Bootstrap-based Inference for Cox's Proportional Hazards Analyses of Clustered Censored Survival Data

YONGLING XIAO
Department of Epidemiology and Biostatistics
McGill University, Montreal

August 2005

A thesis submitted to the Faculty of Graduate Studies and Research
In partial fulfillment of the requirements for the degree of
Master of Science

Your file Votre référence
ISBN: 978-0-494-24831-7
Our file Notre référence
ISBN: 978-0-494-24831-7

ABSTRACT

Background: Clustering of observations occurs frequently in epidemiological and clinical studies of time-to-event outcomes. However, only a few papers addressed the challenge of accounting for clustering while analyzing right-censored survival data. I propose two bootstrap-based approaches to correct standard errors of Cox's proportional hazards (PH) model estimates for clustering, and validate the approaches in simulations.

Methods: Both bootstrap-based approaches involve 2 stages of resampling the original data. The two methods share the same procedure at the first stage but employ different procedures at the second stage. At the first stage of both methods, the clusters (e.g. physicians) are resampled with replacement. At the second stage, one method resamples individual patients with replacement for each physician (i.e. units within-cluster) selected at the 1st stage, while another method picks up all the patients for each selected physician, without resampling. For both methods, each of the resulting bootstrap samples is then independently analyzed with standard Cox's PH model, and the standard errors (SE) of the regression parameters are estimated as the empirical standard deviation, of the corresponding estimates. Finally, 95% confidence intervals (CI) for the estimates are estimated using bootstrap-based SE and assuming normality.

Simulations Design: I have simulated a hypothetical study with N patients clustered within practices of M physicians. Individual patients' times-to-events were generated from the exponential distribution with hazard conditional on (i) several patient-level variables, (ii) several cluster-level (physician's) variables, and (iii) physician's "random

effects". Random right censoring was applied. Simulated data were analyzed using 4 approaches: the proposed two bootstrap methods, standard Cox's PH model and "classic" one-step bootstrap with direct resampling of the patients.

Results: Standard Cox's model and "Classic" 1-step bootstrap under-estimated variance of regression coefficients, leading to serious inflation of type I error rates and coverage rates of 95% CI as low as 60-70%. In contrast, the proposed approach that resamples both physicians and patients-within-physicians, with the 100 bootstrap resamples, resulted in slightly conservative estimates of standard errors, which yielded type I error rates between 2% and 6%, and coverage rates between 94% and 99%.

Conclusions: The proposed bootstrap approach offers an easy-to-implement method to account for interdependence of times-to-events in the inference about Cox model regression parameters in the context of analyses of right-censored clustered data.

RESUMÉ

Contexte : Les données issues d'études épidémiologiques et cliniques nécessitant de l'analyse de survie ont souvent une structure de données conglomérée. Seuls quelques articles discutent du défi de tenir compte de la structure conglomérée des données lors de l'analyse de survie de données censurées à droite. Nous proposons deux méthodes basées sur la technique de bootstrap tenant compte de la structure conglomérée des données pour corriger l'erreur-type des coefficients estimés à l'aide du modèle de risques proportionnels de Cox. Nous validons ensuite ces méthodes à l'aide de simulations.

Méthodes : Les deux approches sont basées sur la technique de bootstrap et comportent deux étapes de rééchantillonnage des données initiales. La première étape est la même pour les deux méthodes: les conglomérats (e.g. les médecins) sont rééchantillonnés avec remise. La seconde étape diffère selon la méthode choisie. Dans la première méthode, chaque patient (e.g. unités à l'intérieur des conglomérats) est rééchantillonné avec remise pour chacun des médecins échantillonnés lors de la première étape. Dans la seconde méthode, tous les patients de chaque médecin échantillonné sont sélectionnés, sans nouveau rééchantillonnage. Dans le cas de chaque méthode, chaque jeu de données créé par le bootstrap est analysé de façon indépendante à l'aide d'un modèle de Cox standard et les erreur-types des paramètres de régression sont estimées à partir de leur écart-type empirique. Les intervalles de confiance 95% des coefficients sont estimés à l'aide des erreur-types basées sur le bootstrap et sous l'hypothèse de normalité.

Devis pour les simulations : Nous avons simulé une étude hypothétique avec N patients

regroupés dans les clientèles de M médecins. Le temps de réalisation de l'événement d'intérêt de chaque patient a été généré à partir d'une distribution exponentielle avec un risque conditionnel aux : (i) variables individuelles des patients, (ii) variables individuelles des médecins et (iii) effets aléatoires associés aux médecins. La censure à droite a été générée de façon aléatoire. Les données simulées ont été analysées de quatre façons différentes: les deux méthodes basées sur la technique de bootstrap décrites ci-haut, le modèle standard de risque proportionnel de Cox, et la technique de bootstrap 'classique' qui consiste en une étape de rééchantillonnage des N patients.

Résultats : Le modèle standard de risque proportionnel de Cox et la technique de bootstrap 'classique' sous-estiment la variance des coefficients de régression, provocant une inflation importante des taux d'erreur de type I et produisent des taux de couverture des intervalles de confiance de 95% aussi bas que 60-70%. En revanche, les deux méthodes que nous proposons ont nécessité 100 rééchantillonnages bootstrap et ont produit des estimés d'erreur-type légèrement conservateurs, produisant des taux d'erreur de type I entre 2-6% et des taux de couverture entre 94% et 99%.

Conclusions : Les deux méthodes basées sur le bootstrap que nous proposons sont faciles à exécuter et tiennent compte de l'interdépendance entre les temps de réalisation des événements dans l'inférence pour les coefficients de régression des modèles de Cox dans le contexte de l'analyse d'événements censurés à droites de données conglomérées.

Acknowledgements

First and foremost I offer my sincerest gratitude to my supervisor, Dr Michal Abrahamowicz, who has supported me throughout my thesis with his patience and knowledge while allowing me the room to work in my own way. I attribute the level of my Master's degree to his encouragement and effort, and without him this thesis would not have been completed.

I would like to thank Dr. Karen Leffondre and Dr. Debbie Feldman as the members of my thesis committee. I thank you for your friendly advices, encouragement and reviewing my thesis.

Thanks are due to Drs Jeannie Haggerty, Pierre Tousignant for their insightful discussion in epidemiology. I would also like to thank Yves Roy for his work on cleaning the real-life data.

I would like to acknowledge the financial support in the form of scholarships from the Fonds Québécois de la recherche sur la nature et les technologies. This research was also partly supported by the CIHR-funded project "Optimal care trajectories in rheumatoid arthritis (RA): the primary-secondary interface".

Last, but not least, I would like to dedicate this thesis to my parents, my husband and my daughters, for their love, patience, and understanding—they allowed me to spend most of the time on this thesis.

Table of Contents

List of Tables

List of Figures

1 Introduction

The Cox's proportional hazards (PH) model is the most commonly used regression model for survival data. One of the important assumptions for the Cox's model is the independence between observations. However, clustered survival data often arise in biomedical field. The observations within the same cluster tend to be correlated. For example, in health care research studies, individual patients (lower level units) are typically clustered within physician's (higher level units) practices. If the outcome of interest is time-to-referral to a specialist, then the outcomes of patients within the same physician's practice may be correlated, since these patients share the physician, whose unobservable (latent) characteristics may affect patients' times-to-referral. Analyzing the correlated survival data, by assuming the standard Cox's PH model, ignores the correlation between observations, and therefore, may incorrectly estimate the variation of the regression coefficients, possibly leading to inaccurate inference.

During the past decade, considerable progress has been made with the analysis of correlated data for binary and continuous (un-censored) outcomes. Sophisticated methods for clustered data, such as the mixed linear models (Laird et al., 1982; Goldstein, 2003) and the generalized estimating equations (Liang et al., 1986) method have become available in most statistical software packages. However, the development of techniques for correlated survival data analysis has proceeded relatively slowly. The random effects Cox model, which incorporates random effects into the conventional Cox's PH model, either requires imposing restrictive, arbitrary assumptions regarding distribution of

random effects and the baseline hazards function, or involves complicated computations.(Clayton et al. 1985; Klein, 1992; Sastry, 1998; McGilchrist, 1993; Yau, 2001; Ma et al. 2003) Although some of the statistics programs for frailty models i.e. random effects models in survival analysis have become recently available (Stata 7 & S-plus 6), most of them are developed for particular parametric frailty models (such as, follow either a gamma or inverse-Gaussian distribution). The marginal modeling is another technique to deal with the correlated data. This involves first modeling the marginal expectation of the hazards across the population (for this reason, "marginal models" are also referred to as "population-average models") and then modeling correlation between lower level units within higher level units (or clusters) by specifying different correlation structures. Again, marginal models for right-censored survival data are less popular and more difficult to implement than for binary or continuous (un-censored) data. (Wei et al. 1989; Lee et al. 1992; Liang et al. 1993; Cai et al. 1995, 1997; Lu et al. 2005)

In this thesis, I propose and evaluate a relatively simple marginal model for correlated survival data analysis. The approach involves first estimating the conventional Cox's PH model, and then using the computer-intensive bootstrap method to estimate the variance of regression coefficients estimates in a way that accounts for within-cluster correlation of time-to-events.

Bootstrap method is a resampling method for statistical inference. It can be applied to any data analysis, no matter how complicated, and requires no assumptions. However, the

resampling algorithm must be very carefully designed, to account for the relevant features of the data-generating process.

In this thesis, two resampling techniques for the hierarchical survival data have been investigated. One is a two-step bootstrap, which requires to randomly resample physicians with replacement at the first step, and then to randomly resample patients for each selected physician with replacement in the second step. Another resampling approach only randomly resamples physicians with replacement at the first step, but picks up all the patients for each selected physician, without resampling, in the second step. Simulations are then used to compare the standard error estimates and the coverage rates of 95% confidence interval for the two proposed bootstrap methods with the conventional Cox's proportional hazards model.

2 Literature review

2.1 Overview of Survival Analysis

In many biomedical studies, the primary outcome of interest is the time it takes for a certain event to occur (referred to as time-to-event). Examples include the time for a patient to respond to a therapy or the time from cancer diagnosis to death. In order to describe the distribution of the time-to-event for a given population, to compare the time-to-events among different groups, or to model the relationship of time-to-event to other factors, we need a special methodology, which is known under the general term of survival analysis.

2.1.1 Survival function and hazard function

In summarizing survival data, there are two functions of central interest, namely the survival function and the hazard function. Let the random variable T denote the time to the event of interest. The cumulative distribution function of T

$$F(t) = P(T < t), \quad t \geq 0 \tag{2.1}$$

represents the probability that the survival time is less than some value t.

The survival function $S(t)$, is defined to be the probability that the survival time T is greater than or equal to t, and so

$$S(t) = P(T \geq t) = 1 - F(t). \tag{2.2}$$

The hazard function is an instantaneous rate of failure at time t and represents the risk that an individual dies at time t, conditional on the individual having survived until that time. For a formal definition of the hazard function, we consider the probability that the random variable T takes a value between t and $t+\delta t$, where $\delta t > 0$ is a very small increment of time, conditional on T being greater than or equal to t: $P(t \le T < t+\delta t \mid T \ge t)$. This conditional probability is then divided by the length of the corresponding time interval (δt) to give the rate. The value of the hazard function at time t $h(t)$ is then the limiting value of this quantity, as δt goes to zero:

$$h(t) = \lim_{\delta t \to 0} \{ \frac{P(t \le T < t+\delta t \mid T \ge t)}{\delta t} \}. \qquad (2.3)$$

From the above definitions, the relationship between the survival and hazard function can be expressed as:

$$h(t) = \frac{\lim_{\delta t \to 0} \frac{P(t \le T < t+\delta t)}{\delta t}}{P(T \ge t)} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d \log(S(t))}{dt}, \qquad (2.4)$$

where $f(t)$ is the probability density function of T.

2.1.2 Modeling survival data

In the analysis of survival data, the survival function and the hazard function are estimated from the observed survival data. Typically, survival data are distinguished from other types of data because they are partly censored. The survival time of an individual is said to be censored when the event of interest has not been observed for that individual, so that the exact time-to-event remains unknown. The most common type of censoring is

the right censoring, which occurs when the subject is followed until a certain time t_c and then is lost to follow-up, without having the event of interest until t_c, so that we know only that the actual time-to-event $T > t_c$. Censoring prevents the use of standard methods of descriptive and inferential statistics. To estimate the hazard function and the survival function, two different general approaches have been developed: parametric methods and non-parametric methods.

2.1.2.1 Parametric models

Parametric methods require an assumption concerning the parametric form of the distribution of the survival times. For example, the most commonly assumed distributions that have been proposed for modeling survival times are the Exponential distribution, the Weibull distribution and the Log-Logistic distribution (Balakrishnan et al. 1987). The survival functions, hazard functions and density functions of these distributions are summarized in Table 2.1.

Table 2. 1 Properties of some popular distributions of time-to-event

	Survivor function	Density function	Hazard	No. of parameters
Exponential	$\exp(-\lambda t)$	$\lambda \exp(-\lambda t)$	λ	1
Weibull	$\exp[-(\lambda t)^{\gamma}]$	$\gamma\lambda(\lambda t)^{\gamma-1}\exp[-(\lambda t)^{\gamma}]$	$\gamma\lambda(\lambda t)^{\gamma-1}$	2
Log logistic	$[1+(t\lambda)^{\gamma}]^{-1}$	$\gamma\lambda^{\gamma}t^{\gamma-1}[1+(\lambda t)^{\gamma}]^{-2}$	$\frac{\gamma t^{\gamma-1}\lambda^{\gamma}}{[1+(\lambda t)^{\gamma}]}$	2

Maximum likelihood estimation is used to estimate the unknown parameters of the parametric distributions. Since survival data are usually partially censored, a special likelihood function is needed. The survival data consist of a pair $\{t_i, \delta_i\}$, where t_i represents the follow-up time for subject $i\,(i = 1,...,n)$, and δ_i represents the survival status. If the *ith* observation is not censored, then $\delta_i = 1$, otherwise $\delta_i = 0$. If t_i is uncensored ($\delta_i = 1$), the *ith* subject contributes $f(t_i)$ to the likelihood; If t_i is censored ($\delta_i = 0$), the *ith* subject contributes Pr (T $> t_i$) to the likelihood.

The full likelihood for all n subjects is then

$$L = \prod_{i:\delta_i=1} f(t_i;\phi) \prod_{p:\delta_p=0} S(t_p;\phi), \tag{2.5}$$

where ϕ is the vector of parameters. The corresponding log likelihood can be written as:

$$l = \sum_{i:\delta_i=1} \log f(t_i;\phi) + \sum_{p:\delta_p=0} \log S(t_p;\phi) \quad . \tag{2.6}$$

Since $f(t) = h(t)S(t)$, (2.6) may be written as:

$$l = \sum_{i:\delta_i=1} \log h(t_i;\phi) + \sum_{i=1}^{n} \log S(t_i;\phi). \tag{2.7}$$

For example, for the exponential distribution with the rate parameter λ, $S(t) = \exp(-\lambda t)$ has a constant hazard function $h(t;\lambda) = \lambda$. The general form of log likelihood (2.7) takes then the form:

$$l = \sum_{i:\delta_i=1} \log\lambda - \lambda \sum_{i=1}^{n} t_i = d\log\lambda - \lambda \sum_{i=1}^{n} t_i, \tag{2.8}$$

where d is the total number of failures. The maximum likelihood estimate (MLE) of λ can be obtained from the equation $\frac{\partial l}{\partial \lambda} = 0$, which yields $\hat{\lambda} = d/\sum t_i$.

The uncertainty associated with the maximum likelihood estimate of a model parameter is assessed by its standard error, which is evaluated in the usual way by calculating the inverse of the Fisher information at the MLE. The asymptotic normal distribution of maximum likelihood estimates allows us to construct confidence intervals for the parameters, based on the MLEs and their standard errors.

2.1.2.2 Non-parametric models

If the assumption of the parametric model is valid, then the inference based on such an assumption will be accurate and the estimation will be efficient. However, in many cases, it is not easy to specify a *priori* correct assumption concerning the nature or shape of the underlying survival distribution. To avoid such difficulties, non-parametric methods have become very popular in survival analysis.

2.1.2.2.1 Single sample non-parametric methods

Life-table estimate and Kaplan-Meier estimate of survival function are two common simple non-parametric methods, often used to describe survival in a single sample drawn from a homogeneous population.

The life-table estimate of the survival function is obtained by first dividing the period of observation into a series of time intervals. For each interval, we can then compute the number and proportion of subjects that entered the respective interval "alive," the number and proportion of subjects that failed in the respective interval and the number of subjects that were lost or censored in the respective interval. Based on those numbers and proportions, we can estimate the survival function and hazard function (Berkson et al., 1950; Cutler et al., 1958; Gehan, 1969).

