LEIDS UNIVERSITAIR MEDISCH CENTRUM

DEPARTMENT OF MATHEMATICS

MASTER THESIS

STATISTICAL SCIENCE FOR THE LIFE AND BEHAVIOURAL SCIENCES

Individual Patient Data Meta-Analysis of Time-to-Event Outcomes: An Application of a Poisson-Gamma-Frailty Model

---

Author:
Razieh Taghavi

Thesis Advisor:
Dr. M. Fiocco
Leiden University Medical Center

Supervisor:
Prof. Dr. A.W. van der Vaart
Leiden University, Mathematical Institute

May 2014

ABSTRACT

The goal of meta-analysis is to combine outcomes of several independent studies all addressing the same or a closely related research question. Traditionally, meta-analysis models combine summary estimates of a single quantitative endpoint, taken from different studies, to produce a single pooled result. Univariate fixed or random effect models are then employed to analyse the data.

In the presence of multiple outcomes (like overall and disease-free survival), multiple time points (e.g. in longitudinal studies) or multiple treatment groups, multiple pooled results are required. In such situations, a pooled result for each endpoint is usually obtained by applying a separate univariate meta-analysis to each endpoint independently. This approach is rather simple and ignores the potential correlation between endpoints. A multivariate meta-analysis model is required to improve efficiency over separate univariate analysis and allow the association between endpoints to be modelled.

A particular situation with multiple endpoints arises when each trial contributing to the meta-analysis provides survival proportions at a series of time-points. Such values are clearly correlated and a multivariate model is required to synthesize them jointly.

A Poisson correlated gamma-frailty model can be employed to account for within-study correlation and heterogeneity between studies. This model was applied before on aggregate survival data extracted from published trial reports. The aim of this thesis is to extend the use of this model to time-to-event meta-analysis at individual patient data (IPD) level. An IPD approach is considered the gold standard in meta-analysis as it can improve the quality of the analysis and therefore the reliability of the conclusions based on the statistical analysis.

The data used in the thesis is provided by the Dutch Children Oncology Group (DCOG) and comes from a retrospective worldwide study. Children suffering from acute myeloid leukemia (AML) were followed since diagnosis of the disease. All analysis has been implemented in R software. The codes and functions are written generic and can be applied to other datasets of similar structure.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my thesis advisor Dr. Marta Fiocco for her generous support and enthusiastic encouragement during the writing of this dissertation. This work would have never been possible without her guidance.

Special thanks should be given to all my professors from the master track of "statistical science for the life and behavioural sciences" for giving me valuable assistance during my study.

I wish to thank my family and friends for their support and encouragement throughout my study.

The Dutch Children Oncology Group (DCOG) is gratefully acknowledged for providing the dataset.

CONTENTS

1. INTRODUCTION

1.1 Individual patient data meta-analysis

Meta-analysis may be defined as statistical methods for combining results of independent research studies all concerning a closely related research question. Meta-analysis can be performed by using aggregate data supplied by original investigators or more commonly extracted from study reports. However, sufficient information is often unavailable and the robustness of some methods are not yet clearly understood [18]. An alternative approach involves collecting the individual patient data (IPD) from the original studies.

Meta-analyses based on original research data on individual participants enrolled in trials have been described as the gold standard of review [17] , as all the relevant data are used. Although IPD can be difficult to obtain and such an approach can be very resource demanding, it allows a more thorough investigation of patient characteristics as potential causes of heterogeneity between trials in the meta-analysis [5, 15, 16]. Such an investigation during the meta-analysis process is important as an interpretation of overall results in the presence of statistical heterogeneity can be misleading and findings from exploring potential causes of heterogeneity can also be clinically informative. Differences across studies in terms of design features and methodology, clinical procedures, and patient characteristics, are factors that can contribute to heterogeneity between studies [15].

1.2 Meta-analysis of time-to-event outcomes

Traditionally, meta-analysis models combine summary estimates of a single endpoint taken from different studies, to produce a single pooled result (for instance, the treatment effects estimated by means of an odds ratio), to aid evidence-based clinical decision making. The data is then analyzed by standard methods, using either a (univariate) fixed-effects or, as preferred by most statisticians, a (univariate) random-effects model, where random effects can account for between study heterogeneity [6].

In some situations the parameter of interest in meta-analysis can be bivariate or even multivariate. For instance in randomized clinical trials, multiple treatment groups can be involved or in longitudinal studies a mea-

surement is repeated over time. The meta-analysis of such data is more complicated than the meta-analysis of simpler, univariate data. A similar situation arises in meta-analysis of survival curves. For each study, the survival proportions are correlated over time [6].

Starting point for meta-analysis of time-to-event outcomes is a set of survival curves obtained either from IPD or from published literature and reports. For each study, the estimates of the survival probabilities at a predetermined set of time-points are known. Table 1.1 shows an example of N studies and M time-points where $S_i(t_j)$ indicates the survival proportion for study i at time t_j.

Study	Survival endpoints					
1	$S_1(t_1)$	$S_1(t_2)$	...	$S_1(t_j)$	...	$S_1(t_M)$
2	$S_2(t_1)$	$S_2(t_2)$	...	$S_2(t_j)$	...	$S_2(t_M)$
.	.	.	...	.	...	.
.	.	.	...	.	...	.
.	.	.	...	.	...	.
i	$S_i(t_1)$	$S_i(t_2)$	...	$S_i(t_j)$	...	$S_i(t_M)$
.	.	.	...	.	...	.
.	.	.	...	.	...	.
.	.	.	...	.	...	.
N	$S_N(t_1)$	$S_N(t_2)$	...	$S_N(t_j)$	...	$S_N(t_M)$

Tab. 1.1: Survival proportions in M pre-determined time-points for N studies

The objective of meta-analysis for survival studies is to obtain an overall survival curve from a set of survival curves usually under heterogeneity. In the following sections available methods to perform time-to-event meta-analysis will be briefly reviewed.

1.2.1 Univariate meta-analysis

This approach also known as classical method considers each outcome measure separately and applies a separate univariate meta-analysis to each endpoint. Perhaps this is the simplest approach to the meta-analysis of survival data. However there are some problems in applying classical method to obtain a pooled result. First of all it ignores the correlation aspect between reported survival proportions. Second, standard errors of the survival proportions need to be known which usually is not the case when meta-analysis is based on published literature. Moreover, the monotonicity of the overall survival obtained from this approach is not guaranteed.

1.2.2 Multivariate meta-analysis

Alternatively, a multivariate meta-analysis model can be used that jointly synthesizes the multiple endpoints [19, 20]. A multivariate meta-analysis model improves efficiency over separate univariate syntheses and allows the association between endpoints to be modelled [6].

Dear (1994) [3] proposed a fixed-effect model to jointly synthesize survival proportions reported at multiple times. The parameters are estimated by generalized least squares (GLS) method. Fitting a GLS model requires that the correlation matrix of the response variables be known. Dear uses the standard error of the survival estimates to estimate the correlation matrix of the response variable.

Arends et al. (2008) [2] proposed a multivariate random-effects model that can be seen as an extension of Dear's model. The method fits in the framework of the linear mixed models with normally distributed errors. This method has to be adapted in this case because the correlations between the different survival estimates of the same curve have to be estimated as well.

1.2.3 Counting process

Another approach to meta-analysis of time-to-event data is to look at the problem as a discrete counting process, indicating the number of patients at risk, the number of events and the number of censored patients in consecutive pre-determined intervals for each study. This data can be computed when meta-analysis is conducted on IPD. If published articles are used for meta-analysis then the data can be reconstructed as described in Parmar(1998) [12] and Fiocco et al. (2009) [6]. All technical details concerning the data reconstruction are given in Appendix A.

Let N and M be the number of studies and the number of time-points respectively. Define d_{ij} and r_{ij} as the number of events and number of patients at risk, respectively for study i at time-point j. Now by ignoring correlations among serial counts within each study (as classical method does) and by assuming homogeneity between the studies, a pool of all patients in all trials can be made and the aggregate number at risk and number of events at each time-point can be computed as follows:

$$r_j = \sum_{i=1}^{N} r_{ij} \quad \text{and} \quad d_j = \sum_{i=1}^{N} d_{ij}.$$

The hazard and the overall cumulative hazard function are estimated as follows:

$$\hat{h}(t_j) = \frac{d_j}{r_j}$$

and

$$\hat{H}(t_j) = \sum_{l \leq j} \hat{h}(t_l).$$

An estimate of the overall survival function at time-point j, $\hat{S}(t_j)$ can be computed as

$$\hat{S}(t_j) = \exp(-\hat{H}(t_j)).$$

The standard error of $\hat{S}(t_j)$ can be estimated by applying Greenwood's formula.

The aforementioned method ignores the correlations among the counts within studies and can be applied only under homogeneity. An alternative method is required under heterogeneity which also considers the association among the counts. To model the correlation structure and the heterogeneity between studies, Fiocco et al. [7] proposed a Poisson-correlated-gamma-frailty model. Heterogeneity between studies and correlation within studies are taken into account by introducing a gamma-distributed frailty vector for each study. A gamma process was inspired by the bivariate frailty models that had been used in modelling genetic survival data for twins [13, 22]. In these models, related individuals have different but dependent frailties. The frailty of each twin can be decomposed in a pair as a sum of two independent frailties, one of which is shared by both twins. The construction of the frailties is then carried out by using independent additive components with a common component for both frailties.

Since the full likelihood is intractable, a composite likelihood procedure was employed based on all pairs of time-points to estimate the unknown parameters. To facilitate the estimation, a two- stage estimation procedure [1, 9–11, 14] has been used where in the first stage the marginal distributions are used to estimate all parameters except the frailty correlation. In the second stage the correlation is estimated from the likelihood based on pairs of observations.

1.3 Aim and contents of the thesis

In this thesis an IPD meta-analysis of time-to-event outcomes is studied by employing a Poisson-correlated gamma-frailty model proposed by Fiocco et al. (2009) [7]. The model has been used before for a meta-analysis of survival curves obtained from published report. The aim of this thesis is to extend this model to an IPD and to write general R functions to fit the

model for any IPD meta-analysis based on survival curves.

The Poisson-correlated gamma-frailty model and the construction of the frailty process are described in Chapter 2. The estimation procedure and the procedure to estimate the standard errors are illustrated in Chapter 3 and Chapter 4 respectively. In Chapter 5 the Poisson-gamma-frailty model is extended to a meta-analysis of IPD. Data description and results are also presented in this chapter.

The R-codes written to fit the model are provided in Appendix C.

