Department of Mathematics

Master Thesis
Statistical Science for the Life and Behavioural Sciences

Statistical methodology for volume-outcome studies

Floor M. van Oudenhoven

Supervisor: Dr. M. Fiocco (LUMC & LU)

November 2014

Universiteit Leiden
The Netherlands

Leiden University Medical Center

Abstract

A growing body of literature studies the association between measures of hospital volume and patient outcomes after a surgical treatment to evaluate whether hospitals with large case volumes are associated with better outcomes. Applying the appropriate statistical methodology to these so-called volume-outcome studies erases several challenges such as the selection of a longitudinal estimation method and the specification of an appropriate measure for hospital volume. In daily practice, difficulties involved in volume-outcome studies are often not recognized. Regularly, hospital volume is analysed as a categorical variable, thereby neglecting its time-dependent nature. In addition, many volume-outcome studies ignore bias that may occur in the estimation process when certain assumptions are violated and traditional methods are used.

In this thesis we use the recurrent marked point process to approach a longitudinal volume-outcome analysis of clustered data. Statistical issues in the selection of both non-aggregate and yearly aggregate measures for hospital volume are considered.

An additional aspect sometimes associated with clustered data concerns the presence of informative cluster size, where outcome depends on cluster size conditional on covariates. The concept of informative cluster size within a volume-outcome study presents a unique situation since hospital volume is both the covariate of primary interest under study and it is closely linked to cluster size. Within cluster resampling (WCR) is an appropriate method to analyse informative cluster size data.

The novelty of this thesis is to apply WCR in the framework of a recurrent marked point process to study a longitudinal volume-outcome association. A simulation study has been performed to asses the performance of the proposed method and to evaluate whether the use of aggregate measures for hospital volume leads to bias in the estimation of the volume-outcome association. Simulations show that when informative cluster size is present, the proposed method estimates the parameter for volume with small bias. In addition simulations suggest that bias might be introduced when an aggregate measure for present hospital volume is used.

Acknowledgement

I wish to express my deepest appreciation and gratitude to my supervisor Dr. Marta Fiocco, who I thankfully like to call my "scientific mother", for her guidance, critical comments and warm encouragement throughout the process of writing this thesis. Working together was a very pleasant experience, both on a professional and personal level.
The department of surgery at the Leiden University Medical Centre (LUMC) is gratefully acknowledged for providing the dataset.
I like to thank my teachers from the master track *Statistical Science for the Life and Behavioural Sciences* for sharing their expertise and enthusiasm about statistics. I would also like to thank people attending the "Survival lunch" for their useful comments on my thesis presentation.
I gratefully like to thank my friends and family for the fun and support they brought me throughout my study. A special thanks for my roommates and boyfriend for making their laptops available for parts of my computational intensive simulation study. At last, I like to thank my parents for their never-ending support and belief in me. They always help me to accomplish what seems improbable in advance. Words cannot express how grateful I am.

Contents

1 Introduction

The size of a hospital may be a measurable variable and it is assumed to have relevant impact on effectiveness of health care [6, Davoli *et al.*]. Large hospitals or surgeon volume activity may for example denote more resources and experience.
Improving the quality and effectiveness of health care is a central goal of health policies. A growing body of literature studies the association between hospital volume and health outcome of patients following a surgical treatment. Results of these so-called volume-outcome studies may have direct policy implications such as regularization of health care into large centres [19, Livingston *et al.*].
Despite the fact that volume-outcome studies have become a hot topic in literature, there is no common method of estimation. Methodological issues have been raised about volume-outcome studies because the association between hospital volume and post-treatment outcome erases several challenges [17, Kulkarni *et al.*]. First of all, this type of study typically collects patients' information concerning all subjects undergoing a certain surgery or treatment at the same hospital over time. Patients treated at the same hospital may be more likely to experience similar outcomes than patients treated at a different hospital. As a consequence, observations within the same hospital might be correlated.
Second, several problems may arise about the specification of an appropriate measure for hospital volume. At the moment, a standard definition of hospital volume has not been established [17]. It is important to make a precise choice between volume measures that have a present or cumulative character also by considering the research question. Present volume may be defined as the number of surgeries per hospital per year whereas cumulative volume may denote the number of surgeries per hospital accumulated over all years of study.
A key issue as mentioned by [9, French *et al.*] is that hospital volume is not a fixed quantity but rather a quantity that changes over time. Both present and cumulative hospital volume may change over the course of the study. Many volume-outcome studies, however, analyse hospital volume as a categorical variable. In this way, the time-dependent character of hospital volume is neglected. In addition, the selection of cut-off points for the different categories may have an impact on the statistical significance of the obtained volume-outcome associations.
The concept of *informative cluster size* represents a third challenge. Informative cluster size is said to exist when the outcome of interest is related to cluster size given the covariates [15, Hoffman *et al.*]. Volume-outcome studies represent a particular statistical problem since hospital volume is both the covariate of primary interest and it is closely linked to cluster size.
Longitudinal data analysis methods may be used to deal with correlation among data. French *et al.* [9] proposed the recurrent marked point process as a general framework to estimate volume-outcome associations from longitudinal data.
The characteristic of a recurrent marked point process data is that the outcome or *mark*

(e.g. post-treatment outcome) exists if and only if an event (e.g. surgery) occurs [10, French and Heagerty]. Commonly used longitudinal data analysis methods such as generalized estimating equations (GEE) and generalized liner mixed models (GLMMs) can be used to provide estimates under the recurrent marked point process setting, taking into account the clustered nature of the data. When cluster size is related to the outcome however, covariance weighted methods do not longer provide unbiased estimates. In this case independence estimating equations (IEE) is the only option that may be used to provide consistent estimation of the regression parameters [9, 10].

1.1 Aims of this thesis

The first goal of this thesis to investigate how changes in hospital volume are associated with better outcomes by employing the appropriate statistical methodology by using data concerning patients with oesophageal cancer surgery. For this purpose, the recurrent marked point process is used. It is explored how alternative measures for hospital volume, both non-aggregate and aggregate, yield to different results.
The second goal is to test for the presence of informative cluster size in the data employed in this thesis and to propose a new method suitable for volume-outcome studies in which informative cluster size is present.