The Kaplan-Meier estimate (Kaplan et al., 1958) of the survival function is obtained by constructing a series of intervals as for the life-table estimate. However, each of these intervals is designed to be such that one failure time is observed in the interval and the failure time is at the start of the interval.

Let $t_1 < t_2 < ... < t_m$ denote the distinct times at which subsequent events were observed, d_i the number of events that occurred at time t_i, and r_i the size of the risk set at time t_i. The risk set at time t_i includes all the subjects in the sample who have not yet been censored and have not had the event. The Kaplan-Meier product-limit estimate for a survival function is given by:

$$\hat{S}(t_i) = \prod_{j=1}^{i} (1 - \frac{d_j}{r_j}) \qquad (2.9)$$

The Kaplan-Meier estimate $\hat{S}(t)$ is a right-continuous step function with jumps at the event times. Censoring times affect the estimate only by reducing the risk set for the next

event. The corresponding estimate of the standard error is computed using Greenwood's formula (Kalbfleish et al., 1980) as

$$\hat{\sigma}\left(\hat{S}(t_i)\right) = \hat{S}(t_i)\sqrt{\sum_{j=1}^{i}\frac{d_j}{n_j(n_j - d_j)}}. \tag{2.10}$$

2.1.2.2.2 Proportional Hazards model

2.1.2.2.2.1 Definition of the Cox proportional hazards model

In most medical studies that give rise to survival data, explanatory factors or covariates are also recorded on each individual. In many cases, we are typically more interested in the effect of those covariates on the survival time than in simply estimating the un-conditional distribution of survival times or assessing how the hazard changes over time. Therefore, multivariable regression modeling of survival data is of particular importance. By far, the most commonly applied regression model for censored survival data is the proportional hazards model introduced by Cox in 1972(Cox, D. R. 1972), which specifies that the hazard, conditional on covariates, is a product of a term depending on time and a term depending on the covariates:

$$h_i(t \mid x_1, x_2, ..., x_p) = h_0(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}), \tag{2.11}$$

where β is the vector of regression coefficients (logarithms of hazard ratios) for the independent variables $x_1, x_2, \cdots, x_p$; $x_{i1}, x_{i2}, ..., x_{ip}$ represent covariate values for the *ith* individual and $h_0(t)$ is the baseline hazard function, which is the hazard function for individuals for whom the values of all the covariates are zero. Actually, the baseline

hazard does not have to be specified for the Cox's model. In the sense of "distribution free", the model is non-parametric with respect to distribution of time-to-events.

Now, consider two observations i and i' that differ in their covariates. The hazard ratio for these two observations

$$\begin{aligned}\frac{h_i(t)}{h_{i'}(t)} &= \frac{h_0(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})}{h_0(t)\exp(\beta_1 x_{i'1} + \beta_2 x_{i'2} + \cdots + \beta_p x_{i'p})} \\ &= \frac{\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})}{\exp(\beta_1 x_{i'1} + \beta_2 x_{i'2} + \cdots + \beta_p x_{i'p})} = \exp\left[\sum_{j=1}^{p} \beta_j (x_{ij} - x_{i'j})\right]\end{aligned} \tag{2.12}$$

is independent of time t; so that the ratio of hazards remains constant over time, regardless of the change in the absolute values of the hazard. That is why the Cox model is called the proportional hazards model. That hazard ratio is constant over time is one of the major assumptions of the Cox proportional hazards model.

Since the proportional hazards model can be re-expressed in the form:

$$\log\left(\frac{h_i(t)}{h_0(t)}\right) = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \tag{2.13}$$

this model may also be regarded as a linear model for the logarithm of the hazard ratio. This linear relationship between the logarithm of the hazard ratio and covariates facilitates the estimation of the Cox proportional hazards model, and the statistical inference about the estimates.

2.1.2.2.2.2 Partial likelihood

Even though the baseline hazard is not specified, the regression parameters in the proportional hazards model can still be estimated by the method of maximum partial likelihood (MPL), developed by Cox (1972; 1975). Although the resulting estimates are not efficient as maximum-likelihood estimates for a specified parametric regression model (Efron, 1977), not having to make arbitrary, and possibly incorrect assumptions about the form of the baseline hazard is an important practical advantage of Cox's model. The partial likelihood function considers the joint probability of the data conditional on the k observed failure times. MPL estimate relies on the concept of "Risk sets". The relative probability of individual i failing at time t is proportional to the hazard of that individual $h_0(t)\exp(x_i{}^T\beta)$. Hence, the probability that it is individual i (rather than any of the other individuals who were at risk at that time) who failed at time t is given by:

$$\frac{\exp(x_i{}^T\beta)h_0(t)}{\sum_{j:t_j\geq t}\exp(x_j{}^T\beta)h_0(t)} = \frac{\exp(x_i{}^T\beta)}{\sum_{j:t_j\geq t}\exp(x_j{}^T\beta)} \tag{2.14}$$

and the partial likelihood is given by:

$$L(\beta) = \prod_{i=1}^{k}\frac{\exp(x_i{}^T\beta)}{\sum_{j:t_j\geq t}\exp(x_j{}^T\beta)} \tag{2.15}$$

which does not depend on the baseline hazard $h_0(t)$. The partial likelihood given by the above equation is correct only when no ties occurred at any of the failure times, i.e. when each failure occurs at a distinct time. If there are ties in the data set, the calculation of the partial log-likelihood function involves permutations and can be time-consuming. In this

case, either the Breslow (1974) or Efron (1977) approximations to the partial log-likelihood can be used.

2.1.2.2.2.3 Statistical inference

Estimates for the regression parameters are obtained by maximizing the partial likelihood. Let $l(\beta) = \log L(\beta)$, then finding β to maximize $L(\beta)$ is equivalent to finding $\hat{\beta}$ that solves the equation $\frac{\partial l(\hat{\beta})}{\partial \beta} = 0$. The partial likelihood has the same asymptotic properties as a standard likelihood (Cox, 1975). Hence, standard errors, confidence intervals can all be routinely calculated. The estimated covariance matrix of $\hat{\beta}$ is given by:

$$\hat{V}(\hat{\beta}) = -\left[\frac{\partial^2 l(\hat{\beta})}{\partial \beta^2}\right]^{-1}. \qquad (2.16)$$

The corresponding confidence intervals are then obtained relying on normal approximation.

2.2 Overview of correlated data

2.2.1 Introduction

Many types of data analyzed in epidemiological and clinical studies have a hierarchical, or clustered structure. For example, in primary care research studies, outcomes of different patients of the same physician may be correlated because of the physician's competence, practice style or subjective preferences affecting all his/her patients (Donner et al., 1994). In longitudinal studies, repeated measurements for the same individual tend to be correlated with one another (Diggle et al., 1994); and in clustered-randomized clinical trials, the patients from the same clinical center may share some unobserved characteristics and their outcomes may be correlated because of the systematic between-centers differences in quality of care or resources. An appropriate statistical analysis of these data must take the correlation into account to avoid incorrect statistical inference and misleading conclusions (Donner et al., 2000).

Most standard statistical techniques assume that each of the observations from a data set is independent of all the others. However, such independence assumption is inappropriate if subsets of observations represent the same cluster, because the observations in a cluster tend to be more similar to one another. The degree of similarity is typically measured by the intra-class correlation coefficient (ICC) (Kerry et al, 1998; Goldstein, 2003). (See section 2.2.2.1). Ignoring the intra-custer correlation in the analysis results in the under estimation of the true variance of the estimated parameters because the amount of the

information contained in N correlated observations is lower than in N independent observations. Indeed, the higher the ICC the more serious becomes the underestimation of the standard error of the estimates (Kish, 1965; Tate et al., 1983). Under-estimation of the variance can lead, in turn, to incorrectly low p-values, inflated type I error rates, too narrow confidence intervals, and biased estimates, all of which can lead to an incorrect interpretation of the associations between variables (Campbell et al., 1998).

There are two distinct approaches used to analyze correlated data. One approach is to use multilevel models, sometimes referred to as "subject specific" models (Goldstein, 2003). Another approach is to use marginal models, such as Generalized Estimating Equations (GEE) model (Liang et al., 1986; 1988).

2.2.2 Modeling of Correlated Data

2.2.2.1 Multilevel modeling

Multilevel modeling is also known in the statistical literature under a variety of names, such as "random coefficient model" (De leeuw et al., 1986; Longford, 1993) and "variance components model" (Longford, 1993). It assumes hierarchical data structure with units at lower level clustered within larger units, at a higher level. The response variable is measured at the lowest level and different explanatory variables are measured at different levels. The multilevel model is viewed as a hierarchical structure of regression equations (Goldstein, 2003). For example, consider a simple two-level data set. There are m clusters at the first level indexed by $i = 1,...,m$. Within the ith cluster, there are n_i

subjects indexed by $j = 1,...,n_i$. On the individual level, we have the response variable y_{ij} and the covariate x_{ij}. Then, the corresponding multilevel model may be defined by the following two equations:

$$y_{ij} = \beta_{0i} + \beta_1 * x_{ij} + e_{ij}, \tag{2.17}$$

$$\beta_{0i} = \beta_0 + u_{0i}, \tag{2.18}$$

where residuals e_{ij} are independent of each other and arise from a normal distribution with parameters

$$E(e_{ij}) = 0, Var(e_{ij}) = \sigma_e{}^2; \tag{2.19}$$

and u_{0i} are cluster-specific "random intercepts", i.e., independent, identically distributed (i.i.d.) values of a random variable at the cluster level, with parameters

$$E(u_{0i}) = 0,$$

$$Var(u_{0i}) = \sigma_{u0}^2. \tag{2.20}$$

Equation (2.17) is similar to the standard linear regression model, except that the intercept β_{0i} is not constant for all the observations, but depends on the clusters. Equation (2.18) models β_{0i} using a linear combination of a constant coefficient β_0 and a random variable u_{0i} which is referred as to "random effects". Combining (2.17) and (2.18), I get:

$$y_{ij} = \beta_0 + \beta_1 * x_{ij} + u_{0i} + e_{ij}. \tag{2.21}$$

Since equation (2.21) contains both fixed-effects parameters β_0, β_1 and the random effects parameter u_{0i}, the multilevel model is also called the mixed linear regression model (Laird et al., 1982).

The objective of analysis of multilevel models is to estimate the fixed coefficients β_0 and β_1, and two covariance parameters, σ_{u0}^2 and σ_e^2. The variance of response y_{ij} given the fixed effects is:

$$\text{var}(y_{ij} \mid \beta_0, \beta_1, x_{ij}) = \text{var}(u_{0i} + e_{ij}) = \sigma_{u0}^2 + \sigma_e^2, \tag{2.22}$$

which is the sum of a level 1 and a level 2 variance. The covariance between two observations in the same cluster is given by:

$$\text{cov}(y_{ij}, y_{ik}) = \text{cov}(u_{0i} + e_{ij}, u_{0i} + e_{ik}) = \text{cov}(u_{0i}, u_{0i}) = \sigma_{u0}^2. \tag{2.23}$$

Given (2.21) and (2.23), the correlation between the two observations in the same cluster, i.e. the ICC, equals:

$$\rho = \frac{\text{cov}(y_{ij}, y_{ik})}{\sigma_{y_{ij}} \sigma_{y_{ik}}} = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_e^2}. \tag{2.24}$$

A general specification for the mixed linear model is written as:

$$y = X\beta + Z\gamma + \varepsilon \tag{2.25}$$

where y denotes the vector of observed response y_{ij}, β is the unknown fixed-effects parameter vector, X is the known design matrix of explanatory variables x_{ij}'s and $X = \{1, x_{ij}\}$, γ is the vector of unknown *random-effects parameters,* Z is the known

design matrix for γ, and ϵ is the unobserved vector of independent and identically distributed Gaussian random errors.

To estimate parameters in the multilevel model, the best approach is to use *likelihood-based* methods, exploiting the assumption that γ and ε are normally distributed with the expected values and variances defined as respectively: (Laird et al, 1982; Jennrich et al, 1986)

$$\mathbf{E}\begin{bmatrix}\gamma\\ \epsilon\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}$$

$$\mathbf{Var}\begin{bmatrix}\gamma\\ \epsilon\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}$$

The estimation procedure is iterative (Goldstein, 2003). It usually starts with the estimates of the fixed parameters fitted by a traditional estimation procedure, such as 'ordinary least squares' (OLS) (assuming independence between observations). From these initial estimates, 'raw' residuals are formed and used to get the initial estimates of variance-covariance parameters. Then, the new estimates of the fixed effects will be obtained by using maximum likelihood methods, and the algorithm alternates between the variance-covariance parameters and the fixed parameters estimation until the procedure converges. An estimate of the variance-covariance matrix of the parameter estimates is obtained from the inverse of the information matrix. The information matrix is defined as the negative of the expectation of the matrix of second-order derivatives. Expressions for the information matrix are given in Engle (1982) and Bollerslev (1986).

Several alternative covariance structures may be assumed depending on the expected pattern of within-cluster correlation of the residuals. The Compound symmetry (CS) structure, often referred to as exchangeable structure (Digger et al., 1994), assumes that variances are homogeneous across clusters and that the correlation between any two observations in the same cluster is constant. The autoregressive structure (AR) has homogeneous variances and correlations that decline exponentially with distance (Digger et al, 1994). The unstructured (UN) structure allows every term in the variance-covariance matrix to be different which allows total flexibility but may have convergence problem due to too many parameters to be estimated.

The mixed linear regression model works only for continuous outcome variable which is approximately normally distributed. To deal with non-normal distributed outcomes, the generalized linear mixed-effects model (GLMM) (Liang et al., 1986; McCulloch et al., 2001; Fahrmeir, 2001) has been proposed. The GLMM is a straightforward extension of the generalized linear model (Nelder et al., 1972). It involves adding random effects to the linear predictor (transformed by an appropriate link function), and expressing the expected value of the response conditional on the random effects.

2.2.2.2 Marginal modeling of correlated data

The generalized estimated equation (GEE) model (Liang et al., 1986) is the most popular example of a marginal model for correlated data. The marginal model separates the modeling of the between-subjects covariate effects from modeling of within-cluster

correlations. The former is modeled through the model for the marginal mean $E(Y_{ij})$, while the latter is modeled through the modeling of the within-cluster covariance $Cov(Y_{ij}, Y_{ik})$ that accounts for inter-dependence of the outcomes within the clusters.

The marginal mean depends on covariates via a link function:

$$E(Y_{ij} \mid X_{ij}) = \mu_{ij},$$

$$g(\mu_{ij}) = \beta_0 + \beta_1 * x_{ij}, \qquad (2.26)$$

Here, in contrast to the multilevel modeling (2.19), β describes how the population averaged (PA) response, rather than one subject's response, depends on the covariates.

The procedure of fitting a marginal model may be viewed as an iterative procedure. The first step is to fit a standard generalized linear model assuming independence. The second step uses the residuals from the regression model, the current estimates of coefficients and the assumed within-cluster correlation structure, to estimate the working correlation structure. The pre-specified correlation structure can be chosen from the same structures described in section 2.2.2.1 for multilevel modeling. The third step uses this working correlation structure to estimate the covariance, and next to update the estimates of the regression coefficients by solving the GEE equation. Then, the second step and the third step are repeated until the estimates stabilize and convergence is achieved. In general, the final GEE estimates of the regression coefficients are quite similar to those obtained from the first step, i.e. from the independent-data marginal model. In contrast, standard errors

of the final regression coefficients are higher than those from the first step, doe to the inter-correlation of outcomes in the same cluster (Hanley, 2003).

Marginal models are most effectively used in population studies, since the population-averaged response is the focus in these studies. One advantage of marginal modeling is that the PA response for a given covariate can be directly estimated from observations without assumptions about the heterogeneity in the parameters across the clusters. Another advantage of marginal modeling is that the GEE methods permit the calculation of robust estimates for the standard errors of the regression coefficients (Liang et al., 1986) and the robust standard errors ensure consistent inferences even if the chosen correlation structure is incorrect.

2.3 Adaptation of Cox's model to correlated data

Correlated survival data arise often in biomedical study. Since the semi-parametric Cox proportional hazards model allows for the estimation of the relative risk without the need to specify a baseline hazard and is very popularly used in biomedical research, the ability to extend its use to the correlated data setting is important. In the literature, two approaches are commonly used to account for the intra-cluster correlation of time-to-events: random effects models and marginal models. In random effects models (same as the multilevel models in section 2.2), the dependence structure is explicitly specified by some unobserved random quantities that are common to observations from the same cluster. In marginal models, the intra-cluster association is left unspecified but adjusted for at the inference step.