2. POISSON-GAMMA-FRAILTY MODEL

As discussed in Chapter 1, time to event endpoints can be seen as a discrete counting problem in pre-determined time intervals. Since the number of events for a study is repeated over time and they are correlated to each other, the meta-analysis problem can be cast in a longitudinal count data framework. By far the most popular model for analyzing this type of count data is the Poisson regression model with the possibility to account for over dispersion and serial correlations. Frailty provides a suitable way to introduce random effects in the model to account for association among the counts and unobserved heterogeneity between studies. In the context of longitudinal count data a common assumption is that event counts are conditionally independent Poisson variables given the value Z of a gamma-distributed subject-specific frailty term. Poisson-gamma frailty models are flexible and there is a closed form for the marginal distribution of the event counts. In this chapter the construction of the frailty process and the use of this process in a Poisson model is illustrated.

2.1 Construction of the frailty process

Dealing with correlated data typically means making some type of assumption about the form of the correlation among observations taken on the same subject. Fiocco et al. [7] proposed a time-varying frailty process $Z(t)$ with marginal gamma distribution $\Gamma(\alpha, \beta)$ and first order autoregressive correlation, $cor(Z_i(s), Z_i(t)) = \rho^{|s-t|}$ for each study i and time-points s and t. This choice of correlation seems realistic since it is plausible that counts in neighbouring time intervals are more strongly correlated than those further apart.

In discrete time, correlated frailties are constructed using sums of independent gamma distributions (all with the same rate parameter). Relying on the infinite divisibility property of the gamma distribution, frailty construction is obtained by defining building blocks as appropriate sums of infinite sequence of independent gamma distributed random variables.

Let X_{ij} be an infinite sequence of independent gamma random variables for $i, j \in \mathbb{Z}$ and $-\infty < i \leq j < \infty$, with distribution

$$X_{ij} \sim \Gamma(\alpha(1-\rho)^2\rho^{j-i}, \beta). \tag{2.1}$$

Let Z_t for $t \in \mathbb{Z}$ be defined as

$$Z_t = \sum_{i=-\infty}^{t} \sum_{j=t}^{+\infty} X_{ij}. \tag{2.2}$$

It is not difficult to show that Z_t has the desired marginal $\Gamma(\alpha, \beta)$-distribution. The construction of the frailty process for Z_t is illustrated in Figure 2.1 in which the black dots indicate X_{ij}'s. The differences between the size and the colour of the dots are to show the different intensity of correlation between the terms .

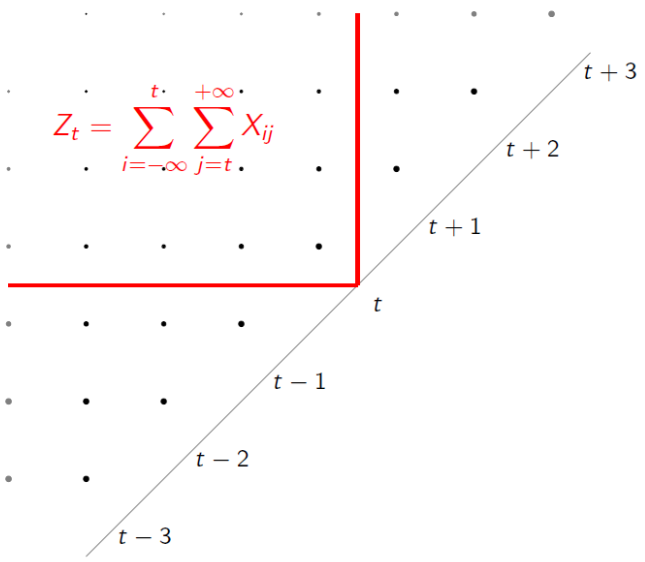

Fig. 2.1: Frailty construction

The correlation between the frailty terms Z_s and Z_t is induced by the fact that they have certain X_{ij} elements in common. In particular, for $s < t$, each term Z_s and Z_t can be decomposed as a sum of two independent components, one of which is shared by both frailty terms. Let X_0, X_s and X_t be respectively the shared component between Z_s and Z_t, the unshared component of Z_s, and the unshared component of Z_t. Then Z_s and Z_t can be written as follow

$$\begin{aligned} Z_s &= \sum_{i=-\infty}^{s} \sum_{j=s}^{+\infty} X_{ij} \\ &= \sum_{i=-\infty}^{s} \sum_{j=s}^{t-1} X_{ij} + \sum_{i=-\infty}^{s} \sum_{j=t}^{+\infty} X_{ij} \\ &= X_s + X_0 \end{aligned} \tag{2.3}$$

and

$$\begin{aligned} Z_t &= \sum_{i=-\infty}^{t} \sum_{j=t}^{+\infty} X_{ij} \\ &= \sum_{i=-\infty}^{s} \sum_{j=t}^{+\infty} X_{ij} + \sum_{i=s+1}^{t} \sum_{j=t}^{+\infty} X_{ij} \\ &= X_0 + X_t. \end{aligned} \tag{2.4}$$

It can be shown that the common term X_0 has a $\Gamma(\alpha\rho^{t-s}, \beta)$-distribution. The pair of (Z_s, Z_t) for each $s, t \in \mathbb{Z}$ has a bivariate-correlated gamma distribution with $\rho_{st} = \rho^{|s-t|}$. Figure 2.2 shows how two frailty terms Z_1 and Z_3 are associating by sharing part of their components.

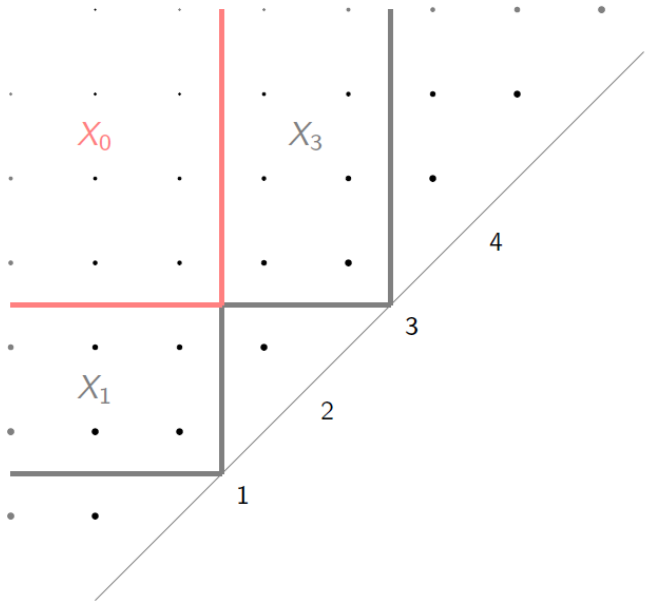

Fig. 2.2: Example of frailty construction at time-point 1 (Z_1) and time-point 3 (Z_3)

2.2 Application of the frailty process in Poisson model

The Poisson distribution is a standard distribution for modelling count data. This distribution is obtained when events occur independently of each other with the same intensity. Count data often shows a larger variability that is exhibited by the Poisson distribution, i.e., the variance is larger than the mean. To account for this possibility, a Poisson–gamma frailty model is suggested.

The new multivariate gamma distribution can be used as the frailty vector in a Poisson model for longitudinal count data. By setting $\alpha = \beta = \theta$ in the gamma process, the marginal gamma distribution of Z_t will have mean 1 and variance $\xi = \theta^{-1}$. Let $\mathbf{Y} = (Y_1, \ldots, Y_T)$ be the vector of event counts and $\mathbf{Z} = (Z_1, \ldots, Z_T)$ be the corresponding gamma-frailty vector in T pre-determined intervals. Given the unobserved frailties, the event counts are assumed to be conditionally independent Poisson variables

$$Y_t | Z_t \sim Po(\mu_t Z_t), \tag{2.5}$$

where $\mu_t = \exp(x_t \boldsymbol{\beta})$ is assumed to be linearly related to a design vector x_t through a log-link and an unknown parameter vector $\boldsymbol{\beta}$.

It is well known that a negative binomial distribution arises for the mixture of a Poisson distribution with a gamma-distributed parameter. Here the resulting marginal distribution of Y_t has a negative binomial distribution with mean μ_t and rate θ, denoted as $Y_t \sim NB(\mu_t, \theta)$. The probability function of the negative binomial distribution is

$$P_{NB}(y; \mu, \theta) = \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)} \left(\frac{\mu}{\theta+\mu}\right)^y \left(\frac{\theta}{\theta+\mu}\right)^\theta \tag{2.6}$$

for $\mu > 0$, and $\theta > 0$ which contributes to the variance of frailty components $\xi = \theta^{-1}$. The marginal mean and variance of Y_t are respectively, $E(Y_t) = \mu_t$ and $var(Y_t) = \mu_t + {\mu_t}^2 \xi$. The association between a pair of event counts Y_s and Y_t is induced by the correlation of the corresponding frailty terms Z_s and Z_t as $cov(Y_s, Y_t) = \rho_{st}\, \xi\, \mu_s\, \mu_t$ where $\rho_{st} = \rho^{|s-t|}$.

3. ESTIMATION

The parameters to be estimated in the proposed Poisson-gamma-frailty model described in Chapter 2 are the regression vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_T)$, the variance of the gamma process $\xi = \theta^{-1}$, and within-subject (within-study) correlation ρ in a longitudinal framework (meta-analysis). As the full likelihood is intractable, composite likelihood and a two-stage estimation procedure will be applied to estimate the parameters. First, the regression parameter vector $\boldsymbol{\beta}$ and the variance $\xi = \theta^{-1}$ are simultaneously estimated from the marginal distributions of the event counts introduced in Section 2.2. In the second stage, the estimated values obtained in the first stage are plugged into the pairwise composite likelihood for estimating the correlation parameter ρ.