1.2 Structure of this thesis

This thesis is organized as follows. In Chapter 2 a detailed description of the data is provided. Chapter 3 starts by introducing the basic concepts of point processes, followed by a description of the recurrent marked point process. In Chapter 4 more technical information concerning GLMMs and GEE is given since they will be used to provide estimates under the recurrent marked point process model. Chapter 5 describes the application of the recurrent marked point process to the dataset employed in this thesis. Technical details and results are provided. In Chapter 6 informative cluster size is introduced. In the same chapter an overview of existing methods concerning marginal inference under informative cluster size is given. In the last section it is examined whether cluster size is informative in the dataset used in this thesis.
Within cluster resampling (WCR) is one of the marginal methods appropriate for inference under informative cluster size. In Chapter 7 it is explored how a different estimation method, based on WCR, can be applied when informative cluster sizes are present. To assess the performance of the new method proposed, a simulation study is performed in Chapter 8. In addition it is evaluated, by means of simulation, whether the use of aggregate measures lead to bias in the estimation of the volume parameter. This thesis ends with a critical appraisal on statistical methodology used in existing volume-outcome studies.
The statistical analyses are performed in the R-software environment. All R code can be found in the appendices.

2 Data description

2.1 Background information

Data from the Netherlands Cancer institute, covering all hospitals in the country, is used in this thesis. Information about all newly diagnosed malignant cancer patients is routinely collected from hospital records between 6-18 months after diagnosis. Topography and morphology are coded according to the International Classification of Diseases for Oncology (ICD-O) [11, Fritz]. Quality and completeness are outstanding [24, Schouten *et al.*]. The data used in this thesis concerns information about patients after oesophageal cancer surgery between 1989 and 2010. Oesophageal cancer surgery is associated with high postoperative mortality rates. To reduce mortality and improve survival, it has been suggested that these high-risk operations should be performed in specialized centres with adequate annual volume [7, Dikken *et al.*]. Earlier studies showed that oesophageal cancer patients have better health outcomes when surgery is performed in hospitals with large case volumes. Since 2006 a minimum volume of 10 oesophagectomies per year is implemented by the Dutch Healthcare Inspectorate. Since 2011, this minimal volume is increased to 20 oesophagectomies per year. More information about the data can be found in [7].

2.2 Data description

As described in Section 2.1, the data contains information about patients with oesophageal cancer diagnosed between 1989 and 2009. All 10,0025 patients in the dataset underwent oesophageal cancer surgery. Oesophagectomies are performed at 148 different hospitals. The dataset does not include patients with carcinoma in situ or patients with distant metastases. For each patient, information on several demographic variables is available, next to information on stage and cancer morphology. Additionally, for each patient an id number identifying in which hospital the surgery was performed is associated. Table 1 gives an overview of all registered patients' characteristics.

Surgery. Oesophageal cancer surgeries are defined as resections for cancers of the oesophagus (C10-15.9) and gastric cardia (C16.0) [7]. Gastric cardia is located at the end of the oesophagus; from here the contents of the oesophagus empty in the stomach. Minimum and maximum cumulative number of surgeries per hospital between 1989 and 2010, are respectively 1 and 1057; mean and median cumulative hospital size are respectively 65 and 165 surgeries.

The majority (76%) of patients in the dataset is male. Patients' age is distributed between 23 and 94 years with mean at 63 years (see Figure 2.1). After 6 months of follow-up since surgery, approximately 13% of the patients died.

Figure 2.1: Patient age distribution.

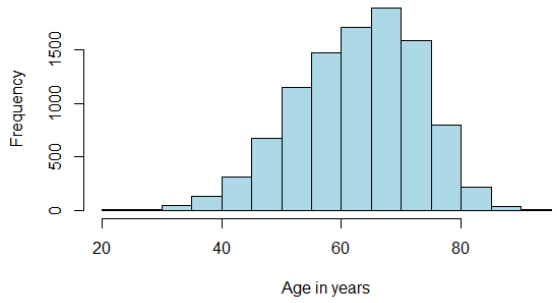


Figure 2.2: Total number of surgeries accumulated over all years of study (1989-2009) corresponding to each hospital.

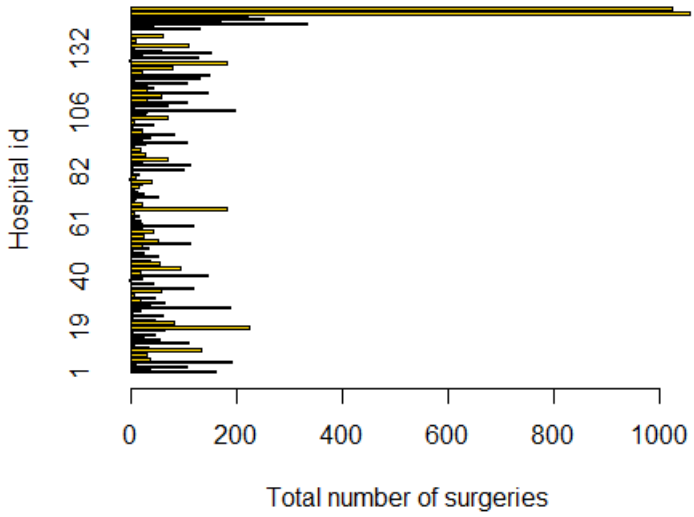


Figure 2.3: Observed patient outcome (dead or alive) after 6 months since surgery for each category of cumulative hospital size (very large, large, medium and small).