2.3.1 Random effects models

The use of random effects modeling in statistics has increased greatly in recent years. However, the introduction of such modeling to the field of survival analysis has proceeded more slowly. The most popular modeling method in survival analysis, Cox proportional hazards regression (Cox, 1972), requires no specification of the baseline hazard function $h_0(t)$. Unfortunately, multilevel modeling (random effects modeling) methods typically require such parameterization (Sargent, 1998).

Clayton and Cuzick (1985) introduce the proportional hazards frailty (i.e. random effect in survival analysis) model, where a cluster of observations is assigned a random effect that acts multiplicatively on the baseline hazard function. Therefore, this cluster specific random effect modeling allows the baseline hazard function to be the same within clusters while to differ between clusters. This model implies independent survival times conditional on the frailty terms. Suppose there are m independent clusters indexed by $i, i = 1,...,m$. Within the ith cluster, there are J_i individuals indexed by $j, j = 1,...J_i$. Let the hazard function for individual (i, j) at time t be denoted by $h_{ij}(t)$. Given the random effects w_i, the proportional hazards frailty model is written as:

$$h(t; x_{ij}, w_i) = h_0(t) w_i \exp(x_{ij}^T \beta), \tag{2.27}$$

where $h_0(t)$ is the baseline hazard function, x_{ij} is the covariates vector and β is the coefficients vector.

Frequentist approaches to frailty survival models have usually been restricted to specific parametric random effects distributions. Klein (1992) and Sastry (1998) considered gamma frailties. McGilchrist (1993) and Yau (2001) considered lognormal frailties.

With the same set-up of the frailty model as equation (2.27), Klein (1992) assumes that random effects w_i's are independent and identically gamma distributed with $E(w_i) = 1$ and $\text{var}(w_i) = \theta$. The association between individuals in the same cluster, measured by Kendall's τ (Hougaard, 2000), is $\frac{\theta}{\theta+2}$, so the strength of the inter-dependence is increased as θ is increased, with $\theta = 0$ corresponding to independence between cluster members.

To get the estimates of the fixed and random effects, Klein (1992) uses an EM algorithm (Dempster et al., 1977) and estimates the baseline hazard function $h_0(t)$ based on a profile likelihood at each iteration. Klein applied this method to the Framingham Heart Study (Dawber, 1980) to examine the risks of death from any cause associated with smoking, adjusting for potential random effects due to within-family clustering. He considered two possible types of clustering, one was the frailty shared by siblings and another one was the frailty shared by a married couple. In the analysis, a total of 25 covariates were considered. The EM algorithm yielded $\hat{\theta}'s$ for sibling effect and marital effect. In both cases $\hat{\theta}$ was significantly different from 0, which showed there were associations between times to death of siblings and married couples. He found that the absolute magnitude of the estimates ($\hat{\beta}$) of fixed effects tended to be smaller under the assumption of independence than under frailties analysis. However, we need to interpret $\beta's$ in a

different manner than in the usual Cox's regression model. For example, in usual Cox's model, we will interpret the coefficient of "smoking" as the logarithm of the hazard ratio for a smoker as compared to a non-smoker, adjusted for other covariates. In contrast, under the model adjusted for the random sibling effect, the coefficient will be interpreted as the logarithm of the hazard ratio of smoking between two brothers (or two sisters), one of them is a smoker and another is a non-smoker, adjusted for other covariates. The EM algorithm approach proposed by Klein (1992) is used only for a single level of random effects, and is limited by the gamma frailty assumption.

Sastry (1997) introduces a nested frailty model for hierarchically clustered survival data. In this model, two independent cluster-specific random effects, v_i and w_{ij} are considered. The higher-level random effect v_i is assigned to each of $i = 1,...,m$ clusters, and the lower-level random effect w_{ij} is assigned to each of $j = 1,...,J_i$ subclusters of cluster v_i. For example, clustered child survival data are correlated at the community level and at the family level (Sastry, 1997). Within each sub-cluster (i, j) there are n_{ij} individuals indexed by $k = 1,...,n_{ij}$. The random effects are assumed to operate multiplicatively on the baseline hazard, therefore the hazard function for individual (i, j, k) at time t is given:

$$h(t; x_{ijk}, v_i, w_{ij}) = h_0(t) v_i w_{ij} \exp(x_{ijk}^T \beta) \qquad (2.28)$$

where $h_0(t)$ is the baseline hazard and x_{ijk} is the vector of covariates for individual (i, j, k).

Sastry (1997) assumes v_i and w_{ij} are mutually independent, and are gamma distributed with variances $1/\alpha$ and $1/\eta$ respectively. Furthermore, to ensure identifiability, he assumes both effects have means equal to 1 at time zero. Finally, he assumes the baseline hazard to be piecewise constant. Sastry applies the EM algorithm to estimate the parameters α, η and β. There are 3 components in the expectation of the log likelihood $Q(\alpha, \eta, \beta)$. Sastry introduces a new approach that essentially applies the EM algorithm independently to the two components $Q_\alpha(\alpha)$ and $Q_\eta(\eta)$ of Q, therefore, resulting in a more rapid convergence. However, the EM algorithm in his approach does not provide correct standard errors (Sastry, 1997). Therefore, he proposes to report the standard errors based on an estimate of the asymptotic covariance matrix constructed from the first derivatives of the incomplete-data log-likelihood function. A potential problem with this method is that with a piecewise baseline hazard, the standard errors based on the score may not be consistent as the number of intervals increases (Sastry, 1997).

The problem of accounting for clusters effects in survival analysis has been also approached from a Bayesian perspective. Recent work includes Clayton (1991) and Sinha (1993), who parameterize the baseline hazard function as an independent-increment gamma process. Aslanidou et al. (1998) uses a piecewise-constant baseline hazard function with correlated pieces.

Sargent (1998) develops a hierarchical Cox model, which is an extension of Cox's original model. He rewrites the proportional hazards frailty model, equation (2.27) as

$$h(t; x_{ij}, z_i) = h_0(t)\exp(x_{ij}^T \beta + z_i) \tag{2.29}$$

where $z_i = \log(w_i)$ and w_i is defined as a random effect in equation (2.27). He then uses the Cox's partial likelihood as the likelihood component for the model parameters and uses Markov Chain Monte Carlo (MCMC) methods (Smith and Roberts, 1993) to compute Bayesian quantities such as means, standard deviations and the marginal posterior density estimates. These methods allow sampling from the joint posterior density of the base model parameters and a set of random effects parameters that capture the dependence between observations in the same cluster. The advantage of this method is that an assumption for the distribution of the baseline hazard function is not needed since the partial likelihood is used. The use of appropriate MCMC methods also eliminates the need for assumptions about the distribution of the random effects. However, this approach is computationally intensive, and thus impractical for massive datasets (Sargent, 1998).

Applying the generalized linear mixed model (GLMM) to analyze multilevel survival data has been also considered in recent years (McGilchrist, 1993; Yau, 2001). The GLMM method starts with the construction of a log likelihood analogous to the likelihood associated with the best linear unbiased prediction (BLUP) (Henderson, 1975) based on the Cox partial likelihood. Parameter estimation is then achieved by maximization of this log likelihood at the initial step of estimation, and is extended to

obtain residual maximum likelihood (REML) (Thompson, 1980) estimators of the variance components.

Ma et al. (2003) also combine the random effects Cox's model with the generalized linear mixed model, but in a different way. They propose characterizing the random effects Cox's model as an auxiliary random effects Poisson regression model. Ma's approach allows an unspecified baseline hazard function and relies only on the first and the second moments of the random effects distributions. Using this approach, a Cox's model with two levels of nested random effects is proposed. There are $i = 1,...,m$ independent clusters, $j = 1,...,J_i$ correlated sub-clusters within the ith cluster, and $k = 1,...,n_{ij}$ individuals within each sub-cluster (i, j). Let U_i represent the cluster level random effect and U_{ij} represent the sub-cluster level random effect within the ith cluster. Instead of incorporating both random effects U_i and U_{ij} into the hazard function as equation (2.28) does, Ma et al. include only the sub-cluster level random effect U_{ij} in the model:

$$h(t; x_{ijk}, u_{ij}) = h_0(t) u_{ij} \exp(x_{ijk}^T \beta), \tag{2.30}$$

where t is the follow-up time and x_{ijk} is the vector of covariates for individual (i, j, k).

On the other hand, the cluster level random effect U_i is used to define the distribution of U_{ij}. It is assumed that U_i's are independent and identically distributed positive random effects with

$$E(U_i) = 1, \qquad \text{var}(U_i) = \sigma^2. \tag{2.31}$$

Furthermore, it is assumed that, given the cluster level random effects $U_* = u_* = (u_1,...,u_m)$, the sub-cluster random effects $U_{11},...,U_{mJ_m}$ are positive and conditionally independent, and that the conditional distribution of U_{ij} depends on $U_i = u_i$ only with:

$$E(U_{ij} \mid U_*) = U_i, \qquad \text{var}(U_{ij} \mid U_*) = \nu^2 U_i. \tag{2.32}$$

Ma et al. then define an auxiliary random effects Poisson regression model (to simplify, I omit the stratum in Ma's formula). Let $\tau_1,...,\tau_q$ denote the distinct failure times, with $m_h, h = 1,...,q$ indicating the number of failures. Let t_{ijk} be the observed survival time for individual (i,j,k) at time τ_h, and let $Y_{ijk,h}$ be 1 if that individual fails at that time, and 0 otherwise. The risk set at time τ_h is then defined as $\Re(\tau_h) = \{(i,j,k) : t_{ijk} \geq \tau_h\}$. Let Y and U denote the vectors of the $Y_{ijk,h}$ and the sub-cluster random effects U_{ij}, respectively. An auxiliary random effects Poisson model is then defined:

$$Y_{ijk,h} \mid u_{ij} \sim Po\{u_{ij} \exp(\alpha_h + x_{ijk}\beta)\} \quad ((i,j,k) \in \Re(\tau_h)), \tag{2.33}$$

where Po means a Poisson distribution. Given the random effects, the maximum conditional Poisson likelihood estimates $(\hat{\alpha}, \hat{\beta})$ for (α, β) satisfy the equation

$$\exp(\hat{\alpha}_h) = \frac{m_h}{\sum_{(i,j,k)\in\Re(\tau_h)} u_{ij} \exp(x_{ijk}^T \hat{\beta})}. \tag{2.34}$$

and a nonparametric estimate of the cumulative baseline hazard function is given by

$$\hat{\Lambda}_0(t) = \sum_{\tau_h \leq t} \exp(\hat{\alpha}_h). \tag{2.35}$$

With some algebra, Ma et al. show that the joint partial likelihood for the random effects Cox model ($l_p(\beta;Y,U)$) is equal to the joint Poisson likelihood ($l(\alpha,\beta;Y,U)$) multiplied by a constant (not dependent on the parameters of interest) term. Therefore, the maximum likelihood estimates and the inferences for β about the random effects Cox models can be made by fitting the auxiliary random effects Poisson model (Ma et al., 2003).

Thus, the random effects Cox Proportional hazards models specified by (2.30), (2.31) and (2.32) can be studied using the equivalent auxiliary random effects Poisson models specified by (2.33), (2.31) and (2.32). First, the random effects are estimated by the orthodox best linear unbiased predictors (Harvey, 1981; Jorgensen et al., 1996). To estimate regression parameters, the estimating equation is established by setting the first derivatives of the joint log likelihood of the auxiliary model for the data and the random effects to be equal to 0. The parameters in the estimating equation are $\gamma = (\alpha,\beta)$, whereα is the vector $\{\alpha_1,...,\alpha_h\}$ and β is the vector of regression coefficients. For detailed description of the equation, see Ma et al. (2003). The Newton scoring algorithm (Jorgensen et al., 1996) is used to solve this equation in order to provide the estimates of the regression parameters. Under mild regularity conditions, this equation can be shown to be asymptotically normal with asymptotic mean and asymptotic variance given by the inverse of the Godambe information matrix (Lele, 1991; Artes et al., 2000). If the dispersion parameters are unknown, then they can be estimated by the adjusted Pearson estimators. An analogue of Wald test is available for testing hypotheses about regression coefficients (Ma et al., 2003).

In the computational procedure, the initial values of the regression parameters are obtained as the estimates of coefficients from the standard Poisson regression model, with independence assumption. Initial estimates of random effects $\hat{U}_i$ and $\hat{U}_{ij}$ are given by the mean of the responses in cluster i divided by the mean of all responses, and the mean of the responses within sub-cluster (i, j) divided by the mean of all responses, respectively. Initial dispersion parameter estimates are calculated from the Pearson estimators. The algorithm then iterates between updating the regression parameters estimates, updating random effects and updating the dispersion parameter estimates (Ma et al., 2003).

Ma et al. use the proposed method to reanalyze the data from a large cohort study of air pollution and mortality (Pope et al., 1995). There are 574,438 subjects, nested in 151 cities, and the cities nested in 44 states. Twelve covariates are considered (sulphate particle level, smoking history, alcohol consumption, education, occupational exposures, body mass index and other potential risk factors for mortality). The corresponding nested random effects Cox model is defined by equation (2.30), and the state random effects U_i and the city random effects U_{ij} are described by (2.31) and (2.32), respectively. The survival time is the time from entry into the study to death or censoring at the end of the study or loss to follow-up. The authors used a C++ program to implement their method, and computation took more than 10 hours to complete the estimation. They compared the results from the standard Cox model, the Cox models with single level random effects at either city or state level, and the two-level nested Cox model. They found that the level of

statistical significance of association between city-level air pollution and mortality depended on the underlying assumptions of the fitted model.

The approach proposed by Ma et al. (2003) deals with an unspecified baseline hazard function and relies on the first and second moments of random effects only, which is very attractive. However, in their paper, the authors didn't mention any evaluation of the method in simulation. Moreover, the time used to analyze the air pollution data was very long. Moreover, the program is not publicly available and the authors were not willing to provide us with their software.

2.3.2 Marginal modeling for correlated survival data

The marginal hazard function for observations specified by the proportional hazards model may take two different forms (Wei et al., 1989; Liang et al., 1993). One assumes that the baseline hazard is the same for all the observations. The alternative approach assumes that the baseline hazard is the same only for the observations within the same cluster. The hazard functions for the $jth\,(j=1,...,n_i)$ observation of the $ith\,(i=1,...,m)$ cluster are written, respectively, as

$$h(t \mid x_{ij}) = h_0(t)\exp(x_{ij}^T\beta) \tag{2.36}$$

or $$h(t \mid x_{ij}) = h_{0j}(t)\exp(x_{ij}^T\beta), \tag{2.37}$$

where x_{ij} is the vector of covariates and β is the vector of regression coefficients.

Wei, Lin and Weissfeld (1989) use the second form (2.37) to model multivariate failure time data, where two or more distinct failures are recorded for each subject. The dependence structure among the distinct failure times of the same subject is unspecified. First, the data are stratified by the type of failure and a separate analysis of all regression coefficients for each stratum is carried out by using the standard Cox proportional hazards model. This gives a number of estimates of each regression coefficient which are consistent if the model is correctly specified. Wei, Lin and Weissfeld show that asymptotically, for large m (the number of subjects or clusters), the estimates of each regression coefficient from all strata are correlated and normally distributed. Next, the robust estimate of the covariance matrix of the regression coefficients estimates (Lin et al., 1989) is then computed. Finally, the last step is to combine the estimates of regression coefficients. Suppose that $\beta_1 = \beta_2 = ... = \beta_K$, where K is the number of strata, then the estimate of β is given by the linear combination of the $\beta_k's, k = 1,...,K$, that is, $\sum_{k=1}^{K} c_k \hat{\beta}_k$ with $\sum_{k=1}^{K} c_k = 1$. The weight $c = (c_1, c_2, ..., c_K)'$ is calculated according to the robust covariance matrix (Wei et al., 1985). If the $\beta_k's$ are unequal but with no qualitative differences among them, the combined $\hat{\beta}$ can be interpreted as the "average effect" of the covariates.

Weighted procedures to estimate regression parameters under both stratified and unstratified marginal proportional hazards models are developed by Cai and Prentice (1995) and Cai (1997). Analogous to the GEE approach developed by Liang and Zeger (1986), weights that account for failure time dependencies are introduced into the partial likelihood score equations. There are many possible choices for the weight matrices. In

particular, the weight matrices can be specified as the inverse of the correlation matrix for the marginal martingale (Cai et al., 1995). The estimates for regression coefficients are then shown to be consistent and asymptotically normally distributed (Cai et al., 1995; Cai, 1997).