3.1 Composite likelihood

For likelihood-based inference one requires to write a joint distribution in the longitudinal set-up. Theoretically, the joint distribution can be obtained through differentiation of the Laplace transform but in practice it is only possible for low-dimensional distributions. Since it is not manageable to obtain the full likelihood for high-dimensional distributions, it may be useful to approximate it by a composite likelihood procedure. The composite likelihood approach helps to reduce the computational complexity of the full likelihood based on the univariate or bivariate marginal distributions.
In the proposed Poisson-gamma-frailty model a composite likelihood estimation procedure based on pairs of observations at all pairs of time-points is applied. The frailty terms Z_s and Z_t are replaced by the sums of appropriate independent additive components $Z_s = X_0 + X_s$ and $Z_t = X_0 + X_t$ as described in (2.3) and (2.4). The bivariate distribution of the counts (Y_s,Y_t) may be derived as follow

$$
\begin{aligned}
P\big(Y_s = y_s, Y_t = y_t\big) &= \frac{\mu_s^{y_s}\mu_t^{y_t}}{y_s!y_t!}.E\big(Z_s^{y_s}Z_t^{y_t}e^{-\mu_s Z_s}e^{-\mu_t Z_t}\big) \\
&= E\bigg(e^{-\mu_s X_s-\mu_t X_t-(\mu_s+\mu_t)X_0}. \\
&\quad \frac{\mu_s^{y_s}\mu_t^{y_t}}{y_s!y_t!}(X_s+X_0)^{y_s}(X_t+X_0)^{y_t}\bigg) \\
&= \sum_{k=0}^{y_s}\sum_{l=0}^{y_t} E\bigg(e^{-\mu_s X_s-\mu_t X_t-(\mu_s+\mu_t)X_0}. \\
&\quad \frac{X_s^k X_0^{y_s-k}}{k!(y_s-k)!}\frac{X_t^l X_0^{y_t-l}}{l!(y_t-l)!}\mu_s^{y_s}\mu_t^{y_t}\bigg) \\
&= \sum_{k=0}^{y_s}\sum_{l=0}^{y_t} E\bigg(e^{-\mu_s X_s}.\frac{(\mu_s X_s)^k}{k!}\bigg). \\
&\quad E\bigg(e^{-\mu_t X_t}.\frac{(\mu_t X_t)^l}{l!}\bigg). \\
&\quad E\bigg(e^{-(\mu_s+\mu_t)X_0}.\frac{((\mu_s+\mu_t)X_0)^{y_s+y_t-k-l}}{(y_s+y_t-k-l)!}\bigg). \\
&\quad \bigg(\frac{(y_s+y_t-k-l)!}{(y_s-k)!(y_t-l)!}(\frac{\mu_s}{\mu_s+\mu_t})^{y_s-k}.(\frac{\mu_t}{\mu_s+\mu_t})^{y_t-l}\bigg) \\
&= \sum_{k=0}^{y_s}\sum_{l=0}^{y_t} P_{NB}\Big(k;\mu_s(1-\rho_{st}),\theta(1-\rho_{st})\Big). \\
&\quad P_{NB}\Big(l;\mu_t(1-\rho_{st}),\theta(1-\rho_{st})\Big). \\
&\quad P_{NB}\Big(y_s+y_t-k-l;(\mu_s+\mu_t)\rho_{st},\theta\rho_{st}\Big). \\
&\quad P_{BIN}\Big(y_s-k;y_s+y_t-k-l;\frac{\mu_s}{\mu_s+\mu_t}\Big). \qquad (3.1)
\end{aligned}
$$

The probability distribution $P_{NB}(y;\mu,\theta)$ is defined in (2.6) and $P_{BIN}(y;n;p)$ is the binomial distribution defined as

$$P_{BIN}(y;n,p) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}. \qquad (3.2)$$

The advantage of the pairwise composite likelihood is that the double sum at the right-hand side of (3.1) is a sum of a product of known probability distributions and hence it is possible to implement it in common statistical software.

Using all pairs of observations in composite likelihood still entails a high-dimensional maximization problem. Therefore a two-stage procedure is pro-

posed to facilitate the estimation procedure.

3.2 First stage

In the first stage, the correlation among the event counts is ignored and the marginal negative binomial distributions from (2.6) are used to estimate the regression parameter vector $\boldsymbol{\beta}$ and the dispersion parameter θ simultaneously. Let $y_{il}, \ldots, y_{iT}$ denote repeated counts over T occasions for the subject i (study i) in a longitudinal set-up (meta-analysis), where $i = 1, \ldots, N$. The marginal distribution of y_{it} is a negative binomial $NB(\mu_{it}, \theta)$ with $\mu_{it} = \exp(\boldsymbol{x}_{it}{}^T\boldsymbol{\beta})$. The log-likelihood corresponding to this stage is given by

$$\begin{aligned}
\ell_1(\eta) &= \sum_{i=1}^{N} \ell_{1i}(\eta) = \sum_{i=1}^{N}\sum_{t=1}^{T} \ell_{1it}(\eta) \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T} \log\, P_{NB}(y_{it}; \mu_{it}, \theta) \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T} \log\Gamma(y_{it}+\theta) - \log\Gamma(\theta) + y_{it}\log\mu_{it} \\
&\qquad - (y_{it}+\theta)\log(\mu_{it}+\theta) + \theta\log\theta,
\end{aligned} \tag{3.3}$$

where $\eta = (\boldsymbol{\beta}, \theta)$ and $\ell_{1it}(\eta)$ is the log of the negative binomial probability function defined in (2.6) for individual i at time-point t. The log-likelihood corresponds to an independent working correlation.

3.3 Second stage

The second stage of the estimation procedure is based on the pairwise composite likelihood (3.1). In this stage the correlation parameter is estimated by fixing the margins in the composite likelihood at the estimates from stage one. For all possible pairs of time-points s and t and all the subjects (studies), the composite log-likelihood is given by

$$\ell_2(\rho; \hat{\eta}) = \sum_{i=1}^{N} \ell_{2i}(\rho; \hat{\eta}) = \sum_{i=1}^{N}\sum_{s=1}^{T-1}\sum_{t=s+1}^{T} \log(P(Y_{is} = y_{is}, Y_{it} = y_{it})), \tag{3.4}$$

where $\hat{\eta} = (\hat{\boldsymbol{\beta}}, \hat{\theta})$ is the estimate of $\eta = (\boldsymbol{\beta}, \theta)$ obtained from the first stage. An estimate of the correlation parameter ρ is found by maximizing (3.4).

4. STANDARD ERROR

Standard errors of the estimates of $\boldsymbol{\beta}$, θ and ρ can be obtained by applying a parametric bootstrap, as it is feasible to simulate from the proposed multivariate gamma distribution. Alternatively, asymptotic theory can be used to obtain the standard errors using sandwiching estimators. In this chapter the parametric bootstrap approach is described. This method will be used later on the motivative example employed in this thesis.
Details concerning asymptotic theory approach are provided in Appendix B.

4.1 Parametric bootstrap

Let $\hat{\boldsymbol{\beta}}$, $\hat{\theta}$ and $\hat{\rho}$ be respectively the estimate of regression vector, dispersion parameter, and the correlation parameter from the original data. Recall that T is the number of time-points and N the number of subjects (studies), one bootstrap dataset (dataset i) can be generated as follows:

1. Given $\hat{\theta}$ and $\hat{\rho}$, generate N independent copies of frailty vector $\mathbf{z}_i^* = (z_{i1}^*, \ldots, z_{iT}^*)$ from the multivariate gamma with marginal $\Gamma(\hat{\theta}, \hat{\theta})$ and correlation $\text{corr}(Z_{is}^*, Z_{it}^*) = \hat{\rho}^{|s-t|}$.

2. Given $\hat{\boldsymbol{\beta}}$, derive $\hat{\mu}_{it}$ and generate N independent vector of event counts $\mathbf{y}_i^* = (y_{i1}^*, \ldots, y_{iT}^*)$, with $y_{it}^* \sim \text{Po}(\hat{\mu}_{it} z_{it}^*)$.

3. From the bootstrap dataset $\mathbf{Y}^* = (\mathbf{y}_1^*, \ldots, \mathbf{y}_N^*)$ estimate $(\boldsymbol{\beta}, \theta, \rho)$ as described in Chapter 3, obtaining bootstrap estimates $(\hat{\boldsymbol{\beta}}^*, \hat{\theta}^*, \hat{\rho}^*)$.

In the first step of bootstrap it is necessary to generate correlated frailties within subject (study) which is the crucial part of the bootstrap. In the next section it is described how to generate correlated frailties. The parameter estimates obtained from the bootstrap method are used to compute standard errors of the parameters of interest in the model.

4.2 Generating correlated frailties

At first sight, this may seem impossible to simulate data from the proposed multivariate gamma distribution due to the fact that infinite sums of X_{ij} are used in the construction of frailty term Z_t, where $Z_t = \sum_{i=-\infty}^{t} \sum_{j=t}^{+\infty} X_{ij}$. However, it becomes feasible by collapsing X_{ij} components in blocks in such

a way that it is possible to generate them. For this purpose the gamma-distributed blocks are defined as follows:

$$X_{i+} = \sum_{j=T+1}^{+\infty} X_{ij} \sim \Gamma(\alpha(1-\rho)\rho^{T+1-i}, \beta),\ i = 1, \ldots, T\ ,$$

$$X_{+j} = \sum_{i=-\infty}^{0} X_{ij} \sim \Gamma(\alpha(1-\rho)\rho^{j}, \beta),\ j = 1, \ldots, T\ ,$$

and

$$X_{++} = \sum_{i=-\infty}^{0} \sum_{j=T+1}^{+\infty} X_{ij} \sim \Gamma(\alpha\rho^{T+1}, \beta).$$

Then Z_t is given by

$$Z_t = \sum_{i=1}^{t} X_{i+} + \sum_{j=t}^{T} X_{+j} + X_{++} + \sum_{i=1}^{t} \sum_{j=t}^{T} X_{ij}. \tag{4.1}$$

Frailties are generated by simulating independent gamma-distributed blocks and components, and summing up the appropriate terms to obtain the frailty terms. The shared components between the simulated frailties assure that the frailties are correlated. It is not difficult to show that for a T-dimensional correlated frailty vector $\boldsymbol{Z} = (Z_1, \ldots, Z_T)$, it is necessary to generate $\frac{1}{2}T^2 + \frac{5}{2}T + 1$ independent gamma variables.

To illustrate the simulation procedure, Figure 4.1 shows how the X_{ij} elements are collapsed into blocks for a 4-dimensional correlated-gamma-distributed $\boldsymbol{Z} = (Z_1, Z_2, Z_3, Z_4)$ as an example.
Figure 4.2 and Figure 4.3 show how frailty terms Z_1 and Z_3 from the vector $\boldsymbol{Z} = (Z_1, Z_2, Z_3, Z_4)$ share some of their components with each other.