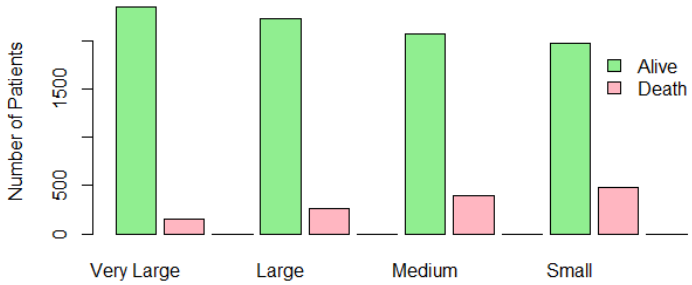
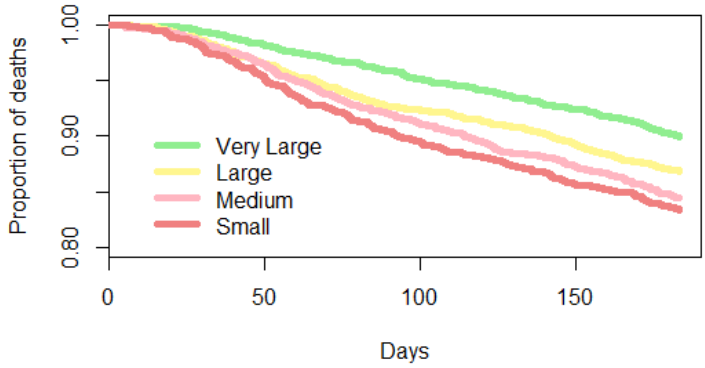


Figure 2.4: Proportion of deaths after 6 months since surgery for each category of cumulative hospital size (very large, large, medium and small).



The outcome of interest is death from any cause within 6 months since surgery. Rather than modelling time to event (e.g. death) a binary outcome variable is modelled, indicating whether or not the patient is still alive, by using logistic models. Figure 2.3 shows observed patients' outcomes after 6 months follow-up since surgery per different categories of cumulative hospital size. Categories are based on the first quartile, median and third quartile of cumulative hospital size so that every category contains the same amount of observations. Figure 2.3 suggests a possible association between cumulative hospital size

and post-treatment outcome. In this figure, four different volume categories are defined from very large to small. In the category defined as small, 25% of the patients died within 6 months after surgery, where in the remaining three categories the percentages are equal to 19%, 12% and 6%. In Figure 2.4 survival curves for each category are shown. This figure shows the same trend as observed in Figure 2.3. As stated in the previous chapter, although many volume-outcome studies analyse hospital volume as a categorical variable, this strategy requires great caution. In this thesis hospital volume is analysed as a continuous variable; Figures 2.3 and 2.4 merely serve for the purpose of illustration.

Table 1: Patients' characteristics.

Variable	Coding
Gender	Male
	Female
Age	< 60
	60-75
	> 75
SES	Low
	Medium
	High
	Unknown
Hospital id (surgHospital)	1-385
Year of surgery (surgYear)	1989-2010
Morphology of cancer	Adenocarcinoma
	Squamous-cell carcinoma
	Other
Stage of cancer	I
	II
	III
	IV
	X
Use of preoperative therapy	No
	Yes
Use of postoperative therapy	No
	Yes

3 Recurrent marked point process

In this thesis the recurrent marked point process is used as an approach to model volume-outcome associations. In Section 3.1 general concepts of point processes are introduced. The recurrent marked point process is illustrated in Section 3.2.

3.1 Point processes

A point pattern is basically a random collection of points. Many real phenomena produce data that can be represented by a point pattern, either in one, two or more dimensions. A *point process* aims to analyse the random structure of such patterns. It is a useful model for the timing or location of points in space.

A point process in one dimension is a sequence of real numbers, e.g. the timing of events, and may be represented as

$$T_1 < T_2 < \cdots < T_i < \dots$$

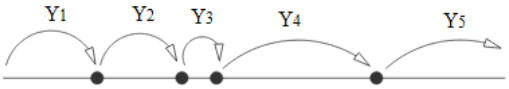
with T_i denoting the time point at which a particular event takes place. Examples include arrival time points of customers at service stations, failure time points of machines, times of earthquakes etc. In Figure 3.1 a one-dimensional (temporal) point process is illustrated.

Figure 3.1: One dimensional (temporal) point process.

A temporal point process can equivalently be represented by its *inter-arrival* or *inter-event* times (see Figure 3.2) defined as

$$\{Y_1, Y_2, \dots\} \text{ with } Y_i = T_i - T_{i-1};\ i = 1, 2, \dots;\ T_0 = 0.$$

Figure 3.2: Inter-arrival times Y_i.

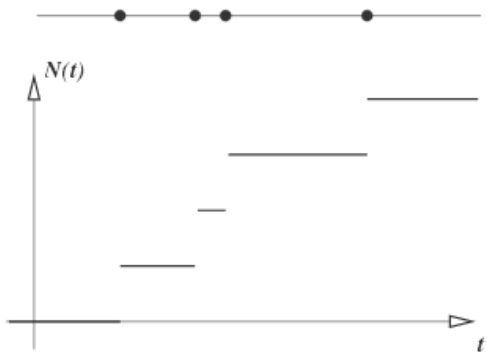


Definition A point process is said to be *recurrent* if its corresponding inter-arrival times $\{Y_1, Y_2, \dots\}$ is a sequence of independent, identically distributed random variables.

A point process may also be represented by a cumulative counting measure N_t (see Figure 3.3) which represents the number of points arriving up to time t

$$N(t) = \sum_{i=1}^{\infty} 1\{T_i \leq t\}, \text{ for } t \geq 0.$$

Figure 3.3: A point process represented by a counting measure N(t).

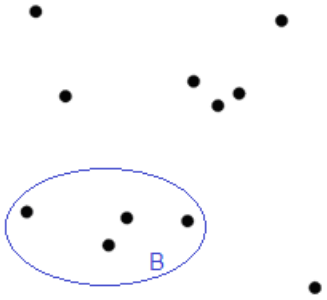


A *spatial point process* is a point process in d-dimensional space, where $d \geq 2$. For example, an earthquake may be represented by a time point, next to a point location. Figure 3.4 shows a two dimensional (spatial) point process.

Figure 3.4: A two dimensional (spatial) point process.

