the time at risk of the ith patient in the jth study and d_{ij} be the event indicator, taking the value of 0 or 1, representing whether the observation is censored or not. $\underline{X}_{ij}$ is a participant-level covariate vector, which could be composed of continuous or binary prognostic factors. A potentially naive approach may assume the all IPD were collected from a single and homogeneous population. This approach ignores the studies of participants and a Royston-Parmar proportional hazard model is then fitted to the one dataset:

$$\ln H(t;\underline{X}) = \ln H_0(t) + \underline{\beta}^T \underline{X} \tag{5.2.1}$$

where $\ln H_0(t) = \gamma_0 + \gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t)$ is the log cumulative hazard baseline function using restricted natural cubic splines with four knots, and $\underline{\beta}$ is the coefficient vector of the covariate vector $\underline{X}$. With reference to Section 1.6.1, extension to 5 or more knots could be easily achieved but as suggested by Section 4.5.1, 4 knots are sufficient to model the breast cancer data. The common intercept γ_0 in (5.2.1) indicate any study-level heterogeneities in baseline risks are being ignored. This type of meta-analysis is too simple to capture the individual characteristics of study and may be biased in the presence of the heterogeneity across studies[1].

An alternative method is to use a random effects flexible parametric model to allow for the heterogeneity occurring in the baseline risk. To this purpose, the intercept term is set to be a random variable following a normal distribution rather than a fixed term, that is

$$\ln H(t;\underline{X}) = \ln H_0(t) + \underline{\beta}^T \underline{X} \tag{5.2.2}$$

where $\ln H_0(t) = \gamma_0 + \gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t)$ with $\gamma_0 \sim N(\gamma, \tau_\gamma^2)$. Here we assume γ to be the average study intercept and τ_γ^2 to be the variance of the heterogeneity between studies. By assuming the random effect, it becomes possible to model heterogeneity in baseline risk with very few parameters[84]; however, in practice it is difficult to justify normal assumption of random effects or to interpret it in the validation study. Further, estimating a random-effect model demands advanced software packages and high computing costs, particularly when dealing with a large number of studies in the IPD meta analysis.

Given the limitations of the above two methods, we propose a third option: Royston-

Parmar proportional hazard model with a stratified intercept for each study. This means each study included in the dataset is assigned a unique intercept to account for its own baseline hazard, and the model can be written as

$$\ln H(t; \underline{X}) = \ln H_{0j}(t) + \underline{\beta}^T \underline{X} \tag{5.2.3}$$

where $\ln H_{0j}(t) = \gamma_{0j} + \gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t)$. γ_{0j} is a study-specified term for each study j where $j = 1, 2, ..., K$. In this model, the normality assumption in (5.2.2) is no longer necessary; however, it still assumes the study baseline risks are proportional to log hazards.

The weakness of (5.2.3) is that its intercept term γ_{0j} only focuses on the studies at hand, which cannot be applied immediately to new individuals as these don't fall in a 'study' purse[31]. Therefore, to extend the model to a new study that is not involved in the development of the prognostic model, we should propose some proper intercept strategies to model the heterogeneity in new populations.

5.2.2 Implement the model in a new population

In this section we propose three different approaches to obtain the study-specified intercept for the new population. The first two methods require only the descriptive statistics from the derivation dataset or new population, while the final one depends on re-estimating the intercept using new IPD which is also known as a recalibration technique[95]. In the validation study, we do not need to estimate the hazard rate function of model (5.2.3) again, and as long as we replace γ_{0j} in (5.2.3) with the new intercept term obtained from either of the three intercept strategies, we can generalize this model to predict the risk of an individual in a new population.

Average intercept

A straightforward method for obtaining a new model intercept may be to pool the K individual intercepts γ_{0j}, $j = 1, ..., K$ to produce a weighted average estimator where the weights can be determined by common fixed effect or random effects meta-analysis (average strategy)[33][123]. In our research, the γ_{0j} are pooled using random effects meta analysis and restricted maximum likelihood method. See the introduction of meta analysis in Section 5.2.4 for the equation.

The average strategy is quite easy to put in practice where the obtained intercept can be used as an approximation to the new dataset. However, this strategy is not flawless: when the baseline risk of the new study is very different from the average in the original datasets, then the discrepancy may lead to a large bias in predicted risks over time (*i.e.* poor calibration of predicted and observed risk).

Neighbour intercept

An alternative to the average intercept is to select an estimated intercept of which study is closest to the new population (nearest neighbour intercept strategy). To determine the most similar dataset to the target study requires in-depth knowledge of the clinical research and careful consideration of all the potential covariates amongst studies. However, if sufficient clinical guidance to the selection of the candidate study is lacking, the statistical measure could help us to make the decision. Debray et al.[31] proposed an algorithm to compare the similarities between the study in the derivation dataset and the validation dataset:

1. For each predictor and/or outcome, calculate difference in mean (continuous variables) or proportion (discrete variables) of observed individuals between each study in the derivation dataset and the validation dataset.

2. For each predictor and/or outcome, assign a rank for the mean or proportion across all the studies in the derivation dataset according to similarity, where increasing ranks indicate increasing differences (*i.e.* , less similarity).

3. For each study in the derivation dataset, determine the median rank of all its predictors and/or outcome, and then select the intercept for the new population from the study with the smallest median rank.

Steyerberg et al.[148] applied this method to develop and validate a risk prognostic model across multiple studies, and then Debray et al.[31]] generalized this method into the setting of binary outcome data. To use this method, we should know the baseline characteristics of the new population to which we want to apply the model.

New intercept

If the baseline characteristics of the new population differ greatly from any study in the derivation dataset, we may find that neither of the previous two strategies could work properly here. To ensure the optimal estimate of the intercept for the new dataset, Debray et al.[31] suggested that one may additionally use data from the new population to re-estimate the intercept term. This new strategy can be treated as a benchmark in intercept estimation as it is able to correctly reflect the heterogeneities in new individuals.

To determine the new intercept term γ_0, we treat the other terms in baseline function $\ln H_0(t)$ and the linear predictor $\underline{\beta}^T \underline{X}$ as the offset terms, and then re-estimate the intercept term γ_{new} in the new population using flexible parametric regression. This algorithm can be implemented using any general linear model software package[37][52]. For example, consider the log likelihood function for the ith individual in the Royston-Parmar proportional hazard model as

$$\begin{aligned}\ln L_i &= d_i\{\ln[\gamma_1 + \gamma_2 z_1'(\ln t) + \gamma_3 z_2'(\ln t)] + \gamma_0 + \gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t) \\ &\quad + \underline{\beta}^T \underline{X}_i\} - \exp(\gamma_0 + \gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t) + \underline{\beta}^T \underline{X}_i)\end{aligned}$$

where $z_j'(\ln t) = \partial z_j(\ln t)/\partial \ln t$ for $j = 1, 2$. For the new intercept strategy, except for the intercept γ_0, all the other terms are treated as the offset, hence we drop the terms that do not depend on γ_0 and then obtain the resulted log likelihood as

$$\ln L_i = d_i\gamma_0 - \exp(\gamma_0 + \gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t) + \underline{\beta}^T \underline{X}_i).$$

Exactly, the above expression is identical to the likelihood function of a Poisson regression with the outcome d_i, and mean $\exp(\gamma_0 + \gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t) + \underline{\beta}^T \underline{X}_i)$. The $\exp(\gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t) + \underline{\beta}^T \underline{X}_i)$ is a constant that can be incorporated into a linear predictor via an offset. We are therefore able to compute the estimator of the new intercept term from a general linear model with outcome d_i, a Poisson error structure with only unknown constant term γ_0, a log link, and an offset of $(\hat{\gamma}_1 \ln t + \hat{\gamma}_2 z_1(\ln t) + \hat{\gamma}_3 z_2(\ln t) + \hat{\underline{\beta}}^T \underline{X}_i)$, where $\hat{\gamma}_1$, $\hat{\gamma}_2$, $\hat{\gamma}_3$, and $\hat{\underline{\beta}}$ are as estimated in the developed model.

5.2.3 Model evaluation with internal-external cross-validation

In the previous section, three methods were proposed for obtaining a unique intercept to validate new individuals when baseline risks are heterogeneous across studies. This step can be regarded as the external validation of the derived model if a new population is available. However, in most trials it is found that the additional dataset is lacking or it requires more costs to collect further data[3][4][149]. This exposes the need to propose some internal validation approaches to assess a prognostic model that does not rely on the external dataset. Specifically with IPD from multiple studies, it requires a framework to maximize the data available in the construction of a model and to support the corresponding model validation within the IPD meta analysis.

The internal-external cross validation scheme first proposed by Royston et al.[123] may help to solve our problem. The main idea of this method is to iteratively use $K - 1$ of the K studies to develop a prognostic model and take the omitted one as the external validation dataset. It is a common method in model selection and data smoothing problems[7][57][139]. In the setting of hazard rate estimation with time-to-event data, we extend this technique to flexible parametric regression as follows:

Step 1. Select $K - 1$ studies from IPD database to form a derivation dataset and treat the omitted one as the validation (new) dataset.

Step 2. Develop a flexible parametric model from the derivation dataset using Model (5.2.3).

Step 3. Choose a suitable intercept for the validation study (for example, using one of the intercept strategies introduced in Section 5.2.2).

Step 4. In the external validation study, utilize the estimated model from Step 2 to predict the risks of new individuals where the intercept term γ_0 of the model is particularly determined in Step 3.

Step 5. Assess the external validation performance of the developed model based on two fundamental aspects, the discrimination and calibration of model predictions.

Step 6. Repeat Step 1-5 for each permutation of $K-1$ studies and determine the final prognostic model according to the model performance in each permutation and by summarising performance across all permutations.

We now outline calibration and discrimination:

Calibration

Calibration refers to whether the predicted probabilities agree with the observed probabilities. For example, in a validation study, a well-calibrated prognostic model should assign the correct hazard rates to each level of risk groups from the validation dataset[121].

First, we may evaluate an overall calibration ability of a prognostic model by calculating its calibration slope in the validation dataset. This is a recognized approach to estimate the regression coefficient on the prognostic index in the validation dataset[10][93]. To make adjustments for the heterogeneity in the validation dataset, the intercept term γ_0 is obtainable from either of the three intercept strategies above while the other terms in the baseline hazards and the coefficients of the prognostic index are given by the developed model. One could utilize any general linear model package to compute the calibration slope b_{new} in practice[37][52]. Let's consider the likelihood function of a Royston Paramar model with the prognostic index,

$$\begin{aligned}\ln L_i &= d_i\{\ln[\gamma_1+\gamma_2 z_1'(\ln t)+\gamma_3 z_2'(\ln t)]+\gamma_0+\gamma_1\ln t+\gamma_2 z_1(\ln t)+\gamma_3 z_2(\ln t) \\ &\quad +b_{new}\mathrm{PI}\}-\exp(\gamma_0+\gamma_1\ln t+\gamma_2 z_1(\ln t)+\gamma_3 z_2(\ln t)+b_{new}\mathrm{PI})\end{aligned}$$

where b_{new} is the calibration slope on the prognostic index to be estimated and PI denotes the prognostic index $\underline{\beta}^T\underline{X}$ which are obtained from the developed model. Since $(\ln[\gamma_1 + \gamma_2 z_1'(\ln t) + \gamma_3 z_2'(\ln t)])$ and $(\gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t))$ are known from the prognostic model, and γ_0 is given by any of three intercept strategies, they can be dropped from the log likelihood function and thus it achieves

$$\ln L_i = d_i(b_{new}\mathrm{PI}) - \exp(\gamma_0 + \gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t) + b_{new}\mathrm{PI}).$$

The above function is identical to the likelihood for a Poisson regression with the outcome d_i, and mean $\exp(\gamma_0 + \gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t) + b_{new}\mathrm{PI})$. Hence the same parameter estimator b_{new} can be obtained from a general linear model with outcome d_i, a Poisson error structure with the predictor, prognostic index but no constant term, a log link, and an offset of $\gamma_0 + \hat{\gamma}_1 \ln t + \hat{\gamma}_2 z_1(\ln t) + \hat{\gamma}_3 z_2(\ln t)$ where PI= $\hat{\underline{\beta}}^T\ \underline{X}$, $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $\hat{\gamma}_3$ are as the estimates from derivation model.

A poor calibration that $b_{new} \neq 1$ may imply unaccounted for differences in baseline risks and/or in predictor-outcome associations[31]. However, the calibration slope can reflect only the overall calibration performance of a model. To detect the potential weakness of the prognostic model at each risk level, it is better for us to draw the calibration curves to illustrate the deviance between predicted and survival probabilities at each risk level[143][145]. To this purpose, we divide the new samples from the validation dataset into four risk groups according to their individual prognostic index, where the cut-off points are at its 20, 50 and 80 centiles. Next, we average the survival curves of the patients in each group, and superimpose the mean survival curves and observed (Kaplan-Meier) survival curves on one graph[120]. From visually inspecting the calibration curves, the predicted survival probability of the derived model and the observed survival probability can be compared directly at each risk level.

Discrimination

Discrimination, also called separation, reflects how well it can distinguish between patients with high and low risk under a prognostic model. As an example, a model discriminates better if it predicts the survival probabilities of individuals at two years ranging between 20% and 80% rather than between 40% and 60%[124].

A general discrimination measure for time-to-event data is Harrell C statistic (see Harrell et al.[54] or Newson[98]). Consider all the possible pairs of patients in the dataset, where one has died and the other is alive at the other person's death time. If the dead person has a higher prognostic index than the alive person, then the prediction of this pair is said to be concordant with the outcome. Then C statistic is the proportion of the concordant ones in all the possible pairs. For C statistics, a value of 0.5 indicates no discrimination beyond chance while 1 means a perfect discrimination of patients.

The standard deviation of the C statistics is computed by a resampling (bootstrapping) technique, namely as jackknife method[159]. For instance, in a validation dataset given n observations, the jackknife algorithm is found by aggregating the C statistics of each permutations of $(n-1)$ observations and then calculate the variance of the obtained n copies of C statistic.

Royston and Sauerbrei[124] proposed an alternative discrimination measurement focusing on the spread of the outcome in Cox or other proportional hazard regression models. It is motivated by the fact that the discrimination of a model may be quantified by the variation in outcome among patients on the proportional hazard scale. For example, a weak model will have difficulty distinguishing between the risks of different patients, and this will tend to be reflected by a narrow spread of prognostic index values. Therefore the D index is proposed by ordering the estimated prognostic index, calculating the corresponding expected normal order statistics (rankits), scaled by a factor $\kappa = \sqrt{8/\pi} \simeq 1.6$ and performing an auxiliary regression on the scaled rankits. The reason to scale the rankits by κ is to ensure the D index has the character of a log hazard ratio between equal-sized prognostic groups.

Intuitively, larger D statistics may represent better discrimination, however, it lacks guidance for us to set a proper threshold for the D index to classify whether the discrimination abilities of a model is satisfied or not. Therefore a monotonic transformation, R_D^2

of D statistics is proposed by Royston and Lambert[120]:

$$R_D^2 = \frac{D^2/\kappa^2}{\sigma^2 + D^2/\kappa^2}$$

where $\sigma^2 = \pi^2/6$ for proportional hazards models. We can interpret D^2/κ^2 as an estimate of the variance of prognostic index across individuals, R_D^2 as a measure of explained variation on the natural scale of the model and the σ^2 as the counterpart to the residual variance in linear regression model. In the real datasets, R_D^2 for prognostic models may vary quite widely from 0% to 60% although theoretically, R_D^2 can reach 100% by little chance[120]. In our research, we will report both Royston D index and R_D^2 index.

5.2.4 Assessment of the model in the internal-external cross validation framework

In the internal-external cross validation scheme, the $K-1$ studies from an available IPD meta analysis are iteratively used to develop a prognostic model, which is validated in the omitted dataset using several important measurements introduced above, for example, calibration slopes, calibration curves, C statistics and D statistics. This produces K estimates of each validation statistic. Snell et al.[142] recommended that meta-analysis can then be used to summarize them[101][162]. A perfect model will have good average performance with small errors, and little or no heterogeneities across studies[32]. The basic idea of univariate meta analysis and multivariate meta analysis are illustrated in the next section.

Meta analysis

Suppose that $j = 1$ to K studies each provide validation estimates, Y_{jl} for $l = 1, 2$ and 3, and associated standard errors, s_{jl}. Each summary statistic Y_{jl} is assumed to be an estimate of a true value θ_{jl} in each study, and in a hierarchical structure (with random effects setting) each θ_{jl} is assumed to be drawn from a distribution with mean value θ_l and between study variance τ_l^2.

In a univariate meta analysis, for each $l = 1$, 2 or 3,

$$Y_{jl} \sim N(\theta_{jl}, s_{jl}^2)$$

where θ_{jl} is a constant for a fixed effect model but it follows a normal distribution $N(\theta_l, \tau_l^2)$ for a random effects model.

With the reference to Riley et al.[103][105] and Snell et al.[142], we could alternatively pool the three statistics jointly using a trivariate meta-analysis:

$$\begin{pmatrix} Y_{j1} \\ Y_{j2} \\ Y_{j3} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_{j1} \\ \theta_{j2} \\ \theta_{j3} \end{pmatrix}, \delta_j\right), \delta_j = \begin{pmatrix} s_{j1}^2 & & \\ s_{j1}s_{j2}\rho_{Wj12} & s_{j2}^2 & \\ s_{j1}s_{j3}\rho_{Wj13} & s_{j2}s_{j3}\rho_{Wj23} & s_{j3}^2 \end{pmatrix}.$$

In a fixed effect model, θ_{jl} for $l = 1, 2$ or 3 is set to be constant. In random effects model, it additionally specifies that

$$\begin{pmatrix} \theta_{j1} \\ \theta_{j2} \\ \theta_{j3} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix}, \Omega\right), \Omega = \begin{pmatrix} \tau_1^2 & & \\ \tau_1\tau_2\rho_{B12} & \tau_2^2 & \\ \tau_1\tau_3\rho_{B13} & \tau_2\tau_3\rho_{B23} & \tau_3^2 \end{pmatrix}$$

where δ_j and Ω are the within-study and between study covariance matrices respectively. Here the within-study correlation represents the association between the summary estimates within a study and the between-study correlation indicates how the true underlying statistics are associated across studies[104].

As usual, the within-study correlations ρ_W are assumed to be known to us. If not, with IPD, Snell et al.[142] notes that it is possible for us to quantify the within-study correlations of multiple endpoint statistics using the bootstrap technique[29][39]. Specifically, a random sample is produced with replacement of patients within each study where the size of the sample equals to the size of the study. Then the resampling procedure is repeated for 1000 times (which ensures adequate precision in estimating the correlation[39]), each time recomputing the corresponding summary estimates. Finally the correlation of each pair of summary statistics Y_1, Y_2 and Y_3 over the 1000 replicates are calculated and treated as their correlation estimates in trivariate meta analysis.

The calculation of the between-study matrix in trivariate meta analysis can be undertaken using restricted maximum likelihood[70][73]. In comparison to univariate meta analysis, trivariate meta analysis has several advantages, for example, it is more elegant to conduct a single IPD meta analysis than many univariate ones. Also the relationship between the multiple effects could be easily illustrated and interpreted. Further, parameter estimates of trivariate meta analysis is often superior to a univariate meta analysis since each summary statistics can ‘borrow strength’ from the other endpoints in estimation[70]. Here, our 3 statistics of interest are C, D and calibration slope.

Model evaluation in the internal-external cross validation framework

In general, if the derived models all calibrate well across the considered permutations, then all the IPD studies at hand can be combined to develop a final prognostic model. However, if any of the derived models do not calibrate well in the omitted study, it indicates less generalizability of the model and its real cause should be cautiously identified. Several reasons may cause this discrepancy in the calibration slope or calibration curves; for example, the failure of the chosen intercept strategy to capture the baseline characteristics in a validation study, unexpected heterogeneities in predictor-outcome association, or overfitting of model[31][121].

In cases of homogeneous predictor-outcome association, if the calibration slope is strongly biased from 1, it may suggest that the intercept strategy applied for the validation dataset is not appropriate, and therefore it may be preferable to re-estimate a more study-specified intercept term from the validation population (new intercept strategy)[133].