For the unstratified marginal proportional hazards model, Liang et al. (1993) develop a pseudo-likelihood approach by taking at most one individual in the risk set for each cluster, thereby removing dependence within clusters. Lu and Wang (2005) apply the same principle, but devise a risk set sampling method to sample new risk sets that are composed of the independent individuals and preserve the marginal risk structure at each distinct failure time. The risk set sampling method consists of two steps. First, among all the individuals in a risk set, randomly choose one individual per cluster that is at risk (not including the cluster from which the failure comes). In the second step, a weight is given through a probability weighting for each individual chosen in step1; the probability weight is proportional to the number of nonfailures (still at risk) in the cluster from which that individual is chosen. Lu and Wang (2005) prove that the resulting estimates $\hat{\beta}$'s are consistent and converge to mean-zero normal distribution. Since this approach involves sampling the data, the precision and the efficiency can be improved by repeating the estimation procedure several times and taking the average of the estimates.

The "independence working model" approach, which takes the form (2.36), is described by Lee et al. (1992). The approach consists of two stages. In the first stage, all the data are analyzed using the standard Cox proportional hazards model, i.e. ignoring the

dependence between observations. The estimates obtained in this stage are used as the final estimates. The second stage attempts to estimate the variability of the estimates for regression coefficients, by means of the robust variance matrix, not relying on the assumption of independence. As long as the marginal model is correctly specified and censoring is independent, the estimates of regression coefficients are consistent and asymptotically normally distributed (Lee et al. 1992).

Lipsitz and Parzen (1996) proposed a "one-step jackknife" estimator of variance for Cox regression for correlated survival data, which is another "independence working model" approach, and accounts for both stratified and unstratified marginal models. Similar to the approach of Lee et al. (1992), "one-step jackknife" approach involves two stages. The estimates of regression coefficients are first obtained by using the standard Cox's model, treating all the observations as independent. The one-step jackknife estimator of variance (Lipsitz et al., 1994) is then used to estimate the variance of the regression coefficients. The one-step jackknife estimator of variance of $\hat{\beta}$ proposed by Lipsitz et al. (1994) is calculated as follows:

$$\left(\frac{m-p}{m}\right)\sum_{i=1}^{m}(\hat{\beta}_{-i}-\hat{\beta})(\hat{\beta}_{-i}-\hat{\beta})', \tag{2.38}$$

where m is the number of clusters, p is the dimension of $\hat{\beta}$, and $\hat{\beta}_{-i}$ is the estimate of β obtained by deleting the all the observations in cluster i, and performing one step of the Newton- Raphson algorithm, using $\hat{\beta}$ as the starting value. Thus, the variance estimator in (2.38) simply applies the leave-one-out approach to jackknife the entire clusters. The one-step jackknife variance estimate is proved to be asymptotically

equivalent to the robust variance estimate proposed by Wei et al. (1989) and Lee et al. (1992). Lipsitz and Parzen (1996) apply this method to analyze the survival data, which were originally analyzed by Wei et al. (1989). The data were obtained from a clinical trial to determine the effects of different doses of ribavirin in preventing the occurrence of HIV-1 positive virus in subjects with AIDS. Three blood samples for each patient were collected at different times. Therefore, the cluster consisted of three blood samples from the same patient. The failure time was defined as the number of days until the virus was detected in the blood sample. The estimates for β were obtained from the Cox's model with independence assumption, and the variance was estimated by the one-step jackknife method. The results show that the jackknife variance estimates are only slightly larger than the robust variance estimates proposed by Wei et al. (1989). The results from a small simulation, with 20 clusters and two observations per cluster, show that the jackknife variance estimates yield coverage rates for the 95% confidence intervals that are quite close to the nominal rate of 95%.

2.4 Overview of bootstrap methodology

Bootstrap is a very general data-based resampling method for statistical inference, which was introduced by Efron (1979; 1983). This method requires no theoretical calculations and can be applied in an automatic way to almost any data analysis, no matter how complicated (Davison et al., 1997). It is commonly used to estimate confidence intervals, but it can also be used to estimate bias and variance of an estimator.

2.4.1 Parametric VS nonparametric bootstrap

The essence of bootstrap is the concept that the distribution of the statistic of interest $\theta = t(F)$ can be approximated by estimates from repeated samples, drawn from an approximation to the unknown population. Here, F is the cumulative distribution function (CDF) of a set of independent, identically distributed (i.i.d.) observations: $y_1, y_2, ... y_n$, θ is called a statistic and t is a real-valued function whose domain includes the sample space of $(y_1, y_2, ..., y_n)$.

The most common approximations lead to parametric and nonparametric bootstraps. The parametric bootstrap assumes that the distribution of the data $F_\psi(y)$ is known except for the unknown parameter vector ψ. Then $F_\psi(y)$ can be approximated by $F_{\hat{\psi}}(y)$, with an estimate $\hat{\psi}$ for ψ. The nonparametric bootstrap is a method where the population distribution function is unknown and is approximated by the empirical distribution function (EDF) which puts equal probabilities $\frac{1}{n}$ at each value in the original sample.

2.4.2 Selecting bootstrap samples

For parametric bootstrap, the resampling procedure is very straightforward (Efron, 1993; Davison, 1997). However, in many applications, the parametric form of the distribution F is unknown. Therefore, we will consider the case of non-parametric bootstrap. Suppose we have i.i.d. data $y_1, y_2, ... y_n$ with an unknown distribution function. We use the EDF to estimate the unknown CDF. Because the EDF puts equal probabilities

on the original data values, each bootstrap sample $y_j^*, j = 1,...,n$, is a simple random sample, with replacement, of the observations.

The computation of the standard error $\sigma(\hat{\theta})$ of an estimator $\hat{\theta}$ can be approximated by Monte Carlo methods:

1. Draw a bootstrap sample $y_1^*, y_2^*,..., y_n^*$ by independent random sampling from the EDF $\hat{F}$. Obtain the estimate of θ from the bootstrap replication $\hat{\theta}^* = t(y_1^*, y_2^*,..., y_n^*)$.
2. Repeat step 1 "B" times, where B is a large number, typically $100 \le B \le 10{,}000$, obtaining independent bootstrap-based estimates $\hat{\theta}^{*1}, \hat{\theta}^{*2},..., \hat{\theta}^{*B}$ and approximate $\hat{\sigma}_B$ by

$$\hat{\sigma}_B = \left[\left(\sum_{b=1}^{B} (\hat{\theta}^{*b} - \hat{\theta}^{**})^2 \right) \Big/ (B-1) \right]^{1/2}, \hat{\theta}^{**} = \frac{\sum_{b=1}^{B} \hat{\theta}^{*b}}{B}. \qquad (2.39)$$

As $B \to \infty$, the bootstrap-based standard error $\hat{\sigma}_B$ converges to $\sigma(\hat{\theta})$ (Efron et al., 1983).

2.4.3 Bootstrap-based confidence intervals

In this section, I review four methods for computing bootstrap-based confidence interval.

1. Standard bootstrap confidence interval (Davison et al., 1997)

   Standard bootstrap confidence interval is based on the assumption that the estimator $\hat{\theta}$ is normally distributed with mean θ and variance σ^2. $\hat{\theta}$ is an estimate

calculated from the original data set, and using the bootstrap estimate of the standard error as in (2.39), an approximate $100(1-\alpha)\%$ confidence interval is given by $\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}_B$, where z follows a standard normal [0, 1] distribution.

2. Biased-corrected percentile method (Efron, 1981)

   The bootstrap estimate of bias is given by $Bias_B(\hat{\theta}) = \hat{\theta}^{**} - \hat{\theta}$, where $\hat{\theta}^{**}$ is defined in (2.39). The bootstrap bias corrected estimate is then given by $\hat{\theta}_B = \hat{\theta} - (\hat{\theta}^{**} - \hat{\theta}) = 2\hat{\theta} - \hat{\theta}^{**}$. Accordingly, using the same approach as above, the bias-corrected $100(1-\alpha)\%$ confidence interval is defined to be

   $$\hat{\theta}_B \pm z_{\alpha/2}\hat{\sigma}_B.$$

3. Bootstrap percentile confidence interval (Efron, 1981)

   This is a more direct approach for constructing a $100(1-\alpha)\%$ confidence interval for θ. The method is based on the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of the empirical distribution of B bootstrap estimates $\hat{\theta}^{*b}, b = 1,...,B$. Thus, the approximate confidence interval is given by $\left[\hat{\theta}_B(\alpha/2), \hat{\theta}_B(1-\alpha/2)\right]$.

4. Studentized bootstrap method (Davison et al., 1997)

   The studentized bootstrap is based on a different bootstrap distribution than the other bootstrap-based estimates of the confidence intervals. The estimate $\hat{\theta}_i^{*b}$ and its standard error $s_{\hat{\theta}_i^{*b}}$ from each bootstrap sample are used to calculate studentized estimates, $t_i^b = (\hat{\theta}_i^{*b} - \hat{\theta})/s_{\hat{\theta}_i^{*b}}$. Then, the $100(\alpha/2)$ and $100(1-\alpha/2)$ quantiles of the distribution of t_i^b are used to calculate the confidence interval for θ.

2.4.4 Extensions to non- i.i.d. data

The bootstrap methods discussed so far are appropriate for a single sample of i.i.d. observations. However, many problems involve observations that are not i.i.d. Examples include regression problems and hierarchical data.

2.4.4.1 Regression data

In regression analysis, the effects of covariates on a response variable are of interest. To estimate regression coefficients, we need the response value and the covariates values. Therefore, there are two general resampling approaches for such data: resampling the observations, also called case resampling, and resampling the residuals, also called error resampling (Efron et al., 1993; Davison et al., 1997).

Case resampling is an approach that considers the data as a sample from some multivariate distribution F of (X,Y). The regression coefficients are viewed as statistical functions of F (Davison et al., 1997). The resampling therefore involves sampling the cases with replacement, where each case is a vector of covariates values and a response value. For each bootstrap resample, the regression model is fitted to get the resample-specific estimates of regression coefficients.

Error resampling involves three steps (Davison et al., 1997). First, the regression model is fitted to the original sample and residuals, $\varepsilon_i = y_i - x_i^T \beta$, are calculated for each observation. Then a series of bootstrap samples of residuals $(\varepsilon_1^*, \varepsilon_2^*, ..., \varepsilon_n^*)$ is drawn with replacement from the observed residuals. The bootstrap sample of observations is

constructed by $(x_1^*, y_1^*), (x_2^*, y_2^*), ..., (x_n^*, y_n^*)$, where $x_i^* = x_i$ and $y_i^* = x_i^T \beta + \varepsilon_i^*$. Finally, the regression model is fitted to the bootstrap sample to give the estimates of coefficients for a given resample.

Resampling observations and resampling residuals are asymptotically equivalent (Efron, 1986). The choice of bootstrap depends on the goal and context of the analysis. Bootstrapping residuals maintains the structure of the covariates, but the inference assumes that the model used to calculate the residuals is appropriate. Bootstrapping observations repeats some covariate values and omits others. It is a useful choice when the analysis involves models selection.

2.4.4.2 Bootstrapping censored data

Right censoring often occurs in survival data. In this case, the failure time observed is $t = \min(T, t_c)$, where t_c is a censoring value, and T is a non-negative failure time, which is known only if $T \leq t_c$. d is used to indicate censoring, which equals one if T is observed and equals zero if t_c is observed.

The simplest model for censoring is random censorship, under which t_c is a random variable independent of T. There are several ways to resample censored data. Under the random censorship model, the simplest way is to apply case sampling i.e. to resample observations from the original data $(x_1, t_1, d_1), (x_2, t_2, d_2), ..., (x_n, t_n, d_n)$ (Davison et al., 1997). Conditional bootstrap is another way to resample the censored data (Davison et al.,

1997). In this method, the expected time-to-event (T_j^*) and the censored time ($t_{c,j}^*$) are generated separately from the corresponding distributions. Then the minimum value is chosen from these two values to be the observed time-to-event. Since $t_{c,j}^*$ is generated conditional on the censoring status d_j, this method is called conditional bootstrap.

2.4.4.3 Hierarchical data

In biomedical science, many collected data have a hierarchical or clustered structure. The most basic structure of such data can be expressed as

$$y_{ij} = x_i + z_{ij}, i = 1,...,m, j = 1,...,n_i,$$

where x_i and z_{ij} are independent random variables, at the cluster and individual level, respectively. The feature of this model that complicates resampling is the correlation between observations within a cluster,

$$\text{var}(y_{ij}) = \text{var}(x_i + z_{ij}) = \text{var}(x_i) + \text{var}(z_{ij}) = \sigma_x^2 + \sigma_z^2,$$

$$\begin{aligned}\text{cov}(y_{ij}, y_{ik}) &= \text{cov}(x_i + z_{ij}, x_i + z_{ik}) \\ &= \text{cov}(x_i, x_i) + \text{cov}(x_i, z_{ik}) + \text{cov}(x_i, z_{ij}) + \text{cov}(z_{ij}, z_{ik}) \\ &= \sigma_x^2, j \neq k.\end{aligned}$$

There are two strategies for nonparametric resampling of such nested data. The first stage for both strategies is to randomly sample groups, with replacement. At the second stage, the individual observations are randomly sampled within the selected groups, either without replacement (strategy 1) (Davison et al., 1997) or with replacement (strategy 2) (Abrahamowicz et al., 1998; Davison et al., 1997). Note that strategy 1 keeps selected

groups intact. Davison et al.(1997) show that strategy 1 underestimates the covariance of observations, while strategy 2 overestimates the covariance. Strategy 1 more closely mimics the variation properties of the data if the number of cluster is moderately large. Both strategies work well if both cluster size and the number of clusters are large (Davison et al., 1997).

3 Objectives

The literature review, presented in Chapter 2, indicates that the random effects Cox's models for analysis of correlated survival data either need make assumptions on the distributions of random effects or involve complicated computations. On the other hand, most methods for estimating the variance in the marginal modeling are also very complicated, especially for right-censored time-to-event data. Thus, in this thesis, I attempt to propose and validate an easy-to-implement method for the right-censored correlated survival data. Specifically, in order to obtain valid confidence intervals for the regression coefficients estimates, I propose to apply the computer-intensive bootstrap procedure, which requires no assumptions, no complicated computations and becomes more feasible with constant progress in computing techniques. I expect that this approach will yield reliable standard errors and confidence intervals for the Cox's proportional hazards analysis of clustered data.

To this aim, the following specific objectives will be addressed:

1) To propose novel resampling algorithms, specially adapted for randomly censored hierarchical survival data.
2) To evaluate the performance of the proposed methods through simulations, in order to:
    a) Validate the methods for inference for the Cox's proportional hazards analysis of clustered data, and, in particular:

i) To evaluate how within-cluster correlation of outcomes affects the accuracy of the regression coefficients estimates and the standard errors estimates from the conventional Cox's proportional hazards model.

ii) To compare the standard errors estimates and 95% CI coverage rates for conventional Cox's PH model, "classical" one-step bootstrap method and the two bootstrap methods proposed in this thesis.

iii) To assess how the performance of the proposed methods varies depending on either the strength of random effects or the cluster size.

iv) To assess the impact of the number of bootstrap resamples on the accuracy of the inference based on the proposed bootstrap-based methods.

b) Further validate the proposed resampling methods by comparing their performance with that of the "classic" GEE model (available in commercial statistical packages such as SAS or S-plus) for a binary un-censored outcome.

4 Methods

4.1 Overview of Simulation Design and Data Generation

I simulated a hypothetical study with N patients clustered within practices of M physicians. The study design and variables were based on the general features of an empirical study that focuses on assessing whether time until referral of patients with Rheumatoid Arthritis (RA) to specialists differs according to geographical, patient or physician-related characteristics. Individual patients' times-to-event were generated from the exponential distribution with hazard conditional on several patient-level variables, several physician-level (cluster-level) variables, and physicians' "random effects (intercepts)". Random right censoring was applied, and the strength of the clustering effect was controlled by manipulating the variance of random effects.

4.1.1 Physicians' variables

I generated M=50 physicians indexed by $j, j = 1,...,50$. For each physician, I generated values of three covariates independent of each other: gender (X_1) from a binomial distribution with $P(X_1 = 1) = 0.6$, where $X_1 = 1$ indicates a man; age (X_2) from a uniform distribution $U[30,65]$, and the geographical region (X_3) from a multinomial distribution with probabilities shown in the table:

Category(X_3)	Probability in each category	Corresponding dummy variables
1	0.4	reference
2	0.3	D_1
3	0.2	D_2
4	0.1	D_3

The fourth cluster-level variable, specialty (X_4) of a physician ($X_4 = 0$ and $X_4 = 1$ indicating, respectively, a general practitioner or a specialist) was generated from a binomial distribution with the probability $P(X_4 = 1)$ conditional on the physician's gender and the geographical region:

$$P(X_{4,j} = 1 \mid X_{1,j}, D_{1,j}, D_{2,j}) = \begin{cases} 0 & \text{if } D_{3,j} = 1 \\ 0.3 + 0.15 * X_{1,j} + (-0.1) * D_{1,j} + (-0.25) * D_{2,j} & \text{if } D_{3,j} = 0 \end{cases}$$

Note that $\Pr ob(X_4 = 1 \mid D_3 = 1) = 0$, which reflects the reality in which all physicians in a remote region are general practitioners.