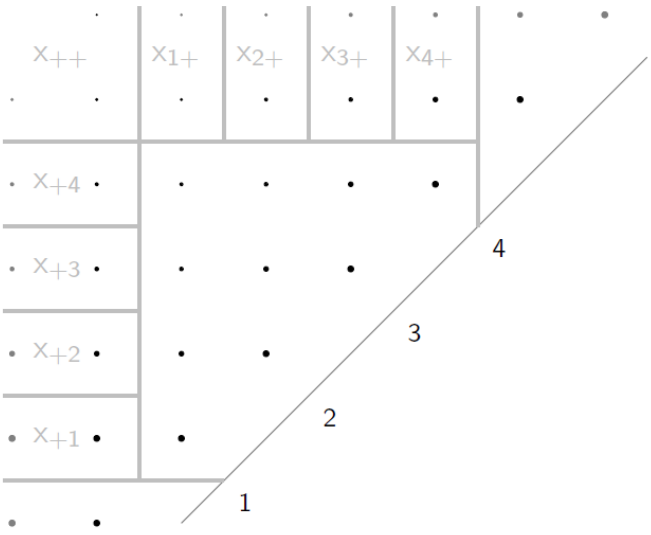

Fig. 4.1: Collapsing X_{ij} elements in appropriate blocks ($T = 4$)

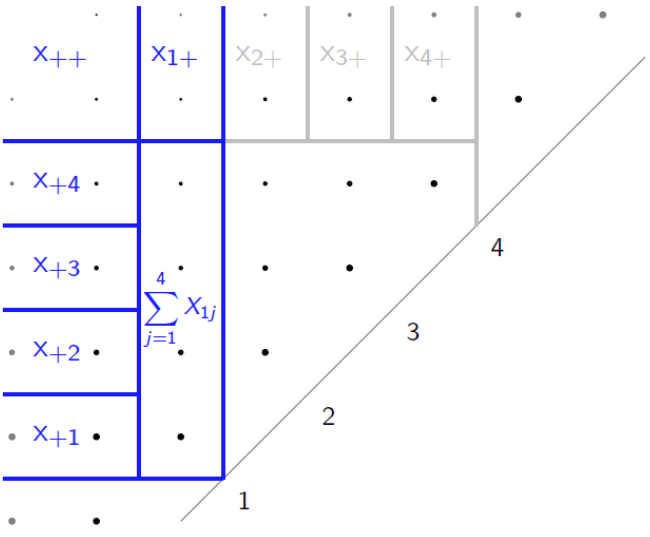

Fig. 4.2: Generation of Z_1 ($T = 4$)

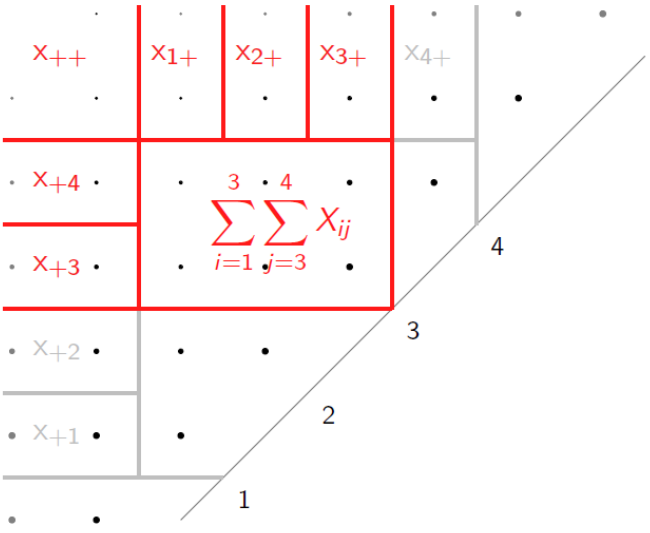

Fig. 4.3: Generation of Z_3 ($T = 4$)

5. INDIVIDUAL PATIENT DATA META-ANALYSIS BY POISSON-GAMMA-FRAILTY MODEL

In this chapter the application of the Poisson-gamma-frailty model to the meta-analysis of survival curves for IPD is discussed. Data description and results are also presented.

5.1 Poisson-gamma-frailty model in meta-analysis

As discussed in Chapter 2, the Poisson-gamma-frailty model was introduced in longitudinal count data to account for within-subjects correlation and between subjects heterogeneity. Meta-analysis of time-to-event data can be cast in a longitudinal set-up by approaching the problem as counting process in pre-determined time intervals. This implies that the Poisson-gamma-frailty model can be also applied in the context of meta-analysis of survival curves.

To apply the model on the meta-analysis of time-to-event data, the number of events and the number of patients at risk as well as the censoring mechanism should be known at a set of pre-determined time intervals. This information is usually not provided when the meta-analysis is applied on the published papers. However, the desired information can be reconstructed by assuming that the patients are censored at a constant rate during the time intervals. All details concerning the data reconstruction are described in Appendix A. For IPD, however, the exact information can be computed and there is no need to reconstruct the data.

For survival data with piecewise constant hazard, the contribution to the likelihood of the j-th interval can be obtained by $D_j \sim Po(\lambda_j \Delta_j r_j)$, where D_j, r_j, λ_j and Δ_j indicate respectively the number of events, the number of individuals at risk, exponential intensity or hazard, and the length of the interval, over interval j. The component Δ_j can be combined with r_j and the model can be written as $D_j \sim Po(\lambda_j \tilde{r}_j)$, with $\tilde{r}_j = \Delta_j r_j$ indicating the number of person-years over interval j. By including a frailty component, the model becomes

$$D_{ij}|Z_{ij} \sim Po(Z_{ij}\lambda_j \tilde{r}_{ij}), \tag{5.1}$$

where i indicates study and Z_{ij} is the j-*th* gamma-frailty component of the

vector $\boldsymbol{Z}_i$ with mean 1, and variance $\xi = \theta^{-1}$ introduced to model the heterogeneity among the studies included in the meta-analysis. The correlation between the time intervals s and t is modelled by assuming first-order autoregressive correlation structure $corr(Z_{is}, Z_{it}) = \rho^{s-t}$. This model is a special case of (2.5) with $\mu_j = \lambda_j \tilde{r}_{ij} = \exp(\log \lambda_j + \log \tilde{r}_{ij})$. The components of the unknown vector of parameters $\boldsymbol{\beta}$ are given by $\beta_j = \log \lambda_j$ and $\log \tilde{r}_{ij}$'s are used as offsets. The vector $\boldsymbol{\beta}$ and the parameter of the marginal gamma distribution θ are estimated in the first stage of the estimation procedure by employing the marginal negative binomial distributions $D_{ij} \sim NB(\lambda_j \tilde{r}_{ij}, \theta)$.

For every fixed θ the negative binomial distribution becomes a special case as exponential family and can be formulated as a generalized linear model. However in this model θ is not known and needs to be estimated. An alternating iteration process can be used to estimate the $\boldsymbol{\beta}$ and θ. The iteration process for fixed θ fits *glm* and estimates means and then uses the estimated means to estimate the θ parameter. The two processes are alternated until convergence of the means and dispersion. This procedure has been implemented in the function `glm.nb` in the **MASS** library by Venables and Ripley (2002) [21] in **R** software. The function can be used on data as follows:

```
glm.nb(formula = Count ~ Interval + offset(log(pyrs)),
data = data, link = "log"),
```

where Count, Interval, and pyrs denotes number of events, index of interval, and number of person-years respectively. The estimation values can be easily extracted from `glm.nb` outputs. These estimates will be used in the second stage of the estimation procedure to estimate parameter ρ as described in Section 3.3.

Once the estimate of vector $\boldsymbol{\beta}$ is obtained, the hazards can be computed. Using the estimated hazards $\hat{\lambda}_j$ as parameters of the piecewise exponential distribution, the estimate of the overall survival function can be obtained as

$$\hat{S}(t) = \hat{S}_{j-1} \,.\, \exp(-\hat{\lambda}_j(t - t_{j-1})) \tag{5.2}$$

for $t_{j-1} < t \leq t_j$ in the intervals defined by $0 = t_0 < t_1 < \cdots < t_M$, where $\hat{S}_j$ is defined recursively as $\hat{S}_0 = 1, \ldots, \hat{S}_j = \hat{S}_{j-1} \,.\, \exp(-\hat{\lambda}_j(t_j - t_{j-1}))$ with $j = 1, \ldots, M$ and using the convention $t_0 = 0$. The estimation of the exponentially distributed survival curve with piecewise-constant rate can be obtained by using **ppexp** function in **msm** library.

The standard error of the meta-analytic overall survival curve can be obtained by the delta method, if the covariance matrix of the estimates of hazards $(\hat{\lambda}_1, \ldots, \hat{\lambda}_M)$ is available. Alternatively, the bootstrap data can be

used to construct the bootstrap estimates of the overall survival probabilities and they can be used to compute the standard error of the survival probabilities. The standard errors may be used in the standard way to construct the 95% confidence interval of the survival curve by using the central limit theorem.

5.2 Motivative example

The proposed Poisson-gamma-frailty model is applied on an international IPD collected by the Dutch Children Oncology Group (DCOG). The dataset comes from a large worldwide retrospective study where children have been diagnosed with Acute Myeloid Leukemia (AML) and are followed since the diagnosis of the disease. Leukemia is a type of cancer of the blood or bone marrow due to an abnormal increase of immature white blood cells and is characterized by a rapid progress of the disease. After receiving treatment, patients may achieve a phase called complete remission (CR). A patient is considered to be in complete remission if the disease has disappeared (using criteria developed by the International Working Group).

In total, 838 children from 10 collaborative study groups suffering from AML were included in the study. Interest is only on children who achieved complete remission. Therefore only 770 children have been included in the analysis. A patient in complete remission could experience relapse or death at later stages. Table 5.1 shows how the events of relapse and death are distributed among the patients who achieved complete remission, where 1 and 0 indicate whether the corresponding event occurs or does not occur, respectively.

		Death		
		0	1	Sum
Relapse	0	485(63%)	67(9%)	552(72%)
	1	94(12%)	124(16%)	218(28%)
	Sum	579(75%)	191(25%)	770(100%)

Tab. 5.1: Distribution of events

Survival analysis was already performed for each single study and the results are available in the cited paper [4]. Event-free survival, overall survival and cumulative incidence of relapse were already studied for each study. Since data is available for each individual in every study it is also of interest to study the problem in the context of meta-analysis at IPD level.

Since the disease presents several stages, the interest might be in differ-

ent phases as the starting point to estimate the survival. In this context, diagnosis and complete remission are the phases of interest to be considered as the starting point. However, due to the high percent (40%) of unknown time to complete remission, the time of diagnosis is considered as the starting point.

5.3 Model estimation

In this thesis the meta-analysis is performed on event-free survival (EFS) curves. EFS is analyzed from the date of diagnosis to the first event which can be relapse or death, or to the date of last follow-up in case a child has not experienced an event by the end of the study. Patients who do not experience an event of relapse or death are censored at the time of last follow-up. Kaplan-Meier methodology is used to estimate the 10-year probabilities of EFS (pEFS) for each study group. Figure 5.1 shows the estimated EFS curves for each study group.