However, when strong heterogeneities also exist in predictor-outcome associations across studies, the intercept estimator may capture too much unexplained risk from both baseline risks and predictor-outcome associations and thus returns the biased estimator[56]. This problem may be alleviated by adding more covariates or nonlinear effects in the model; however, it might also incur a risk of overfitting. Therefore in most scenarios, it is suggested to exclude the study that gives poor validation statistics from the final model and treat it separately[31].

Harrell C index and Royston D index measure the discrimination ability of a model which are known to reach the optimality at 1. It is worth highlighting that both C and D measurements are only sensitive to the ranks of prognostic index. A subtle change in the model that alters prognostic index but leaves its rank order unchanged does not affect the two statistics at all. That is to say, the C or D measure does not depend on the choice of intercept strategy and thus can not be easily improved unless additional predictors are included[95][119][161].

The last point to emphasize is that the internal-external cross validation framework requires sufficient individuals in each study and also enough numbers of studies in the dataset to guarantee well generalizability ability of the model. In particular, if some studies in the IPD meta analysis contains too few samples, the resulted prognostic model may behavior poor and the corresponding confidence interval may become unstable. If the IPD meta analysis has too few studies, it becomes difficult for the prognostic model to identify individual characteristics of new population in validation study[146]. For this reason, Debray et al.[31] suggested that, to develop a meta-analytical prognostic model, it should include at least four or five individual studies that each has a reasonable large effective samples and number of events.

5.3 Application to the breast cancer data

To demonstrate the potential value of the internal-external cross validation scheme for model development and validation study, we conducted a study using the internal-external cross validation framework on a real data sample. The dataset utilized is same as that introduced in Chapter 4, which is about the patients with breast cancer from 15 laboratories in Europe. Specifically, according to registry country, the 15 studies formed 8 individual countries including Netherland, Ireland, Sweden, Slovenia, Austria, France, Switzerland

and Denmark. From log-log plot given in Section 4.5.2, it was seen that the proportional hazard assumption roughly holds across countries. Hence a proportional hazard model with the stratified intercept term ($d.f. = 3$) could be applied to model the underlying database (see Model (5.2.3)) and the aforementioned variables of patient age, tumour type, tumour grade, number of lymph nodes, menopausal status, tumour size, adjuvant treatment and hormone receptor status were included into the model as the prognostic factors.

We summarize the important results as follows: Firstly, the parameter estimators for each rotation within the internal-external cross validation framework are given in Section 5.3.1. The calibration and discrimination performance of the model are assessed in Section 5.3.2. Then the selection of the intercept strategy is discussed in Section 5.3.3 using meta analysis. In the last section, a final prognostic model is developed for the purpose to predict individuals in new population.

5.3.1 Parameter estimators in internal-external cross validation framework

We summarize the stratified intercept and parameter estimates for each permutation of the internal-external cross validation approach in Table 5.1, Table 5.2 and Table 5.3 respectively.

Omitted	**Intercept**	Ned	Irl	Swe	Slo	Aut	Fra	Sui	Den	Average	Neighbour	New
Ned	Estimate		-3.354	-4.332	-3.867	-3.017	-3.568	-3.548	-2.847	-3.492	-2.847	-3.296
	(s.e.)		(0.418)	(0.398)	(0.411)	(0.362)	(0.328)	(0.358)	(0.366)	(0.231)	(0.366)	(0.029)
Irl	Estimate	-3.207		-4.260	-3.678	-2.892	-3.497	-3.475	-2.848	-3.398	-4.260	-3.244
	(s.e.)	(0.217)		(0.246)	(0.333)	(0.275)	(0.218)	(0.249)	(0.230)	(0.188)	(0.246)	(0.131)
Swe	Estimate	-3.247	-3.305		-3.697	-2.926	-3.538	-3.552	-2.929	-3.304	-2.929	-4.329
	(s.e.)	(0.222)	(0.274)		(0.336)	(0.278)	(0.223)	(0.254)	(0.235)	(0.191)	(0.235)	(0.083)
Slo	Estimate	-3.172	-3.221	-4.230		-2.860	-3.470	-3.455	-2.824	-3.321	-3.455	-3.658
	(s.e.)	(0.217)	(0.270)	(0.247)		(0.275)	(0.218)	(0.250)	(0.230)	(0.186)	(0.250)	(0.258)
Aut	Estimate	-3.196	-3.247	-4.252	-3.690		-3.487	-3.480	-2.847	-3.441	-3.480	-2.903
	(s.e.)	(0.222)	(0.274)	(0.251)	(0.337)		(0.223)	(0.254)	(0.235)	(0.190)	(0.254)	(0.186)
Fra	Estimate	-2.851	-2.865	-3.880	-3.354	-2.562		-3.103	-2.478	-3.001	-3.103	-3.138
	(s.e.)	(0.248)	(0.297)	(0.275)	(0.354)	(0.297)		(0.279)	(0.260)	(0.201)	(0.279)	(0.060)
Sui	Estimate	-3.122	-3.152	-4.163	-3.621	-2.854	-3.442		-2.758	-3.292	-2.854	-3.422
	(s.e.)	(0.218)	(0.271)	(0.248)	(0.334)	(0.276)	(0.218)		(0.231)	(0.190)	(0.276)	(0.115)
Den	Estimate	-3.196	-3.239	-4.248	-3.655	-2.867	-3.495	-3.472		-3.453	-3.196	-2.855
	(s.e.)	(0.221)	(0.273)	(0.250)	(0.336)	(0.279)	(0.221)	(0.254)		(0.192)	(0.221)	(0.062)

Table 5.1: The parameter estimates of the stratified country intercepts for each permutation within the internal-external cross validation framework and the intercept for the omitted study using three different intercept strategies respectively (average, nearest neighbour and new strategies). All the estimates in this table are strongly significant (P value< 0.001).

Omitted	**Coefficient**	**Age**	**Tumour type**						
study			ilc	colloid	tubul	medull	papil	other	unknown
Ned	Estimate	0.015^{**}	0.051	-0.661	-0.168	-0.897^{**}	-0.390	0.099	0.079
	(s.e.)	(0.005)	(0.153)	(1.012)	(0.713)	(0.362)	(0.729)	(0.222)	(0.146)
Irl	Estimate	0.012^{***}	-0.099	-0.307	-0.530	-0.774^{***}	-0.420	-0.076	0.023
	(s.e.)	(0.003)	(0.095)	(0.711)	(0.502)	(0.222)	(0.506)	(0.131)	(0.066)
Swe	Estimate	0.011^{***}	-0.122	-0.310	-0.539	-0.742^{**}	-0.419	-0.086	0.025
	(s.e.)	(0.003)	(0.096)	(0.711)	(0.502)	(0.222)	(0.506)	(0.131)	(0.066)
Slo	Estimate	0.011^{***}	-0.101	-0.252	-0.529	-0.772^{**}	-0.662	-0.073	0.023
	(s.e.)	(0.003)	(0.096)	(0.711)	(0.502)	(0.222)	(0.583)	(0.131)	(0.066)
Aut	Estimate	0.011^{***}	-0.119	-0.296	-0.496	-0.744^{**}	-0.257	-0.068	0.024
	(s.e.)	(0.003)	(0.097)	(0.711)	(0.502)	(0.227)	(0.580)	(0.133)	(0.066)
Fra	Estimate	0.008^{*}	-0.121	0.260	-0.512	-0.763^{*}	-0.375	-0.113	0.010
	(s.e.)	(0.003)	(0.100)	(1.005)	(0.502)	(0.222)	(0.507)	(0.140)	(0.074)
Sui	Estimate	0.012^{***}	-0.077	-0.285	-0.347	-0.733^{**}	-0.389	-0.043	0.023
	(s.e.)	(0.003)	(0.099)	(0.711)	(0.502)	(0.227)	(0.507)	(0.135	(0.066)
Den	Estimate	0.012^{***}	-0.100	-0.333	-1.060	-0.821^{**}	-0.422	-0.121	0.023
	(s.e.)	(0.003)	(0.107)	(0.711)	(0.709)	(0.262)	(0.506)	(0.140)	(0.066)

Table 5.2: The parameter estimates of the predictors in the prognostic model in each permutation in the internal-external cross validation framework where ‘idc’ for tumour type is treated as the reference level. N.B. * for P value< 0.05, ** for P value< 0.01 and *** for P value<0.001.

Omitted	**Coefficient**	**Tumour Grade**			**Lymph nodes**			**Menopasual**	**Tumour**	**Adjuvant**	**Hormone**
study		moderate	poor	unknown	$np1-3$	np3-10	np10+	**status**	**size**	**treatment**	**receptor**
Ned	Estiamte	0.437*	0.861***	0.583*	0.632***	1.408***	1.858***	-0.152	0.308***	-0.113*	-0.744***
	(s.e.)	(0.221)	(0.226)	(0.268)	(0.104)	(0.108)	(0.135)	(0.114)	(0.060)	(0.049)	(0.077)
Irl	Estiamte	0.426**	0.786***	0.608***	0.715***	1.241***	1.716***	-0.071	0.326***	-0.122***	-0.576***
	(s.e.)	(0.159)	(0.157)	(0.163)	(0.066)	(0.068)	(0.084)	(0.077)	(0.038)	(0.029)	(0.053)
Swe	Estimate	0.445**	0.808***	0.668***	0.739***	1.242***	1.671***	-0.020	0.326***	-0.106**	-0.486***
	(s.e.)	(0.162)	(0.160)	(0.167)	(0.068)	(0.070)	(0.088)	(0.078)	(0.038)	(0.031)	(0.055)
Slo	Estimate	0.439**	0.789***	0.617***	0.709***	1.253***	1.697***	-0.052	0.328***	-0.114***	-0.573***
	(s.e.)	(0.161)	(0.158)	(0.165)	(0.065)	(0.067)	(0.084)	(0.076)	(0.037)	(0.029)	(0.052)
Aut	Estimate	0.468**	0.830***	0.654***	0.704***	1.248***	1.724***	-0.056	0.331***	-0.113***	-0.567***
	(s.e.)	(0.170)	(0.166)	(0.172)	(0.065)	(0.067)	(0.084)	(0.076)	(0.037)	(0.029)	(0.052)
Fra	Estiamte	0.301	0.652**	0.469*	0.803***	1.324***	1.786***	0.001	0.315***	-0.144***	-0.596***
	(s.e.)	(0.196)	(0.190)	(0.194)	(0.070)	(0.071)	(0.089)	(0.082)	(0.039)	(0.031)	(0.056)
Sui	Estimate	0.423**	0.748***	0.552**	0.689***	1.231***	1.683***	-0.059	0.317***	-0.104***	-0.572***
	(s.e.)	(0.159)	(0.157)	(0.164)	(0.066)	(0.068)	(0.086)	(0.077)	(0.038)	(0.029)	(0.053)
Den	Estimate	0.430**	0.795***	0.617***	0.656***	1.163***	1.620***	-0.079	0.340***	-0.114***	-0.538***
	(s.e.)	(0.159)	(0.157)	(0.163)	(0.068)	(0.070)	(0.087)	(0.081)	(0.039)	(0.030)	(0.056)

Table 5.3: The parameter estimates of the predictors in the prognostic model in each permutation in the internal-external cross validation framework where ‘good’ for tumour grade and ‘np=0’ for lymph nodes are treated as the reference levels. N.B. * for P value< 0.05, ** for P value< 0.01 and *** for P value<0.001.

Regardless of which permutation of the internal-external cross validation framework and regardless of intercept strategy, the coefficient estimates of the predictors and the intercept estimates for each included study are very similar. Of all the permutations, menopasual status is not significant at all, hence this factor may not be considered in our final established model.

5.3.2 Calibration and discrimination

The resulted C statistic, D statistic and calibration slopes in each omitted study are summarized in Table 5.4 for each permutation of the internal-external cross validation approach for each intercept strategy.

Omitted	C	D	R_D^2	**Calibration Slope**		
study	**Index**	**Index**	**Index**	Average	Neighbour	New
Netherland	0.697 (0.008)	0.493 (0.027)	0.055 (0.006)	1.049 (0.012)	0.805 (0.012)	0.977 (0.012)
Ireland	0.701 (0.036)	0.420 (0.117)	0.040 (0.022)	1.066 (0.057)	1.414 (0.056)	1.002 (0.057)
Sweden	0.715 (0.023)	0.106 (0.056)	0.003 (0.003)	0.578 (0.037)	0.405 (0.037)	1.026 (0.036)
Slovenia	0.735 (0.068)	0.326 (0.187)	0.025 (0.028)	0.870 (0.098)	0.919 (0.097)	0.991 (0.097)
Austria	0.666 (0.050)	0.238 (0.168)	0.013 (0.019)	1.168 (0.086)	1.184 (0.086)	0.946 (0.088)
France	0.682 (0.017)	0.182 (0.041)	0.008 (0.004)	0.896 (0.038)	0.951 (0.037)	0.969 (0.037)
Switzerland	0.781 (0.027)	0.280 (0.063)	0.018 (0.008)	0.996 (0.053)	0.794 (0.054)	1.054 (0.052)
Denmark	0.722 (0.016)	0.541 (0.058)	0.065 (0.013)	1.315 (0.029)	1.197 (0.030)	1.035 (0.030)

Table 5.4: The estimators of C, D statistics and calibration slopes in the internal-external cross validation when dataset ID is used for validation and the remaining studies for derivation.

The C statistics indicated that the discriminative abilities of the prognostic model across studies are very stable around 0.7, suggesting that the prognostic model correctly discriminate the risks of approximately 70% paired participants in the dataset. Compared with C, the D measure, which focuses on estimating the variance of prognostic index is broadly unstable, and the average explained variation, R_D^2 is around 0.04 across studies.

In views of calibration slopes, the average strategy yields quite consistent performance in 6 countries ($0.8 < b < 1.2$) whereas the calibration slope is too low for Sweden ($b = 0.578$, s.e.=0.037) and too high for Denmark ($b = 1.315$, s.e.=0.029). In contrast, the nearest neighbour intercept seems somewhat less competitive where the slope estimators of Ireland ($b = 1.414$ s.e.=0.056) is upward biased and of Sweden ($b = 0.405$, s.e.=0.037) and Switzerland (Sui, $b = 0.794$, s.e.=0.054) are downward biased. To recalibrate the bias, we could utilize the new intercept strategy which validates perfect across studies as its calibration slopes are close to the optimal value of 1 in all the omitted studies.

It is noticed that, the worst estimate of the calibration slope comes from the study in Sweden for either average or nearest neighbour strategies. It is not clear whether the bias is caused by its unique baseline characteristic or predictor-outcome associations, but the former is more plausible as the new intercept strategy returns a reasonable calibration slope estimator ($b = 1.012$, s.e.=0.025). Because in practice, re-estimating the intercept in the new population is not always available to us, to be cautious, in Section 5.3.4 we establish the final model excluding the study from Sweden.

We also created four prognostic groups by dividing the prognostic index at 20, 50, 80 centiles and then compared the Kaplan-Meier survival curves and predicted mean survival curves to display the group-specific prognosis.

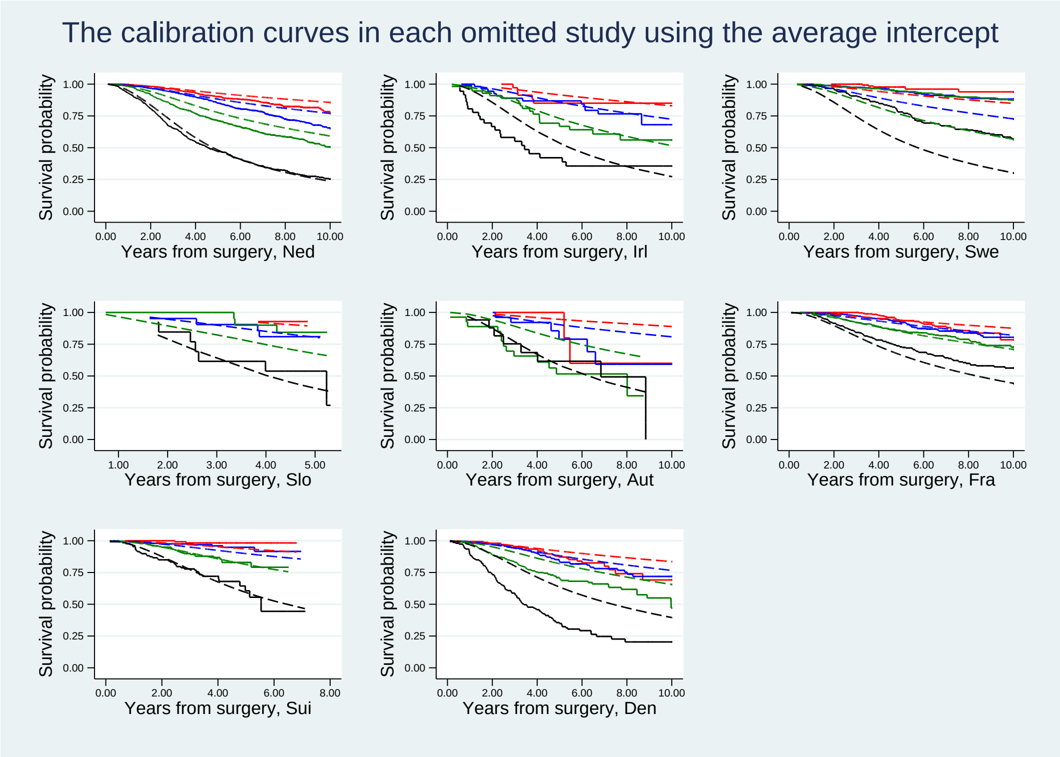

Figure 5.1: Average intercept: Kaplan-Meier curves (jagged lines) and mean survival curves (dashed lines) in four prognostic groups.

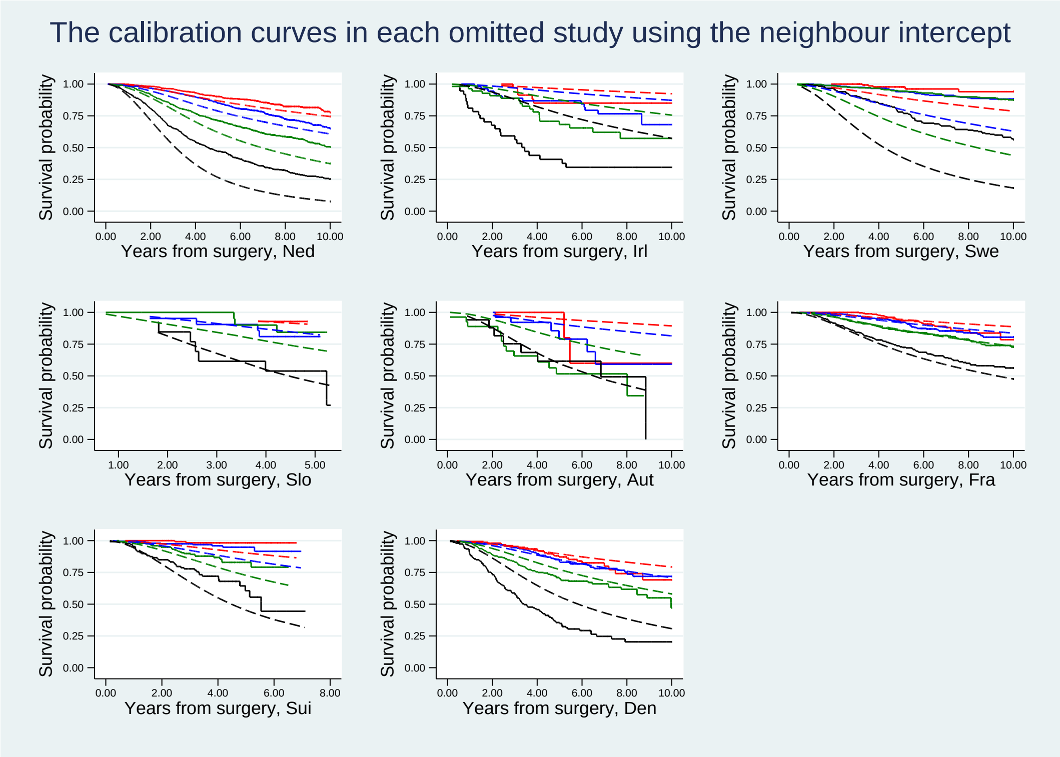

Figure 5.2: Neighbour intercept: Kaplan-Meier curves (jagged lines) and mean survival curves (dashed lines) in four prognostic groups.