A latent variable (i.e. random effect) (ε) for a physician was generated from a normal distribution with mean =0, and variance = log (1.5), which was used to create clustering effect.

Finally, the number of patients (i.e. the cluster size) n_j for the *jth* physician was generated from a normal distribution with the expected value conditional on the characteristics of the physician. This involved two steps:

1. Calculate the "expected number of patients" M_j of the *jth* physician, conditional on physician's characteristics:

$$M_j = 40 + X_{1,j} * 10 - 5 * \left| X_{2,j} - 40 \right| / 10 - 5 * D_{1,j} - 10 * D_{2,j} - 20 * D_{3,j} \quad (4.1)$$

Specifically, to mimic a realistic scenario, I assumed that men have higher number of patients, and that the number of patients is the highest for physicians aged 40 years, and decreases for both younger and older physicians.

2. Generate n_j from a normal distribution with the expected value:

$$\begin{cases} n_j \sim N[M_j,20], & if\ X_{4,j} = 0 \\ n_j \sim N[1.5 * M_j,15], & if\ X_{4,j} = 1 \end{cases}$$

This implied that the mean number of patients (with a given doctor) is higher for specialists ($X_{4,j} = 1$) and the variance is lower.

If n_j is less than 10, then n_j is set to 10.

4.1.2 Patients' variables

Three characteristics: gender, age and severity of disease, were generated for the ith $(i = 1,...,n_j)$ patient within the practice of the $jth(j = 1,...,50)$ physician.

4.1.2.1 Patients' characteristics

The patient gender (Z_1) was generated from a binary distribution. The probability of $Z_1 = 1$ (patient is a man) depends on the gender of the physician (X_1). E.g. female patients are more likely to have female physician. So, I generated $Z_{1,ij}$ in two steps:

1) If $X_{1,j}$ = 0 (physician is a woman) then the expected proportion of $Z_{1,j}=1$, $E_j(Z_1)$, was generated from a uniform distribution: U [0.2, 0.4]. Otherwise, $E_j(Z_1)$ was generated from U [0.3, 0.6].
2) The gender was generated from a binary distribution with $P[Z_{1,ij}=1]=E_j(Z_1)$.

The age (Z_2) was generated from a log-normal distribution in three steps:

1) Generate the expected mean of patients' age for the *jth* physician: $E_j(Z_2)\sim$ Normal [60, 3];
2) Generate r_{ij} from a normal distribution, $N\left[\log\left(E_j(Z_2)\right), \log(3)\right]$;
3) Finally, generate $Z_{2,ij}=\exp(r_{ij})$.
4) Next, the patient's age was restricted to [20,100] interval, if $Z_{2,ij}$<20 then $Z_{2,ij}$=20; if $Z_{2,ij}$>100 then $Z_{2,ij}$=100.

The disease severity (Z_3) was generated from a normal distribution conditional on the specialty (X_4) of the physician. In particular, based on empirical evidence, I assumed that mean severity of disease is higher for patients seen by a specialist ($X_{4,j}=1$). First, the expected mean of patients' disease severity $E_j(Z_3)$ was generated from a normal distribution N [20, 3] if $X_{4,j}=0$, or from N [30, 4] if $X_{4,j}=1$. Individual patient's severity $Z_{3,ij}$ is then generated from normal distribution with the physician-specific mean generated in the first step: $\left(N\left[E_j(Z_3),5\right]\right)$. $Z_{3,ij}$ was set to 1 if $Z_{3,ij}$<1.

4.1.2.2 Patients' times-to-event

The expected time to event t_{ij} was generated from an exponential distribution with the patient-specific hazard rate λ_{ij}. The hazard rate λ_{ij} depends on the *jth* physician's and the *ith* patient's characteristics, as well as the *jth* physician's random effect ε_j. In addition, random right censoring times, independent of patients or physicians' characteristics and of the outcome, were then generated.

4.1.2.2.1 Generation of the expected times-to-events

First, I generated the hazard rate λ_{ij} assuming the Cox proportional hazards model

$$\lambda_{ij} = \lambda_0 * \exp[\beta^T X], \tag{4.2}$$

where $\lambda_0 = 1$ is the baseline hazard, corresponding to all patient's and physician's covariates equal to 0; X is the vector of covariates $(1, X_{1,j}, X_{2,j}, X_{3,j}, X_{4,j}, Z_{1,ij}, Z_{2,ij}, Z_{3,ij}, \varepsilon_j)$, and β is the vector of the corresponding regression coefficients $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \gamma_1, \gamma_2, \gamma_3, \theta)$. Here, $X_{3,j}$ was changed from the defined four categories variable to a binomial variable by setting $X_{3,j} = 1$ for region 3 and 4, and setting $X_{3,j} = 0$ for region 1and 2. The values of regression parameters: β = {0, log (0.8), 0, log (0.6), 0, 0, log (1.02), log (1.05), 1} were used first. Thus, I assumed that physician's age, physician's specialty, and patient's gender have no effect on hazards. λ_{ij} was set to 0.01 if λ_{ij} <0.01.

With the generated hazard λ_{ij}, I then generated the expected time-to-event t_{ij}. Since t_{ij} was assumed to follow an exponential distribution with the parameter λ_{ij}, the survival function is: $S(T = t_{ij}) = \exp(-\lambda_{ij} * t_{ij})$. Let $u_{ij} = S(T = t_{ij}) = P\left[t > t_{ij} \mid \lambda_{ij}\right]$. Then, $u_{ij} = \exp(-\lambda_{ij} * t_{ij})$ and the correspondingly time-to-event t_{ij} is, therefore, calculated by the formula:

$$t_{ij} = \frac{-\log(u_{ij})}{\lambda_{ij}} . \qquad (4.3)$$

Therefore, the individual times-to-event were generated in two steps:

1) Generate u_{ij} from Uniform [0, 1] distribution.

2) Calculate the corresponding t_{ij} using equation (4.3).

4.1.2.2.2 Generation of losses to follow-up

The possible time of loss-to-follow-up (c_{ij}) was also generated from an exponential distribution with λ_c = exp (1.2). This value was selected so that the censoring rate would be around 30%. Then the same method as above was used:

1) Generate $d_{ij} \sim$ Uniform [0, 1].

2) Calculate c_{ij} = -log (d_{ij})/ λ_c .

4.1.2.2.3 Generation of the " observed" event or censoring times

Finally, the final "observed" data, i.e. follow-up time τ_{ij} and the censoring status δ_{ij} were generated. I assumed that all subjects will be censored at $T = 1.0$, correspondingly to a

hypothetical "administrative" end of the study. The algorithm consists of the following steps:

1) If $t_{ij} < c_{ij}$ and $t_{ij} < 1.0$ then $\tau_{ij} = t_{ij}$, $\delta_{ij} = 1$.

2) If $t_{ij} \geq c_{ij}$ and $c_{ij} < 1.0$ then $\tau_{ij} = c_{ij}$, $\delta_{ij} = 0$.

3) If $t_{ij} > 1$ and $c_{ij} > 1$ then $\tau_{ij} = 1$, $\delta_{ij} = 0$.

The SAS program, written to generate the above data and to implement the bootstrap algorithms described in section 3.2, is included in Appendix A.

4.2 Bootstrap algorithm

According to the general principles of bootstrap methodology, discussed in section 2.4, my bootstrap algorithm consists of three steps. The first step involves data resampling. The second step focuses on the estimation of the standard errors of regression coefficients estimates. The last step involves estimation of the 95% confidence intervals.

Three different bootstrap methods are studied. The three methods share the procedures of stage 2 and stage 3, but differ in the first stage. The resampling step (stage 1) accounts for the hierarchical structure of my data, with individual patients nested within practices of their physician. I used three different strategies to resample the hierarchical data.

Strategy 1 limits resampling to a random resampling of the data at the lower level (i.e. patient-level) with replacement. This strategy assumes that all the observations are independent of each other and, thus, completely ignores the clustering effects. *PROC*

SURVEYSELECT in SAS was used to directly, randomly sample the patients, using simple individual sampling with replacement.

Strategy 2 includes two steps. First, I randomly resample individual physicians (the higher level), with replacement. Then, the bootstrap sample includes all the patients of each selected physician without any resampling of individual patients within physicians' practices. *PROC SURVEYSELECT* was used to resample the physicians. *PROC SQL* in SAS was then used to pick up all the patients of the selected physicians from the original data.

Strategy 3 also includes two steps. However, in contrast to strategy 2, here both physician and patients-within-practices are bootstrapped. The first step is the same as in strategy 2, i.e. involves resampling of individual physicians, with replacement. The second step is to randomly resample patients of each selected physician, with replacement. *PROC SURVEYSELECT* was used to implement resampling at the two levels and *PROC SQL* was used to pick up observations from the databases.

In all three strategies, the second stage is to estimate the standard errors ($\hat{\sigma}_B$) of the regression coefficients estimates. As discussed in the section 2.4.2, each of B resulting bootstrap samples was first independently analyzed with the standard Cox PH model. Next, the bootstrap-based standard errors (SE) of the regression parameter was estimated as the empirical standard deviation of B corresponding estimates $\hat{\theta}^{*1}, \hat{\theta}^{*2}, ..., \hat{\theta}^{*B}$, by using the equation (2.39)

$$\hat{\sigma}_B = \left[\left(\sum_{b=1}^{B}(\hat{\theta}^{*b} - \hat{\theta}^{**})^2\right)\Big/(B-1)\right]^{1/2}, \hat{\theta}^{**} = \frac{\sum_{b=1}^{B}\hat{\theta}^{*b}}{B}.$$

At the third step, I used the standard bootstrap-based confidence interval (See section 2.4.3) to construct the 95% confidence interval. This method is based on the normality assumption of the distribution of regression coefficients estimates. I used the estimate ($\hat{\theta}$) of the regression coefficient from the standard Cox's proportional hazards analysis of the original sample and the bootstrap-based standard error ($\hat{\sigma}_B$) to construct the 95% confidence intervals as $\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}_B$, where Z follows a standard normal [0, 1] distribution.

4.3 Data Analysis

One hundred independent random samples were generated, using assumption and methods described in section 4.1. Each dataset was analyzed using 4 approaches: (a) the standard Cox's PH model, (b) the bootstrap with the strategy 1 resampling method, (c) the bootstrap with the strategy 2 resampling method, and (d) the bootstrap with the strategy 3 resampling method. For each approach, the following covariates were included in the multivariable Cox's PH model, regardless of their statistical significance: the patient-level covariates: age, gender and disease severity, and the physician-level covariates: age, gender, specialty and working region.

In approach (a), the regression coefficients, the corresponding standard errors, and the 95% confidence interval were estimated directly from the standard Cox's PH model. In

approaches (b), (c) and (d), the regression coefficients estimates were also obtained from the standard Cox's PH model, while the standard errors were estimated using the respective bootstrap strategy (section 4.2), and the confidence intervals were estimated using bootstrap-based SE and assuming normality.

I first calculated for each physician-level and patient-level covariate, the mean $\overline{\hat{\beta}}$ of the 100 regression coefficients, estimated from each of 100 independent simulated samples. The bias of the estimates was calculated as the difference between the mean of 100 estimates $\overline{\hat{\beta}}$ and the true β . The root mean square error (RMSE), that combines the bias and the variance of the estimates, was calculated as the square root of the mean of the squared errors $\left(\hat{\beta}_i - \beta\right)^2$. Notice that the mean estimate $\overline{\hat{\beta}}$, the bias and RMSE, were, by definition, the same across all models. The main focus was on assessing the accuracy of the variance estimates. To this end, I compared the mean of the 100 standard errors, estimated using approaches (a)-(d), to the empirical standard error, calculated as the observed standard deviation of the $\hat{\beta}$ estimates from 100 individual samples. Then, the coverage rate of the 95 percent confidence interval was estimated as the proportion of simulation samples in which the confidence interval, calculated according to a given approach (a)-(d), included the true β .

5 Results

Three different scenarios for data generation were considered. The data generation described in Chapter 4 (Methods) corresponded to scenario 1. The motivation for using additional scenarios 2 and 3 was to assess the robustness of the conclusions with respect to the strength of the clustering effect. With all the other parameters the same as those in scenario 1, the standard deviation of the random effects was reduced from log(1.5) to log(1.25) in scenario 2. The goal was to assess to what extent the clustering affects the results even if the random effects were relatively weak. Finally, in scenario 3, the parameter that controlled the cluster size, "expected number of patients" M_j, was reduced by half.

5.1 The intra-correlation of the generated data

Censoring of survival times makes it difficult to measure the intra-class correlation. To give a sense of the strength of the within-cluster correlation of the generated survival times, I estimated the intra-class correlation coefficient (ICC) of the individuals' hazards, used to generate these times. It should be noticed, however, that this approach is feasible only in simulations. In real life, individual patients' hazards remain unknown.

5.1.1 The effect of the variance of random effects

In data generation, the random effects (ε) of physicians were used to produce the intra-class correlation. The strength of the clustering effect was controlled by manipulating the

variance of the random effects. I generated 100 data sets in scenario 1 (with $SD(\varepsilon) = \log(1.5)$) and additional 100 data sets in scenario 2 (with $SD(\varepsilon) = \log(1.25)$). Table 5.1 compares the means of the 100 ICCs for the two scenarios. As expected, ICC is stronger when variance of random effects is larger.

Table 5. 1The effect of the variance of random effects on the intra-class correlation

$Sd(\varepsilon) = \log(1.5)$			$Sd(\varepsilon) = \log(1.25)$		
$\overline{ICC}$	SD (ICC)	range	$\overline{ICC}$	SD (ICC)	Range
0.3235	0.0531	0.2333-0.5082	0.2760	0.0408	0.1967-0.3758

5.1.2 The effect of the cluster size

The cluster size n_j of jth physician can be controlled by changing the "expected number of patients" M_j of that physician (See Chapter 4). Since the number of patients for each physician (i.e. the cluster size) may affect the impact of the intra-cluster correlation, I compared the means of the 100 ICCs in scenario 1 and in scenario 3. Table 5.2 shows that $\overline{ICC}$ is decreased only slightly when the cluster size is reduced about 50%.

Table 5. 2The effect of the cluster size on the intra-class correlation

$M_j^{'}$ (original cluster sizes)			$M_j^{''}$ (cluster sizes reduced by 0.5)		
$\overline{ICC}$	SD (ICC)	range	$\overline{ICC}$	SD (ICC)	Range
0.3235	0.0531	0.2333-0.5082	0.3018	0.0578	0.1828-0.5807

5.2 Conventional Cox's proportional hazards model

This section summarizes the results of the three simulation scenarios obtained with the conventional Cox PH model, which ignored clustering of patients within physicians' practices.

First, for each covariate and each scenario, the mean of the 100 estimates yielded by the standard Cox's model was calculated. Then the relative bias was estimated as the ratio $(\bar{\hat{\beta}} - \beta)/\beta$ when true $\beta \neq 0$. The results are shown in table 5.3.

Table 5. 3 Bias and relative bias of log hazard ratios for the standard Cox's PH model

Variables		True effect β	$\bar{\hat{\beta}} - \beta$			Relative bias[1] $(\bar{\hat{\beta}} - \beta)/\beta$		
			Scen.1	Scen.2[3]	Scen.3[4]	Scen.1[2]	Scen.2[3]	Scen.3[4]
Physician-level	Gender	-0.223	-0.04012	-0.00375	0.00095	0.1799	0.0168	-0.0043
	Age	0	0.00050	0.00018	0.00134	NA	NA	NA
	Region	-0.5108	0.04101	-0.00786	-0.01222	-0.0803	0.0154	0.0239
	Specialty	0	-0.05053	-0.00215	-0.00977	NA	NA	NA
Patient-level	Gender	0	-0.00057	0.00302	0.00193	NA	NA	NA
	Age	0.0198	-0.00147	-0.00055	-0.00139	-0.0742	-0.0278	-0.0702
	Severity	0.0488	-0.00299	-0.00160	-0.00430	0.06127	0.03279	0.08811

[1] Relative bias cannot be estimated when the true $\beta = 0$.

[2] The data generation for Scenario 1 is described in Chapter 4.

[3] With all the other parameters the same as those in Scenario 1, the standard deviation of the random effect was reduced from log (1.5) to log (1.25) in Scenario 2.

[4] With all the other parameters the same as those in Scenario 1, the parameter that controlled the cluster size, "expected number of patients", was reduced by half in Scenario 3.