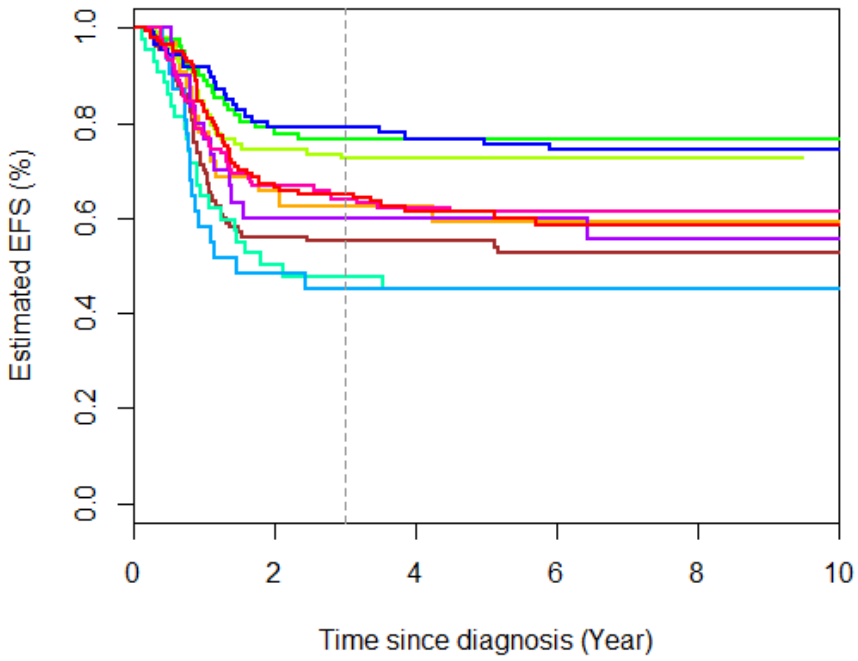

Fig. 5.1: Event Free Survival curves for each study

Since not many events occur after the first three years of follow-up, for further analysis only the first three years will be considered. To perform the meta-analysis by applying Poisson-gamma-frailty model on the IPD introduced in Section 5.2, the time intervals for the analysis should be determined. The time-points are selected at every three months which means that the pre-determined time-points contain in the set of (0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75, and 3.00 years). The number of events and the number of person-years for each interval in each study must be known in order to fit the model. Table 5.2 shows the number of events, and the number of patients at risk for each time interval in each study. By having access to IPD, the exact number of person-years for each interval can be computed, considering the exact censoring time in the intervals.

Intervals		Studies									
		1	2	3	4	5	6	7	8	9	10
1	(0.00,0.25]	0/107	0/32	0/98	0/81	2/43	0/32	0/86	0/30	0/117	3/144
2	(0.25,0.50]	7/107	1/32	2/98	2/81	4/41	2/32	5/86	0/30	8/117	2/141
3	(0.50,0.75]	9/100	3/31	5/96	4/79	3/36	4/29	2/81	3/30	10/109	4/137
4	(0.75,1.00]	15/91	3/28	10/91	3/75	6/33	7/25	0/79	4/27	8/99	16/132
5	(1.00,1.25]	11/76	3/25	6/81	3/72	2/27	2/18	4/79	2/23	5/91	7/116
6	(1.25,1.50]	4/65	0/22	1/75	4/69	2/25	1/16	4/75	2/21	5/86	10/108
7	(1.50,1.75]	1/61	0/22	1/74	1/65	1/23	0/15	2/71	1/19	3/81	2/98
8	(1.75,2.00]	0/60	1/22	0/73	1/64	1/22	0/14	1/69	0/18	0/77	2/94
9	(2.00,2.25]	0/60	1/21	0/73	0/63	1/21	0/14	0/68	0/18	0/77	2/91
10	(2.25,2.50]	1/60	0/20	1/73	1/63	0/20	1/14	0/68	0/18	0/76	1/88
11	(2.50,2.75]	0/59	0/20	0/72	0/62	0/20	0/13	0/68	0/18	1/75	0/86
12	(2.75,3.00]	0/58	0/20	1/71	0/61	0/19	0/13	0/68	0/18	2/74	0/83

Tab. 5.2: Number of patients at risk and number of events in each interval and each study. Each element a/b in the table refers to number of events (a) and number of patients at risk (b).

Before fitting the model the data are transformed in a long form where each row contains the number of events and the number of person-years for each time interval ($T = 12$) at each study ($N = 10$). In this example the long form data includes 120 rows in total. Table 5.3 shows part of the data for study 1 where the columns indicate the study number, time interval index, number of events, and number of person-years respectively.

Study	TimeInt	N.Event	PYRS
1	1	0	26.75
1	2	7	25.71
1	3	9	23.79
1	4	15	20.65
1	5	11	17.34
1	6	4	15.68
1	7	1	15.02
1	8	0	15.00
1	9	0	15.00
1	10	1	14.95
1	11	0	14.60
1	12	0	14.50

Tab. 5.3: The first 12 rows of data in long form for the first study

Poisson-gamma-frailty model can be fitted on the long data as described in Section 5.1. Let $(\lambda_1, \ldots, \lambda_T, \xi, \rho)$ be the vector of parameters to be estimated where λ_j, ξ and ρ represent the hazard at time interval j, the variance of the gamma distribution and the correlation parameter respectively. The estimates and the corresponding SE's obtained from parametric bootstrap technique, are presented in Table 5.4. The results have been rounded to three decimals.

The low estimate of frailty variance (0.1) indicates that the heterogeneity between the studies is negligible. This is to be expected since the studies are following a strict clinical protocol proposed to treat this disease. The estimated correlation is equal to 0.57 which indicates that variation in hazard at the beginning of the study is correlated with variations at later intervals.

The estimation of the overall survival curve based on all studies is obtained as described in Section 5.1. Its point-wise confidence interval is constructed by using the parametric bootstrap. In Figure 5.2 the estimate of the overall survival and its 95 percent point-wise confidence interval plotted along with the survival curves of each study are shown.

The variation in survival implied by this model is illustrated in Figure 5.3,

Interval	Hazard (SE)
1	0.026 (0.012)
2	0.179 (0.037)
3	0.279 (0.052)
4	0.473 (0.076)
5	0.318 (0.058)
6	0.240 (0.049)
7	0.092 (0.029)
8	0.047 (0.020)
9	0.032 (0.016)
10	0.040 (0.019)
11	0.008 (0.008)
12	0.025 (0.014)
Heterogeneity(ξ)	0.104 (0.057)
Correlation(ρ)	0.570 (0.137)

Tab. 5.4: Estimation of the parameters and their SE based on 1000 bootstrap samples.

where 1000 survival curves have been randomly drawn from the Poisson model using the estimates of Table 5.4.

Results based on univariate meta-analysis are usually illustrated in the classical forest plot. An extension of the forest plot to a multivariate case is illustrated in Figure 5.4. The survival proportion estimates and their confidence intervals at each time-point for all studies along with those of the overall survival obtained by the analysis are shown. As it can be seen in the forest plot the confidence interval of the studies are relatively wide and they overlap with each other. This may also explain the presence of low estimation of heterogeneity between the studies.

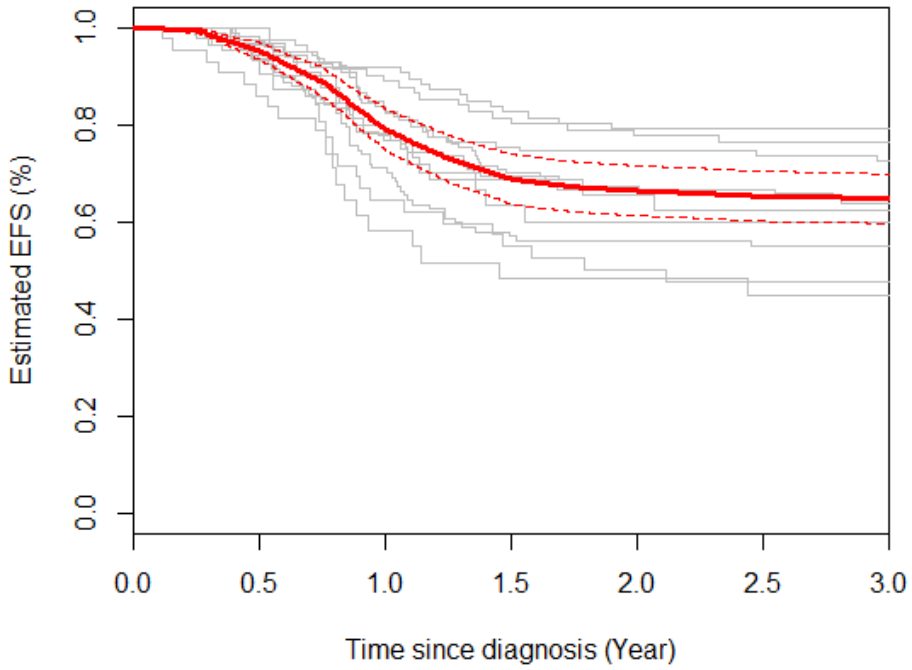

Fig. 5.2: Estimated overall survival curve (red line) along with the survival lines for each study (gray lines). The dashed red lines represent confidence intervals of the meta-analytic survival curve.

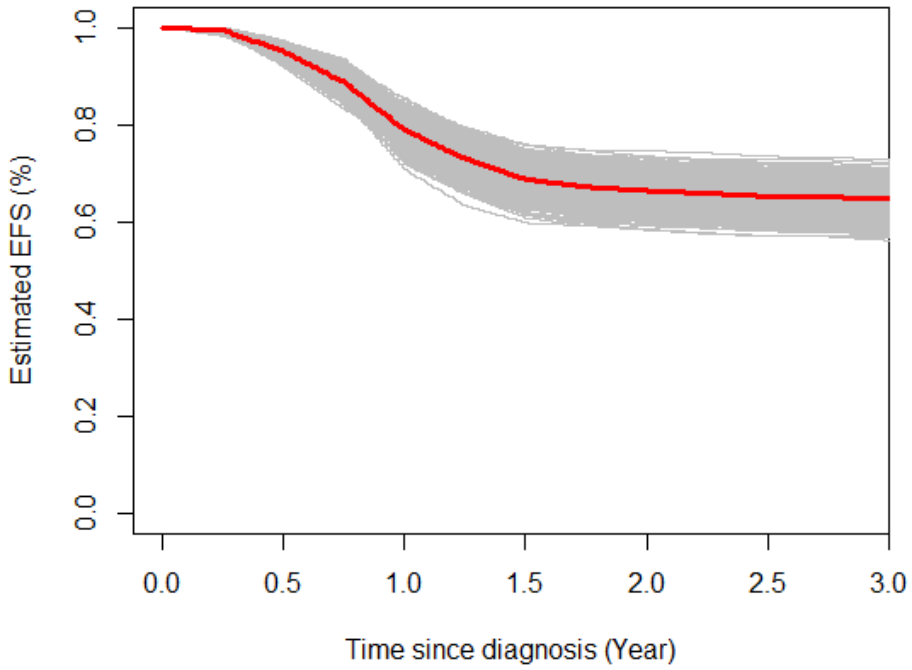

Fig. 5.3: Estimated overall survival curve (thick red line) along with 1000 survival curves estimated from data generated from the estimated Poisson-correlated gamma-frailty model.