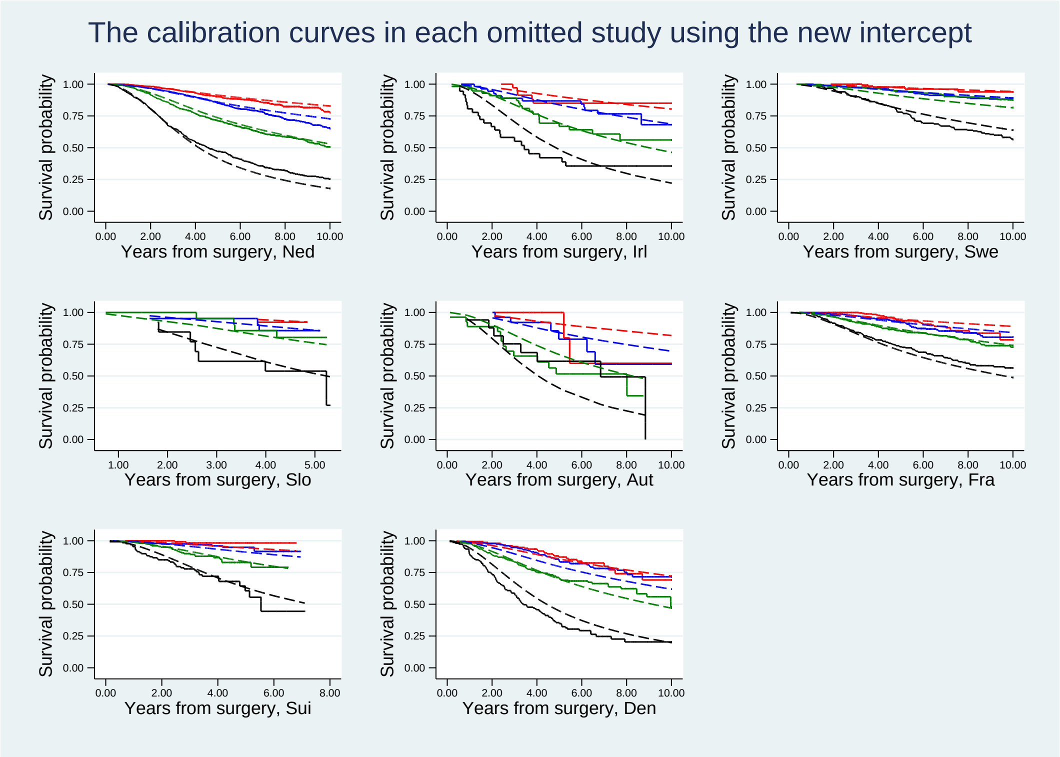

Figure 5.3: New intercept: Kaplan-Meier curves (jagged lines) and mean survival curves (dashed lines) in four prognostic groups.

The calibration curves of the average intercept strategy were plotted in Fig 5.1. The model clearly does not calibrate perfectly in every single omitted study due to unexplained heterogeneities in the dataset. The predicted curves in Netherland, Ireland, France and Switzerland roughly coincided with the reference curves, however, the others are less accurate.

As for the nearest neighbour strategy, from Fig 5.2, we find that the returned calibration curves performed visually worse than its average counterpart. It clearly shows that except for France and Denmark, the big deviations could be found between the predicted and observed survival curves in the rest of 6 countries. It might be because neighbour s-trategy which was determined by the statistical measure, is unlikely to explain real clinical difference amongst studies.

Finally, Fig 5.3 shows that almost perfect agreement in predicted curves were achieved when the intercept using the new strategy was applied to the model. The only exception was found in Austria which is probably due to its small effective sample size ($n = 88$).

5.3.3 Selection of intercept strategy

Of the three intercept strategies involved, we could use the approach of Snell et al.[142] to meta analyse the achieved calibration slopes to quantify the best one. Here trivariate meta analysis allows us to to 'borrow the strength' from C measure and D measure (R^2_D statistic) to summarize the calibration slope estimate for each strategy. To this purpose, we first employed the bootstrap technique to obtain the correlations between C measure or R^2_D measure and calibration slopes (see Appendix B). Then we conducted trivariate meta analysis using the correlations, variance and estimates of the performance statistics and summarized results in Table 5.5.

	Summary results from meta analysis			**Between-study Correlation**				
Intercept Strategy	C Index estimate	D Index estimate	Calibration slope	C& Slope	D& Slope	C&D	Multi I^2 test	Between study τ^2
Average	0.711 (0.011)	0.028 (0.009)	0.993 (0.080)	0.019	0.784	-0.148	97.0%	0.046
Neighbour	0.709 (0.012)	0.028 (0.009)	0.960 (0.112)	-0.315	0.532	-0.194	98.1%	0.074
New	0.710 (0.011)	0.027 (0.008)	1.001 (0.017)	0.984	-0.302	-0.127	90.6%	< 0.001

Table 5.5: The trivariate meta analysis for three different intercept strategies: The pooled estimators, the between-study correlations, the multivariate I^2 test and heterogeneity chi-square τ^2 in calibration slopes across 8 studies. N.B. the numbers in brackets represent standard errors.

The multivariate I^2 statistics (see Jackson et al.[71]) are in good agreement to support random effects model for all three strategies (rather than fixed effect). All the between-study correlation estimators were well defined and interpretable[105].

Of the three strategies, the new intercept strategy achieved the optimal estimator of the calibration slope (b = 1.001, s.e.=0.017) and has the smallest heterogeneity across 8 studies ($\tau^2 < 0.001$). The second best one is the average intercept strategy since its calibration slope estimator (b = 0.993, s.e.=0.080) is secondly closest to the best value 1 and the heterogeneity τ^2 is in the second place being 0.046. The nearest neighbour strategy is demonstrated to be the worst amongst the 3 candidates no matter with respect to the point estimator ($b = 0.960$, s.e.=0.112) or heterogeneity statistic ($\tau^2 = 0.074$).

5.3.4 Final model to predict individuals in new population

Recall that in Section 5.3.2, we have pointed out that the achieved calibration slope in Study Sweden is poorest of 8 countries. Therefore, we established a prognostic model excluding the Study Sweden and repeated the internal-external cross validation framework for the entire process. Notice that the factor, menopausal status was excluded from the model as it has no significant impact on individual risks. We summarize the obtained estimators of C, D statistics and calibration slopes within each permutation in Table 5.6.

Omitted	C	D	R_D^2	**Calibration Slope**		
study	**Index**	**Index**	**Index**	**Average**	**Neighbour**	**New**
Netherland	0.699 (0.008)	0.492 (0.027)	0.055 (0.006)	1.019 (0.011)	0.913 (0.011)	0.984 (0.011)
Ireland	0.702 (0.037)	0.423 (0.117)	0.041 (0.022)	1.006 (0.056)	0.852 (0.056)	1.005 (0.056)
Slovenia	0.730 (0.070)	0.314 (0.187)	0.023 (0.027)	0.834 (0.097)	0.941 (0.096)	0.992 (0.096)
Austria	0.661 (0.050)	0.230 (0.168)	0.012 (0.018)	1.126 (0.085)	1.205 (0.084)	0.951 (0.086)
France	0.682 (0.017)	0.181 (0.041)	0.008 (0.003)	0.832 (0.036)	0.975 (0.035)	0.978 (0.035)
Switzerland	0.782 (0.027)	0.282 (0.063)	0.019 (0.008)	0.919 (0.051)	0.779 (0.052)	1.053 (0.051)
Denmark	0.721 (0.017)	0.540 (0.058)	0.065 (0.013)	1.226 (0.029)	1.173 (0.029)	1.033 (0.029)

Table 5.6: The estimators of C, D statistics and calibration slopes in the internal-external cross validation when dataset ID is used for validation and the remaining studies for derivation. In this case, the study from Sweden is excluded from the model.

In comparison to the aforementioned prognostic model using all the 8 studies, no significant improvement is found in the fitting of the model while Sweden being excluded from the dataset. However, when we summarized the pooled estimate of calibration slope using trivariate meta analysis in Table 5.7, it is interesting to find that for the average and nearest neighbour strategy, the corresponding heterogeneities τ^2 of calibration slope estimates were reduced from 0.046 and 0.074 to 0.017 and 0.018 respectively. It implied that the prognostic model did yield homogeneous estimates in calibration slopes by excluding the study from Sweden. Therefore, we may recommend to exclude the study in Sweden while developing a final prognostic model for future use.

	Summary results from meta analysis				
Intercept Strategy	C Index estimate	D Index estimate	Calibration slope	Multi I^2 test	Between study τ^2
Average	0.710 (0.013)	0.033 (0.009)	0.996 (0.057)	95.7%	0.017
Neighbour	0.707 (0.015)	0.031 (0.009)	0.976 (0.062)	97.2%	0.018
New	0.710 (0.014)	0.031 (0.009)	1.000 (0.019)	89.9%	< 0.001

Table 5.7: The trivariate meta analysis for the three different intercept strategies excluding study from Sweden: The pooled estimators, the multivariate I^2 test and heterogeneity τ^2 in calibration slopes across 8 studies. N.B. the numbers in brackets represent standard errors.

In summary, the cumulative hazard rate function of the final model is given by

$$H(t; \underline{X}) = H_{0j}(t) \exp(\underline{\beta}^T \underline{X})$$

where $H_{0j}(t) = \gamma_{0j} + \gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t)$. The estimates of the coefficient vector $\underline{\beta}$ for the prognostic factor vector $\underline{X}$ and the coefficients γ_1, γ_2 and γ_3 for the baseline hazards are given in Table 5.8. For a patient from one of the country included in our study, the estimate of the study-specified intercept γ_{0j} is also listed in Table 5.8 and the resulted baseline survival functions of each included study in the final model are plotted in Fig 5.4.

So to obtain $S(t)$ for a new individual we need

$$S(t; \underline{X}) = \exp(-H_0(t) \exp(\underline{\beta}^T \underline{X})).$$

For any patient from the included study in the final model, the baseline hazards is from the appropriate country in Figure 5.4. However for any external patient, we could estimate the baseline hazards from the new dataset directly. If the dataset is not available at hand, the average or the neighbour strategy could be utilized. For all patients, $\underline{\beta}^T \underline{X}$ is obtained using $\underline{\beta}^T$ from Table 5.8.

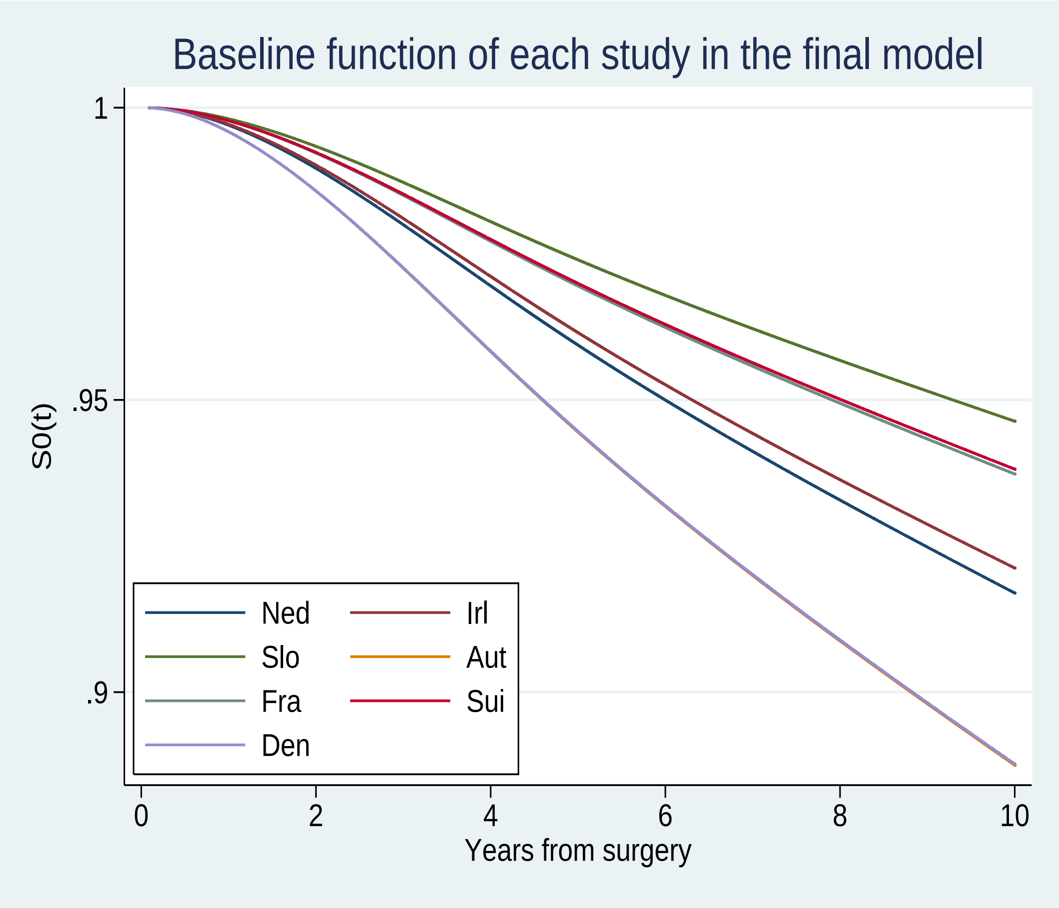

Figure 5.4: The baseline function of each included study in the final model by excluding the study from Sweden and the predictor, menopausal status.

5.4 Discussion

Prognostic models developed from IPD meta analysis are increasingly popular in clinical research[109]. However, very little research has been devoted to how to identify the potential heterogeneities across the studies and how to apply the prognostic model in practice[1].

Variable	**Coef.**	**Std. Err.**	P **value**	**95% CI for HR**
Age	0.010	0.002	< 0.001	(0.006, 0.014)
Tumour type	Reference level:idc			
ilc	-0.122	0.095	0.202	(-0.309, 0.065)
colloid	-0.309	0.711	0.664	(-1.702, 1.085)
tubul	-0.540	0.502	0.282	(-1.524, 0.444)
medull	-0.741	0.222	0.001	(-1.176, -0.305)
papil	-0.422	0.506	0.404	(-1.413, 0.569)
other	-0.086	0.131	0.512	(-0.342, 0.170)
unknown	0.026	0.066	0.698	(-0.104, 0.155)
Tumour grade	Reference level:good			
moderate	0.445	0.162	0.006	(0.126, 0.763)
poor	0.807	0.160	< 0.001	(0.493, 1.121)
unknown	0.668	0.167	< 0.001	(0.340, 0.995)
Lymph nodes	Reference level:np=0			
np< 3	0.738	0.068	< 0.001	(0.606, 0.871)
3 <np< 10	1.241	0.070	< 0.001	(1.103, 1.379)
np> 10	1.670	0.088	< 0.001	(1.498, 1.842)
Tumor size	0.326	0.038	< 0.001	(0.252, 0.401)
Adjuvant treatment	-0.106	0.030	0.001	(-0.166, -0.046)
Hormone receptor	-0.485	0.055	< 0.001	(-0.594, -0.377)
Baseline hazards				
γ_1	0.883	0.019	< 0.001	(0.846, 0.921)
γ_2	0.133	0.018	< 0.001	(0.097, 0.169)
γ_3	0.033	0.010	0.002	(0.012, 0.053)
Study intercept				
Netherland	-3.279	0.207	< 0.001	(-3.684, -2.874)
Ireland	-3.334	0.260	< 0.001	(-3.843, -2.825)
Slovenia	-3.732	0.329	< 0.001	(-4.376, -3.087)
Austria	-2.959	0.268	< 0.001	(-3.485, -2.434)
France	-3.571	0.210	< 0.001	(-3.983, -3.159)
Switzerland	-3.585	0.242	< 0.001	(-4.060, -3.110)
Denmark	-2.961	0.222	< 0.001	(-3.397, -2.526)

Table 5.8: The coefficient and intercept estimates of the final model excluding the study from Sweden and the predictor, menopausal status.

5.4.1 Key findings

In this chapter, a prognostic model was constructed where the stratified intercepts were introduced to account for the heterogeneities in the baseline risks. Particularly when the derived model is generalized into new individuals, three different intercept strategies were proposed to adjust for the heterogeneities in new population. The first method is to meta analyse the intercepts from each individual study in the derivation dataset to produce a pooled estimate for the new population. The second strategy is to utilize the intercept from a study in the derivation dataset that is closest in proximity to the new study. The third method, which always offer the best validation statistics, is to re-estimate the intercept term from the new population directly.

In the cases lacking additional dataset to validate the derived model, the internal-external cross validation framework was proposed to identify the generalizability of our derived model[123]. The additional validation studies provided by the framework make it possible for researchers to gain insight into future calibration and discrimination ability of the newly constructed model and help researchers to identify which study, predictor or intercept strategy (if any) is not suitable to be utilized to construct an empirical model for future use[31].

In this process, to assess generalizability of the derived model in each cycle of the internal-external cross validation approach, several important validation statistics are summarized, *i.e.* calibration measures (calibration slopes and calibration curves), and discrimination measures (C and D indices)[4][25]. The calibration slopes can represent an overall assessment of the derived model in each omitted study but to further identify the potential prediction problems, calibration curve at each risk level should be visually inspected[31][123]. Both C and D statistics describe the ability of the model to discriminate the patients with high risk and low risk but C statistic is more recommended due to its easy application and interpretation[54].

To find the best intercept strategy, the (multivariate) meta-analysis framework of Snell et al.[142] has been used to summarize the estimate of calibration slopes for each strategy[101][162]. Our example demonstrated that the new intercept strategy did outperform the other two in terms of precision accuracy and between-study heterogeneity[133]. Considering the poor calibration slope estimates in Sweden, our final model excludes Sweden. By comparing the validation performance of the final model excluding Sweden to the one including Sweden, we demonstrated that clinically, our final model was more recommended since the heterogeneses of calibration slope estimates across different validation studies were decreased significantly while the data from Sweden being excluded.

5.4.2 Limitations

A potential limitation in our work is that our model does not account for the heterogeneities in predictor-outcome associations. Although the case study implied that the prediction of our model was satisfied in some of countries, the researcher still need to observe the resulting calibration slopes and curves carefully to detect the potential problems[31].

5.4.3 Conclusion

In summary, a framework to develop, implement and evaluate a prognostic model with IPD is introduced in this research. The internal-external cross validation method, which was first proposed by Royson[123], was extended to validate the time-to-event data from multiple studies using Royston-Parmar models. Further, we utilized the meta analysis to summarize the prediction performance of the proposed model within the internal-external cross validation framework and thus determine the final model for future validation and prediction[142].

The main issues and key findings of this chapter are summarized in Table 5.9 and the future work will be discussed in the final chapter. In the next chapter, we will discuss how to evaluate the interaction effects between treatment and certain subgroup of patients in survival analysis.

What is already known on this topic:

- Prognostic models developed from IPD meta analysis are increasingly popular in clinical research and it is important to validate the derived model in new population before it enters practice.

- Calibration measures (calibration slope and calibration curve) and discrimination measures (C and D statistics) are very important statistics to assess generalizability of the derived model.

- Multivariate meta analysis usually provides better parameter estimates than univariate meta analysis since it borrows the strength from other correlated measure to improve the precision of the pooled estimate.

What this study adds

- A prognostic model using the Royston-Parmar framework is constructed from multiple studies where the stratified intercept is utilized to account for the heterogeneities across studies.

- To validate the model in a new study, we propose three different intercept strategies to adjust for the heterogeneities in new individuals.

- A detailed guidance is given to illustrate how to apply the internal-external cross validation framework, derive validation statistics, and how to interpret the validation results in this procedure.