For both physician-level variables and patient-level variables, the bias is very close to 0, as the mean values of the estimates $\bar{\hat{\beta}}$ are very close to the true values of the

corresponding regression parameter β . Accordingly, most of relative biases are very small except for the log hazard ratio of physician's gender in scenario 1.

5.3 Comparison of the standard errors obtained with the conventional Cox's model and the three bootstrap methods

This section starts by summarizing the results of Scenario 1, with strong clustering and larger cluster sizes. For each independent variable, Table 5.4 compares the means of the 100 estimated standard errors, for the conventional Cox's model and the three bootstrap methods (with B=100 bootstrap resamples) described in Chapter 4. All of these four methods use the standard Cox's model to estimate the regression coefficients but employ different ways to estimate the variance of the regression coefficients. First, for each covariate, the empirical standard error ($SD(\hat{\beta})$) is shown, as the standard deviation of 100 $\hat{\beta}'s$, each from an independent random sample. To enhance the comparability and the interpretability, the results related to estimated standard errors (SE) are shown as the ratio of the mean of 100 model=specific SE estimates, from independent random samples, to the observed standard deviation of the estimates $SD(\hat{\beta})$.

As shown in table 5.4, the $\bar{\hat{\sigma}}$ for the standard Cox's model is systematically lower than the corresponding empirical standard deviation of the estimates ($SD(\hat{\beta})$). In particular, for each physician-level covariate, the conventional estimate of SE is below 50% of the corresponding $SD(\hat{\beta})$. The $\bar{\hat{\sigma}}$ for the bootstrap method with strategy 1, in which patients (lower-level units) were directly resampled, using individual random sampling with

replacement, is also much lower than the corresponding empirical standard deviation (Table 5.4). This was expected as the bootstrap strategy 1 did not account for the clustering. In contrast, the last two columns of Table 5.4 show that the $\overline{\hat{\sigma}}$ for the bootstrap strategy 2 and 3 are generally similar to the $SD(\hat{\beta})$. However, the estimated standard errors are systematically lower than the empirical standard errors for strategy 2 and systematically higher for strategy 3.

Table 5. 4 Comparison of mean of the 100 estimated standard errors for the standard Cox's model, and the three bootstrap methods, with the empirical standard error of the estimates

Variable		True effect β	$SD(\hat{\beta})$ [1]	$\overline{\hat{\sigma}} / SD(\hat{\beta})$ [2]			
				Standard Cox	Bootstrap (Strategy1)[3]	Bootstrap (strategy2)[4]	Bootstrap (Strategy3)[5]
Physi-	gender	-0.223	0.19927	0.3987	0.2876	0.8752	1.1348
cian	age	0	0.01058	0.4036	0.2977	0.8856	1.1815
	region	-0.5108	0.19256	0.4060	0.3136	0.8369	1.1106
	specialty	0	0.15227	0.4384	0.3303	0.8995	1.1960
Patient	gender	0	0.05099	0.9553	0.7221	0.7996	1.6493
	age	0.0198	0.00083	0.9518	0.7108	0.8072	1.6627
	severity	0.04581	0.00690	0.5841	0.4217	0.7652	1.1986

[1] $SD(\hat{\beta})$ indicates the actual standard deviations of the 100 standard Cox's model's estimates, each from an independent sample, of the corresponding regression coefficient, which approximates the empirical standard error of a given coefficient.

[2] Last 4 columns of Table 5.4 show the ratio of the mean of 100 estimates of $SE(\hat{\beta})$, obtained from the corresponding method, to the empirical standard error $SD(\hat{\beta})$. Accordingly, ratio lower than 1.0 indicates that a given method under-estimates the true variance of the coefficient, where ratio>1.0 indicates over-estimation of the true variance.

[3] In bootstrap strategy 1, the patients (lower-level units) were directly resampled using individual random sampling with replacement.

[4] In bootstrap strategy 2, only physicians (clusters) are randomly resampled with replacement..

[5] In bootstrap strategy 3, both physicians (clusters) and patients-within-physician are randomly resampled with replacement.

Since the variation of the regression coefficients for the correlated data comes from both between-cluster and within-cluster variation, the standard errors will be under-estimated if either variation is not taken into account, such as in the standard Cox's model and in the conventional bootstrap strategy 1 that both ignore between-physicians variation. In contrast, the first step for both strategy 2 and 3 is to randomly resample physicians with replacement, therefore mimicking the variability between physicians (i.e. clusters). For each selected physician in step 1, in strategy 2, all the original patients are selected. In contrast, strategy 3 randomly resamples the patients of each selected physician with replacement, thus, inducing additional variability, at the patient level. This difference led to the different estimates of the standard error. Strategy 2 slightly underestimated the standard error, while strategy 3 slightly overestimated the standard error (Table 5.4).

For each method, the coverage rates of the nominal 95 percent confidence interval for each regression coefficient are shown in Table 5.5. Due to under-estimation of the standard errors (Table 5.4), the conventional Cox's model and the bootstrap method 1, which resamples only patients, yielded very low coverage rates. This is especially evident for the physician-level covariates, for which all the coverage rates are below 60%. However, even for patient level covariates, the coverage rates may be as low as 58% to 73% (Table 5.5). The fact that under-estimation of the variance of regression coefficients is more dramatic for physician-level covariates suggests ignoring the cluster effect affects more cluster-level standard errors. In contrast, the bootstrap method with strategy 2 and strategy 3 yielded much better coverage rates. As described above, strategy 2 slightly underestimated the standard error and, therefore, the corresponding coverage rates vary

between 80% and 90%. Strategy 3 slightly overestimated the standard error; therefore, the coverage rates are between 92% and 100%. There is one exception for strategy 2, related to the very low coverage rate for patient's age (only 44%). This can be explained by the ratio of the bias of the corresponding estimate relative to its estimated standard error. From tables 5.3 and 5.4, we can see that the bias of the estimate for patient's age is -0.00147, while the $\overline{\hat{\sigma}}$ is only 0.00067. The absolute bias, even if small, is two times higher than the mean of the estimated standard errors and, therefore, the probability of the 95 percent confidence interval including the true effect was very low.

Table 5. 5 Comparison of the coverage rates for the standard Cox's model, and the three bootstrap methods

Variable		Coverage rate (%)[1]			
		Standard Cox	Bootstrap (Strategy1)	Bootstrap (strategy2)	Bootstrap (Strategy3)
Physician	gender	52 (41.83-62.01)[2]	36 (26.82-46.27)	89 (80.78-94.11)	96 (89.49-98.71)
	age	49 (38.94-59.13)	39 (29.56-49.3)	91 (83.17-95.54)	99 (93.76-99.95)
	region	56 (45.74-65.8)	47 (37.04-57.2)	87 (78.44-92.62)	95 (88.17-98.14)
	specialty	59 (48.7-68.6)	46 (36.09-56.22)	88 (79.6-93.37)	95 (88.17-98.14)
Patient	gender	93 (85.62-96.9)	83 (73.89-89.51)	85 (76.15-91.09)	100 (95.39-100)
	age	58 (47.71-67.67)	35 (25.91-45.26)	44 (34.2-54.26)	92 (84.39-96.23)
	severity	73 (63.04-81.16)	61 (50.7-70.44)	83 (73.89-89.51)	95 (88.17-98.14)

[1] The coverage rate of the 95 percent confidence interval was estimated as the proportion of simulation samples in which the confidence interval included the true β .

[2] The exact 95% confidence interval (including continuity correction) for the coverage rate (out of 100) was calculated.

5.4 Assessing the effect of the number of bootstrap resamples

The standard error estimates from the bootstrap methods are empirical approximations. Accordingly, their asymptotic properties are justified by the law of large numbers (Efron, 1979). The estimate of the standard error by a bootstrap method converges to the true standard error as the number of the bootstrap resamples (B) increases. Therefore, in additional simulations, I have varied the number of bootstrap resamples and investigated its impact on the standard errors and coverage rates. Table 5.6 compares the mean of the 100 standard error estimates $\overline{\hat{\sigma}}$ and the coverage rates for the double-resampling bootstrap method (strategy 3) with B=100, 300 and 500. As shown in the table, no material differences in either $\overline{\hat{\sigma}}$ or the coverage rate were observed with increasing number of resamples. This implies that B=100 bootstrap resamples are sufficient to obtain reasonably accurate variance estimates, which substantially reduces computing time compared to larger B values.

Table 5. 6 Impact of the number of bootstrap resamples on the standard error estimates and the coverage rates from the double bootstrap method (strategy 3)

Variable		$SD(\hat{\beta})$	B=100		B=300		B=500	
			$\overline{\hat{\sigma}}$	Coverage rate (%)[1]	$\overline{\hat{\sigma}}$	Coverage rate (%)	$\overline{\hat{\sigma}}$	Coverage rate (%)
Physician -level	Gender	0.19927	0.22614	96 (89.49-98.71)[2]	0.22787	97 (90.85-99.22)	0.22621	96 (89.49-98.71)
	Age	0.01058	0.01250	99 (93.76-99.95)	0.01250	99 (93.76-99.95)	0.01243	99 (93.76-99.95)
	Region	0.19256	0.21386	95 (88.17-98.14)	0.21547	95 (88.17-98.14)	0.21519	95 (88.17-98.14)
	Specialty	0.15227	0.18212	95 (88.17-98.14)	0.18299	96 (89.49-98.71)	0.18332	96 (89.49-98.71)
Patient -level	Gender	0.05099	0.08410	100 (95..39-100)	0.08372	100 (95..39-100)	0.08380	100 (95..39-100)
	Age	0.00083	0.00138	92 (84.39-96.23)	0.00139	93 (85.62-96.9)	0.00138	92 (84.39-96.23)
	Severity	0.00690	0.00827	95 (88.17-98.14)	0.00830	96 (89.49-98.71)	0.00827	96 (89.49-98.71)

[1] The coverage rate of the 95 percent confidence interval was estimated as the proportion of simulated samples in which the confidence interval included the true β .

[2] The exact 95% confidence interval (including continuity correction) for the coverage rate (out of 100 simulated samples) is shown in brackets.

5.5 Assessing the normality of the distribution of bootstrap estimates of regression coefficients

As described in Chapter 4, I used the standard bootstrap confidence interval, which is based on the assumption of normality of the distribution of the regression coefficients estimates. To verify this assumption, I tested the normality of the distribution of 100 bootstrap–based estimates of each regression coefficient. All **p**-values from the alternative tests (Kolmogorov-Smirnov statistic, Anderson-Darling statistic and Cramer-von Mises statistic) for normality are >0.15, indicating lack of evidence to reject the null hypothesis of normality. The corresponding histograms also support the assumption that the estimates are normally distributed. Since histograms for all the regression coefficients

estimates are similar, I arbitrarily chose, for illustration, the histograms for one physician-level covariate (gender) and one patient-level covariate (severity) (figure 5.1 and figure 5.2, respectively).

Figure 5. 1 Distribution of 100 bootstrap estimates of the log hazard ratio for physicians' gender

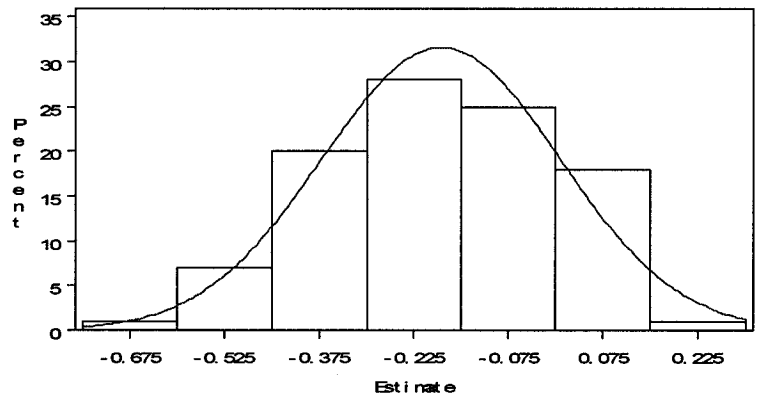
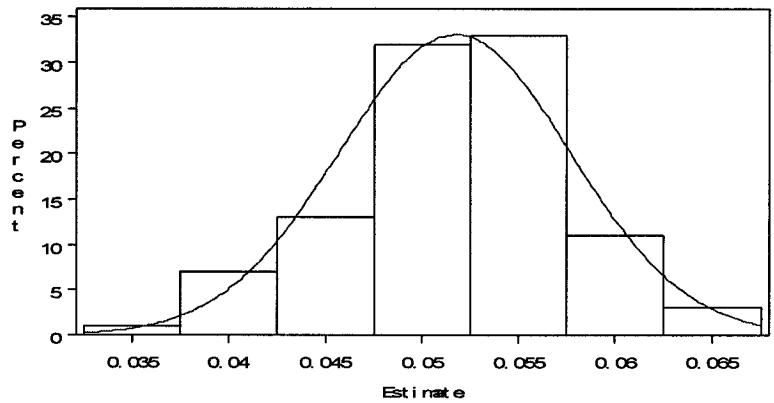


Figure 5. 2 Distribution of 100 bootstrap estimates of the log hazard ratio for patients' severity



5.6 Comparison of bootstrap-based standard errors with Generalized Estimating Equations (GEE) for a binary outcome

The approach I propose in this thesis relies on standard Cox's model to get parameter estimates and then uses bootstrap to adjust the standard errors for clustering. In this sense, the approach is similar to marginal models such as GEE (Liang et al., 1986), and it would be interesting to use simulations to demonstrate the similarity of the results of the two approaches. However, there is no generally available software for GEE analysis of clustered right-censored time-to-event data. Therefore, I have carried out an additional simulation experiment, in which I generate binary outcomes and then compare the standard errors for bootstrap with strategy 2 and 3 against the conventional GEE results, obtained from PROC GENMOD in SAS.

5.6.1 Brief description of the additional simulations

First, I used the same data generation algorithm as described in Chapter 4 for my original simulations, to generate all the physician-level covariates (i.e. gender, age, region, specialty and number of patients) and all the patient-level covariates (i.e. gender, age and severity). Next, instead of generating the times-to-event as the outcome for each patient, I generated a binary outcome (y_i) for each patient. First, I calculated $\log(\frac{p_i}{1-p_i})$, where p_i is the probability of $y_i = 1$, assuming the multiple logistic regression model :

$$\log\left(\frac{p}{1-p}\right) = \beta^T X \cdot$$

The same vector of covariates X as in the original simulation was considered, and the same values for regression coefficients: $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \gamma_1, \gamma_2, \gamma_3) = \{0, \log(0.8), 0, \log(0.6), 0, 0, \log(1.02), \log(1.05)\}$ were assumed. However, $\beta's$ represent now logOR's rather than logHR's. Next, to induce clustering within physicians' practices, a random intercept ε_j was generated for each physician $j = 1,...,50$ from N [0, log(1.5)] and added to the logit for each patient of this physician, to obtain:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta^T X_{ij} + \varepsilon_j \ .$$

Then, y_{ij} was generated from a binary distribution with the probability of $y_{ij} = 1$ equal to $\exp(\log it_{ij}) \Big/ (1 + \exp(\log it_{ij}))$, where $\log it_{ij} = \log\left(\frac{p_{ij}}{1-p_{ij}}\right)$.

One hundred independent random samples of size N (varied around 2000), were simulated using the above algorithm. Then, each sample was analyzed using 4 alternative methods: standard multiple logistic regression, GEE extension of logistic model with exchangeable covariance structure (Digger et al., 1994), and bootstrap methods with strategy 2 and 3 (see Chapter 4).

5.6.2 Brief summary of results of the additional simulations

Table 5.7 and table 5.8 show that, as expected, the standard logistic model underestimated the standard errors of the estimates and, accordingly, yielded very low coverage rates. In contrast, GEE and the two proposed bootstrap methods gave acceptably accurate results.

Compared with the GEE model, the bootstrap method with strategy 2 (resampling with replacement limited to physicians) gave very similar but slightly lower coverage rates, while the bootstrap method with strategy 3 (resampling both physicians and patients-within-physician with replacement) gave a somewhat higher coverage rates. These results confirm that the proposed bootstrap approaches are similar to GEE modeling.

Table 5. 7 Simulations with a binary outcome: comparison of mean of the 100 estimated standard errors for the standard logistic model, GEE, bootstrap methods with strategy 2 and with strategy 3.