To give a more technical definition of a spatial point process a region specific counting measure $N(B)$ is needed. The region specific counting measure $N(B)$ denotes the number of points falling in B defined for each bounded closed set, so-called borel set $B \subset \mathbb{R}^2$ (see Figure 3.5).

Figure 3.5: Region specific counting measure $N_X(B) = 4$ for a spatial point process X.



Definition A spatial point process is a random variable X with an observed pattern x. Let N and (Ω, F, P) be the set of all counting measures on X and some probability space respectively.
The random variable X may then be regarded as a measurable map $N : \Omega \to N$ from (Ω, F, P) into an outcome space $(N, \mathcal{N})$, where x is a single realisation of X.

There are several kind of point processes. One of the simplest point processes is the so-called Poisson process, which is often used as a model for counting problems. In the Poisson process the number of events in successive intervals is assumed to be independent, as is the time between events (i.e. inter-arrival times). Another point process, typically used to model arrival times of customers at a service station, is the renewal process. It approximates the inter-arrival times by independent and identically distributed random variables and it is therefore a recurrent point process.
Point processes can also be applied in clinical research. They may be used in case frequency of recurrent events is the focus of research. For example, the frequency of hospitalisation can be used as a measure of medical costs in health economics [16, Huang *et al.*]. In medical studies, the frequency of recurrent events is often an indication of the severity of a disease. Figures in this chapter are based on [1, Baddeley *et al.*].

3.2 Marked point process

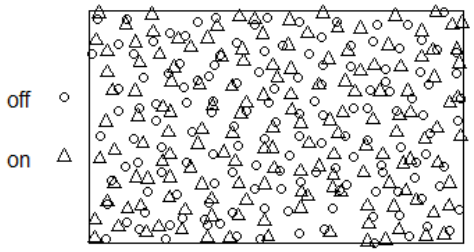
Marked point processes form another class of point processes, arising when the point process is not the primary object of study, but part of a more complex model [5, Daley *et al.*]. In a marked point process additional information is associated to each point, defined as *mark*. The mark is usually, but not necessarily, the outcome of interest. From a mathematical perspective, a mark can be considered as an extra coordinate for each point of a pattern.

A marked point process Y on a space S with marks in a space M may be represented as

$$Y = \{x_i, m_i\}$$

where x_i represents the point location and m_i is the corresponding mark (see Figure 3.6).

Figure 3.6: A realisation of a marked point process in the unit square with a binary mark space M= {off, on}.



In marked point processes, marks are observed only at point locations. Phrased in other words, an outcome exists if and only if an event occurs [9, 10]. Consider for example a point process that models arrival times of customers; a customer will spend a certain amount of money (i.e. the mark), at a specific time point. This amount of money will only be observed if the event exists (i.e. customers arrives at the particular service station).
A spatial point process could equivalently be seen as a marked point process where each time point is labelled with a mark. Interpreting the example with earthquakes as a marked point process may be done when the location is of primary interest. For example in case particular interest is in identifying the most dangerous hotspots of earthquakes in a certain region.

A marked point process consists of two parts:

- *Intensity measure*; the average number of points per unit area. The intensity measure of a point process is comparable with the expected value of a random variable.
- *Conditional mark distribution*; given the intensity, the marks corresponding to the points have a specific probability distribution.
  The mark space M can assume different forms such as a finite set, a binary set or a continuous interval. The mark space in Figure 3.6 is binary. Figure 3.9 shows an example of a marked point process with a finite mark space M={1, 1.5, 2, 2.5, 3}.

Both intensity measure and mark distribution may depend on a vector of covariates.

Definition Let X be a point process on $S = \mathbb{R}^2$, the intensity measure is given by

$$\Lambda(B) = E[N_X(B)], \quad B \subset S. \tag{1}$$

A binomial point process has intensity $\Lambda(B) = np$. The uniform Poisson process has intensity $\Lambda(B)$, proportional to the volume of the region B_{vol}; $\Lambda(B) = B_{vol}\lambda(B)$.

Definition If the intensity measure $\Lambda(B)$ satisfies

$$\Lambda(B) = \int_B \lambda(x)dx \tag{2}$$

for some function λ, then λ is called the *intensity function* of a random point process X.

Definition When the intensity is constant, i.e. the events occur at a constant rate, the point process X is *homogeneous*.

Figure 3.7: Realisation of an homogeneous Poisson process in the unit square, with intensity equal to 25.

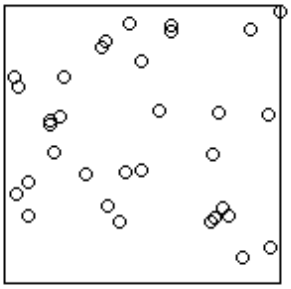
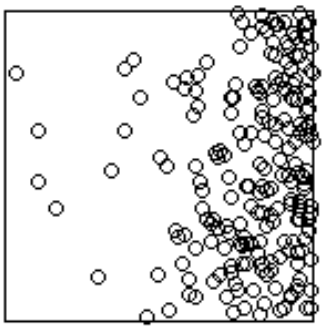


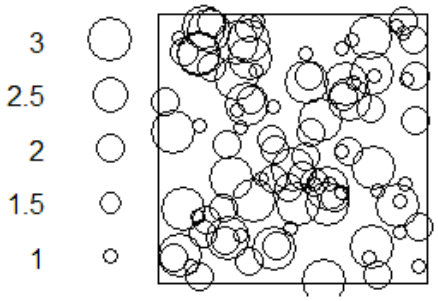
Figure 3.8: Realisation of an inhomogeneous Poisson process in the unit square, with intensity function $\beta(x, y) = exp(2 + 5x)$.