- Based on the internal-external cross validation results, we introduce how to establish an appropriate prognostic model to predict mortality risk in breast cancer patients across a range of countries.

Table 5.9: Summary of the main issues and key findings in Chapter 5.

CHAPTER 6

ESTIMATION OF TREATMENT-COVARIATE INTERACTIONS IN A ONE-STAGE IPD META ANALYSIS WITH TIME-TO-EVENT DATA

6.1 Introduction

The previous chapter looked at the use of survival modelling to predict individual outcome risk. However, another key use of survival analysis is to estimate and predict the effect of new treatments.

In clinical trials, a new medicine may be more effective in some patients than others depending on patients' individual biological or risk characteristics[115][158]. See Section 1.7 for the definition of clinical randomized trials. For example, imatinib was found to be more effective in treating the chronic myeloid leukaemia problem for the patients with epidermal growth factor receptor mutations and trastuzumab improved the outcome of breast cancer in the subgroup of patients with positive human epidermal growth factor receptor[61].

Research to analysis the interaction effect between a patient-level variable and the treatment on outcome is known as stratified medicine and it is one of the fundamental issues in a prognosis research[61]. Fully understanding this topic will benefit clinical decision making[58]. For example, it will help clinical decision makers to determine which patient strata is of the highest priority in clinical trials and more importantly, it could provide important guidance to customize healthcare for each individual patient, which refers to the term 'personalized medicine', to maximize clinical benefits or reduce side-effects.

Now consider a time-to-event dataset from multiple trials, for the purpose to identify the association between the treatment effect and one patient level factor, the conventional approach is to utilize aggregate data meta analysis to produce a pooled estimate of the interaction effect from multiple trials[32]. However, it is well known that the only availability of the trial-level values in aggregate data meta analysis is insufficient to assess the patient level association directly unless the interaction itself is reported[48][130], which is unfortunately rare.

For that, we may overcome this problem by using the IPD from multiple trials[109]. This idea had been investigated mainly in the setting of continuous or binary outcome and here we consider survival outcomes. For example, some researchers developed a single model with an interaction term to investigate the association between the treatment and covariate[112]. Alternatively, others treated the within and across trial interaction effects separately to avoid the well documented problem, 'ecological bias'[151]. However, the estimation of the interaction effect had not been considered in much detail within one-stage IPD meta analysis of time to event outcome. Therefore, to address this issue using time-to-event data, we propose three main objectives in this chapter:

- Examine if separation of within and across terms is important in one-stage IPD meta analysis of time-to-event outcomes.

- Consider situations with and without ecological bias respectively when estimating the treatment-covariate interaction effects.

- Apply the proposed one-stage IPD meta analysis model to a epilepsy data.

This chapter is structured as follows. In the next section, we review the past literature on the evaluation of the treatment-covariate interaction effect in meta analysis. Then in Section 6.3, we describe a range of hierarchical Cox models to assess the interaction effect in IPD meta analysis. We show that the models could be further divided into two groups according to whether they treat the within and across trial interaction terms separately. In Section 6.4, we describe a series of simulation studies where the two different groups of the Cox models were utilized to fit the simulated data. In the process, we highlight the advantages in parameter estimation for the models treating the within and across trials effects individually, especially when confounding is introduced in the simulated data. Section 6.5 applies the various models to a real case study on the effects of two anti-epileptic drugs, and carefully investigates the treatment-covariate interaction for the potential ecological bias. In the last section, several key findings and limitations of the work are discussed.

6.2 Literature review

Many papers have been devoted to investigating the treatment-covariate interaction effects in IPD meta analysis. One common method is to estimate the interaction effect within each trial and then combine the results assuming a common effect across trials[137]. See, for example, Mccleary et al.[91] who explored the impact of older age on the efficacy of newer adjuvant therapies by calculating their interaction effects in each of 6 individual trials first and then pooled them using standard meta-analysis. In general, this two-stage process is same as a traditional aggregate data meta analysis which obtains the effect estimate from each trial with their variance and then calculates a pooled average result across trials using fixed or random effects meta analysis[110].

An alternative method is to meta analyse the IPD from the multiple trials in a one-stage approach. In comparison to the two-stage method, it is more flexible to incorporate all trials simultaneously and estimate the treatment effects and interactions all in one analysis. One typical model includes the treatment, covariate and their interaction terms as the predictor factors. See for example, Tudur Smith et al.[141] investigated a Cox model with the interaction effects between the anti-epileptic drug effects and patients's age by directly including their interaction term into their model. However in binary and continuous outcome settings, other researches suggested this approach may be prone to ecological bias, as it mixed within and between trial interactions[112]. For example, Riley et al.[106] utilized a simulation study to demonstrate that the models ignoring the difference between the within and across trial treatment-covariate effects may give a biased estimator of the patient level effects in IPD meta analysis of binary outcomes. An alternative model can carefully separate within trial and across trial treatment-covariate effects to reduce the problem[130]. For example, Riley et al.[106] and Simmonds et al.[138] specified models to separate patient-level and trial-level interaction terms in one-stage IPD meta analysis of continuous outcomes. Riley et al.[110] also used this model to analyze interactions in a one-stage meta analysis of binary outcomes and Riley et al.[106] generalized the idea into the setting of a binary meta-analysis where the association between the two outcomes, sensitivity and specificity of a diagnostic test and a participant-level covariate was investigated.

However, there is little consideration of how to treat the within and across trial interaction effects separately in the framework of time-to-event data and whether our estimation benefits from it. Hence in this chapter, we consider the issue with IPD from multiple trials with survival outcomes where treatment effects and interactions are of interest.

6.3 Treatment-covariate interaction model

Consider the IPD meta analysis of time-to-event data across $j = 1$ to J trials. Let x_{ij} be a participant-level covariate of interest which can be continuous, such as age, or binary, such as sex, and z_{ij} denotes whether the ith patient in the jth trial is in the experimental group or in the control group (1=Experiment group, 0=Control group).

There are a variety of models available to fit the time-to-event data to randomised trials[141]. The Cox proportional hazards model is a popular one which does not require any assumption regarding the baseline hazard rate. Given in this chapter, we are only interested to identify the treatment-covariate interaction effects on individual risks, the Cox regression model is therefore a reasonable choice to solve the problem[141], as we do not need the baseline hazard.

6.3.1 Cox regression merging interaction terms

The Cox models could include the treatment-covariate interaction term as a prognostic factor to predict individual risks in the IPD meta analysis. For the ith individual in the jth trial, an initial model to predict the hazard rate function at time t can be written as

$$\lambda_{ij}(t) = \lambda_{0j}(t) \exp(\beta_1 z_{ij} + \beta_2 x_{ij} + \beta_T x_{ij} z_{ij}) \tag{6.3.1}$$

where $\lambda_{0j}(t)$ denotes the unique baseline hazard function in the jth trial and $x_{ij}z_{ij}$ represent the interaction term between the treatment and covariate of interest[21]. The constant coefficients β_1 is the change in the log hazard for patients in the treatment group rather than control group where $x_{ij} = 0$, β_2 is the change in the log hazard for a 1 unit increase in the patient level covariate where $z_{ij} = 0$ and β_T denotes the additional change in the log hazard for patients in the new treatment group compared with the control group for one unit increasing values of x_{ij}.

The unobserved trial confounding may arise from an underlying causal mechanism or may be due to artificial difference in measurements or methods (*i.e.* chance, bias, or confounding) across trials. Therefore we could further introduce a random treatment effect into model (6.3.1) to make allowance for potential excessive variation in treatment effects as

$$\lambda_{ij}(t) = \lambda_{0j}(t) \exp(\beta_{1j} z_{ij} + \beta_2 x_{ij} + \beta_T x_{ij} z_{ij}) \tag{6.3.2}$$

where $\beta_{1j} = \beta_1 + b_{1j}$ and $b_{1j} \sim N(0, \tau^2)$. The fixed coefficient β_1 is the average log hazard ratio for a population of possible treatment effects in individuals where $x_{ij} = 0$ and the random variable b_{1j} following a normal distribution $(0, \tau^2)$ is to describe the heterogeneities in the treatment effects across trials[154].

6.3.2 Cox regression separating interaction terms

As discussed previously, when we include the interaction as in (6.3.1) and (6.3.2), it may amalgamate within and across trial interactions[141]. Alternatively, we can model the within-trial relationship by centering the covariate x_{ij} about the mean $\bar{x}_j$ in each trial j and model the across-trial relationship by the mean $\bar{x}_j$ in each trial j[106][110]. In this setting, the hazard rate model with the assumption of the stratified baseline function across trials can be written as

$$\lambda_{ij}(t) = \lambda_{0j}(t)\exp(\beta_1 z_{ij} + \beta_2 x_{ij} + \beta_W(x_{ij} - \bar{x}_j)z_{ij} + \beta_A\bar{x}_j z_{ij}) \qquad (6.3.3)$$

where $\lambda_{0j}(t)$ is the baseline function in the jth trial. As with model (6.3.1), the baseline hazard for each trial is not assumed to be same. β_1 and β_2 are interpreted in the similar way to (6.3.1). Additionally, the within trial coefficient β_W denotes the change in the log hazard rate for individuals who receives the new treatment rather than control for each one unit change in x_{ij} and the across trial coefficient β_A denotes the change in the log hazard rate of individuals who receive the new treatment for a one unit change in $\bar{x}_j$.

The final model considered in this section is about a generalization from (6.3.3) to account for the heterogeneities in trial treatment effects. For the ith patient in the jth trial, it can be written as

$$\lambda_{ij}(t) = \lambda_{0j}(t)\exp(\beta_{1j} z_{ij} + \beta_2 x_{ij} + \beta_W(x_{ij} - \bar{x}_j)z_{ij} + \beta_A\bar{x}_j z_{ij}) \qquad (6.3.4)$$

where $\beta_{1j} = \beta_1 + b_{1j}$ and $b_{1j} \sim N(0, \tau^2)$. The coefficient β_1 is the average log hazard ratio for a population of possible treatment effects in those where $x_{ij} = 0$ and $\bar{x}_j = 0$, and b_{1j} is the deviation of the relative treatment effect in the jth trial from this population average. β_2 is defined same as model (6.3.1) and β_W and β_A are defined same as (6.3.3).

In models (6.3.3) and (6.3.4), the difference between the within trial interaction effect β_A and the across trial interaction effect β_W can be explained as the ecological bias[106][108][110][112]. In practice, if there is truly no ecological bias, then β_A and β_W obtained from model (6.3.3) or (6.3.4) should be same as the amalgamated effect β_T obtained from (6.3.1) or (6.3.2)[104]. Subsequently, models (6.3.3) and (6.3.4) could simplify to (6.3.1) and (6.3.2) respectively and β_T can explain both the within-trial and between-trial variations across trials[106]. However, as suggested by Riley and Steyerberg[112], even when no ecological bias is present, models (6.3.3) and (6.3.4) are more recommended so as to make clear about the potential ecological bias in the underlying database. However, it must be emphasised again that these finding have not yet been confirmed for time-to-event studies (trials).

6.3.3 Modelling fitting

To fit the stratified Cox regression for model (6.3.1) and (6.3.3), many standard statistical packages are available such as coxph in R[46] and stcox in STATA[21]. To estimate the random effects Cox regression in model (6.3.2) and (6.3.4), the coxme package in R could be utilized[154]. In particular, the value of a fixed effect or random effects parameter could be estimated from its usual Cox partial likelihood function using a maximum likelihood algorithm where the random effects could be integrated by partial likelihood function[113][156].

6.3.4 Reasons for ecological bias

We have emphasized the potential importance to separate the within and across trial treatment-covariate terms to account for the potential ecological bias in modelling individual risks[11]. There is rich literature being devoted to explain why we should consider the ecological bias in the meta-analysis[12][111] and why individual risks both depend on the covariate mean and individual covariate value[49]. In this section, to further clarify this idea, we elaborate two main reasons which are very likely to cause the ecological bias[49][112].

If the treatment effect depends on both the patient level covariate and the trial covariate mean with different degrees, then the ecological bias would be more likely to be found in model fitting. For example, for patients with depression, younger people may have a better treatment effect than elders due to better health overall allowing for a better treatment response. But it may also depend on the mean age of the study (population) as well. The clinical trials from a population with an elder mean age might be better for elder patients since they may not feel as lonely, surrounded by others of a similar age.

Another reason for ecological bias is trial-level confounding in treatment. For example, for a new drug firstly applied in clinical trials, the dose may differ across trials. Subsequently, the trials with the larger proportion of male patients may be more likely to give their patients larger dosage of the drug. This may induce a trial-level association between the drug effect and proportion male in the study, but it is due to confounding by dose. That is being male appears to be associated with a better response to treatment at the trial-level, but this is due to male receiving a higher dose, not because they are male.

6.4 Simulation study

We now describe two simulation studies to assess the performance of the models treating the treatment-covariate interaction term as a whole (*i.e.* (6.3.1) and (6.3.2)) and the models separating the within and across trial interaction terms (*i.e.* (6.3.3) and (6.3.4)).

In the first simulation study, we exclude any trial level confounding factor ('No confounding' simulation study). In the second simulation study, we include a confounding factor ('Confounding' simulation study). In each simulation study, we consider binary (sex) or continuous (age) variables and their interaction with treatment. The SURVSIM package in STATA is utilized to simulate a survival data (see Crowther and Lambert[28] for the details) and the main steps of the simulation study are summarized as follows[19],

Step 1. Each simulated IPD meta analysis dataset consists of J trials. The number of patients in each trial was randomly determined by the normal distribution with the mean N and standard error, $N/5$.

Step 2. In each individual trial, each patient has an equal chance to be assigned to the experimental group $z_{ij} = 1$ or the control group $z_{ij} = 0$ randomly.

Step 3a. If the covariate x is the binary variable, such as sex (1=Male, 0=Female), then for the ith patient in the jth trial, we firstly sample the mean of x_j in the jth trial from a uniform distribution $(0.5 - V1, 0.5 + V1)$ where $V1$ is chosen to be between 0 and 0.5 and next randomly sample x_{ij} from a binomial distribution with the obtained mean x_j.

If the covariate x is the continuous variable, age, then for the ith patient in the jth trial, the mean of x_j in the jth trial is firstly sampled from a uniform distribution $(50 - V1, 50 + V1)$ where $V1$ is chosen to be between 0 and 35 and then x_{ij} is sampled from a normal distribution truncated at 15 and 85 with the obtained mean x_j and a standard error $V2$, where $V2$ is chosen to be a positive number.

Step 3b. In addition, for the simulation study with confounding, we define y_j to indicate whether the extra dose of drugs is given to the patients in the experimental group in the jth clinical trial (1=yes, 0=no). In this framework, the trials with the mean of the binary covariate (sex) above 0.5 or the mean of continuous covariate (age) above 50 are given an extra drug effect β_4 ($y_j = 1$).

Step 4. We can generate the survival data for the 'no confounding' simulation study using (6.4.1a) and the 'confounding' simulation study using (6.4.1b) respectively:

$$\lambda_{ij}(t) = \lambda_0(t) \exp(\beta_{0j} + \beta_1 z_{ij} + \beta_2 x_{ij} + \beta_3 x_{ij} z_{ij}) \tag{6.4.1a}$$

$$\lambda_{ij}(t) = \lambda_0(t) \exp(\beta_{0j} + \beta_1 z_{ij} + \beta_2 x_{ij} + \beta_3 x_{ij} z_{ij} + \beta_4 y_j z_{ij}) \tag{6.4.1b}$$

where the baseline hazards within each trial are proportional to the same common hazard function $\lambda_0(t)$, which is taken to be the exponential distribution with mean 0.1. The fixed term β_{0j} for $j = 1, 2, ..., J$ represents the uniqueness of the baseline hazards within each trial where β_{0j} is sampled from the uniform distribution $U(0, 0.5)$ and β_1, β_2 and β_3 are chosen to be the fixed constants. The fixed term β_4 is defined for the confounding factor in the 'confounding' simulation which is chosen to be a positive constant.

Step 5. By choosing the proper values of J, N, V_1 and V_2, Step 1- Step 4 are repeated 1000 times to generate 1000 meta-analysis datasets for the following scenarios:

- 'No confounding' simulation study: Binary variable (sex)
- 'No confounding' simulation study: Continuous variable (age).
- 'Confounding' simulation study: Binary variable (sex).
- 'Confounding' simulation study: Continuous variable (age)

Step 6. To each 1000 meta-analysis IPD generated, we fit to the simulated patient data two types of Cox models, one of which ignores the ecological bias (*i.e.* (6.3.1) or (6.3.2)) and the other accounts for the bias (*i.e.* (6.3.3) or (6.3.4)). Then to evaluate and compare the 1000 achieved parameter estimates from the two different types of models, we look at the mean bias, standard error, mean squared errors and coverage probability of 95% confidence interval of the estimates respectively. Of course, we are especially interested in the interaction terms.

6.4.1 Parameter choices

We need to define β_1, β_2, β_3 and β_4 in (6.4.1) to generate the meta-analysis database for each simulation study. To obtain a simulated database with the reasonable proportion of events and censored cases, β_1 was set to be 1. β_2 and β_3 were defined to be 0.5 for the binary covariate (sex) and 0.01 for the continuous covariate (age). β_4 was set to be 0 in the first simulation study and 0.75 in the second study to account for the unobservable trial confounding.

Besides, the selection of the values of J, N, V_1 and V_2 also influences the power of interaction estimates in meta-analysis[69][138]. To investigate the possible associations between the sample size of the database and the performance of the Cox models, two settings were defined according to the number of trials and the number of observations per trial, that is, $J = 10$ and $N = 500$ for the 'large' setting, and $J = 5$ and $N = 250$ for the ‘small’ setting. To explore the association between the scale of the covariate x and interaction effects, we also varied $V1$ and $V2$ in our study: For the binary case, $V1$ was chosen to be 0.4 or 0.2 and for the continuous case, $V1$ was set to be 20 or 10 and $V2$ was set to be 5 or 10. In summary, under each scenario listed in Step 6, we simulated 1000 meta analysis datasets for each combination of $V1$, $V2$ and the sample size (J and N) and then fitted the achieved data with the proper Cox models.

6.4.2 Result 1: binary covariate, no trial confounding

In the ‘no confounding’ study with the binary covariate, sex, we applied models (6.3.1) and (6.3.3) to fit the data respectively. Since no confounding factor was considered in this case, there is no unexplained heterogeneity across trials and so the random effects models (6.3.2) and (6.3.4) are not considered. The estimated results for the covariate sex with the different combinations of $V1$ and the sample size are given in Table 6.1.

Sample Size	V_1	**Equation (6.3.1) (amalgamated interaction)**					**Equation (6.3.3) (separate interaction)**							
		Mean (s.d.)			**MSE**	**Coverage**	**Mean (s.d.)**				**MSE**		**Coverage**	
		β_1	β_2	β_T	β_T	β_T	β_1	β_2	β_W	β_A	β_W	β_A	β_W	β_A
Large	0.4	1.001	0.501	0.500	0.005	0.939	1.002	0.501	0.500	0.500	0.007	0.027	0.945	0.957
		(0.053)	(0.059)	(0.072)			(0.093)	(0.064)	(0.082)	(0.164)				
Large	0.2	1.001	0.501	0.501	0.005	0.946	1.006	0.501	0.502	0.490	0.005	0.112	0.953	0.956
		(0.053)	(0.058)	(0.070)			(0.173)	(0.058)	(0.071)	(0.335)				
Small	0.4	1.004	0.503	0.494	0.020	0.958	0.998	0.505	0.492	0.517	0.023	0.235	0.964	0.945
		(0.111)	(0.118)	(0.143)			(0.271)	(0.121)	(0.153)	(0.484)				
Small	0.2	0.998	0.502	0.505	0.019	0.953	1.001	0.502	0.505	0.497	0.020	0.707	0.948	0.959
		(0.103)	(0.108)	(0.138)			(0.445)	(0.11)	(0.143)	(0.841)				

Table 6.1: The estimates of the treatment-sex interaction effects in the simulated data without trial level confounding. N.B. the true values of $\beta_1 = 1$, $\beta_2 = 0.5$, $\beta_T = \beta_W = \beta_A = 0.5$, coverage is for the 95% confidence interval and the numbers in the brackets denote standard deviation of repeated 1000 parameter estimates.