Variable		True effect β	$SD(\hat{\beta})$ [1]	$\overline{\hat{\sigma}} / SD(\hat{\beta})$ [2]			
				logistic model	GEE[3]	Bootstrap (strategy2)[4]	Bootstrap (Strategy3)[5]
Physi-cian	gender	-0.223	0.2872	0.276602	0.98085	0.905641	1.174791
	age	0	0.0168	0.254167	0.910714	0.85119	1.089286
	region	-0.5108	0.3397	0.230144	0.964086	0.916691	1.180748
	specialty	0	0.333	0.20045	0.917718	0.885285	1.153453
Patient	gender	0	0.0988	0.493016	0.804656	0.758097	1.571862
	age	0.0198	0.00151	0.523179	0.860927	0.721854	1.529801
	severity	0.04581	0.0147	0.27415	0.537415	0.748299	1.108844

[1] $SD(\hat{\beta})$ indicates the actual standard deviations of the 100 standard logistic model's estimates, each from an independent sample, of the corresponding regression coefficient, which approximates the empirical standard error of a given coefficient.

[2] Last 4 columns of Table 5.7 show the ratio of the mean of 100 estimates of $SE(\hat{\beta})$, obtained from the corresponding method, to the empirical standard error $SD(\hat{\beta})$. Accordingly, ratio lower than 1.0 indicates that a given method under-estimates the true variance of the coefficient, where ratio>1.0 indicates over-estimation of the true variance.

[3] Generalized Estimated Equations are used to analyze binary data with logit link function.

[4] In bootstrap strategy 2, only physicians (clusters) are randomly resampled with replacement.

[5] In bootstrap strategy 3, both physicians (clusters) and patients-within-physician are randomly resampled with replacement.

Table 5. 8 Simulations with a binary outcome: comparison of coverage rate for the standard logistic model, GEE, bootstrap methods with strategy 2 and with strategy 3.

Variable		95% Coverage rate (%)[1]			
		Standard logistic model	GEE	Bootstrap (strategy2)	Bootstrap (Strategy3)
Physician	gender	55 (44.75-64.86)[2]	93 (85.62-96.23)	92 (84.39-96.23)	99 (93.76-99.95)
	age	46 (36.09-56.22)	91 (83.17-95.54)	91 (83.17-95.54)	95 (88.17-98.14)
	region	49 (38.94-59.13)	93 (85.62-96.23)	92 (84.39-96.23)	97 (90.85-99.22)
	specialty	47 (37.04-57.2)	95 (88.17-98.14)	91 (83.17-95.54)	98 (92.26-99.65)
Patient	gender	91 (83.17-95.54)	90 (81.96-94.84)	85 (76.15-91.09)	99 (93.76-99.95)
	age	51 (40.87-61.06)	40 (30.48-50.3)	33 (24.12-43.21)	87 (78.44-92.62)
	severity	62 (51.71-71.36)	80 (70.57-87.08)	75 (65.16-82.88)	93 (85.62-96.23)

[1] The coverage rate of the 95 percent confidence interval was estimated as the proportion of simulation samples in which the confidence interval included the true β .

[2] The exact 95% confidence interval (including continuity correction) for the coverage rate (out of 100) was calculated.

6 Real-life applications

A real-life application of the proposed methods uses data collected within the project "Optimal care trajectories in rheumatoid arthritis (RA): the primary-secondary interface" (OPTRA) (Feldman et al, 2004) (The assumptions underlying data generation for my simulations were, to a large extent, derived from this project). The OPTRA project aims to evaluate parameters potentially associated with the primary-secondary care interface, which is characterized by the shared care that is initiated when a primary care physician requests a consultation with a rheumatologist for a RA patient. RAMQ (Régie d'Assurance Maladie du Québec) and Med-Echo database were used to identify RA patients, the characteristics of patients and physicians, and the place of residence.

One of the main objectives of the OPTRA project is to assess whether time until consultation with a rheumatologist differs according to geographical region, and/or patient or physician-related characteristics. Time to consultation is measured since the diagnosis of RA, and its analysis requires using time-to-event methodology. Since patients are nested within physicians' practices, *clustered* censored survival data analysis needs to be considered. Therefore, I applied the proposed bootstrap-based methods to the analyses of relevant data from the OPTRA project.

A total of 13,244 incident RA patients nested within practices of 3,866 physicians with specialty different from rheumatology, were identified from the database. Time 0 for each RA patient is defined as the first non-rheumatologist physician claim for RA, and the "event" of interest is the first-time consultation with a rheumatologist. Subjects who had

not seen a rheumatologist are censored at the end of follow-up, death or relocation out of Quebec. Patients' characteristics (age, gender and comorbidity), physicians' characteristics (gender, specialty and years from graduation), and an indicator of the urban (vs rural) residence were included in the regression model. Since most clusters (physicians) have very small cluster sizes (number of patients), it is not possible to resample patients in the second stage (one of the proposed bootstrap approaches). Thus, only the bootstrap method involving resampling only physicians was applied. Table 6.1 compares the results from the bootstrap method with those from the standard Cox's PH model.

Table 6. 1 Analysis of OPTRA time-to-events data with the standard Cox's PH model and the proposed bootstrap method

Variable		$\hat{\beta}$ [1]	Standard Cox[2]		Bootstrap-based Cox[3]	
			stderr	p-value	Bootstrap-based stderr	p-value
Patient -level	age	-0.00381	0.00124	0.0021	0.00188	0.0427
	gender	0.26103	0.04170	<0.001	0.05049	<0.001
	comorbidity	0.03434	0.01161	0.0031	0.01504	0.02245
Physician -level	gender	-0.86324	0.05611	<0.001	0.13006	<0.001
	specialty	-0.75017	0.05900	<0.001	0.13255	<0.001
	Grad. years	0.0005574	0.0001375	<0.001	0.0001415	<0.001
rural		-0.11963	0.04870	0.0140	0.11111	0.0734

[1] $\hat{\beta}$ is the estimate of regression coefficient, which is the log of hazard ratio of the corresponding covariate.
[2] The standard error and the p-value of the standard Cox's model are given by SAS procedure: PHREG.
[3] For the bootstrap method, the standard error is bootstrap-based standard error, which was estimated by using the empirical standard deviation of bootstrap-based estimates of regression coefficients. The p-value was calculated by finding the probability of $|t| \geq \hat{\beta}/SE_{boot}$

Since the OPTRA project is at the preliminary stage, and the data set has not been properly cleaned and validated, I would not interpret the estimated effects of covariates (The analysis presented here is just a tentative analysis, to illustrate the proposed bootstrap approach). As expected, the standard errors estimated from the standard Cox's model are systematically smaller than the corresponding bootstrap-based standard errors. Accordingly, the significance levels for the effects of some covariates have changed. For example, the effect of the geographical variable is significant, at the $\alpha = 0.05$ level, in the standard Cox's model but not significant for the bootstrap-based method. This confirms that the standard Cox's PH model, that fails to account for the correlation between observations on the patients of the same physician, may underestimate the standard error and, accordingly, may lead to incorrect inference about the effects of particular covariates. Similar to the results of simulations reported in Chapter 5, the results for physician-level characteristics are more affected by the failure to account for the clustering, as the standard errors are often twice smaller for the conventional analyses than for the bootstrap.

7 Discussion and Conclusion

In this thesis, I first used simulation to assess the regression coefficients estimates and the standard errors estimates of the conventional Cox PH model for correlated survival data. As expected, conventional Cox PH model yielded unbiased estimates of the effects of both individual-level and cluster-level covariates. However, due to ignoring the correlations between observations within the same cluster, the standard errors estimates are much smaller than the actual standard deviation of the estimates, "directly" observed in simulations. Thus, the corresponding 95% confidence intervals are too narrow. And furthermore, the coverage rates of the 95% confidence interval (CI), especially for cluster-level covariates, may be as low as 40%-60%, which suggests that ignoring the clustering effect affects more inference about cluster-level characteristics. The "classical" bootstrap method, which directly resampled patients with replacement, also ignored the correlation between observations within the same cluster and therefore, underestimated the standard errors, yielded very narrow 95% confidence interval, and gave very low 95% CI coverage rates (35% - 50%).

To estimate such inaccuracies while avoiding complicated random-effects extension of the Cox's model (Ma et al., 2003), I have proposed an easy-to-implement bootstrap-based approach with two variants. I then investigated the performance of the two proposed variants in simulations. The bootstrap method with strategy 3, which first randomly resampled physicians with replacement and then randomly resampled patients for each selected physician with replacement, took both between-cluster variance and within-cluster variance into account. Indeed, simulations suggested that

the resulting standard errors estimates were very close to, but slightly *higher*, than the true standard errors. Accordingly, the 95% CI coverage rates were between 92% and 100%. The bootstrap method with strategy 2 only randomly resampled physicians with replacement, and then included all the individual patients of each selected physician into bootstrap samples. This method accounted for the between-cluster variation and kept the original set of individual observations for each selected cluster (physician). In simulations, the standard errors estimates based on this method were also very close to, although slightly *lower* than, the empirical standard errors, observed in simulations. The 95% CI coverage rates varied between 80% and 90%.

The latter bootstrap method was applied in a real-life example, involving the assessment of the determinants of the time between the RA diagnosis and the first consultation with a rheumatologist. The results confirmed the importance of accounting for clustering as the conventional Cox's model under-estimated the standard errors, compared to bootstrap-based standard errors, by a factor of two or more, especially for physicians' characteristics.

The bootstrap-based method is an empirical approximation method, which asymptotic properties are based on the law of large numbers. Therefore, I investigated the impact of the number of bootstrap resamples on the standard errors and coverage rates. When I increased the number of bootstrap resamples from 100 to 300, and then to 500, the standard errors estimates and the coverage rates were not significantly changed. This

suggests that 100 bootstrap resamples are sufficient to obtain reliable variance estimates.

The assumption of normality of the empirical distribution of the bootstrap-based regression coefficients estimates, on which the 95% confidence interval inference was based, was also verified in simulations. The histograms showed that the bootstrap-based estimates of the effects of both individual-level and cluster-level covariates were approximately, normally distributed.

I also investigated the impact of (i) the cluster size, and (ii) of the strength of random effect on the clustering effect. The intra-class correlation coefficient (ICC) of the individuals' hazards became weaker when the variance of random effects was smaller or when the cluster size was reduced. As expected, the coverage rates for conventional Cox PH model were improved as the ICC decreased, but remained too low, in contrast to accurate coverage for the proposed bootstrap methods.

To gain further insight into the performance of the proposed approach, I applied the two bootstrap-based approaches to correlated data with binary outcomes and compared the standard errors estimates and 95% CI coverage rates with those of GEE method, which is available in SAS package. As expected, conventional multiple logistic regression for independent observations under-estimated the true variance of regression coefficients. In contrast, GEE and the two proposed bootstrap methods gave acceptably accurate and quite similar results. These additional simulations

confirmed my expectations that the proposed bootstrap-based methods are similar, in both their spirit and their results, to marginal models such as those underlying the GEE approach. Indeed, both GEE and my bootstrap-based methods first estimate regression coefficients and then correct their standard errors for the within-clusters correlations (Liang et al., 1986).

The proposed bootstrap methods were implemented in the SAS programming language. It took 5 minutes for bootstrap method with strategy 3 to analyze one simulation sample with 50 clusters and about 2,500 individual observations, when using 100 bootstrap resamples. It took much less time for the bootstrap method with strategy 2 to analyze the same data, because only physicians were re-sampled in this method.

Overall, the above results suggest that the proposed bootstrap-based methods provide a reasonable, accurate and easy-to-implement approach for estimating standard errors and confidence intervals in the context of the Cox's proportional hazards analysis of correlated survival data.

Clearly, further studies and more computer simulations are needed to better understand the performance of the two proposed bootstrap methods and the difference between these two methods. So far, I just investigated the relatively simple scenarios. For example, I assumed that survival times follow an exponential distribution (i.e. constant hazards rates). Moreover, only random intercepts are considered, and the

effects of covariates are assumed constant over time, as in proportional hazards model (Cox, 1972). Future research should consider more complicated situations, such as using different distributions of survival times, taking into account random slopes, investigating how to adapt the proposed methods to hierarchical data with more than two levels, and extending them to non-proportional hazards models.

Some limitations of this study have to be acknowledged. First, I assumed that censoring was independent of covariates and therefore, I did not need to consider whether individual observations are censored or not when resampling the data. Yet, the resampling methods must take the censoring into account if the censoring is differential (Davison et al., 1997).

Another limitation is that when I calculated the 95% confidence interval, I assumed that the estimates of regression coefficients follow a normal distribution and did not consider possible bias in point estimates. However, in my simulations, normality was never significantly violated and most regression coefficients showed little or no bias. Thus, from the pragmatic point of view, such possibly over-simplifying assumptions do not seem to materially affect the accuracy of the proposed methods. Still, the empirical coverage rate of one covariate was quite low even when I used the proposed bootstrap methods with strategy 2 to estimate the standard errors. This suggests that other methods for bootstrap-based confidence interval estimation should be perhaps considered, such as biased-corrected percentile method (Efron, 1981).

Moreover, in my thesis, the number of clusters was fixed and pretty large (m=50). However, the number of clusters may also affect the performance of the proposed bootstrap methods and this issue needs to be investigated further.

As mentioned in the literature review, the "one-step" jackknife approach for correlated survival data, proposed by Lipsitz et al. (1996), is similar to my methods except that they used different resampling technique – jackknife (vs. bootstrap) to estimate the variance. The different performance of the two approaches should be compared in the future studies.

Lastly, I proposed two alternative bootstrap-based methods to resample the hierarchical survival data, and both of them were found to yield reasonable accurate standard errors so far. However, as suggested by Davison and Hinkley (1997), the two-stage method (strategy 3) tended to slightly over-estimate the true variance of regression coefficients, whereas, the method 2 (strategy 2), that resample physician only, under-estimated the true variance. Thus, future studies should assess which strategy is more robust in more complicated situations.

Overall, the results suggest that the proposed bootstrap methods, yield reasonable estimates of the standard errors and very good coverage rates. The programs implementing these methods are simple, and run fast and therefore, may be preferred by the applied statisticians to analyze correlated, censored survival data.