There are different possible structures in marked point processes. The marks may both be dependent and independent from each other. Further, marks may depend locally on point intensity. Earthquakes may for example be heavier in areas with high earthquake densities. Software have been written to simulate (marked) point process data. The package **spatstat** can be used for the statistical analysis of spatial point patterns in the R-software environment [2, Baddeley and Turner]. The following R code is used to generate and plot simulations of the point processes shown in Figures 3.7, 3.8 and 3.9.

```
library(spatstat}
X <- rpoispp(25)
plot(X)
X <- rpoispp(function(x, y) { exp( 2 + 5 * x) })
plot(X)
X <- rpoispp(100)
M <- sample(1:3, X$n, replace=TRUE)
plot(X %mark% M, main="n")
```

Figure 3.9: A marked point process with a finite mark space M={1, 1.5, 2, 2.5, 3}. The points are a realisation of an homogeneous Poisson process in the unit square, with intensity equal to 100.



4 Overview of statistical methods

In this thesis patients are clustered within hospitals. The outcome of interest is binary (i.g death or alive). Therefore, we are dealing with clustered binary data. Three leading methods for analysing clustered binary data include marginal models (GEE), random-effects models (GLMMs) and conditional models.
In this thesis GEE and GLMMs are used to provide estimates under the recurrent marked point process setting. Conditional models are not used here.
This chapter gives an overview of GEE (Section 4.2) and the GLMM (Section 4.2). In Section 4.1, the GLM is discussed since GEE and the GLMM may be viewed as extensions of a GLM.

4.1 Generalized linear models

Generalized estimating equations (GEE) are an extension of the generalized linear model (GLM), used to analyse longitudinal data. GLMs allow for situations with a non-normal error distribution, that cannot be handled with a linear model. Binary or count data lead to a non-normal error distribution due to the restricted range of possible outcomes. The GLM allows the linear model (i.e. linear predictor) to be related on the outcome via a link function.

Definition A generalized linear model consists of

- a *stochastic component*, specifying the conditional distribution of the outcome variable, Y_i (for the $i = 1, \ldots, n$ independent observations), given the values of the covariates X_{ip} in the model. Often, the outcome variable Y_i follows a distribution from the exponential family (e.g. Gaussian, binomial, Poisson).

- *linear predictor*,
$$\eta_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}.$$

  Two functions:

- a *link function*, transforming the expectation of the outcome variable, $E(Y_i) = \mu_i$, to the linear predictor
$$g(\mu_i) = \eta i,$$

- a *variance function* V that describes how the variance, $var(Y_i)$ depends on the mean
$$var(Y_i) = \phi V(\mu_i),$$

  with ϕ the so-called constant *overdispersion parameter*.

Various link functions can be used, such as the identity, logit, complementary log-log, or probit link function. The most common link function for binary data is the logit link, resulting in logistic regression, represented as

$$g(u_i) = logit(u_i) = log\left(\frac{u_i}{1-u_i}\right) = \eta_i.$$

The logit link transforms the range of the binary outcome variable (0,1) to a range of $(-\infty, +\infty)$ for the linear predictor. The interpretation of regression parameters is as log odds-ratios associated with a unit change in the covariate.
Equating the partial derivatives of the log likelihood with respect to $\beta_0, \beta_1, \ldots, \beta_p$, to zero and summing over all observations produces the set of GLM estimating equations, given by

$$\sum_{i=1}^{N}\left(\frac{\partial u_i}{\partial \beta}\right) var(Y_i)^{-1}(Y_i - u_i) = 0. \tag{3}$$

The maximum likelihood estimates are obtained by solving the set of equations in (3).

4.2 Generalized estimating equations

Longitudinal data analysis generally involves repeated measurements on the same subject over time. The dependence between observations on the same subjects must be taken into account. A GLM, however, ignores the dependent structure within subjects.
There are several approaches to extend generalized linear models to longitudinal data analysis. Examples include the generalized linear mixed model (GLMM) [20, McCulloch & Neuhaus], and generalized estimating equations (GEE) [18, Liang and Zeger]. The latter accounts for the correlation between observations on the same subject, or generally speaking, between members of the same cluster, through the use of a working correlation matrix and sandwich variance estimates.

Suppose that for each subject $i = 1, \ldots, n$ there are observations at time points t ($t = 1, \ldots, T$) with corresponding outcome $\boldsymbol{Y_i} = (Y_{i1}, \ldots, Y_{it}, \ldots, Y_{iT})$. Each individual i can be seen a cluster with T observations. In the GEE setting, the T observations within the same cluster (e.g. subject) are allowed to be correlated while measurements in different clusters are assumed to be independent.
The generalized estimating equations are derived without a full specification of the joint distribution of a cluster's observations Y_i. Instead a likelihood for the marginal distribution at each time point (i.e. Y_{it}) is specified, next to a *working* correlation matrix for the intra-cluster correlation.

A link function relates the expectation of the outcome $E(Y_{it}) = \mu_{it}$ to the linear predictor $g(\mu_i t) = \eta it$. A variance function

$$var(Y_{it}) = \phi V(\mu_{it})$$

is also specified. The covariance matrix $var(Y_{it})$ in the estimating equations is replaced by the working covariance matrix V_i of y_i given by

$$V_i = \phi A_i^{1/2} R_i(a) A_i^{1/2},$$

where ϕ is the overdispersion parameter and A_i is a diagonal matrix with entries $V(u_{it})$. $R_i(a)$ is a working model for the intra-cluster correlation of the $Y_{it}s$, possibly depending on the parameter vector a of length m. The number of measurements may vary across subjects, but the dependence structure R_i on a must be invariant. Note that the working covariance matrix V_i consists of a model for the intra-cluster correlation, next to a diagonal matrix with elements $var(Y_{it})$.
The generalized estimating equations are then given by

$$\sum_{i=1}^{N} \left(\frac{\partial u_i}{\partial \beta} \right) V_i^{-1} (Y_i - u_i) = 0. \tag{4}$$