In all settings, $\hat{\beta}_T$ from (6.3.1) and $\hat{\beta}_W$ and $\hat{\beta}_A$ from (6.3.3) were approximately unbiased estimates of the true treatment-sex interaction effect and their coverage probabilities of 95% confidence interval were also very close to the true value 0.95. For equation (6.3.3), the mean squared errors of $\hat{\beta}_W$ were generally much smaller than those of $\hat{\beta}_A$. This highlights that the within trial interaction term usually has greater power than its across trial counterpart, and this difference becomes to be bigger as the sample size or $V1$ decreases. However, $\hat{\beta}_T$ from equation (6.3.1) has the smallest mean squared errors, as it is essentially a weighted combination of $\hat{\beta}_W$ and $\hat{\beta}_A$.

6.4.3 Result 2: continuous covariate, no trial confounding

In the similar way, we generated the data with the continuous variable, age for the different combinations of $V1$, $V2$ and the sample size, and then fitted the simulated data by models (6.3.1) and (6.3.3) respectively. The corresponding results are summarized in Table 6.2.

Sample Size	V_1	V_2	Equation (6.3.1) (amalgamated interaction)					Equation (6.3.3) (separate interaction)							
			Mean (s.d.)			MSE	Coverage	Mean (s.d.)				MSE		Coverage	
			β_1	β_2	β_T	β_T	β_T	β_1	β_2	β_W	β_A	β_W	β_A	β_W	β_A
Large	20	10	0.996 (0.127)	0.010 (0.002)	0.010 (0.002)	<0.001	0.952	0.989 (0.179)	0.010 (0.003)	0.010 (0.003)	0.010 (0.004)	<0.001	<0.001	0.948	0.949
Large	20	5	0.994 (0.148)	0.010 (0.004)	0.010 (0.003)	<0.001	0.958	0.992 (0.171)	0.010 (0.005)	0.010 (0.006)	0.010 (0.003)	<0.001	<0.001	0.946	0.959
Large	10	10	1.006 (0.144)	0.010 (0.002)	0.010 (0.003)	<0.001	0.965	0.994 (0.350)	0.010 (0.002)	0.010 (0.003)	0.010 (0.007)	<0.001	<0.001	0.961	0.949
Large	10	5	0.996 (0.234)	0.010 (0.004)	0.010 (0.005)	<0.001	0.949	0.992 (0.346)	0.010 (0.005)	0.010 (0.006)	0.010 (0.007)	<0.001	<0.001	0.959	0.948
Small	20	10	1.008 (0.274)	0.010 (0.005)	0.010 (0.005)	<0.001	0.930	0.988 (0.469)	0.010 (0.005)	0.010 (0.007)	0.010 (0.009)	<0.001	<0.001	0.954	0.935
Small	20	5	1.014 (0.320)	0.010 (0.007)	0.010 (0.006)	<0.001	0.967	1.013 (0.421)	0.011 (0.010)	0.009 (0.013)	0.010 (0.008)	<0.001	<0.001	0.953	0.960
Small	10	10	0.991 (0.316)	0.010 (0.005)	0.010 (0.006)	<0.001	0.945	0.943 (0.921)	0.010 (0.005)	0.010 (0.007)	0.011 (0.018)	<0.001	<0.001	0.947	0.953
Small	10	5	0.993 (0.491)	0.010 (0.009)	0.010 (0.010)	<0.001	0.949	0.984 (0.890)	0.010 (0.010)	0.010 (0.013)	0.010 (0.018)	<0.001	<0.001	0.949	0.952

Table 6.2: The estimates of the treatment-age interaction effects in the simulated data without trial level treatment confounding. N.B. the true values of $\beta_1 = 1$, $\beta_2 = 0.01$, $\beta_T = \beta_W = \beta_A = 0.01$, coverage is for the 95% confidence interval and the numbers in the brackets denote standard deviation of repeated 1000 parameter estimates.

The amalgamated effect, $\hat{\beta}_T$ from model (6.3.1) and the within and across trial effects, $\hat{\beta}_W$ and $\hat{\beta}_A$ from model (6.3.3) were generally unbiased as they were close to 0.01 across all settings. In addition, the resulted coverage in each setting was also very close to true value, 0.95.

In the database with large sample size when using (6.3.3), the standard error of the within and across trial estimators were very similar, for example, see the cases for $V1 = 10, V2 = 5$ or $V1 = 20, V2 = 10$. However, when $V1$ was large relative to $V2$, the standard error of $\hat{\beta}_W$ appeared slightly larger than $\hat{\beta}_A$. For example, given $V1 = 20$ and $V2 = 5$, the standard error of $\hat{\beta}_A$ was always around 0.006 where the standard error of $\hat{\beta}_W$ was around 0.003. Conversely, when $V1$ was small relative to $V2$, the standard error of $\hat{\beta}_W$ turned out to be smaller than $\hat{\beta}_A$. For example, given $V1 = 10$ and $V2 = 10$, the standard error of $\hat{\beta}_A$ was always around 0.003 while $\hat{\beta}_W$ was almost 0.007. These findings showed a very important property in IPD meta analysis, that is, the power to detect the patient level interaction effects using $\hat{\beta}_W$ increases when $V2$ increases, and when using $\hat{\beta}_A$ it increases when $V1$ increases[138]. As for the simulations with small sample size, findings were similar except standard errors were of a larger magnitude throughout.

6.4.4 Result 3: binary covariate, trial-level confounding

We extended the simulation study to incorporate the trial-level confounding factor. From (6.4.1b), the coefficient of the confounding factor, β_4 was set to be 0.75 indicating that the extra dose increased the log hazards of patients who received the treatment by 0.75. The confounding factor was assumed to be unobservable here (*i.e.* an unknown trial-level confounder). We fitted the simulated data by the fixed effect stratified Cox regression models (6.3.1) and (6.3.3) and then models (6.3.2) and (6.3.4) with random treatment effects to account for the heterogeneities across trials (caused by the unobserved confounder). Using the different combinations of $V1$, $V2$ and the sample size, we generated 1000 IPD meta analysis repeatedly for the covariate, sex, and then fitted the data by models (6.3.1) and (6.3.2) and models (6.3.3) and (6.3.4) respectively. The obtained results are summarized in Table 6.3.

Consider the fixed and random effects models (6.3.3) and (6.3.4) which treated the within and across trial interaction terms separately. The patient level interaction estimators $\hat{\beta}_W$ were still approximately unbiased for all settings as they were very close to the true value, 0.5. However, due to the confounding, $\hat{\beta}_A$ were clearly biased in every setting. For example, given $V1 = 0.2$ and the small sample size, the within trial interaction estimator $\hat{\beta}_W = 0.502$ (s.e.=0.136) from the random effects model (6.3.3) was close to the truth, whereas its across trial interaction estimator $\hat{\beta}_A = 3.47$ was seriously upwards biased with a very big uncertainty, s.e.=1.537. The stark differences between the estimators of $\hat{\beta}_W$ and $\hat{\beta}_A$ demonstrated that there was strong ecological bias in the simulated datasets, as expected due to the unaccounted for trial level confounder of dose. Interestingly, the bias was not reduced in $\hat{\beta}_A$ when using the random effects model (6.3.4) rather than the fixed effect model (6.3.3). In views of the MSE and coverage of $\hat{\beta}_W$ and $\hat{\beta}_A$, it is more clear that the estimation of models (6.3.3) and (6.3.4) reached a high precision in $\hat{\beta}_W$ where the MSE were always small then 0.03 and coverage were very close to 0.95. However, very poor MSE and coverage of $\hat{\beta}_A$ were reported due to the presence of ecological bias.

Models (6.3.1) and (6.3.2) also gave estimates of β_T, those were upwardly biased compared to 0.5. The random effects model (6.3.2) performed better in terms of the coverage of which were close to 0.95, but $\hat{\beta}_T$ was still upwards biased in most settings. For example, given the large sample size and $V1 = 0.4$, the estimate of β_T was 0.528 (s.e.=0.079, coverage=0.927) for the random effects model and 0.721 (s.e.=0.091, coverage=0.192) for the fixed effect model. This is due to the amalgamation of the unbiased within trial interaction with upwardly biased across trial interaction. The MSE of $\hat{\beta}_T$ were still acceptable which were always below 0.07, however, they were based on the biased estimates of the interactions[106].

Sample Size	V_1	**Model**	**Mean (s.d.)**				**MSE**	**Coverage**	**Model**	**Mean (s.d.)**					**MSE**		**Coverage**	
			β_1	β_2	β_T	τ_1	β_T	β_T		β_1	β_2	β_W	β_A	τ_2	β_W	β_A	β_W	β_A
Large	0.4	(6.3.1)	1.245 (0.121)	0.350 (0.069)	0.721 (0.091)		0.057	0.192	(6.3.3)	0.634 (0.174)	0.499 (0.062)	0.495 (0.078)	1.970 (0.320)		0.006	2.263	0.944	0.003
Large	0.4	(6.3.2)	1.362 (0.126)	0.482 (0.062)	0.528 (0.079)	0.125 (0.034)	0.007	0.927	(6.3.4)	0.640 (0.172)	0.500 (0.062)	0.499 (0.078)	1.967 (0.317)	0.028 (0.019)	0.006	2.252	0.947	0.008
Large	0.2	(6.3.1)	1.310 (0.132)	0.425 (0.058)	0.599 (0.076)		0.016	0.696	(6.3.3)	-0.074 (0.321)	0.500 (0.056)	0.496 (0.071)	3.400 (0.638)		0.005	8.816	0.943	0.000
Large	0.2	(6.3.2)	1.373 (0.130)	0.493 (0.056)	0.510 (0.071)	0.127 (0.034)	0.005	0.947	(6.3.4)	-0.071 (0.320)	0.501 (0.056)	0.499 (0.071)	3.401 (0.634)	0.028 (0.017)	0.005	8.815	0.946	0.006
Small	0.4	(6.3.1)	1.273 (0.211)	0.371 (0.134)	0.694 (0.176)		0.069	0.683	(6.3.3)	0.645 (0.428)	0.505 (0.128)	0.490 (0.162)	1.977 (0.749)		0.026	2.741	0.949	0.107
Small	0.4	(6.3.2)	1.363 (0.209)	0.471 (0.132)	0.547 (0.168)	0.105 (0.069)	0.03	0.934	(6.3.4)	0.647 (0.426)	0.506 (0.128)	0.492 (0.162)	1.978 (0.747)	0.019 (0.028)	0.026	2.742	0.949	0.196
Small	0.2	(6.3.1)	1.310 (0.201)	0.424 (0.112)	0.597 (0.139)		0.029	0.897	(6.3.3)	-0.110 (0.801)	0.493 (0.108)	0.501 (0.135)	3.463 (1.543)		0.018	11.159	0.962	0.086
Small	0.2	(6.3.2)	1.363 (0.198)	0.480 (0.109)	0.524 (0.136)	0.111 (0.065)	0.019	0.958	(6.3.4)	-0.111 (0.796)	0.494 (0.108)	0.502 (0.136)	3.470 (1.537)	0.018 (0.027)	0.018	11.184	0.963	0.157

Table 6.3: The estimators of the treatment-sex interaction effects in the simulated data considering trial-level treatment confounding. N.B. the true values of $\beta_1 = 1$, $\beta_2 = 0.5$, $\beta_T = \beta_W = \beta_A = 0.5$, coverage is for the 95% confidence interval and the numbers in the brackets denote standard deviation of repeated 1000 parameter estimates.

6.4.5 Result 4: continuous covariate, trial-level confounding

The results of IPD meta analysis for the continuous variable, age, are summarized in Table 6.4.

In all settings, the mean and coverage of $\hat{\beta}_W$ were close to 0.01 and 0.95 respectively indicating that the estimates of the within trial interaction effects from models (6.3.3) and (6.3.4) were unbiased. On the contrary, the across-trial association estimates from the two models were significantly larger than the within-trial counterparts and the coverage were far away from the optimal value 0.95, highlighting again the ecological bias. Now consider that we wrongly fit the data with (6.3.1) and (6.3.2) that did not account for ecological bias. The pooled estimator and coverage of $\hat{\beta}_T$ achieved were again biased by the unobservable across-trial confounding factor, especially when the fixed effect model was utilized. Though the standard error of $\hat{\beta}_T$ was sometimes smaller than $\hat{\beta}_W$, this only arose by utilizing the biased $\hat{\beta}_A$. In a sense, gain in standard error comes at the expense of bias[106] and thus the MSE of $\hat{\beta}_T$ was always larger than $\hat{\beta}_W$.

6.4.6 Summary of simulation findings

In conclusion, our simulation study has demonstrated that to understand how patient-level covariate interacts with the treatment effect, it is better to examine the patient level interaction effect β_W rather than either the trial level interaction effect β_A, or the amalgamated interaction effect β_T, because ecological bias can seriously bias β_T and β_A[106][151]. Therefore separation into β_W and β_A is important for one-stage IPD meta analysis of time-to-event outcomes, as otherwise important predictors of treatment response may be missed or false predictors of treatment identified wrongly.

Sample Size	V_1	V_2	Model	Mean (s.d.)				MSE	Coverage	Model	Mean (s.d.)					MSE		Coverage	
				β_1	β_2	β_T	τ	β_T	β_T		β_1	β_2	β_W	β_A	τ	β_W	β_A	β_W	β_A
Large	20	10	(6.3.1)	0.687 (0.163)	0.002 (0.002)	0.023 (0.003)		<0.001	0.018	(6.3.3)	-0.158 (0.317)	0.010 (0.003)	0.010 (0.003)	0.040 (0.006)		<0.001	0.001	0.958	0.003
Large	20	10	(6.3.2)	1.239 (0.204)	0.008 (0.003)	0.013 (0.003)	0.112 (0.040)	<0.001	0.848	(6.3.4)	-0.155 (0.312)	0.010 (0.003)	0.010 (0.003)	0.041 (0.006)	0.027 (0.019)	<0.001	0.001	0.958	0.010
Large	20	5	(6.3.1)	0.270 (0.199)	-0.003 (0.004)	0.032 (0.004)		<0.001	0.003	(6.3.3)	-0.094 (0.310)	0.010 (0.005)	0.010 (0.006)	0.039 (0.006)		<0.001	0.001	0.965	0.002
Large	20	5	(6.3.2)	0.820 (0.339)	0.003 (0.005)	0.021 (0.006)	0.068 (0.043)	<0.001	0.418	(6.3.4)	-0.092 (0.307)	0.010 (0.005)	0.010 (0.006)	0.039 (0.006)	0.025 (0.018)	<0.001	0.001	0.964	0.005
Large	10	10	(6.3.1)	0.851 (0.199)	0.004 (0.003)	0.020 (0.003)		<0.001	0.131	(6.3.3)	-1.523 (0.591)	0.010 (0.002)	0.010 (0.003)	0.068 (0.012)		<0.001	0.003	0.949	0.003
Large	10	10	(6.3.2)	1.308 (0.204)	0.009 (0.003)	0.011 (0.003)	0.121 (0.037)	<0.001	0.927	(6.3.4)	-1.529 (0.587)	0.010 (0.002)	0.010 (0.003)	0.068 (0.012)	0.026 (0.018)	<0.001	0.004	0.946	0.005
Large	10	5	(6.3.1)	0.103 (0.298)	-0.005 (0.005)	0.035 (0.006)		0.001	0.018	(6.3.3)	-1.518 (0.576)	0.010 (0.005)	0.010 (0.007)	0.068 (0.012)		<0.001	0.003	0.945	0.001
Large	10	5	(6.3.2)	1.121 (0.370)	0.007 (0.005)	0.015 (0.007)	0.109 (0.041)	<0.001	0.826	(6.3.4)	-1.520 (0.576)	0.010 (0.005)	0.010 (0.007)	0.068 (0.012)	0.026 (0.019)	<0.001	0.003	0.945	0.004
Small	20	10	(6.3.1)	0.764 (0.377)	0.003 (0.005)	0.022 (0.007)		<0.001	0.38	(6.3.3)	-0.183 (0.891)	0.010 (0.006)	0.010 (0.007)	0.041 (0.017)		<0.001	0.001	0.942	0.121
Small	20	10	(6.3.2)	1.128 (0.437)	0.007 (0.006)	0.015 (0.008)	0.085 (0.078)	<0.001	0.778	(6.3.4)	-0.180 (0.889)	0.010 (0.006)	0.010 (0.007)	0.041 (0.017	0.015 (0.026)	<0.001	0.001	0.943	0.184
Small	20	5	(6.3.1)	0.340 (0.508)	-0.003 (0.009)	0.030 (0.010)		0.001	0.173	(6.3.3)	-0.100 (0.881)	0.009 (0.010)	0.010 (0.013)	0.039 (0.017)		<0.001	0.001	0.939	0.115
Small	20	5	(6.3.2)	0.658 (0.620)	0.001 (0.009)	0.024 (0.012)	0.050 (0.073)	<0.001	0.464	(6.3.4)	-0.099 (0.875)	0.009 (0.010)	0.010 (0.013)	0.039 (0.017)	0.017 (0.028)	<0.001	0.001	0.939	0.185
Small	10	10	(6.3.1)	0.898 (0.375)	0.005 (0.005)	0.019 (0.007)		<0.001	0.665	(6.3.3)	-1.693 (1.619)	0.010 (0.005)	0.010 (0.006)	0.071 (0.032)		<0.001	0.005	0.953	0.103
Small	10	10	(6.3.2)	1.246 (0.379)	0.009 (0.005)	0.012 (0.007)	0.106 (0.069)	<0.001	0.917	(6.3.4)	-1.697 (1.611)	0.010 (0.005)	0.010 (0.006)	0.071 (0.032)	0.019 (0.030)	<0.001	0.005	0.952	0.179
Small	10	5	(6.3.1)	0.235 (0.639)	-0.003 (0.010)	0.032 (0.012)		0.001	0.354	(6.3.3)	-1.577 (1.597)	0.011 (0.010)	0.009 (0.013)	0.069 (0.031)		<0.001	0.004	0.961	0.109
Small	10	5	(6.3.2)	0.926 (0.703)	0.005 (0.010)	0.019 (0.014)	0.086 (0.075)	<0.001	0.808	(6.3.4)	-1.581 (1.595)	0.011 (0.010)	0.009 (0.013)	0.069 (0.031	0.016 (0.027)	<0.001	0.004	0.961	0.185

Table 6.4: The estimators of the treatment-age interaction effects in the simulated data considering trial-level treatment confounding. N.B. the true values of $\beta_1 = 1$, $\beta_2 = 0.01$, $\beta_T = \beta_W = \beta_A = 0.01$, coverage is for the 95% confidence interval and the numbers in the brackets denote standard deviation of repeated 1000 parameter estimates.

6.5 Real case study

6.5.1 Background

Consider now evaluating treatment-covariate interaction effects in a real case study. Tudur Smith et al.[141] conducted a systematic review of randomized controlled trials about the effects of two anti-epileptic drugs, Sodium Valproate (drug=1) and Carbamazepine (drug=0) which were mainly used as monotherapy in patients with partial onset seizures or generalized onset seizures[90]. In the review, IPD for 1225 patients from 5 clinical trials were collected totally. See the summary of the dataset in Table 6.5.

Var	Type	Description
Drug	Categorical	Anti-epileptic drugs: Sodium Valproate or Carbamazepine
Sezure time	Continuous	Time to first seizure post-randomization (in days)
Scens	Categorical	Indicator whether the first seizure time was censored
Remission time	Continuous	Time to 12 month remission (in days)
Rcens	Categorical	Indicator whether 12 month remission time was censored
Trial No.	Categorical	The number of the trial
Age	Continuous	Age at randomisation (in years)
Epilepsy type	Categorical	Seizure type of epilepsy (generalized or partial)
Log seizures	Continuous	log number of seizures in 6 months before randomisation

Table 6.5: Variables in the epilepsy dataset with brief description.