Appendix A

1 Generate Data

```
%macro simul;

%do  k=1 %to 100; *k is the number of simulations;
/*******************************************
 Generate physician's characteristics
*******************************************/
%macro physicians(index);
/*Physician's gender x1, binomial distribution with p=0.6*/
      data gender;
      lable x1= "phys_gender";
      do j=1 to 50;
      x1=ranbin(0,1,0.6);
      output;
      end;
      run;

      /* Add physician's age x2, continous variable, U[30,65]*/
      data physicians;
      set gender;
      lable x2="phys_age";
      x2=35*ranuni(0)+30; * since ranuni(0):[0,1];
      x2=round(x2,1); * Round age to an integer;
      run;


      /* Add region x3 where a physician belongs to. Categorical data
         with p(x3=1,2,3,4)=0.4, 0.3, 0.2, 0.1 */
      data physicians;
      set physicians;
      label x3="phys_region";
      x3=rantbl(0,0.4,0.3,0.2,0.1);
      if(x3=2) then do; d1=1; d2=0; d3=0; end;
      else if(x3=3) then do; d1=0; d2=1; d3=0; end;
      else if(x3=4) then do; d1=0; d2=0; d3=1; end;
      else do; d1=0; d2=0; d3=0;end;
      run;

      /* Add physician's specialty x4, binary data. The
         probability depends on x1 and x3 */
      data physicians;
      set physicians;
      lable x4="phys_specialty";
      if (x3=4) then x4=0;
      else do;
      p=0.3+0.15*x1-0.1*d1-0.25*d2;
      x4=ranbin(0,1,p);
      end;
      drop p;
      run;

      /* Add number of patients for each physician n: depends on
```

```
         each physician's characteristics.*/
      data physicians;
      set physicians;
      m=(40 + x1*10 +abs(x2-40)/2 - 5*d1 -10*d2 -20*d3);
      /*n: number of patients, normal distr.*/
      if(x4=0) then n=20*rannor(0)+m;
      else n=15*rannor(0) + 1.5*m;
      n=round(n,1);
      if n<10 then n=10;
      drop m;
      run;

      /* Add individual physician's random effect-- epsilon: normal
         distr.) */
      data physicians;
      set physicians;
      epsilon=rannor(0)*log(1.5);
      run;

      libname phys "c:\xiao\project_thesis\library\physicians";
      data phys.physicians&index;
      set physicians;
      run;
%mend;

%physicians(&k);


/*********************************************************
  Generate patient's characteristics
 *********************************************************/
%macro patients;
%do j=1 %to 50;
      data physician&j;
      set physicians(firstobs=&j obs=&j);
      run;

      /* patient's gender, binary data, depends on gender of
         physician*/
      data exp_gender;
      set physician&j;
      if x1=0 then exp_z1=0.2*ranuni(0)+0.2;
      else exp_z1=0.3*ranuni(0)+0.3;
      call symput("exp_z1", exp_z1);
      call symput("n",n);
      run;

      /* Add patient's gender z1: binomial distr.*/
      data patients;
      label z1="pati_gender";
      %do i= 1 %to &n;
      id=&i;
      z1=ranbin(0,1,&exp_z1);
      output;
      %end;
      run;
```

```
      /* Add patient's age z2,(not depend on physician), lognormal
         distr.*/
      data exp_age;
      exp_age=rannor(0)*3+60;
      call symput("exp_age",exp_age);
      run;

      data patients;
      set patients;
      label z2="pati_age";
      /* r:normal distr.with mean=log(exp_age)*/
      r=log(3)*rannor(0)+log(&exp_age);
      z2=exp(r);
      z2=round(z2,1);
      if z2<20 then z2=20;
      if z2>100 then z2=100;
      drop r;
      run;

      /* Add patient's disease severity z3, depends on physician's
         specialty(x4)*/
      data exp_severity;
      set physician&j;
      if x4=0 then exp_z3=3*rannor(0)+20;
      else exp_z3=4*rannor(0)+30;
      call symput("exp_z3", exp_z3);
      run;

      data patients&j;
      set patients;
      label z3="pati_severity";
      z3=5*rannor(0)+&exp_z3;
      z3=round(z3,1);
      if z3<=1 then z3=1;
      run;
%end;
%mend;
%patients;

/*****************************************************
  Generate time_to_event for each patient
 *****************************************************/
%macro time(coef1,coef2,coef3,coef4,coef5,coef6,coef7,coef8,coef9);
%do j=1 %to 50;
      data _null_;
      set physicians(firstobs=&j obs=&j);
      call symput("x1",x1);
      call symput("x2",x2);
      call symput("x4",x4);
      If x3=1 or x3=2 then r=0;
      else r=1;
      call symput("x3",r);
      call symput("epsilon", epsilon);
      run;

      /*Generate hazard rate (lamda) specific for each patient
        depending on the characteristics of physicain and patient,
```

```
        as well as phyiscian's random effects*/
      data hazard;
      set patients&j;
      lamda=exp(&coef1+&coef2*&x1+&coef3*&x2+&coef4*&x3+&coef5*&x4+
               &coef6*z1+&coef7*z2+&coef8*z3+&coef9*&epsilon);
      if lamda<0.01 then lamda=0.01;
      run;

      /* Generate time-to-event with losses to follow-up, the time
         follows exponential distribution. */
      data time;
      set hazard;
      /* Generate t: the time when an event occurre */
      u=ranuni(0);
      t=-log(u)/lamda;
      /* genreate c: the follow-up time for each patient;*/
      lamda_c=exp(1.2);
      d=ranuni(0);
      c=-log(d)/lamda_c;
      /* Compare c & t to see if a datum is censored
         censor=0 -- no event&censored censor=1--event occurred*/
      if c<t then do;time=c; censor=0; end;
      else do;time=t; censor=1; end;
      /* Since the study lasts 1 year, all the patients who
         don't have an event will be censored */
      if time>1 then do; time=1; censor=0;end;
      drop u d lamda_c;
      run;

      data patients&j;
      set time(drop=t c);
      run;
%end;
%mend;
%time(0,-0.223,0,-0.5108,0,0,0.0198,0.0488,1);

%macro merge_patients(index);
libname pati "c:\xiao\project_thesis\library\patients";
%do j=1 %to 50;
      proc append base=pati.patients&index data=patients&j;
      run;
%end;
%mend;
%merge_patients(&k);
%end;
%mend;

%simul;

/* Merge physicians with patients;*/
%macro add_physicians;
%do k=1 %to 100;
      data physicians&k(drop=l);
      set phys.physicians&k;
      do l=1 to n;
            output;
      end;
```

```
      run;

      data pati.patients&k;
      merge pati.patients&k physicians&k;
      run;

      %macro cal_ICC;
      proc mixed data=pati.patients&k;
      class j;
      model lamda=/solution;
      random int/subject=j type=un;
      ods output CovParms=table(keep=covparm estimate);
      run;

      proc transpose data=table out=table1;
      run;

      data icc;
      set table1;
      ICC=col1/(col1+col2);
      keep ICC;
      run;

      proc append base=pati.correlation data=icc; run;
      %mend;
      %cal_ICC;
%end;
%mend;

options nolabel;
%add_physicians;
```

2 SAS program for bootstrap method: to resample physicians first, then to resample patients of each selected physician.

```
/*     For each simulation, 100 bootstraps are done (First on
       physicians then on patients).
       For each bootstrap sample, run Cox model.
       --- I get Beta for each variable
       --- I calaculate mean and sd of 100 betas
*/


/*************************************************************
      Function: resample---Generate random ID for resample
*************************************************************/

%macro resample(out=, num=,id=);
      data id;
      do &id=1 to &num;
            output;
      end;
```

```
      run;

      proc surveyselect data=id method=urs n=&num out=sample noprint;
      run;

      data &out(keep=&id);
      set sample;
      if numberhits^=0 then do;
            do count=1 to numberhits;
                  output;
            end;
      end;
      run;
%mend;

/*************************************************************
      Function: meanCoef---calculate mean and standard deviation
                of 100 bootstap-based coefficients
*************************************************************/
/* Mean of coefficients of 100 bootstraps and standard deviation */
%macro meanCoef(bb=,name=);
      /* Calculate mean of coefficients and standard errors*/
      %do j = 1 %to 7;
            data new;
            set coef&j(firstobs=1 obs=&bb);
            run;

            proc means data= new;
            output out=mean&j(drop=_type_ _freq_);
            run;

            data mean&j(rename=(estimate=estimate&j));
            set mean&j;
            if _stat_ eq "MEAN" or _stat_ eq "STD";
            run;
      %end;

      data b_estimate;
      merge mean1 mean2 mean3 mean4 mean5 mean6 mean7;
      run;

      proc transpose data=b_estimate
      out=b_estimate(rename=(col1=meanOfBeta col2=sd));
      run;

      /*Put the mean and sd of beta's from each simulation together*/
      %do j=1 %to 7;
            data rowData;
            set b_estimate(firstobs=&j obs=&j);
            run;

            proc append base=&name&j data=rowData;
            run;
      %end;
%mend meanCoef;

/*************************************************************
```

```
     Bootstraps
************************************************************/

libname pati "c:\xiao\project_thesis\library\patients";
libname analysis "c:\xiao\project_thesis\library\analysis\b_analysis";
libname phys "c:\xiao\project_thesis\library\physicians";
%macro bootstraps;
%let b=100;   /* 100 bootstraps */
%do sim= 1 %to 100;/* 100 simulations */
      /* resample physicians and patients: 100 times ;*/
       %do res_num=1 %to &b;
            %resample(out=phys_id, num=50,id=j); *sample physicians;

            proc sql;
            create table physicians as
            select phys_id.j, n
            from phys.physicians&sim s, phys_id t
            where s.j=t.j;

            /* resample patients */
            %do count=1 %to 50;
                  data _NULL_;
                  set physicians(firstobs=&count obs=&count);
                  call symput('pati_num',n);
                  call symput('phys_id',j);
                  run;

                  /*sample patients for each chosen physician*/
                  %resample(out=pati_id,num=&pati_num,id=id);
                  data pati_id;
                  set pati_id;
                  j=&phys_id;
                  run;

                  proc append base=total_pati_id data=pati_id;
                  run;
            %end;
            /* Pick up patients according to the resample ID of
               patients */
            proc sql;
            create table b_patients as
            select *
            from pati.patients&sim s, total_pati_id t
            where s.j=t.j and s.id=t.id;

            /* Use Cox model on the sample*/
            data b_patients;
            set b_patients;
            If x3=1 or x3=2 then x3=0;
            else x3=1;
            run;

            proc phreg data=b_patients;
            model time*censor(0)=x1-x4 z1-z3;
            ODS SELECT ParameterEstimates;
            ODS OUTPUT ParameterEstimates=
                        b_summary(keep=variable estimate);
```

```
            run;

          /* Put the results of 100 bootstraps together by variable*/
            %do j=1 %to 7;
                  data rowData;
                  set b_summary(firstobs=&j obs=&j);
                  run;

                  proc append base=coef&j data=rowData;
                  run;
            %end;
            proc datasets nolist;
            delete total_pati_id;
            quit;
      %end;
%meanCoef(bb=100, name=analysis.coef100_);
%end;
%mend;

%bootstraps;
```

Bibliography

1. Abrahamowicz, M., Mackenzie, T. and Esdaile, J.M. (1996). Time-dependent hazard ratio: modeling and hypothesis testing with application in Lupus Nephritis, *Journal of the American Statistical Association*, **91**, 1432-9
2. Abrahamowicz, M., Fortin, P.R., Berger, R., Nayak, V., Neville, C. and Liang, M.H. (1998). The relationship between activity and expert physician's decision to start major treatment in active systemic lupus erythematosus: a decision aid for development of entry criteria for clinical trials. *The Journal of Rheumatology* **25**, 277-84
3. Artes, R. and Jorgensen, B. (2000). Longitudinal data estimating equations for the dispersion models. *Scand. J. Statist.* **27**, 321-34
4. Aslanidou, H., Dey, D.K., and Sinha, D. (1998). Bayesian analysis of multivariate survival data using Monte Carlo methods. *Canadian Journal of Statistics* **26**, 33-48
5. Balakrishnan, N., and Malik, H.J. (**1987**): Moments of order statistics from truncated log-logistic distributions, *Journal of Statistical Planning and Inference*, 17, 251-267.
6. Berkson, J. and Gage, R.P. (1950) Calculation of survival rates for cancer. Mayo Clin Proc Staff Meeting, **25**, 270–86.
7. Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89-99
8. Cai, J. and Prentice, R.L., (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika* **82**, 151-64

9. Cai, J. and Prentice, R.L. (1997). Regression estimation using multivariate failure time data and a common baseline hazard function model. *Lifetime Data Analysis* **3**, 197-213.

10. Campbell, M.K. and Grimshaw, J.M.. (1998). Cluster randomized trials: time for improvement. The implications of adopting a cluster design are still largely being ignored [editorial]. *BMJ.* ; **317**, 1171–2.

11. Clayton, D.G. (1991): A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* **47**: 467-85.

12. Cox, D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc.* **B58**, 657-9.

13. Cox, D.R. (1975). Partial likelihood. Biometrika, 62, 269-75.

14. Cutler, S.J. and Ederer, F. (1958). Maximum utilization of the life table method in analyzing survival. *J Chron Dis* **8**, 699–712.

15. Davison, A.C. and Hinkley, D.V.(1997). *Bootstrap Methods and their application.* Cambridge university press.

16. Dawber, T.R. (1980). *The Framingham study: the epidemiology of Atherosclerotic disease.* Cambridge, Massachusetts: Havard University press.

17. De Leeuw, J., Kreft, and Ita, G.G. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics*, **11**, 57-85.

18. Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1-38

19. Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994). Analysis of longitudinal data. Oxford, United Kingdom: Oxford University Press.

20. Donner, A. and Klar, N. (1994). Methods for comparing event rates in intervention studies when the unit of allocation is a cluster. *Am J Epidemiol* **140**,279–89

21. Donner A and Klar N. (2000). *Design and analysis of cluster randomization trials in health research.* London, UK; Arnold.

22. Efron, B. (1977). Efficiency of Cox's likelihood function for censored data, *Journal of the American Statistical Association* **72**: 557–65.

23. Efron, B. (1979). Bootstrap methods: another look at the Jackknife. *Annals of Statistics* **7**, 1-26

24. Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics* **9**, 139-72

25. Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife and cross-validation. *The American Statistician* **37**, 36-48

26. Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap.* Chapmen and Hall, New York.

27. Fahrmeir, L. and Tutz, G.T. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd Edition, Springer-Verlag, New York.

28. Gehan, E.A. (1969). Estimating survival functions for the life table. *Journal of Chronic Diseases,* **21**, 629-44

29. Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd edition. London, Edward Arnold: New York, Wiley.

30. Grambsch, P.M. and Therneau, T.M. (1994). Proportional hazards tests in diagnostics based on weighted residuals. *Biometrika*, **81**, 515-26

31. Hanley, J.A., Negassa, A., Edwrdes, M. and Forrester, J. (2003). Statistical analysis of correlated data using generalized estimating equations: an orientation, *American journal of Epidemiology*, **157**: 364-75.

32. Harvey, A.C. (1981). *Time Series Models.* Oxford: Allan.

33. Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**,423-47.

34. Hougaard P. (2000). *Analysis of multivariate survival data.* Springer-Verlag New York, inc.

35. Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42**, 805-820

36. Jorgensen, B., Lundbye-Christensen, S., Song, X.K. and Sun, L. (1996). A longitudinal study of emergency room visits and air pollution for Prince George, British Columbia. *Statist. Med.* **15**, 823-836.

37. Kaplan E.L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association,* **53**, 457-81

38. Kalbfleish JD and Prentice RL (1980). *The Statistical Analysis of Failure Time Data*, John Wiley and Sons, New York, 321 pp.

39. Kerry S.M. and Bland J.M. The intracluster correlation coefficient in cluster randomization. *BMJ.* 1998; 316, 1455–1460.

40. Kish, L. (1965). *Survey sampling,* Wiley, New York.

41. Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **48**, 795-806.

42. Laird, N.M. and Ware, J.H. (1982). Random-effects Models for Longitudinal Data. *Biometrics,* **38**, 963-74.

43. Lee, E.W., Wei, L.J. and Amato, D.A. (1992). Cox-type regression for large numbers of small groups of correlated failure time observations. *Survival Analysis: State of the Art*, Klein, J.P., Goel, P.K. (ed.) Kluwer Academic: Dordrecht, 237-247

44. Lele, S. (1991). Resampling using the estimating equations. In *Theory of Estimating Equations*, Ed.V.P.Godambe, pp295-304. Oxford: Clarendon Press.

45. Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73,** 13-22.

46. Liang, K.Y., Self, S.G. and Chang, Y. (1993). Modeling marginal hazards in multivariate failure time data. *Journal of the Royal Statistical Society B* **55**,441-53

47. Lin, D.Y. and Wei, L.J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* **84**, 1074-8

48. Lipsitz S. and Parzen M. (1996). A jackknife estimate of variance for Cox regression for correlated survival data. *Biometrics* **52**, (291-8).

49. Longford, N. (1993). *Random Coefficients Models*. Amold, London

50. Lu, S.E. and Wang, M.C. (2005). Marginal analysis for clustered failure time data. *Lifetime Data Analysis* **11**, 61-79

51. Ma, R., Krewski, D. and Burnett, R.T. (2003). Random effects Cox models: a Poisson modeling approach. *Biometrika* **90**, 157-69

52. McGilchrist,C.A. (1993). REML estimation for survival models with frailty. *Biometrics* **49**, 221-5

53. McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*, Wiley, New York.

54. Nelder, J. A. and Wedderburn, R.W.M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A* **135**, 370–384.

55. Oahes,D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84**, 487-93

56. Pope, C. A., Thun, M. J., Namboodiri, M. M., dockery, D.W., Evans, J.S., Speizer, F.E. and Heath, C.W. (1995). Particulate air pollution as a predictor of mortality in a prospective study of US adults. *Am. J. Respir. Crit. Care Med.* **151**,669-74

57. Sargent, D.J. (1998). A general framework for random effects survival analysis in the Cox proportional hazards setting, *Biometrics*, **54**:1486-97

58. Sastry, N. (1997). A multilevel hazards model for hierarchically clustered data: Model estimation and an application to the study of child survival in northeast Brazil. *Journal of the American Statistical Association* **92**, 426-35.

59. Sinha, D. (1993). Semiparametric Bayesian analysis of multiple event time data. *Journal of the American Statistical Association* **88**, 979-83

60. Smith, A.F.M. and Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *Journal of the Royal statistical society, series B* **55**, 3-23

61. Tate, R. and Wongbundhit, Y. (1983): Random versus Nonrandom coefficient models for multilevel analysis, *Journal of Educational Statistics*, **18**, 273-289.

62. Thompson, R. (1980). Maximum likelihood estimation of variance components. *Math. Operations-forsch. Statist., Ser. Statistics* **11**, 546-61.

63. Wei, L.J., and Johnson, W.E. (1985). Combining dependent tests with incomplete repeated measurements. *Biometrika* **72**, 359-64.

64. Wei, L.J., Lin, D.Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the Amarican Statistical Association* **84**, 1065-1073

65. White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, **48(4)**, 817–838.

66. Yau, K.K.W. (2001). Multi-level models for survival analysis with random effects. *Biometrics* **57**, 96-102.

67. Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049-60.

68. Zeger, S.L. and liang, K.Y. (1992). An overview of methods for the analysis of longitudinal data, *Statistics in Medicine*, **11**, 1825-1839