The GEE approach is a marginal method and it is most suitable when the emphasis is on the marginal means in relation to the regression parameters rather than in the intra-cluster correlation structure.
The main advantage of GEE is that even under misspecification of the correlation structure, the model gives consistent estimates of the regression parameters and their estimated standard errors [12, Ghisletta and Spini], [13, Halekoh *et al.*], [29, Zeger *et al.*]. For this reason the specified intra-cluster correlation structure is referred to as working correlation. Correct specification of the correlation structure improves efficiency and leads to smaller standard errors. Two common choices are an *independent* correlation structure ($m = 3$),

$$R_i = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and an *exchangeable* correlation structure ($m = 3$),

$$R_i(a) = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix}.$$

With the independence working correlation matrix, the observations within the same cluster $Y_{i1}, \ldots, Y_{iT}$ are assumed to be independent. The estimating equations under an independent correlation structure are called independence estimating equations (IEE). If working independence is assumed, V_i in (4) is a diagonal matrix. This implies that IEE do not fall under covariance-weighting methods.
The exchangeable correlation structure assumes constant time dependency, with all off-diagonal elements being equal to a.
The GLM is identical to GEE if the working correlation matrix is specified as independent and if the model-based standard error estimator is chosen [12]. The model-based standard error is one of the two variance estimators offered by the GEE approach. The other variance estimator is generally called *robust* or *sandwich* estimator and it is robust to misspecification of the working correlation.

4.3 Generalized linear mixed models

The generalized linear mixed model (GLMM) is an extension of the GLM in which the linear predictor additionally contains cluster-specific random effects, providing inference specific to each cluster. The random effects are assumed to have a (multivariate) normal distribution. The extension of the linear predictor with a random cluster-specific intercept γ_{i0} follows

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \gamma_{i0} \text{ with } \gamma_{i0} \sim N(0, \sigma^2).$$

Next to a random intercept one or more random slopes may be included in the model. Random effects have a (multivariate) normal distribution. The vector of random effects $\boldsymbol{\gamma}$ is distributed as $\boldsymbol{\gamma} \sim N(0, G)$ with

$$G = \begin{pmatrix} \sigma^2_{int} & \sigma^2_{int,slope} \\ \sigma^2_{int,slope} & \sigma^2_{slope} \end{pmatrix},$$

for the situation with a random intercept and one random slope. Random effects have mean equal to zero because they are modelled as deviations from the fixed effects.
Alternatively, the GLMM may be viewed as an extension of the linear mixed model, allowing response variables from different distributions. Inference in GLMMs is based on the standard likelihood method. The likelihood involves integration over the random effects distribution which may be numerically very difficult.
The regression coefficients for the fixed effects of a GLMM measure the change in expected value of the response while holding constant other covariates and the random effects.

5 Application of a recurrent marked point process

The first goal of this thesis is to employ the appropriate statistical methodology to investigate the volume-outcome associations between hospital volume and patient outcome after oesophageal cancer surgery. This chapter describes why and how the recurrent marked point process may be applied for this purpose.
Volume-outcome analysis requires the specification of a measure for hospital volume. Statistical issues involved in the selection of measures for hospital volume are discussed in Section 5.2.1. Information about the statistical model and mathematical notation is given in Section 5.2.2. In Section 5.3, results concerning the volume-outcome analyses are presented.

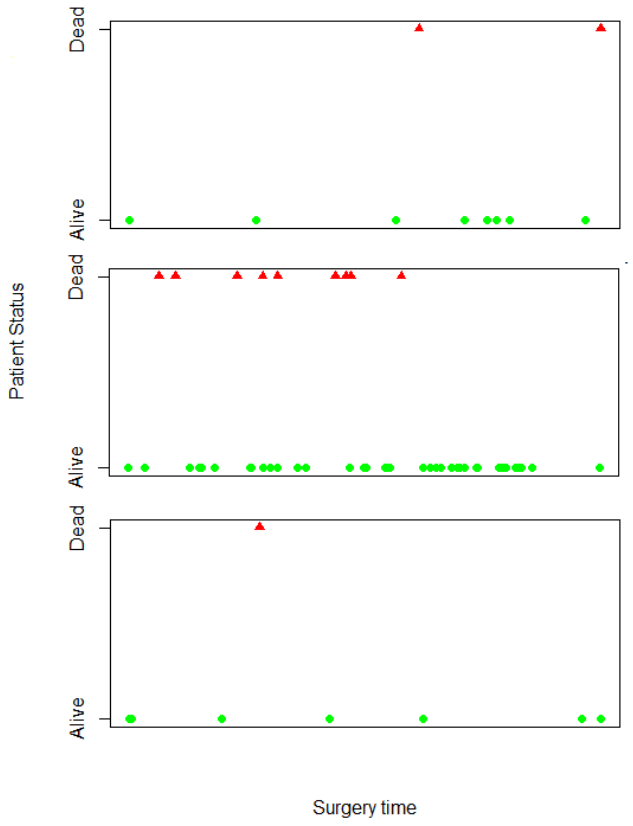
5.1 Why a recurrent marked point process in our situation?

In this thesis the scientific interest lies in the association between hospital volume and patient outcome after oesophageal cancer surgery. For this purpose, a model that describes post-treatment outcome as a function of hospital volume and accounts for dependence between patients is proposed. Next to hospital volume, some covariates are included in the model, in order to adjust for patient's risk and other characteristics.
French *et al.* [9] proposed the recurrent marked point process as a general framework for estimating volume-outcome associations from longitudinal data. This process may be a suitable approach since patient outcome (e.g. dead or alive) are only observed for patients that underwent surgery. This means that marks are only observed at point locations.
Since an outcome only exists when an event of surgery takes place, measurement times differ considerably between hospitals. A recurrent marked point process can cope with different time points, whereas observations in traditional longitudinal data analysis are usually at fixed time points.

Use of a recurrent marked point process enables us to interpret the case under study as a point process in one dimension (i.e. time). In this context, the surgeries can be seen as 'points' and the corresponding point locations capture information about time the surgery is performed.
The intensity measure defined in (1) is the average number of surgeries for a certain hospital per unit of time interval. In this context the outcome of interest (i.e. mark) is death from any cause within 6 months since surgery, giving a binary mark space M={Alive, Dead}. In this thesis the focus is on marks rather than point locations. Figure 5.1 represents marked point processes for three specific hospitals in the population under study.