Three patient-level characteristics variables of interest were age at randomisation (in years), seizure type of epilepsy (generalized or partial) and the log number of seizures in 6 months before randomisation. Two outcomes, time to 12 month remission and time to first seizure post-randomization were also of interest. Thus, in total there were six different pairs of the outcome and covariate to be investigated in this study. *i.e.* 6 treatment-covariate interaction terms were of interest. For each covariate separately, we applied models (6.3.1) and (6.3.2) which ignore potential ecological bias, and models (6.3.3) and (6.3.4) which remove ecological bias. In their original analysis of this data, Tudur Smith et al. only considered models like (6.3.1) and (6.3.2). Thus our work adds important new evaluation of this data as ecological bias may have been affecting the previous conclusions. Table 6.6 summarizes the descriptive statistics of the underlying data and for further details, please see Tudur Smith et al.[141]. The fitting of the Cox models in this section were conducted by Coxme package in R using the maximum likelihood method introduced in Section 6.3

Variable	**Mean**	**s.d.**	**median**	**Min**	**Max**	**Censored**
Trial 1	n=122					
Drug	0.50	0.50	0.50	0.00	1.00	
Seizure time	609.71	976.64	145.50	7.00	4520.00	36
Remission time	691.73	828.82	367.50	7.00	4614.00	41
Age	30.65	14.85	24.91	13.06	69.76	
Epilepsy type	0.40	0.49	0.00	0.00	1.00	
Log seizure	1.27	1.31	0.69	0.00	5.87	
Trial 2	n=103					
Drug	0.48	0.50	0.0	0.00	1.00	
Seizure time	452.53	824.60	63.0	6.00	4070.00	8
Remission time	1001.32	966.17	559.0	281.00	4544.00	15
Age	10.19	3.73	10.19	2.86	15.95	
Epilepsy type	0.52	0.50 1.0	0.00	1.00		
Log seizure	1.79	1.66	1.1	0.00	6.80	
Trial 3	n=288					
Drug	0.50	0.50	0.00	0.00	1.00	
Seizure time	478.15	585.57	135.92	0.35	2348.00	90
Remission time	563.03	338.57	387.75	22.00	2164.00	64
Age	33.32	15.06	31.00	16.00	79.00	
Epilepsy type	0.49	0.50	0.00	0.00	1.00	
Log seizure	1.67	1.01	1.39	0.69	4.62	
Trial 4	n=246					
Drug	0.48	0.50	0.00	0.00	1.00	
Seizure time	371.47	438.32	134.58	0.17	1520.00	59
Remission time	596.65	289.24	463.00	60.00	1400.00	63
Age	10.09	2.91	10.11	4.94	15.96	
Epilepsy type	0.42	0.49	0.00	0.00	1.00	
Log seizure	1.56	1.17	1.10	0.00	4.64	
Trial 5	n=466					
Drug	0.51	0.50	0.00	0.00	1.00	
Seizure time	266.19	446.14	49.00	0.12	1832.00	168
Remission time	357.92	302.97	365.00	4.00	1833.00	275
Age	47.21	16.20	44.48	18.32	83.33	
Epilepsy type	1.00	0.00	1.00	1.00	1.00	
Log seizure	3.01	2.08	2.48	0.00	7.72	

Table 6.6: Descriptive statistics of the epilepsy data by clinical trial (s.d.=standard deviation).

6.5.2 Summary of results

We summarize the parameter estimates in Table 6.7 and focus in great detail on the interaction effect estimates in Table 6.8.

Outcome	Covariate	Model	Parameter Estimate (s.e.)				Model	Parameter Estimate (s.e.)				
			β_{drug}	β_{cov}	β_T	τ		β_{drug}	β_{cov}	β_W	β_A	τ
Time to 12 month remission	Age at randomisation	(6.3.1)	0.199 (0.129)	0.008^{**} (0.003)	-0.011^{***} (0.004)		(6.3.3)	0.269^{*} (0.158)	0.006 (0.004)	-0.007 (0.006)	-0.013^{***} (0.005)	
		(6.3.2)	0.199 (0.129)	0.008^{**} (0.003)	-0.011^{***} (0.004)	0.004	(6.3.4)	0.269^{*} (0.158)	0.006 (0.004)	-0.007 (0.006)	-0.013^{***} (0.005)	0.004
	Epilepsy type	(6.3.1)	-0.035 (0.113)	-0.238^{**} (0.110)	-0.128 (0.147)		(6.3.3)	0.168 (0.197)	-0.287^{**} (0.118)	-0.026 (0.168)	-0.467 (0.307)	
		(6.3.2)	-0.039 (0.132)	-0.256^{**} (0.113)	-0.090 (0.156)	0.136	(6.3.4)	0.186 (0.239)	-0.287^{**} (0.118)	-0.025 (0.168)	-0.479 (0.376)	0.106
	Log number of seizures	(6.3.1)	-0.037 (0.121)	-0.166^{***} (0.040)	-0.025 (0.056)		(6.3.3)	0.112 (0.245)	-0.171^{***} (0.041)	-0.014 (0.058)	-0.100 (0.122)	
		(6.3.2)	-0.030 (0.131)	-0.168^{***} (0.040)	-0.020 (0.057)	0.104	(6.3.4)	0.134 (0.285)	-0.171^{***} (0.041)	-0.013 (0.058)	-0.105 (0.142)	0.096
Time to first seizure post-randomization	Age at randomisation	(6.3.1)	-0.194 (0.123)	-0.013^{***} (0.003)	0.009^{***} (0.003)		(6.3.3)	-0.205 (0.152)	-0.013^{***} (0.004)	0.009^{*} (0.005)	0.010^{**} (0.004)	
		(6.3.2)	-0.194 (0.123)	-0.013^{***} (0.003)	0.009^{***} (0.003)	0.003	(6.3.4)	-0.205 (0.152)	-0.013^{***} (0.004)	0.009^{*} (0.005)	0.010^{**} (0.004)	0.003
	Epilepsy type	(6.3.1)	-0.140 (0.120)	0.332^{***} (0.110)	0.339^{**} (0.146)		(6.3.3)	-0.176 (0.184)	0.344^{***} (0.119)	0.316^{*} (0.170)	0.396 (0.264)	
		(6.3.2)	-0.142 (0.125)	0.336^{***} (0.112)	0.337^{**} (0.151)	0.070	(6.3.4)	-0.184 (0.205)	0.347^{***} (0.119)	0.316^{*} (0.171)	0.405 (0.301)	0.070
	Log number of seizures	(6.3.1)	-0.031 (0.107)	0.242^{***} (0.028)	0.040 (0.036)		(6.3.3)	-0.248 (0.223)	0.254^{***} (0.030)	0.018 (0.041)	0.147 (0.104)	
		(6.3.2)	-0.037 (0.120)	0.246^{***} (0.029)	0.034 (0.038)	0.110	(6.3.4)	-0.259 (0.247)	0.254^{***} (0.030)	0.019 (0.041)	0.150 (0.117)	0.070

Table 6.7: Summary of the parameter estimates in the epilepsy data. N.B. * for P value< 0.1, ** for P value< 0.05 and *** for P value<0.01.

Outcome	Covariate	Model	$\hat{\beta}_T$	CI of $\hat{\beta}_T$	Model	$\hat{\beta}_W$	$\hat{\beta}_A$	CI of $\hat{\beta}_W$	CI of $\hat{\beta}_A$
Time to 12 month remission	Age at randomisation	(6.3.1)	-0.011*** (0.004)	-0.019 to -0.003	(6.3.3)	-0.007 (0.006)	-0.013*** (0.005)	-0.019 to -0.005	-0.023 to -0.003
		(6.3.2)	-0.011*** (0.004)	-0.019 to -0.003	(6.3.4)	-0.007 (0.006)	-0.013*** (0.005)	-0.019 to -0.005	-0.023 to -0.003
	Epilepsy type	(6.3.1)	-0.128 (0.147)	-0.416 to 0.160	(6.3.3)	-0.026 (0.168)	-0.467 (0.307)	-0.355 to 0.303	-1.069 to 0.135
		(6.3.2)	-0.090 (0.156)	-0.396 to 0.216	(6.3.4)	-0.025 (0.168)	-0.479 (0.376)	-0.354 to 0.304	-1.216 to 0.258
	Log number of seizures	(6.3.1)	-0.025 (0.056)	-0.135 to 0.085	(6.3.3)	-0.014 (0.058)	-0.100 (0.122)	-0.128 to 0.100	-0.339 to 0.139
		(6.3.2)	-0.020 (0.057)	-0.132 to 0.092	(6.3.4)	-0.013 (0.058)	-0.105 (0.142)	-0.127 to 0.101	-0.383 to 0.173
Time to first seizure post-randomization	Age at randomisation	(6.3.1)	0.009*** (0.003)	0.003 to 0.015	(6.3.3)	0.009* (0.005)	0.010** (0.004)	-0.001 to 0.019	0.002 to 0.018
		(6.3.2)	0.009*** (0.003)	0.003 to 0.015	(6.3.4)	0.009* (0.005)	0.010** (0.004)	-0.001 to 0.019	0.002 to 0.018
	Epilepsy type	(6.3.1)	0.339** (0.146)	0.053 to 0.625	(6.3.3)	0.316* (0.170)	0.396 (0.264)	-0.017 to 0.649	-0.121 to 0.913
		(6.3.2)	0.337** (0.151)	0.041 to 0.633	(6.3.4)	0.316* (0.171)	0.405 (0.301)	-0.019 to 0.651	-0.185 to 0.995
	Log number of seizures	(6.3.1)	0.040 (0.036)	-0.031 to 0.111	(6.3.3)	0.018 (0.041)	0.147 (0.104)	-0.062 to 0.098	0.351 to -0.057
		(6.3.2)	0.034 (0.038)	-0.040 to 0.108	(6.3.4)	0.019 (0.041)	0.150 (0.117)	-0.061 to 0.099	-0.057 to 0.351

Table 6.8: Summary of the treatment-covariate effect estimates in the epilepsy data. N.B. s.e. represent the standard error of the parameter estimate. * for P value< 0.1, ** for P value< 0.05 and *** for P value<0.01.

Results for time to 12 month remission

First consider results for the outcome of time to 12 months remission. In all settings, the amalgamated effect estimator $\hat{\beta}_T$ was larger in absolute magnitude than the patient level estimator $\hat{\beta}_W$ suggesting that ecological bias may be present. For example, in the random effects model with the binary covariate 'Epilepsy type' the amalgamated interaction effect $\hat{\beta}_T = -0.09$ (s.e.=0.156) was much larger than the within trial estimator $\hat{\beta}_W = -0.025$ (s.e.=0.058) due to $\hat{\beta}_W$ being combined with an extremely large $\hat{\beta}_A = -0.479$ (s.e.=0.376).

The associations between the treatment effect and the two covariates, Epilepsy type and the log time of seizures were not statistically significant for any of the interactions in any models (P-values> 0.1).

Fitting random effects model (6.3.4) with age at randomization as the covariate, the between trial variance estimator $\hat{\tau} = 0.004$ showed that only a small amount of unexplained heterogeneity of trial treatment effects remained. An interesting finding was that the within trial effect $\hat{\beta}_W = -0.007$ (P value> 0.1) was not statistically significant whereas the amalgamated effect estimator $\hat{\beta}_T = -0.011$ (P value< 0.01) was strongly significant. The significant P value for $\hat{\beta}_T$ was due to amalgamating $\hat{\beta}_W$ with $\hat{\beta}_A$, which increased precision but might be biased as $\hat{\beta}_A$ was susceptible to ecological bias.

Results for time to first seizure post-randomization

Now consider the outcome of time to first seizure. In the models using age at randomisation as the covariate, no ecological bias was found as the within trial estimate $\hat{\beta}_W = 0.009$ (P value< 0.1) and the across trial estimate $\hat{\beta}_A = 0.01$ (P value< 0.05) were identical. As for the log number of seizures, no significant association was found in the models.

The models examining Epilepsy type detected the weak heterogeneities across the five included trials as $\hat{\tau} = 0.07$. In the random effects models, the amalgamated estimate $\hat{\beta}_T = 0.337$ (P value < 0.05) was slightly higher than the patient level estimate $\hat{\beta}_W = 0.316$ (P value< 0.1) due to the trial level confounding from the across trial estimate $\hat{\beta}_A = 0.405$ (P value > 0.1). Interestingly, the across trial interaction is large but non-significant, whereas the within-trial interaction is smaller but close to significance. This illustrates why the general power of within-trial estimates is greater[80].

6.6 Discussion

With IPD from multiple trials, there is greater flexibility in the meta analysis to identify the association between the treatment effect and patient-level factors on the survival probabilities of patients. In this chapter, the work builds on a number of hierarchy formulations of the Cox models appropriate for the analysis of the time-to-event data[155]. Inclusion of the treatment effect, patient level covariate and their interaction term in the functional form provides a straightforward approach for exploring treatment-covariate interaction effects using IPD[44].

In the settings of our Cox models, we could define the treatment effects to be fixed or random. But the key component of our work is to understand how an IPD meta analysis can correctly estimate the treatment-covariate interaction effect at patient level[130][151]. The simulation study shows that without any trial confounding, whether patient level and trial level interaction effects are treated separately or combined make no differences in terms of bias of estimates[80][138]. Further, the across trial estimates appear more powerful when the across trial variation of the covariate is large, as in such scenarios the across estimates may be the major source of information available[138]. However, ecological bias is often a threat due to trial level confounding by unobservable factors[12][49][81]. When the simulated data is confounded by any unobservable trial factor, then the across trial estimates are subject to ecological bias. This leads to bias in the across-trial interaction term, and thus potential bias in the amalgamated interaction estimate. This was seen in the epilepsy example, where age was not identified as important within trials but only across trials. For this reason, we suggest that practitioners should be more cautious when interpreting the across trial estimates as ecological bias could be an issue[106][157], and should generally separate within-trial and across-trial interactions in a one-stage IPD meta analysis.

The findings in this chapter reach the similar conclusions to the other researches in this field. For example, Riley and Steyeberg[112] emphases that the within trial level factor is more important than the amalgamated factor in the presence of ecological bias using IPD meta analysis with binary outcome. Riley et al.[108] and Schmid et al.[130] pointed out that the meta regression approaches to estimate the treatment-covariate interaction at trial level have low power and are subject to ecological bias. The simulations in Riley et al.[106] also gave the similar result that the amalgamated effect was seriously biased by the trial confounding but the within patient level factor still performed well in the framework of binary meta analysis.

Limitations

Though we used the hierarchy Cox models (6.3.1)-(6.3.4) to identify interaction effects between the treatment and patient-level factor on the survival probabilities, there is still possibility to further elaborate the setting of the multi-level Cox models in the study[141]. For example, we could further allow for the random effects in the baseline hazards across trials or consider an interaction term with random effects rather than being the fixed constant.

Another limitation is that the real data we utilized is about the patients from two active treatment groups rather than one treatment group and a placebo group[90]. Consequently, the difference in the effects between the two treatments themselves were not always significant, let alone the interaction effects with any other patient level factor. For example, consider the case for the time to seizure/remission being outcome and the log number of seizures being covariate, no statistically significant parameters were reported at all.

6.6.1 Conclusion

This chapter has shown the importance of separating the within-trial and across-trial treatment-covariate interactions in a one-stage IPD meta analysis of time-to-event outcomes. Key finds are summarized in Table 6.9. The thesis concludes in the next chapter with discussion about the whole thesis and intended further work.

What is already known on this topic:

- IPD meta analysis enables us to have more power to analyze the association between the treatment effect and the covariate at patient level. However, practical guidance is limited in the framework of time-to-event data analysis for how to specify interactions in a one-stage model.

- Ecological bias is a well known problem in estimating the treatment-covariate interaction effect and the estimate of the across trial interaction effect is very prone to confounding. Hence for binary and continuous outcomes, existing work suggested to estimate the within and across trial interactions separately in IPD meta analysis. However, this requires evaluation for time-to-event data.

What this study adds

- IPD meta analysis models treating the within and across trial interaction effects separately are recommended for survival outcomes.

- Such separation removes the threat of ecological bias, and enables interactions based solely on patient-level responses. If these are no trial-level confounders, the across trial interaction estimates may be more powerful than the within trial interaction when the across trial variation of the covariate is large.

- However, the across trial estimates are potentially biased in the presence of unobservable confounding factors, and thus gain in precision from using them (in addition to within-trial interactions) comes at expense of increased bias, and worse coverage.

- Applications to IPD for 5 epilepsy trials illustrate these issues, and show that age may not actually be a true predictor of treatment response, unlike previously thought.

Table 6.9: Summary of the main issues and key findings in Chapter 6.

CHAPTER 7

DISCUSSION AND FUTURE WORK

To examine or predict the outcome risk of individuals is one of the fundamental aspects of survival analysis, and modelling the hazard rate function is a critical part of this process. The thesis has considered survival models in two parts, mathematical research and medical research. In mathematical research, we provided theoretical appraisal of nonparametric modelling of hazard rate, and showed that the kernel method is an important tool in hazard rate estimation[53][166]. In medical research, we focused on the novel situation of individual participant data from multiple studies, and wanting to use flexible hazard modelling to compare mortality rates of breast cancer across countries and develop a prognostic model for new individuals, and use Cox modelling to examine predictors of treatment response. The key findings of the thesis are now summarised.

7.1 Key findings of the thesis

In Chapter 2, we extended Naito[97]'s idea in the setting of estimating density function to the hazard rate case, and proposed a semiparametric hazard rate estimator which depends on what we referred to as the the shape parameter α. We illustrated approximation error as a function of α pictorially, thereby showing the role α plays in the proposed methodology. The asymptotic bias and variance of the resulting estimator showed that in practice, this semiparametric estimate could provide us with an almost unbiased estimate when the true hazard rate function is very close to our prior parametric assumption.

Even if no partial knowledge is available, it performs as good as the usual nonparametric estimates.

In Chapter 3, the standard kernel hazard rate estimate was analysed through L_1 error criterion. We also derived the asymptotic expression of the L_1 optimal bandwidth for the kernel estimate and showed that the bandwidth is of the same order as the usual L_2 optimal bandwidth. We then discussed how to utilize the Newton method to derive the bandwidth since no explicit functional form is available for L_1 bandwidth. Since the optimal estimate still depends on the unknown hazard rate function itself, the data-driven version of L_1 optimal bandwidth was introduced using the plug-in method[53]. In practice, the proposed L_1 optimal estimator permits the minimization of the L_1 errors which provides a possible way to measure the absolute difference between the estimator and the true hazard rate function[51].

In Chapter 4, we developed a prognostic model for examining whether individual risks of breast cancer patients depends on where they live. Consequently, our research revealed the statistical difference in survival time for different countries using a flexible parametric model via the Royston-Parmar scheme[120][122]. In addition, it was found that the comparison of countries was biased if the confounding factors such as patient characteristics were not accounted for. Indeed, additional confounding (such as lead time bias) may still be affecting the finding that Sweden does best and Denmark worst.

In Chapter 5, we discussed several issues for developing and validating a prognostic model using IPD from multiple studies. We firstly considered to utilize the stratified intercept to represent the unique baseline for each included study and then proposed new strategies to extend the derived model to a new population. An internal and external validation method offered us an exciting opportunity to perform external validation on multiple studies[123]. We also utilized the multivariate meta-analysis framework of Snell et al.[142] to conclude and assess the model performance and then determined to exclude the study in Sweden and prognostic factor, menopausal status from our final model. By using the internal-external cross validation method to our final model again, we demonstrate that the calibration ability of the final model did improve after dropping the data from Sweden and menopausal status.

In Chapter 6, we illustrated how to identify the interaction between a treatment effect and patient-level factors in a one-stage IPD meta analysis. In particular, we showed the importance of separating the within and across trial interaction effects in the presence of ecological bias. Previously this had not been considered for time-to-event outcomes, only binary or continuous[106][110]. The across-trial interaction effect is prone to trial-level confounding and subsequently may induce a bias in the amalgamated estimator of within and across-trial interactions[109][151].