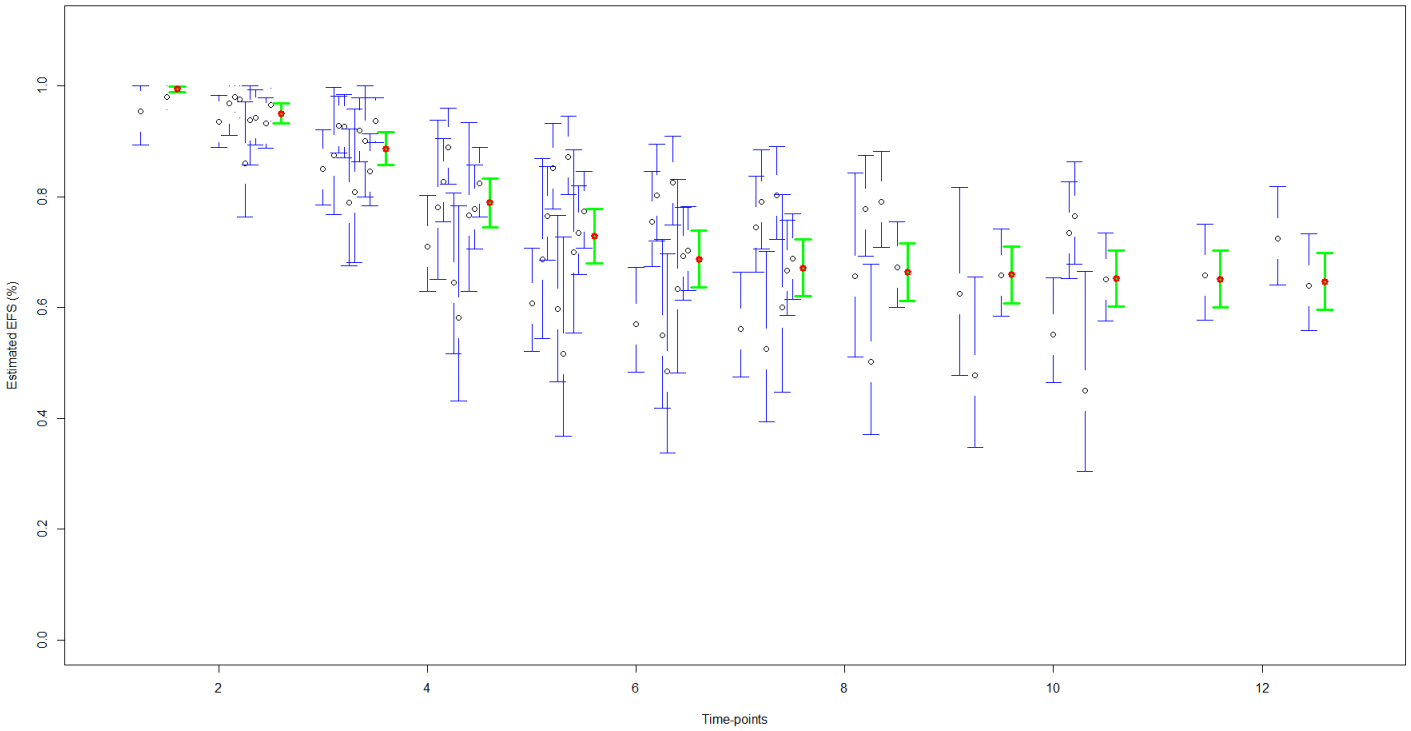

Fig. 5.4: Forest plot
The circles and the bars represent the estimates of survival proportions and their confidence intervals in all pre-determined time-points. The black circles and blue bars indicate each single study and the red circles and green bars are associated with the survival proportions obtained by meta-analysis.

There is no specific rule to determine the cut-off time-points or to specify the number of years to be included in the meta-analysis. They can be determined on the basis of clinical information, however, this information is not always available. If this is the case, they will be determined based on the intensity of the event occurrences and the survival curves. Of course by changing the time intervals and the length of follow-up the estimates of hazards will change too but the interest is on the affect of this change on the heterogeneity estimate and correlation estimate. Therefore, two other choices of time intervals and follow-up years have been investigated to study how different choices would have affected the heterogeneity and correlation estimates. Results are shown in Table 5.5.

	Heterogeneity	Correlation
2YEARS-3MONTHS	0.108	0.360
3YEARS-3MONTHS	0.104	0.570
3YEARS-4MONTHS	0.089	0.563

Tab. 5.5: Comparison between the estimations by different follow-up years and time intervals

The estimation of heterogeneity did not change by fixing the interval length to 3 months and following the studies 2 years instead of 3 years. However, the heterogeneity dropped from 0.104 to 0.089 by widening the interval to 4 months in the same number of follow-up years. The decrease in correlation estimate is considerable (drop from 0.570 to 0.360) by following the studies 2 years instead of 3 years with the same time interval of 3 months. While this variation is negligible if the follow-up length remains the same and the time interval length increases to 4 months.

R-codes developed to estimate Poisson-gamma-frailty model on the IPD is provided in Appendix C. The codes are general and can be applied on any dataset with the similar structure.

6. DISCUSSION

Poisson-gamma-frailty model is a new method to conduct meta-analysis on time-to-event data accounting for the correlation within studies and potential heterogeneity between studies. In this thesis, the focus was on the meta-analysis for a single EFS curve for each study. The methodology can be extended to a meta-analysis where each study presents two arms, by including also between-arm correlation parameter.

It is also feasible to employ covariates in the meta-analysis to investigate the causes of heterogeneity at the individual level. However, this was less interesting in our case as the estimation of heterogeneity was relatively low.

The meta-analysis was performed on 3-years EFS curves by looking at the events in consecutive intervals with the length of 3 months. Further research is needed to study the aspect concerning the choice of the time interval and its impact on the estimation. Further research is required to fit this model to continuous survival data by extending frailty processes to continuous time. This aspect is very interesting but is beyond the scope of this thesis.

Correlation structure is another aspect that might be investigated in future. The first-order autoregressive correlation structure was proposed in [7] and was employed in this thesis. This correlation structure seems realistic since it is expected that the number of events in neighbouring time intervals are more dependent than those further apart. However, it might be interesting to study different correlation structures.

It should be mentioned that the IPD meta-analysis could be performed using other methods. However, the aim of this thesis was to extend the correlated-Poisson gamma-frailty model at IPD level and develop R-codes to fit the model.

REFERENCES

[1] Elisabeth Wreford Andersen. Composite likelihood and two stage estimation in family studies. *Biostatistics*, 5(1):15–30, 2004.

[2] Lidia R. Arends, M. G. Myriam Hunink, and Theo Stijnen. Meta-analysis of summary survival curve data. *Statistics in Medicine*, 27 (22):4381–4396, 2008.

[3] Keith B. G. Dear. Iterative generalized least squares for meta-analysis of survival data at multiple times. *Biometrics*, 50(4):pp. 989–1002, 1994.

[4] Balgobind et al. Novel prognostic subgroups in childhood 11q23/mll-rearranged acute myeloid leukemia: results of an international retrospective study. *Blood*, 114(12):2489–2496, 2009.

[5] Crowther et al. Individual patient data meta-analysis of survival data using poisson regression models. *BMC Medical Research Methodology*, 12:34, 2012.

[6] M. Fiocco, H. Putter, and J. C. van Houwelingen. Meta-analysis of pairs of survival curves under heterogeneity: A poisson correlated gamma-frailty approach. *Statistics in Medicine*, 28(30):3782–3797, 2009.

[7] M. Fiocco, H. Putter, and J.C. Van Houwelingen. A new serially correlated gamma-frailty process for longitudinal count data. *Biostatistics*, 10(2):245–257, 2009.

[8] Marta Fiocco, Theo Stijnen, and Hein Putter. Meta-analysis of time-to-event outcomes using a hazard-based approach: Comparison with other models, robustness and meta-regression. *Computational Statistics and Data Analysis*, 56(5):1028 – 1037, 2012.

[9] C. Genest, K. Ghoudi, and L.-P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552, 1995.

[10] DavidV. Glidden. A two-stage estimator of the dependence parameter for the clayton-oakes model. *Lifetime Data Analysis*, 6(2):141–156, 2000.

[11] Philip Hougaard. A class of multivanate failure time distributions. *Biometrika*, 73(3):671–678, 1986.

[12] Mahesh K. B. Parmar, Valter Torri, and Lesley Stewart. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Statistics in Medicine*, 17(24):2815–2834, 1998.

[13] J. H. Petersen. An additive frailty model for correlated life times. *Biometrics*, 54(2):pp. 646–661, 1998.

[14] Joanna H. Shih and Thomas A. Louis. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51 (4):pp. 1384–1399, 1995.

[15] Catrin Tudur Smith, Paula R. Williamson, and Anthony G. Marson. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Statistics in Medicine*, 24(9):1307–1319, 2005.

[16] Lesley A. Stewart. Practical methodology of meta-analyses (overviews) using updated individual patient data. *Statistics in Medicine*, 14(19): 2057–2079, 1995.

[17] Lesley A. Stewart and Jayne F. Tierney. To ipd or not to ipd?: Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation and the Health Professions*, 25(1):76–97, 2002.

[18] Catrin Tudur, Paula R. Williamson, Saboor Khan, and Lesley Y. Best. The value of the aggregate data approach in meta-analysis with time-to-event outcomes. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(2):357, 2001.