Regression methods such as GEE and GLMM can be used to provide estimates under the recurrent marked point process setting by taking into account the dependence within hospitals. However, an assumption of independence between previous outcome and future

Figure 5.1: Marked point process for three specific hospitals in the population under study with mark space M={Alive, Dead} for patient outcome after 6 months since surgery.



number of events is required for covariance weighted methods to provide unbiased estimates.

5.2 Fitting a recurrent marked point process model

5.2.1 Hospital volumes

In this thesis different measures for hospital volume are considered, all providing measures for the number of oesophagectomies per hospital. Both non-aggregate and yearly aggregate measures are used. Aggregate measures are available at each surgery time, whereas yearly aggregate measures are only available at the end of each year.
In Section 3.1 a region specific counting measure was introduced, denoting the number of points falling in each borel set B. Hospital volume may be compared with a region specific counting measure, where each hospital forms a borel set. In case of yearly aggregate measures, each year forms a borel set. In the sequel it is shown how hospital id and time

of surgery are used to calculate measures for hospital volume.

Non-aggregate specification for hospital volume. In this thesis non-aggregate hospital volume is defined as the cumulative number of surgeries performed at hospital i through time t and may be represented as

$$N_{0i}(t) = \sum_{s=1}^{t} \triangle N_i(s). \tag{5}$$

Yearly aggregate specifications for hospital volume. Three different aggregate specifications of hospital volume are used, all of them as yearly aggregate measures:

1. Yearly total volume (6)
2. Cumulative yearly total volume (7)
3. Running average volume (8)

Yearly total volume is a present measure for hospital volume, denoting the total number of surgeries per hospital per year, given by

$$N_{1i}(j) = N_i(T_j) - N_i(T_{j-1}) \tag{6}$$

for hospital i ($i = 1, \ldots, n$), year j ($j = 1, \ldots, J$), where T_j denotes the end of each year. Yearly total volume is an appropriate measure for present experience.

Cumulative yearly total volume is defined as the cumulative sum of surgeries per hospital including the current year. It is calculated as

$$N_{2i}(T_j) = \sum_{s=1}^{Tj} dN_i(s), \tag{7}$$

for hospital i and year j. A cumulative measure for hospital volume reflects hospital size or experience accumulated over all study period.

Running average volume is an average including the cumulative number of surgeries through the previous year. It may be represented as

$$N_{3i}(j) = \frac{N_{2i}(T_{j-1}) + N_{2i}(T_j)}{2}. \tag{8}$$

Figure 5.2 shows non-aggregate cumulative volume, yearly total volume, cumulative yearly total volume and running average volume over time for a specific hospital in the population under study. Note that aggregate volume measures stay the same throughout a year, giving small horizontal stripes. It can be seen from Figure 5.2 that cumulative yearly total volume and running average volume are quite good approximations of non-aggregate cumulative volume.
Figure 5.3 shows yearly total volume as a single measure over time for three specific hospitals in the population under study.

Figure 5.2: Non-aggregate cumulative volume, yearly total volume, cumulative yearly total volume and running average volume over time for a specific hospital in the population.

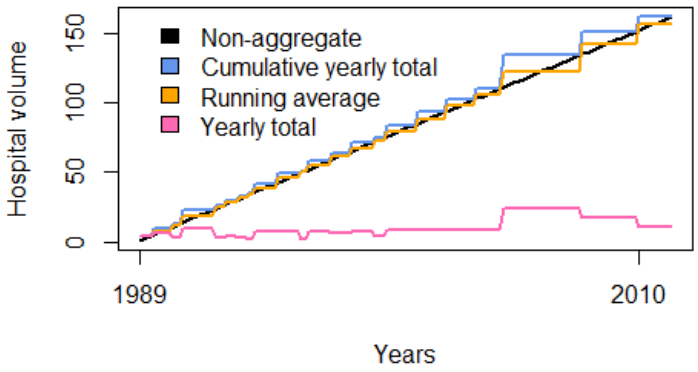


Figure 5.3: Yearly total volume over time for three specific hospitals in the population.

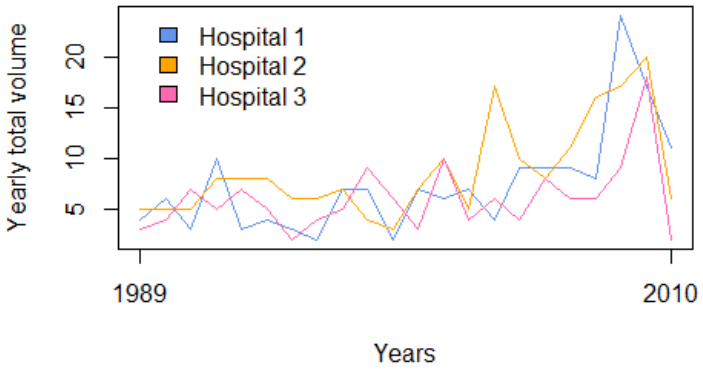
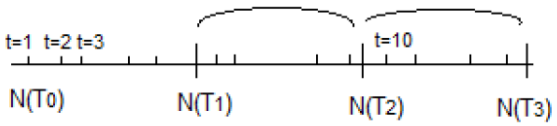


Figure 5.3 shows yearly total volume as a single measure over time for three specific hospitals in the population under study. Yearly aggregate measures for hospital volume for a patient with surgery at time t may be biased when hospital volume before time t is considerably different from hospital volume after t. However, when hospital volume is roughly constant within a year, yearly aggregated measures may represent hospital volume correctly. Cumulative total volume is less sensitive for the potential bias due to the use of aggregate measures than yearly total volume since the former is based on all years of study, whereas the latter is based only on the current year.
Figure 5.4 provides an illustration of the difference between non-aggregate and aggregate measures for hospital volume. Non-aggregate specifications are available at a fine grid of time points; at each surgery time, whereas yearly aggregate specifications are only available at the end of each year T_j.
Note that hospital volume is indeed a time-dependent covariate, changing during the 3 calender years. As Figure 5.4 shows, there are considerable differences between the three measures for hospital volume. Volume specification therefore is a primary challenge of a volume-outcome study [9].