7.2 Implication of findings for medical research

The findings represented in Chapter 4 unequivocally showed that the survival time of European patients with breast cancer was closely associated with the country they lived. Amongst 8 countries, the overall survival probability of patients from Denmark was predicted to be lowest after adjusting for the related confounding factors while comparably, the patients of Sweden were in the least risks of death. No significant association (if any) between two biomarkers, upa and pai1 and geographical factor was detected through our analysis. It is worth mention that all the estimates achieved might be further impacted by additional confounding factors such as lead time bias and thus the bias might still remain in the developed model[15].

In chapter 5, the achieved C statistics in each cycle of the internal-external cross validation framework reflected that the discrimination abilities of the prognostic model were very stable across studies that approximately the risks of 70% paired participants were discriminated correctly. The resulting parameter estimates proved that the prognostic factor, menopausal status had no significant impact on the mortality rate of patients at all. Further due to the poor performance of the calibration slopes of the data from Sweden, it was suspectable that the baseline characteristics of patients from Sweden were very different from other countries. Therefore, in the final model we would not include the data from Sweden nor the predictor, menopausal status. As for the choice of the intercept strategy, in views of the pooled calibration slope estimates across 8 studies, we more recommended the new intercept strategy as it perfectly adjusted the model for the heterogeneity in new individuals. If the new dataset is not available, the average intercept strategy is preferred in terms of its more accurate pooled estimator of the calibration slope as well as the smaller standard error and between-study heterogeneities in comparison to the nearest neighbour strategy.

Chapter 6 looked into the association between the effects of two anti-epileptic drugs, Sodium Valproate and Carbamazepine and three patient-level characteristics factors, age at randomisation, epilepsy type and the log number of seizures. Two outcomes, time to 12 month remission and time to first seizure post-randomization were of our main concern. The hierarchy Cox models in the research indicated that although the amalgamated interaction effect between age and treatment effect was statistically significant on the remission time, it was potentially due to certain unobservable trial confounding (ecological bias) as the within-trial effect was of a smaller magnitude and non-significant. Conversely, the interaction effects of epilepsy type and drug was demonstrated to be positively significant on the seizure time even after adjusting for the ecological bias. Besides that, no ecological bias was found for interaction effect between age and drug on the seizure time of patients where Carbamazepine appeared more effective on elder patients in comparison to Sodium Valproate.

7.3 Implications of methodology

We summarize the methodologies utilized in the thesis.

7.3.1 Kernel type methods

In statistics, kernel estimation is the most favoured nonparametric approach to estimate the unknown hazard rate function. The choice of bandwidth of the kernel, which has a strong influence on the resulting estimate is an important issue in this approach. Generally, L_2 error criterion (Mean (integrated) squared error) is a popular measure to determine the value of the bandwidth. However, the L_1 view of nonparametric kernel hazard rate estimator is probably more appropriate to assess its performance. In Chapter 3, we developed an asymptotic L_1 optimal bandwidth for the standard kernel estimator and also constructed a simple algorithm to calculate it. In addition, we showed that both theoretical and data-driven versions of this L_1 optimal estimator did minimize the asymptotic L_1 distance between the estimator and true function.

Besides that, we investigated a semiparametric kernel-type estimate in Chapter 2. In this case, the parametric estimator can be seen as a crude guess which is then modified by the nonparametric factor. Theoretical research revealed that when our prior assumption is very close to the true function, then the bias of our estimator will vanish. That is to say, in practice, if we could correctly figure out the parametric function of the underlying data, the bias of the semiparametric estimator would be totally removed. This estimator also unifies several multiplicative hazard rate estimators and could be implemented in practice easily using a plug-in method[166].

7.3.2 Flexible parametric regression

Rather than using tradition Cox regression, we chose to use a more flexible parametric method to model the time-to-event data from clinical studies in Chapter 4 and 5[120][122]. One of the key advantage of flexible parametric regression is that the baseline hazards could be estimated using restricted cubic spline functions. This improvement makes the following works possible in research. For example, the population average survival curve for the subgroup of observations in the database could be plotted now. The absolute survive probability of observations could be calculated adjusting for confounding factors. Further the differences in the baseline hazards across studies could be accounted in IPD meta analysis.

7.3.3 Internal-external cross validation and multivariate analysis

Internal and external validation method offers us an exciting opportunity to perform external validation without using any new population[123]. We could develop and validate the model simultaneously with the maximum external validation studies. Meanwhile, to determine the best strategy to develop a model, multivariate meta-analysis helps us to obtain the joint inferences using the statistics from each cycle within the internal-external cross validation framework[70][142].

7.3.4 Treatment-covariate interaction effects

Based on our simulation and real case study, ecological bias can have a big impact on interactions in a one-stage IPD meta analysis of time-to-event outcomes. An amalgamated model with random effects is better than without random effects. But the most effective way is to focus solely on the patient level interaction effect only. In this regard, our research showed how to separate the within-trial and across-trial effects using hierarchy Cox models and demonstrated its necessity in the presence of ecological bias.

7.4 Future work

Chapters 2 and 3 assume all event times are known, but in real life, many real data are censored. Therefore one foreseen work in the future is to extend our theoretical research in Chapters 2 and 3 to the situation of censoring. For example, we would like to see if the censored case is considered, whether the asymptotic property of the generalized estimator will be changed[96].

As for the model development in medical research, in Chapter 4, we may consider to add the non-linear terms into the model to adjust for any time-dependent factors[122]. In Chapter 5, we may consider the following work: Firstly, we may conduct more case studies. In our breast cancer example, the average intercept strategy was proved to be better than the nearest neighbour strategy. However, as suggested by Debray et al.[31], the nearest neighbour strategy was better in the setting of the binary outcome. It raised our concern to collect more evidence to investigate whether this conflicting conclusion was caused by chance or other reasons. Secondly, in our current work, we only considered the perfect data, *i.e.* IPD meta analysis dataset without missing values, however, in practice, IPD can be costly and time consuming to collect and may not be always available. Therefore, we are considering to combine individual participant data and aggregate data together to develop a prognostic model[106][111]. Finally, at this stage, we always assume that the between-study heterogeneities exist in the baseline function and proportional to hazards. However it is a strong assumption in practice. A more flexible model incorporating the time-dependent or nonlinear terms may thus be developed in future research[120].

In Chapter 6, future work could also consider any patient-level confounding across trials. For example, in the clinical trial, the male patients are more likely to be given the extra dose of drugs than females which may not be accounted for in the final data. Then we are interested to see whether and how this unobservable patient-level confounding factor may bias the estimate of within and across trial interactions.

7.5 Conclusion

This thesis has covered many important areas of mathematical and medical research for survival models. Though further work is needed, the thesis has contributed to new findings about kernel estimation, breast cancer mortality, the development and validation of prognostic models, and the estimation of treatment-covariate interactions for time-to-event data. The work is therefore informative to both statisticians and medical researchers, and the chapters will be written for publication in scientific journals in the coming year.

APPENDIX A

MULTIPLE IMPUTATION

In Appendix A, we briefly introduce how to conduct multiple imputation to estimate the missing values in a dataset (see Buuren et al.[160] and Royston[117][118]). Let $\underline{Y}$ be an vector of the covariates with missing values and $\underline{Z}$ is the set of other covariates and response variables. $\underline{Y}_{obs}$ and $\underline{Y}_{mis}$ denote the observed and missing parts of $\underline{Y}$, so $\underline{Y} = (\underline{Y}_{obs}, \underline{Y}_{mis})^T$.

For a single incomplete variable Y, this involves constructing an imputation model which regresses Y on a set of variables $Z = (z_1, ..., z_k)$ with complete data. Set $(\underline{\beta}, \mathbf{V})$ to be the set of estimated regression parameters and their corresponding covariance matrix from fitting the imputation model. We draw a value of $\underline{\beta}^*$ from the posterior distribution, commonly approximated by $\underline{\beta}^* \sim MVN(\hat{\underline{\beta}}, \mathbf{V})$ and then draw a value of Y^*_{mis} from its conditional posterior distribution using $\underline{\beta}^*$ and the probability distribution. Repeating these steps m times yields m samples from the posterior distribution of Y_{mis} which are actually the m imputed samples.

For a set of incomplete variables $\underline{Y}_{mis}$ in the dataset, usually it is convenient to split the multivariate problem into a sequence of univariate problems, and solve the multivariate case by iteration. With the proper initial values of $\underline{Y}_{mis}$, the algorithm using the idea of Gibbs sampling is to estimate each incomplete variable in turn with fixing other variables and the iterations stop when the imputed values of $\underline{Y}_{mis}$ converges.

The choice of the number m of imputations is still in need of research. Rubin[125], van Buuren et al.[160] suggested that for variables with 20 percentage of missing entries, m can be as low as 3-5. But recently, some statisticians proposed that larger m may perform better in imputation[169]. For our breast cancer data from Chapter 4, m is set to be 10 since the estimates appeared stable in 10 iterations.

APPENDIX B

BOOTSTRAP CORRELATION MATRIX FOR CALIBRATION SLOPES, C INDEX AND D INDEX

	C	R^2_D	slope.ave	slope.nei	slope.new
C	1.000				
R^2_D	0.842	1.000			
slope.ave	0.325	0.662	1.000		
slope.nei	0.349	0.703	0.982	1.000	
slope.new	0.334	0.677	0.998	0.991	1.000

Table B.1: Bootstrapped correlation for Netherland.

	C	R^2_D	slope.ave	slope.nei	slope.new
C	1.000				
R^2_D	0.803	1.000			
slope.ave	0.307	0.656	1.000		
slope.nei	0.324	0.661	0.991	1.000	
slope.new	0.303	0.653	1.000	0.987	1.000

Table B.2: Bootstrapped correlation for Ireland.

	C	R^2_D	slope.ave	slope.nei	slope.new
C	1.000				
R^2_D	0.671	1.000			
slope.ave	0.237	0.679	1.000		
slope.nei	0.240	0.682	0.998	1.000	
slope.new	0.218	0.653	0.985	0.974	1.000

Table B.3: Bootstrapped correlation for Sweden.

	C	R_D^2	slope.ave	slope.nei	slope.new
C	1.000				
R_D^2	0.697	1.000			
slope.ave	0.294	0.748	1.000		
slope.nei	0.297	0.751	1.000	1.000	
slope.new	0.300	0.754	0.998	0.999	1.000

Table B.4: Bootstrapped correlation for Slovenia.

	C	R_D^2	slope.ave	slope.nei	slope.new
C	1.000				
R_D^2	0.770	1.000			
slope.ave	0.468	0.643	1.000		
slope.nei	0.468	0.642	1.000	1.000	
slope.new	0.469	0.649	0.996	0.996	1.000

Table B.5: Bootstrapped correlation for Austria.

	C	R_D^2	slope.ave	slope.nei	slope.new
C	1.000				
R_D^2	0.798	1.000			
slope.ave	0.451	0.761	1.000		
slope.nei	0.442	0.750	0.999	1.000	
slope.new	0.438	0.746	0.998	1.000	1.000

Table B.6: Bootstrapped correlation for France.

	C	R_D^2	slope.ave	slope.nei	slope.new
C	1.000				
R_D^2	0.605	1.000			
slope.ave	0.276	0.816	1.000		
slope.nei	0.274	0.825	0.996	1.000	
slope.new	0.276	0.811	1.000	0.993	1.000

Table B.7: Bootstrapped correlation for Switzerland.

	C	R_D^2	slope.ave	slope.nei	slope.new
C	1.000				
R_D^2	0.810	1.000			
slope.ave	0.166	0.573	1.000		
slope.nei	0.183	0.599	0.995	1.000	
slope.new	0.199	0.625	0.974	0.992	1.000

Table B.8: Bootstrapped correlation for Denmark.

LIST OF REFERENCES

[1] Ghada Abo-Zaid, Boliang Guo, Jonathan J Deeks, Thomas Debray, Ewout W Steyerberg, Karel GM Moons, and Richard David Riley. Individual participant data meta-analyses should not ignore clustering. *Journal of clinical epidemiology*, 66(8):865–873, 2013.

[2] Douglas G Altman. *Practical statistics for medical research.* CRC Press, 1990.

[3] Douglas G Altman and Patrick Royston. What do we mean by validating a prognostic model? *Statistics in medicine*, 19(4):453–473, 2000.

[4] Douglas G Altman, Yvonne Vergouwe, Patrick Royston, and Karel GM Moons. Prognosis and prognostic research: validating a prognostic model. *Bmj*, 338, 2009.

[5] Stian Anderson. Nonparametric hazard rate estimation on a parametric start, unpublished. 1998.

[6] Peter A Andreasen, Lars Kjøller, Lise Christensen, and Michael J Duffy. The urokinase-type plasminogen activator system in cancer metastasis: a review. *International Journal of Cancer*, 72(1):1–22, 1997.

[7] Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

[8] Dimitrios Ioannis Bagkavos. *Bias reduction in nonparametric hazard rate estimation.* PhD thesis, University of Birmingham, 2003.

[9] Richard E Barlow and Frank Proschan. Statistical theory of reliability and life testing: probability models. Technical report, DTIC Document, 1975.

[10] Giorgio Bedogni. Clinical prediction modelsa practical approach to development, validation and updating. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(4):944–944, 2009.

[11] Melissa D Begg and Michael K Parides. Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in medicine*, 22(16):2591–2602, 2003.

[12] Jesse A Berlin, Jill Santanna, Christopher H Schmid, Lynda A Szczech, and Harold I Feldman. Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statistics in medicine*, 21(3):371–387, 2002.

[13] RJ Black, F Bray, J Ferlay, and DM Parkin. Cancer incidence and mortality in the european union: cancer registry data and estimates of national incidence for 1990. *European Journal of Cancer*, 33(7):1075–1107, 1997.

[14] WC Black and HG Welch. Screening for disease. *AJR. American journal of roentgenology*, 168(1):3–11, 1997.

[15] William C Black. Overdiagnosis: an underrecognized cause of confusion and harm in cancer screening. *Journal of the National Cancer Institute*, 92(16):1280–1282, 2000.

[16] SE Bleeker, HA Moll, EW Steyerberg, ART Donders, G Derksen-Lubsen, DE Grobbee, and KGM Moons. External validation is necessary in prediction research:: A clinical example. *Journal of clinical epidemiology*, 56(9):826–832, 2003.

[17] Ørnulf Borgan. Nelson–aalen estimator. *Encyclopedia of Biostatistics*, 1998.

[18] Freddie Bray, Peter McCarron, and D Maxwell Parkin. The changing global patterns of female breast cancer incidence and mortality. *childhood*, 4:5, 2004.

[19] Andrea Burton, Douglas G Altman, Patrick Royston, and Roger L Holder. The design of simulation studies in medical statistics. *Statistics in medicine*, 25(24):4279–4292, 2006.

[20] Dennis Albert Casciato and Mary C Territo. *Manual of clinical oncology*. Lippincott Williams & Wilkins, 2009.

[21] Mario Cleves. *An introduction to survival analysis using Stata*. Stata Press, 2008.

[22] Michel P Coleman, Jacques Esteve, Philippe Damiecki, Annie Arslan, Helene Renard, et al. Trends in cancer incidence and mortality. *IARC scientific publications*, (121):1, 1993.

[23] Michel P Coleman et al. Opinion: why the variation in breast cancer survival in europe. *Breast Cancer Res*, 1(1):22–26, 1999.

[24] David Collett. *Modelling binary data*. CRC press, 2002.

[25] Gary S Collins, Joris A de Groot, Susan Dutton, Omar Omar, Milensu Shanyinde, Abdelouahid Tajar, Merryn Voysey, Rose Wharton, Ly-Mee Yu, Karel G Moons, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC medical research methodology*, 14(1):40, 2014.

[26] David Roxbee Cox and David Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984.

[27] Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403, 1978.

[28] Michael J Crowther and Paul C Lambert. Simulating biologically plausible complex survival data. *Statistics in medicine*, 32(23):4118–4134, 2013.

[29] Michael J Daniels and Michael D Hughes. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in medicine*, 16(17):1965–1982, 1997.

[30] Cox R David. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34:187–220, 1972.

[31] Thomas Debray, Karel GM Moons, Ikhlaaq Ahmed, Hendrik Koffijberg, and Richard David Riley. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in medicine*, 2013.

[32] Jonathan J Deeks, Douglas G Altman, and Michael J Bradburn. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. *Systematic Reviews in Health Care: Meta-Analysis in Context, Second Edition*, pages 285–312, 2001.

[33] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.

[34] Jane F Desforges, William L McGuire, and Gary M Clark. Prognostic factors and treatment decisions in axillary-node-negative breast cancer. *New England Journal of Medicine*, 326(26):1756–1761, 1992.

[35] Luc Devroye. The kernel estimate is relatively stable. *Probability Theory and Related Fields*, 77(4):521–536, 1988.

[36] Luc Devroye and László Györfi. *Nonparametric Density Estimation: The L1 View.* New York: John Wiley & Sons, 1985.

[37] Annette J Dobson. *An introduction to generalized linear models.* CRC press, 2001.

[38] Sylvain Durrleman and Richard Simon. Flexible regression models with cubic s-plines. *Statistics in medicine*, 8(5):551–561, 1989.

[39] Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75, 1986.

[40] Regina C Elandt-Johnson and Norman Lloyd Johnson. *Survival models and data analysis.* Wiley New York, 1980.

[41] Anders Engeland, Tor Haldorsen, Paul W Dickman, Timo Hakulinen, Torgil R Möller, Hans H Storm, and Hrafn Tulinius. Relative survival of cancer patients: a comparison between denmark and the other nordic countries. *Acta Oncologica*, 37(1):49–59, 1998.

[42] Randall L Eubank. *Nonparametric regression and spline smoothing.* CRC press, 1999.

[43] J Ferlay, E Steliarova-Foucher, J Lortet-Tieulent, S Rosso, JWW Coebergh, H Comber, D Forman, and F Bray. Cancer incidence and mortality patterns in europe: estimates for 40 countries in 2012. *European Journal of Cancer*, 2013.

[44] DJ Fisher, AJ Copas, JF Tierney, and MKB Parmar. A critical review of methods for the assessment of patient-level interactions in individual participant data meta-analysis of randomized trials, and guidance for practitioners. *Journal of clinical epidemiology*, 64(9):949–967, 2011.

[45] John A Foekens, Maxime P Look, Harry A Peters, Wim LJ van Putten, Henk Portengen, and Jan GM Klijn. Urokinase-type plasminogen activator and its inhibitor pai-1: predictors of poor response to tamoxifen therapy in recurrent breast cancer. *Journal of the National Cancer Institute*, 87(10):751–756, 1995.

[46] John Fox. Cox proportional-hazards regression for survival data. *See Also*, 2002.

[47] Giampietro Gasparini, Franco Pozza, and Adrian L Harris. Evaluating the potential usefulness of new prognostic and predictive indicators on node-negative breast cancer patients. *Journal of the National Cancer Institute*, 85(15):1206–1219, 1993.

[48] Sander Greenland. A review of multilevel theory for ecologic analyses. *Statistics in medicine*, 21(3):389–395, 2002.

[49] Sander Greenland and Hal Morgenstern. Ecological bias, confounding, and effect modification. *International journal of epidemiology*, 18(1):269–274, 1989.

[50] EARLY BREAST CANCER TRIALISTS'COLLABORATIVE GROUP et al. Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy: 133 randomised trials involving 31 000 recurrences and 24 000 deaths among 75 000 women. *The Lancet*, 339(8784):1–15, 1992.

[51] Peter Hall and Matthew P Wand. Minimizing l1 distance in nonparametric density estimation. *Journal of Multivariate Analysis*, 26(1):59–88, 1988.

[52] James William Hardin, Joseph M Hilbe, and Joseph Hilbe. *Generalized linear models and extensions*. Stata Press, 2007.

[53] Wolfgang Hardle. *Applied nonparametric regression*, volume 27. Cambridge Univ Press, 1990.

[54] FE Harrell, Kerry L Lee, and Daniel B Mark. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15:361–387, 1996.

[55] Frank E Harrell, Kerry L Lee, Robert M Califf, David B Pryor, and Robert A Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 3(2):143–152, 1984.