[19] Hans C. Van Houwelingen, Koos H. Zwinderman, and Theo Stijnen. A bivariate approach to meta-analysis. *Statistics in Medicine*, 12(24): 2273–2284, 1993.

[20] Hans C. van Houwelingen, Lidia R. Arends, and Theo Stijnen. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, 21(4):589–624, 2002.

[21] W. N. Venables and B. D. Ripley. Modern applied statistics with s. *Springer*, 2002.

[22] Anatoli I. Yashin and Ivan A. Iachine. Genetic analysis of durations: Correlated frailty model applied to survival of danish twins. *Genetic Epidemiology*, 12(5):529–538, 1995.

APPENDICES

A. RECONSTRUCTION OF COUNT DATA

Count data can be reconstructed from a survival curve and follow-up information. A single survival curve from one study is considered on which the survival probabilities are observed at a pre-determined set of time-points j, $(0 < t_1 < \cdots < t_M)$. The corresponding disjoint time intervals are defined $I_j = (t_{j-1}, t_j]$ for $j = 1, \ldots, M$ with the convenient that $t_0 = 0$ Time t refers to follow-up time and time j indicates the index of the time intervals I_j. A model for the censoring mechanism based on the minimum and the maximum follow-up is assumed here for computing number at risk, number of events and person-years for each time. Let $C(t)$ be the function that models the censoring mechanism. Assuming that the censored observations are distributed uniformly over the intervals, $C(t)$ is defined as follows

$$C(t) = \begin{cases} 1 & \text{if } t \leq min_{FUP}; \\ 1 - \frac{t - min_{FUP}}{max_{FUP} - min_{FUP}} & \text{if } min_{FUP} < t < max_{FUP}; \\ 0 & \text{if } t \geq max_{FUP} \end{cases} \tag{A.1}$$

where min_{FUP} and max_{FUP} indicate minimum and maximum follow-up time respectively. Function $C(t)$, called the completeness function, expresses the proportion of patients at time t that have at least t time units of follow-up. Let define $C_j = C(t_j)$, $S_j = \hat{S}(t_j)$ and r_j as the completeness, estimated survival and the number of patients at risk at time j, respectively. Given the number of eligible patients (n), the effective number at risk and the number of events at time j and the number of censored are estimated respectively as

$$r_j = nS_jC_j$$
$$d_j = n(S_{j-1} - S_j)\frac{C_{j-1} + C_j}{2}$$
$$c_j = n(C_{j-1} - C_j)\frac{S_{j-1} + S_j}{2}$$

the number of person-years over interval I_j can be defined as $\tilde{r}_j = \Delta_j(r_j - c_j/2)$ where $\Delta_j = t_j - t_{j-1}$ is the length of I_j.

B. ASYMPTOTIC THEORY

Asymptotic theory can be used to obtain standard errors of the estimate of the parameters $(\boldsymbol{\beta}, \theta, \rho)$. The first-stage estimate $\hat{\eta} = (\hat{\boldsymbol{\beta}}, \hat{\theta})$ is the maximizer of

$$\ell_1(\eta) = \sum_{i=1}^{N} \ell_{1i}(\eta) = \sum_{i=1}^{N} \sum_{t=1}^{T} \ell_{1it}(\eta)$$

as defined in (3.3). Define the score functions

$$\frac{\partial}{\partial \eta} \ell_{1it}(\eta) = \sum_{t=1}^{T} \left(\frac{\partial}{\partial \boldsymbol{\beta}} \ell_{1it}(\eta), \frac{\partial}{\partial \theta} \ell_{1it}(\eta) \right)^{\top},$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} \ell_{1it}(\eta) = \frac{\theta}{\mu_{it} + \theta} (y_{it} - \mu_{it}) \mathbf{x}_{it} \ ,$$

$$\frac{\partial}{\partial \theta} \ell_{1it}(\eta) = \gamma_2(y_{it} + \theta) - \gamma_2(\theta) - \frac{y_{it} - \mu_{it}}{\mu_{it} + \theta} + \log\left(\frac{\theta}{\mu_{it} + \theta} \right) \ ,$$

and the Hessian matrix

$$\frac{\partial^2}{\partial \eta \partial \eta^{\top}} \ell_{1it}(\eta) = \sum_{t=1}^{T} \begin{pmatrix} \dfrac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\top}} \ell_{1it}(\eta) & \dfrac{\partial^2}{\partial \boldsymbol{\beta} \partial \theta} \ell_{1it}(\eta) \\ \left(\dfrac{\partial^2}{\partial \boldsymbol{\beta} \partial \theta} \ell_{1it}(\eta) \right)^{\top} & \dfrac{\partial^2}{\partial \theta^2} \ell_{1it}(\eta) \end{pmatrix}.$$

Standard asymptotic theory states that the first stage estimator $\hat{\eta} = (\hat{\boldsymbol{\beta}}, \hat{\xi})$ behaves asymptotically as

$$\hat{\eta} = \eta + \frac{1}{N} \sum_{i=1}^{N} \psi_{1i} \quad , \psi_{1i} = \mathbf{B}_1^{-1} \frac{\partial}{\partial \eta} \ell_{1i} \ ,$$

and hence is asymptotically unbiased with covariance matrix $\approx \mathbf{B}_1^{-1} \mathbf{M}_1 \mathbf{B}_1^{-1}$ with

$$\mathbf{B}_1 = -\frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \ell_{1i}(\eta)}{\partial \eta \partial \eta^{\top}}, \quad \mathbf{M}_1 = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\partial \ell_{1i}(\eta)}{\partial \eta} \right) \left(\frac{\partial l_{1i}(\eta)}{\partial \eta} \right)^{\top}.$$

The function ψ_{1i} is called the influence function of $\hat{\eta}$. The asymptotic covariance matrix $\mathbf{B}_1^{-1}\mathbf{M}_1\mathbf{B}_1^{-1}$ is a standard sandwich estimator.

For the asymptotic distribution of the second stage estimation $\hat{\rho}$ we also need to account for the fact that $\boldsymbol{\beta}$ and θ are random rather than fixed. Similar to the first-stage estimator, the influence function for $\hat{\rho}(\hat{\eta})$ for η known, is given by $\psi_{2i} = \mathbf{B}_2^{-1}\dfrac{\partial}{\partial\rho}\ell_{2i}$, with $\mathbf{B}_2 = -\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\frac{\partial^2}{\partial\rho^2}\ell_{2i}(\eta,\rho)$. The influence function of $\hat{\rho} = \hat{\rho}(\hat{\eta})$ is then given by $\psi_{2i} - \mathbf{B}_2^{-1}\mathbf{B}_{12}^{\top}\psi_{1i}$, where $-\dfrac{1}{N}\displaystyle\sum_{i=1}^{n}\frac{\partial^2}{\partial\rho\partial\eta}\ell_{2i} \to \mathbf{B}_{12}$. Define

$$\mathbf{M}_2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial\ell_{2i}(\eta,\rho)}{\partial\rho}\right)^2, \quad \mathbf{M}_{12} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial\ell_{1i}(\eta)}{\partial\eta}\right)\left(\frac{\partial\ell_{2i}(\rho)}{\partial\rho}\right) .$$

It follows that

$$\begin{aligned}
\operatorname{var}(\hat{\rho}) &\approx \frac{1}{N^2}\sum_{i=1}^{n} E(\psi_{2i} - \mathbf{B}_2^{-1}\mathbf{B}_{12}^{\top}\psi_{1i})(\psi_{2i} - \mathbf{B}_2^{-1}\mathbf{B}_{12}^{\top}\psi_{1i})^{\top} \\
&\approx \frac{1}{N}\cdot\Big[\mathbf{B}_2^{-1}\mathbf{M}_2\mathbf{B}_2^{-1} - 2\mathbf{B}_2^{-1}\mathbf{B}_{12}^{\top}\mathbf{B}_1^{-1}\mathbf{M}_{12}\mathbf{B}_2^{-1} \\
&\qquad\qquad + \mathbf{B}_2^{-1}\mathbf{B}_{12}^{\top}\mathbf{B}_1^{-1}\mathbf{M}_1\mathbf{B}_1^{-1}\mathbf{B}_{12}\mathbf{B}_2^{-1}\Big]
\end{aligned}$$

and

$$\begin{aligned}
\operatorname{cov}(\hat{\eta},\hat{\rho}) &\approx \frac{1}{n^2}\sum_{i=1}^{n} E\psi_{1i}(\psi_{2i} - \mathbf{B}_2^{-1}\mathbf{B}_{12}^{\top}\psi_{1i})^{\top} \\
&\approx \frac{1}{n}\cdot\left[\mathbf{B}_1^{-1}\mathbf{M}_{12}\mathbf{B}_2^{-1} - \mathbf{B}_1^{-1}\mathbf{M}_1\mathbf{B}_1^{-1}\mathbf{B}_{12}\mathbf{B}_2^{-1}\right]
\end{aligned}$$