[56] Frank E Harrell Jr. Unexpected predictor–outcome associations in clinical prediction research: causes and solutions. 2013.

[57] Douglas M Hawkins, Subhash C Basak, and Denise Mills. Assessing model fit by cross-validation. *Journal of chemical information and computer sciences*, 43(2):579–586, 2003.

[58] Harry Hemingway, Peter Croft, Pablo Perel, Jill A Hayden, Keith Abrams, Adam Timmis, Andrew Briggs, Ruzan Udumyan, Karel GM Moons, Ewout W Steyerberg, et al. Prognosis research strategy (progress) 1: a framework for researching clinical outcomes. *BMJ: British Medical Journal*, 346, 2013.

[59] Kenneth R Hess. Graphical methods for assessing violations of the proportional hazards assumption in cox regression. *Statistics in medicine*, 14(15):1707–1723, 1995.

[60] Ralf Hildenbrand, Heike Allgayer, Alexander Marx, and Philipp Stroebel. Modulators of the urokinase-type plasminogen activation system for cancer. *Expert opinion on investigational drugs*, 19(5):641–652, 2010.

[61] Aroon D Hingorani, Daniëlle A van der Windt, Richard D Riley, Keith Abrams, Karel GM Moons, Ewout W Steyerberg, Sara Schroter, Willi Sauerbrei, Douglas G Altman, and Harry Hemingway. Prognosis research strategy (progress) 4: stratified medicine research. *BMJ: British Medical Journal*, 346, 2013.

[62] Nils Lid Hjort and Ingrid K Glad. Nonparametric density estimation with a parametric start. *The Annals of Statistics*, pages 882–904, 1995.

[63] Nils Lid Hjort and MC Jones. Locally parametric nonparametric density estimation. *The Annals of Statistics*, pages 1619–1647, 1996.

[64] Nils Lid Hjort, Mike West, and Sue Leurgans. Semiparametric estimation of parametric hazard rates. In *Survival Analysis: State of the Art*, pages 211–236. Springer, 1992.

[65] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[66] DW Hosmer, S Lemeshow, and S May. Applied survival analysis. regression modeling of time to event data, john wiley&sons. *New York*, 1999.

[67] Peter J Huber. Robust statistics. 1981.

[68] Alan Julian Izenman. Review papers: recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86(413):205–224, 1991.

[69] Christopher Jackson, Nicky Best, and Sylvia Richardson. Improving ecological inference using individual-level data. *Statistics in medicine*, 25(12):2136–2159, 2006.

[70] Dan Jackson, Richard Riley, and Ian R White. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*, 30(20):2481–2498, 2011.

[71] Dan Jackson, Ian R White, and Richard D Riley. Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Statistics in medicine*, 31(29):3805–3820, 2012.

[72] Alejandro R Jadad, R Andrew Moore, Dawn Carroll, Crispin Jenkinson, D John M Reynolds, David J Gavaghan, and Henry J McQuay. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Controlled clinical trials*, 17(1):1–12, 1996.

[73] Robert I Jennrich and Mark D Schluchter. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, pages 805–820, 1986.

[74] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

[75] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[76] David G Kleinbaum and Mitchel Klein. *Survival analysis.* Springer, 1996.

[77] David G Kleinbaum and Mitchel Klein. *Logistic regression: a self-learning text.* Springer, 2010.

[78] Djokouri A Kouassi and Jagbir Singh. A semiparametric approach to hazard estimation with randomly censored observations. *Journal of the American Statistical Association*, 92(440):1351–1355, 1997.

[79] Paul C Lambert and Patrick Royston. Further development of flexible parametric models for survival analysis. *Stata Journal*, 9(2):265, 2009.

[80] Paul C Lambert, Alex J Sutton, Keith R Abrams, and David R Jones. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of clinical epidemiology*, 55(1):86–94, 2002.

[81] Joseph Lau, John Ioannidis, and Christopher H Schmid. Summing up evidence: one answer is not always enough. *The lancet*, 351(9096):123–127, 1998.

[82] Jerald F Lawless. Statistical models and methods for lifetime data. 1982. *Wiely, New York.*

[83] Elisa T Lee and John Wenyu Wang. *Statistical methods for survival data analysis.* John Wiley & Sons, 2013.

[84] Katherine J Lee and Simon G Thompson. Flexible parametric models for random-effects distributions. *Statistics in medicine*, 27(3):418–434, 2008.

[85] Shaw-Hwa Lo and Kesar Singh. The product-limit estimator and the bootstrap: some asymptotic representations. *Probability Theory and Related Fields*, 71(3):455–465, 1986.

[86] J Scott Long and Jeremy Freese. *Regression models for categorical dependent variables using Stata.* Stata press, 2006.

[87] Maxime P Look, Wim LJ Van Putten, Michael J Duffy, Nadia Harbeck, Ib Jarle Christensen, Christoph Thomssen, Ronald Kates, Frédérique Spyratos, Mårten Fernö, Serenella Eppenberger-Castori, et al. Pooled analysis of prognostic impact of urokinase-type plasminogen activator and its inhibitor pai-1 in 8377 breast cancer patients. *Journal of the National Cancer Institute*, 94(2):116–128, 2002.

[88] David Machin and PM Fayers. Randomized clinical trials design, practice and reporting. *The Annals of Pharmacotherapy*, 44:2045, 2010.

[89] Susan Mallett, Patrick Royston, Rachel Waters, Susan Dutton, and Douglas G Altman. Reporting performance of prognostic models in cancer: a review. *BMC medicine*, 8(1):21, 2010.

[90] Anthony G Marson, Paula R Williamson, Jane L Hutton, Helen E Clough, and David W Chadwick. Carbamazepine versus valproate monotherapy for epilepsy. *Cochrane Database Syst Rev*, 3, 2000.

[91] NA Jackson McCleary, J Meyerhardt, E Green, G Yothers, A de Gramont, E Van Cutsem, et al. Impact of older age on the efficacy of newer adjuvant therapies in¿ 12,500 patients (pts) with stage ii/iii colon cancer: Findings from the accent database. *J Clin Oncol*, 27(15S):4010, 2009.

[92] K McPherson, CM Steel, and JM Dixon. Breast cancerepidemiology, risk factors, and genetics. *BMJ*, 321(7261):624–628, 2000.

[93] Michael E Miller, Carl D Langefeld, William M Tierney, Siu L Hui, and Clement J McDonald. Validation of probabilistic predictions. *Medical Decision Making*, 13(1):49–57, 1993.

[94] Melvin L Moeschberger and John P Klein. *Survival analysis: Techniques for censored and truncated data.* Springer, 2003.

[95] Karel GM Moons, Douglas G Altman, Yvonne Vergouwe, and Patrick Royston. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ: British Medical Journal*, 338(7709):1487–1490, 2009.

[96] Hans-Georg Muller and Jane-Ling Wang. Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, pages 61–76, 1994.

[97] Kanta Naito. Semiparametric density estimation by local l2-fitting. *The Annals of Statistics*, 32(3):1162–1191, 2004.

[98] Roger Newson. Confidence intervals for rank statistics: Somers' d and extensions. *Stata Journal*, 6(3):309, 2006.

[99] Ingram Olkin and Clifford H Spiegelman. A semiparametric approach to density estimation. *Journal of the American Statistical Association*, 82(399):858–865, 1987.

[100] D Max Parkin, Freddie Bray, J Ferlay, and Paola Pisani. Global cancer statistics, 2002. *CA: a cancer journal for clinicians*, 55(2):74–108, 2005.

[101] Lisa Pennells, Stephen Kaptoge, Ian R White, Simon G Thompson, Angela M Wood, Robert W Tipping, Aaron R Folsom, David J Couper, Christie M Ballantyne, Josef Coresh, et al. Assessing risk prediction models using individual participant data from multiple studies. *American journal of epidemiology*, 179(5):621–632, 2014.

[102] MJ Quinn, C Martinez-Garcia, and F Berrino. Variations in survival from breast cancer in europe by age and country, 1978–1989. *European Journal of Cancer*, 34(14):2204–2211, 1998.

[103] RD Riley, KR Abrams, PC Lambert, AJ Sutton, and JR Thompson. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in medicine*, 26(1):78–97, 2007.

[104] Richard D Riley. Multivariate meta-analysis: the effect of ignoring within-study correlation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(4):789–811, 2009.

[105] Richard D Riley, Keith R Abrams, Alexander J Sutton, Paul C Lambert, and John R Thompson. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology*, 7(1):3, 2007.

[106] Richard D Riley, Susanna R Dodd, Jean V Craig, John R Thompson, and Paula R Williamson. Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Statistics in medicine*, 27(29):6111–6136, 2008.

[107] Richard D Riley, Jill A Hayden, Ewout W Steyerberg, Karel GM Moons, Keith Abrams, Panayiotis A Kyzas, Núria Malats, Andrew Briggs, Sara Schroter, Douglas G Altman, et al. Prognosis research strategy (progress) 2: prognostic factor research. *PLoS medicine*, 10(2):e1001380, 2013.

[108] Richard D Riley, Iram Kauser, Martin Bland, Lutgarde Thijs, Jan A Staessen, Jiguang Wang, Francois Gueyffier, and Jonathan J Deeks. Meta-analysis of randomised trials with a continuous outcome according to baseline imbalance and availability of individual participant data. *Statistics in medicine*, 32(16):2747–2766, 2013.

[109] Richard D Riley, Paul C Lambert, and Ghada Abo-Zaid. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ: British Medical Journal*, pages 521–525, 2010.

[110] Richard D Riley, Paul C Lambert, Jan A Staessen, Jiguang Wang, Francois Gueyffier, Lutgarde Thijs, and Florent Boutitie. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in medicine*, 27(11):1870–1893, 2008.

[111] Richard D Riley, Mark C Simmonds, and Maxime P Look. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. *Journal of clinical epidemiology*, 60(5):431–e1, 2007.

[112] Richard D Riley and Ewout W Steyerberg. Meta-analysis of a binary outcome using individual participant data and aggregate data. *Research Synthesis Methods*, 1(1):2–19, 2010.

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

[114] William F Rosenberger and John M Lachin. *Randomization in clinical trials: theory and practice*. John Wiley & Sons, 2004.

[115] Peter M Rothwell, Ziyah Mehta, Sally C Howard, Sergei A Gutnikov, and Charles P Warlow. From subgroups to individuals: general principles and the example of carotid endarterectomy. *The Lancet*, 365(9455):256–265, 2005.

[116] Patrick Royston. Flexible parametric alternatives to the cox model, and more. *Stata Journal*, 1(1):1–28, 2001.

[117] Patrick Royston. Multiple imputation of missing values. *Stata Journal*, 4:227–241, 2004.

[118] Patrick Royston. Multiple imputation of missing values: update of ice. *Stata Journal*, 5(4):527, 2005.

[119] Patrick Royston and Douglas G Altman. External validation of a cox prognostic model: principles and methods. *BMC medical research methodology*, 13(1):33, 2013.

[120] Patrick Royston and Paul C Lambert. Flexible parametric survival analysis using stata: beyond the cox model. *Stata Press books*, 2006.

[121] Patrick Royston, Karel GM Moons, Douglas G Altman, and Yvonne Vergouwe. Prognosis and prognostic research: Developing a prognostic model. *Bmj*, 338, 2009.

[122] Patrick Royston and Mahesh KB Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in medicine*, 21(15):2175–2197, 2002.

[123] Patrick Royston, Mahesh KB Parmar, and Richard Sylvester. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. *Statistics in medicine*, 23(6):907–926, 2004.

[124] Patrick Royston and Willi Sauerbrei. A new measure of prognostic separation in survival data. *Statistics in medicine*, 23(5):723–748, 2004.

[125] Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 307. Wiley. com, 2009.

[126] M Sant, R Capocaccia, MP Coleman, F Berrino, G Gatta, A Micheli, A Verdecchia, J Faivre, T Hakulinen, JWW Coebergh, et al. Cancer survival increases in europe, but international differences remain wide. *European journal of cancer*, 37(13):1659–1667, 2001.

[127] Milena Sant, Claudia Allemani, Riccardo Capocaccia, Timo Hakulinen, Tiiu Aareleid, Jan Willem Coebergh, Michel P Coleman, Pascale Grosclaude, Carmen Martinez, Janine Bell, et al. Stage at diagnosis is a key explanation of differences in breast cancer survival across europe. *International Journal of Cancer*, 106(3):416–422, 2003.

[128] Milena Sant, Riccardo Capocaccia, Arduino Verdecchia, Jacques Estève, Gemma Gatta, Andrea Micheli, Michel P Coleman, Franco Berrino, et al. Survival of women with breast cancer in europe: variation with age, year of diagnosis and country. *International journal of cancer*, 77(5):679–683, 1998.

[129] Willi Sauerbrei, Norbert Holländer, Richard D Riley, and Douglas G Altman. Evidence-based assessment and application of prognostic markers: the long way from single studies to meta-analysis. *Communications in Statistics-Theory and Methods*, 35(7):1333–1342, 2006.

[130] Christopher H Schmid, Paul C Stark, Jesse A Berlin, Paul Landais, and Joseph Lau. Meta-regression detected associations between heterogeneous treatment effects

and study-level, but not patient-level, factors. *Journal of clinical epidemiology*, 57(7):683–697, 2004.

[131] M Schmitt, N Harbeck, C Thomssen, O Wilhelm, V Magdolen, U Reuning, K Ulm, H Höfler, F Jänicke, and H Graeff. Clinical impact of the plasminogen activation system in tumor invasion and metastasis: prognostic relevance and target for therapy. *Thrombosis and haemostasis*, 78(1):285–296, 1997.

[132] David Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241, 1982.

[133] P Schuetz, M Koller, M Christ-Crain, E Steyerberg, D Stolz, C Müller, Heiner Claudins Bucher, R Bingisser, M Tamm, and B Müller. Predicting mortality with pneumonia severity scores: importance of model recalibration to local settings. *Epidemiology and infection*, 136(12):1628–1637, 2008.

[134] Steve Selvin. *Survival analysis for epidemiologic and medical research.* Cambridge University Press New York, NY, 2008.

[135] Robert J Serfling. Approximation theorems of mathematical statistics. 1980, 2000.

[136] Jun Shao. Second-order differentiability and jackknife. *Statistica Sinica*, 1:185–202, 1991.

[137] Mark C Simmonds, Julian PT Higginsa, Lesley A Stewartb, Jayne F Tierneyb, Mike J Clarke, and Simon G Thompson. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clinical Trials*, 2(3):209–217, 2005.

[138] MC Simmonds and JPT Higgins. Covariate heterogeneity in meta-analysis: Criteria for deciding between meta-regression and individual patient data. *Statistics in medicine*, 26(15):2982–2999, 2007.

[139] Jeffrey S Simonoff. *Smoothing methods in statistics.* Springer, 1996.

[140] Nozer D Singpurwalla and Man-Yuen Wong. Estimation of the failure rate-a survey of nonparametric methods part i: Non-bayesian methods. *Communications in Statistics-Theory and Methods*, 12(5):559–588, 1983.

[141] Catrin Tudur Smith, Paula R Williamson, and Anthony G Marson. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Statistics in medicine*, 24(9):1307–1319, 2005.

[142] Debray T Debray T Ensor J Look MP Moons KM Snell K, Hua H and Riley RD. A meta-analysis framework for summarising and comparing the performance of clinical prediction models across multiple studies (submitted). *Journal of Clinical Epidemiology*, 2014.

[143] David J Spiegelhalter. Probabilistic prediction in patient management and clinical trials. *Statistics in medicine*, 5(5):421–433, 1986.

[144] Spotswood L Spruance, Julia E Reid, Michael Grace, and Matthew Samore. Hazard ratio in clinical trials. *Antimicrobial agents and chemotherapy*, 48(8):2787–2792, 2004.

[145] Ewout W Steyerberg. *Clinical prediction models*. Springer, 2009.

[146] Ewout W Steyerberg, Sacha E Bleeker, Henriëtte A Moll, Diederick E Grobbee, and Karel GM Moons. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *Journal of clinical epidemiology*, 56(5):441–447, 2003.

[147] Ewout W Steyerberg, Karel GM Moons, Danielle A van der Windt, Jill A Hayden, Pablo Perel, Sara Schroter, Richard D Riley, Harry Hemingway, Douglas G Altman, PROGRESS Group, et al. Prognosis research strategy (progress) 3: prognostic model research. *PLoS medicine*, 10(2):e1001381, 2013.

[148] Ewout W Steyerberg, Nino Mushkudiani, Pablo Perel, Isabella Butcher, Juan Lu, Gillian S McHugh, Gordon D Murray, Anthony Marmarou, Ian Roberts, J Dik F Habbema, et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS medicine*, 5(8):e165, 2008.

[149] Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128, 2010.

[150] Charles J Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, pages 1285–1297, 1984.

[151] Alex J Sutton, Denise Kendrick, and Carol AC Coupland. Meta-analysis of individual-and aggregate-level data. *Statistics in medicine*, 27(5):651–669, 2008.

[152] CG Sweep, J Geurts-Moespot, N Grebenschikov, JH De Witte, JJ Heuvel, M Schmitt, MJ Duffy, F Jänicke, MD Kramer, JA Foekens, et al. External quality assessment of trans-european multicentre antigen determinations (enzyme-linked immunosorbent assay) of urokinase-type plasminogen activator (upa) and its type 1 inhibitor (pai-1) in human breast cancer tissue extracts. *British journal of cancer*, 78(11):1434, 1998.

[153] Martin A Tanner and Wing Hung Wong. The estimation of the hazard function from randomly censored data by the kernel method. *The Annals of Statistics*, pages 989–993, 1983.

[154] Terry Therneau. Mixed effects cox models. *R-package description. URL: http://cran. r-project. org/web/packages/coxme/vignettes/coxme. pdf*, 2012.

[155] Terry M Therneau. *Modeling survival data: extending the Cox model*. Springer, 2000.

[156] Terry M Therneau, Patricia M Grambsch, and V Shane Pankratz. Penalized survival models and frailty. *Journal of computational and graphical statistics*, 12(1):156–175, 2003.

[157] Simon G Thompson and Julian Higgins. Can meta-analysis help target interventions at individuals most likely to benefit? *The Lancet*, 365(9456):341–346, 2005.

[158] Mark R Trusheim, Ernst R Berndt, and Frank L Douglas. Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. *Nature Reviews Drug Discovery*, 6(4):287–293, 2007.

[159] John W Tukey. Bias and confidence in not-quite large samples. In *Annals of Mathematical Statistics*, volume 29, pages 614–614. INST MATHEMATICAL STATISTICS IMS BUSINESS OFFICE-SUITE 7, 3401 INVESTMENT BLVD, HAYWARD, CA 94545, 1958.

[160] Stef Van Buuren, Hendriek C Boshuizen, Dick L Knook, et al. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, 18(6):681–694, 1999.

[161] Hans C van Houwelingen. Validation, calibration, revision and combination of prognostic survival models. *Statistics in medicine*, 19(24):3401–3415, 2000.

[162] David van Klaveren, Ewout W Steyerberg, Pablo Perel, and Yvonne Vergouwe. Assessing discriminative ability of risk models in clustered data. *BMC medical research methodology*, 14(1):5, 2014.

[163] M Vercelli, R Capocaccia, A Quaglia, C Casella, A Puppo, and JWW Coebergh. Relative survival in elderly european cancer patients: evidence for health care inequalities. *Critical reviews in oncology/hematology*, 35(3):161–179, 2000.

[164] Brani Vidakovic. *Statistical modeling by wavelets*, volume 503. Wiley-interscience, 2009.

[165] Grace Wahba. *Spline models for observational data*, volume 59. Society for industrial and applied mathematics, 1990.

[166] Matt P Wand and M Chris Jones. *Kernel smoothing*, volume 60. Chapman & Hall/CRC, 1995.

[167] GS Watson and MR Leadbetter. Hazard analysis. i. *Biometrika*, 51(1/2):175–184, 1964.

[168] GS Watson and MR Leadbetter. Hazard analysis ii. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 101–116, 1964.

[169] Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine*, 30(4):377–399, 2011.