C. R-CODES

C.1 Function to transform survival data to count data

```
#Convert survival data to count data

CountData<-function(SurvDat,Int,FU){

  #SurvDat: Survival data with columns "ID","study","event","time_
      event"

  #Int: Indicates desired length of time interval

  #FU: Indicates desired length of follow-up time to be included in
      the analysis

  #Note that the unit of time interval and follow-up should be the
      same as time_event

  #time-points
  points<-unique(c(seq(0,max(SurvDat$time_event),by=Int),max(SurvDat$
      time_event)))

  #Interval labels in the set of integer numbers
  IntLabel<-cut(eventdat$time_event, points,labels = 1:(length(points)
      -1), include.lowest = TRUE, right = TRUE)

  SurvDat$IntLabel<-as.numeric(IntLabel)

  #Subset of data related to the desired follow-up time
  Dat<-SurvDat[SurvDat$time_event<=FU,]

  SurvDatSplit<-split(SurvDat,SurvDat$study)

  DatSplit<-split(Dat,Dat$study)

  TDat<-max(Dat$IntLabel)

  N<-length(split(Dat,Dat$study))

  #Add number of patients at risk to data
  AddatRisk<-list()
  for(j in (1:length(DatSplit))){
    AddatRisk[[j]]<-DatSplit[[j]][order(DatSplit[[j]]$time_event), ]
    AddatRisk[[j]]$risk<-seq(nrow(SurvDatSplit[[j]]),(nrow(
        SurvDatSplit[[j]])-((nrow(DatSplit[[j]]))-1)),by=-1)
  }

  IntSplit<-list()
  for(i in(1:length(AddatRisk))){
    IntSplit[[i]]<-split(AddatRisk[[i]],AddatRisk[[i]]$IntLabel)
  }
```

```
#Nunber at risk in intervals with non-zero number of events
NonZeroInt<-list()
for(i in 1:length(IntSplit)){
  NonZeroInt[[i]]<-as.numeric(names(IntSplit[[i]]))
}

#Matrix of number of events (T x N)
Eventmat<-matrix(0,nrow=TDat,ncol=N)
for(j in(1:length(IntSplit))){
  for(i in 1:length(IntSplit[[j]])){
    Eventmat[NonZeroInt[[j]][i],j]<- sum(IntSplit[[j]][[i]]$event)
  }
}

#Matrix of number censorings (T x N)
Censoringmat<-matrix(0,nrow=TDat,ncol=N)
for(j in(1:length(IntSplit))){
  for(i in 1:length(IntSplit[[j]])){
    Censoringmat[NonZeroInt[[j]][i],j]<- nrow(IntSplit[[j]][[i]])
  }
}

#Matrix of number at risk (T x N)
atRiskmat<-matrix(NA,nrow=1,ncol=N)
atRiskmat[1,]<-sapply(a,FUN=function(x)max(x$risk))
for(i in 2:TDat){
  NextRow<-atRiskmat[i-1,]-Censoringmat[i-1,]
  atRiskmat<-rbind(atRiskmat,NextRow)
}

#Compute number of person-years
for(i in(1:length(IntSplit))){
  for(j in(1:length(IntSplit[[i]]))){
    IntSplit[[i]][[j]]$dtt<-diff(c((IntSplit[[i]][[j]]$IntLabel
        [1]-1)*Int,IntSplit[[i]][[j]]$time_event))
    IntSplit[[i]][[j]]$pyrs<-IntSplit[[i]][[j]]$risk*IntSplit[[i]][[
        j]]$dtt
  }
}

#Matrix of number of person-yeas (T x N)
PYRSmat<-matrix(NA,nrow=TDat,ncol=N)
for(j in(1:length(IntSplit))){
  for(i in 1:length(IntSplit[[j]])){
    PYRSmat[NonZeroInt[[j]][i],j]<- sum(IntSplit[[j]][[i]]$pyrs)+
      ((NonZeroInt[[j]][i]*Int-max(IntSplit[[j]][[i]]$time_event))*
         (IntSplit[[j]][[i]]$risk[nrow(IntSplit[[j]][[i]])]-1))
  }
}

for(j in (1:length(AddatRisk))){
  for(i in (1:TDat)){
    PYRSmat[i,j]<-ifelse(is.na(PYRSmat[i,j])==TRUE,Int*atRiskmat[i,j
        ],PYRSmat[i,j])

  }
}

#Write count data (number of events and person-years at each
    interval for each study)
```

```
  CountData<-data.frame(rep(1:N,each=TDat),rep(c(1:TDat),N),as.vector(
      Eventmat),as.vector(PYRSmat))

  colnames(CountData)<-c("Study","Interval","N.Event","PYRS")

   return(list(CountData=CountData,Eventmat=Eventmat,atRiskmat=
       atRiskmat,PYRSmat=PYRSmat))
}
```

./CountDataFun.R

C.2 Poisson-Gamma-Frailty model function

```
library(MASS)

#PoissonCOrrelatedFunction

corr.AR <- function(rho,T){
  # Function to create correlations matrix for autoregressive models
  #Input:
  #  rho: correlation parameter
  #  T: number of time points
  R <- diag(T)
  for (s in 1:T)
    for (t in 1:T)
      R[s,t] <- rho^(abs(s-t))
  return(R)
}
##############################################
colrep <- function(v,n){
  #function to repeat vector v n times
  # Input:
  #   v: vector
  #  n: number of columns in the matrix
  # Output:
  # matrix of repeated columns, dimension: length(v)X n
  return(matrix(rep(v,n),length(v),n))
}

##############################################

rowrep <- function(v,n){
  #function to repeat vector v n times
  # Input:
  #   v: vector
  # n: number of rows in the matrix
  # Output:
  # matrix of repeated rows, dimension: length(v)X n
  return(t(colrep(v,n)))
}

##############################################

PoisCorrGammFrailty<-
  function (formula, data, rho = TRUE)
  {
    data$Study<-as.factor(data$Study)
    data$Time<-as.factor(data$Time)
    formula <- as.formula(formula)
```

```
    glmnb <- glm.nb(formula, data = data, link = "log")
    bbeta <- glmnb$coef
    b <- c(bbeta[1],bbeta[-1]+bbeta[1])
    th <- glmnb$theta
    xi <- 1/th
    p <- length(bbeta)
    n <- nrow(data)
    T <- length(unique(data$Time))
    N <- n/T
    y <- glmnb$y
    mu <- glmnb$fit
    r <- y - mu
    X <- model.matrix(glmnb)
    mumat <- matrix(mu, N, T , byrow=TRUE)
    if (!rho)
      return(list(b = b, th = th, xi = xi, glmnb = glmnb, mumat =
          mumat))
    else {
      ymat <- matrix(y, N, T)
      opt <- optimize(f = loglikrho, interval = c(0, 1), lower = 0,
                      upper = 1, maximum = TRUE, tol = .Machine$double
                          .eps^0.25,
                      y = ymat, mu = mumat, th = th)
      rho <- opt$maximum
      return(list(b = b, th = th, xi = xi, rho = rho, glmnb = glmnb,
                  mumat = mumat))
    }
  }
loglikrho<-
  function (rho, y, mu, th)
  {
    res <- pcgf.loglik(y, mu, th, rho)
    return(res)
  }

pcgf.loglik<-
  function (ymat, mumat, th, rho)
  {
    xi <- 1/th
    N <- nrow(ymat)
    T <- ncol(ymat)
    loglik <- 0
    R <- corr.AR(rho, T)
    for (i in 1:N) {
      for (s in 1:(T - 1)) {
        yis <- ymat[i, s]
        muis <- mumat[i, s]
        for (t in (s + 1):T) {
          yit <- ymat[i, t]
          muit <- mumat[i, t]
          rhost <- R[s, t]
          loglik <- loglik + loglik1.ist(yis, yit, muis,
                                         muit, xi, rhost)
        }
      }
    }
    return(sum(loglik))
  }
loglik1.ist<-
  function (y1, y2, mu1, mu2, xi, rhost)
  {
    mu12 <- mu1 + mu2
```

```
    theta <- 1/xi
    rho <- rhost
    Pist1 <- dnbinom(y1:0, size = theta * (1 - rho), mu = mu1 *
                     (1 - rho))
    Pist2 <- dnbinom(y2:0, size = theta * (1 - rho), mu = mu2 *
                     (1 - rho))
    P1 <- colrep(Pist1, y2 + 1)
    P2 <- rowrep(Pist2, y1 + 1)
    outerm <- outer(0:y1, 0:y2, "+")
    outerv <- as.vector(outerm)
    helpv <- as.vector(colrep(0:y1, y2 + 1))
    P3 <- matrix(dnbinom(outerv, size = theta * rho, mu = mu12 *
                         rho), y1 + 1, y2 + 1)
    P4 <- matrix(dbinom(helpv, outerv, mu1/mu12), y1 + 1, y2 +
                 1)
    P <- P1 * P2 * P3 * P4
    return(logP = log(sum(P)))
  }
##########################################

#Run the function on data

PoisCorrResult<-PoisCorrGammFrailty(N.Event ~ Time + offset(log(PYRS))
    , data=longdata)

#Extract the results

mumat<-PoisCorrResult$mumat
b<- PoisCorrResult$b
lambda<-exp(b)
rho<- PoisCorrResult$rho
xi<- PoisCorrResult$xi
```

./PoissonGammaFrailtyFun.R

C.3 Bootstrap function

```
#Bootstrap Function

bootNG <- function(mumat,nrep,T,xi,rho,pyrs){
  # Function to simulate correlated poisson-frailty data
  # Input:
  #   mumat: matrix estimated exp(X*beta)
  #   p: length of vector of parameters (p=T+ number of covariates)
  #   nrep: number of simulations
  #   T: number of time points
  #   b: vector of parameters (beta)
  #   xi: variance of Gamma
  #   rho: subject correlation corr(Z[i,s],Z[i,t])=rho^|s-t|
  # Output:
  #   A matrixe with estimated beta, xi,rho
  N <- nrow(mumat)
  n <- N*T
  estmat2stage <- matrix(NA,nrep,T+2)
  # estmat1stage <- matrix(NA,nrep,p+2)
  # th: parameter of the multivariate marginal Gamma(th,th)
      distribution
  th <- 1/xi
```

```
  for (irep in 1:nrep) {
    cat("Replication",irep,"\n")
    flush.console()
    events <-Z <- matrix(NA,N,T)
    # generate correlated frailty
    for (i in 1:N) {
      # generate vector of elements X_{i+} ~ Ga(th(1-rho)rho^{T+1-i}
      Xk <-  rgamma(T, th*(1-rho)*rho^(T:1),rate = th)
      # generate vector of elements X_{+j} ~ Ga(th(1-rho)rho^j,th), l
          =1,...,T
      Xl <-  rgamma(T, th*(1-rho)*rho^(1:T),rate = th)
      # generate  X_{++} ~ Ga(th*rho^{T+1},th)
      XX <-  rgamma(1, th*rho^{T+1},rate = th)
      # generate matrix X_{ij}
      Xkl <- matrix(0,T,T)
      for (k in 1:T)
        for(l in k:T)
          Xkl[k,l] <- rgamma(1,th*(1-rho)^2*rho^(l-k),rate=th)
      # generate frailties Z_t
      for(t in 1:T)
        Z[i,t] <- sum(Xk[1:t]) + sum(Xl[t:T]) +XX +
        sum(Xkl[1:t,t:T])
      # generate number of events from a Poisson(lambda) with
      # lambda=mu[i,t]*Z[i,t]  mu=exp(X*beta)
      events[i,] <- rpois(T,mumat[i,]*Z[i,])
    }

    # create data in a long format
    datalong <- data.frame(Time=rep(1:T,N),st=rep(1:N,each=T),count=as
        .vector(t(events)),pyrs=pyrs)
    datalong$Time <- factor(datalong$Time)
    # Estimate vector beta,xi,rho with two stage procedure
    ests<-PoisCorrGammFrailty(count ~ Time + offset(log(pyrs)), data=
        datalong)
    estmat2stage[irep,1:T] <- exp(ests$b)
    estmat2stage[irep,T+1] <- ests$xi
    estmat2stage[irep,T+2] <- ests$rho
  }

  return(estmat2stage)
}
```

./BootstrapFun.R