Figure 5.4: Yearly total volume, cumulative yearly total volume and running average volume for an hypothetical hospital at at time t where $t = 10, \ldots, 13$.



Time t	1 2 3 4 5	6 7 8 9	10	11	12	13
Year	1	2	3			
Volume specification	*Non-aggregate*	Total	10	11	12	13
	Aggregate					
	Present	Yearly total	4	4	4	4
	Cumulative	Cumulative yearly total	13	13	13	13
		Running average	11	11	11	11

The R code shown below is used to calculate the three alternative aggregate measures for hospital volume. The function `count` from the package **plyr** is used first to obtain counts for each combination of hospital id and year of surgery. The package **caTools** is needed for the calculation of the running average.

```
library(plyr)
library(caTools)

# Yearly total
temp1 <- count(data.order ,c("surgHospital", "surgYear"))
head(temp1)
  surgHospital surgYear freq
1            1     1989    4
2            1     1990    6
3            1     1991    3
4            1     1992   10
5            1     1993    3
6            1     1994    4

yearly.tot <- rep(temp1$freq, temp1$freq)
```

```
head(yearly.tot,10)
     4 4 4 4 6 6 6 6 6 6

# Cumulative yearly total
temp2 <- tapply(temp1$freq, temp1$surgHospital, cumsum)
temp3 <- unname(unlist(temp2))
cum.yearly.tot <- rep(temp3, temp1$freq)
head(cum.yearly.tot, 10)
     4  4  4  4 10 10 10 10 10 10

# Running average
temp4 <- count(data.order, c("surgHospital", "surgYear",
                             "cum.yearly.tot")
temp5 <- tapply(temp4$cum.yearly.tot, temp4$surgHospital, runmean, k=2)
temp6 <- unname(unlist(temp5))
run.ave <- rep(temp6, temp4$freq)
head(run.ave, 10)
     4 4 4 4 7 7 7 7 7 7
```

5.2.2 Statistical model and notation

Let $Y_i(t)$, $N_i(t)$ denote patient outcome (i.e. mark) and hospital volume, respectively, for hospital i ($i = 1, \ldots, n$) at time point $t = 1, \ldots, T$. For the specification of non-aggregate hospital volume, see (5) and for the specification of yearly aggregate measures of hospital volume $N_{1i}(j)$, $N_{2i}(j)$ or $N_{3i}(j)$, see (6), (7), and (8) respectively.
Let $X_i(t)$ be a vector of patient-level covariates at time point $t = 1, \ldots, T$ and hospital i.

As introduced in Section 3.2, according to a marked point process approach a surgery event must occur for an outcome to exist. A surgery which took place at time point t is denoted by $\triangle N_i(t) = N_i(t) - N_i(t-1) = 1$.

Let

$$\begin{aligned}\mathcal{X}_i(t) &= \{X_i(s) | s \leq t\},\\ \mathcal{N}_i(t) &= \{N_i(s) | s \leq t\},\\ \mathcal{Y}_i(t) &= \{Y_i(s) | s \leq t\}.\end{aligned}$$

denote the complete history of each variable.

Statistical model. The first aim of this thesis is to study the association between hospital volume $N_i(t)$ and post-treatment outcome $Y_i(t)$ among patients undergoing surgery. For this purpose, a marginal regression model is fitted, which describes the association between hospital volume and the *average* outcome among patients satisfying the criteria $\triangle N_i(t) = 1$, given by

$$\mu_i(t) = \quad E[Y_i(t) \mid \triangle N_i(t) = 1,\ \mathcal{X}_i(t),\ \mathcal{N}_i(t)] = x_{it}\beta. \tag{9}$$

The expectation of $Y_i(t)$ is modelled conditional on a relevant subset of the covariate and event-time process histories up to time t. The marginal model in (9) may also be called a partly conditional regression model because it does not always condition on the complete covariate history or on the entire event-time process until time t, and it does not condition on past outcomes [23, Pepe and Cooper].
In this thesis the partly conditional model is used, quantifying the marginal association between the complete history of the event-time process and the mark process after adjusting for a full history of the covariate process. In the fitted mean model for $Y_i(t)$ a proper link function must be used because the outcome is binary (i.e. death or alive). This may be represented as

$$\begin{aligned} &E[Y_i(t) \mid \triangle N_i(t) = 1,\ \mathcal{X}_i(t),\ \mathcal{N}_i(t)] \\ &\quad = g^{-1}(\beta_0 + \beta_1 X_i(t) + \beta_2 N_i(t)). \end{aligned} \tag{10}$$

The parameters β_1 and β_2 quantify the association between the covariate and event-time processes and the average outcome among patients for whom an event of surgery takes place. To take into account the dependent nature of the data, the model in (19) is fitted by IEE and GEE by assuming an exchangeable correlation structure. Both IEE and GEE are fitted by using the package **gee**.
A GLMM with hospital specific random intercepts (GLMM-RI), and a GLMM with hospital specific random intercepts and slopes for hospital volume (GLMM-RS) are fitted by using the package **lme4**. The latter model is defined as

$$\begin{aligned} &E[Y_i(t) \mid \triangle N_i(t) = 1,\ \mathcal{X}_i(t),\ \mathcal{N}_i(t)] \\ = \quad &g^{-1}(\beta_0 + \beta_1 X_i(t) + \beta_2 N_i(t) + \gamma_{i0} + \gamma_{i1} N_i(t)). \end{aligned} \tag{11}$$

Assumptions of independence. To ensure consistency of the GEE estimator, the GLMM-RI and the GLMM-RS, two assumptions should hold for all $t' > t$, given by

Assumption 1′.

$$Y_i(t) \perp N_i(t') \mid \triangle N_i(t) = 1,\ \mathcal{X}_i(t),\ \mathcal{N}_i(t),$$

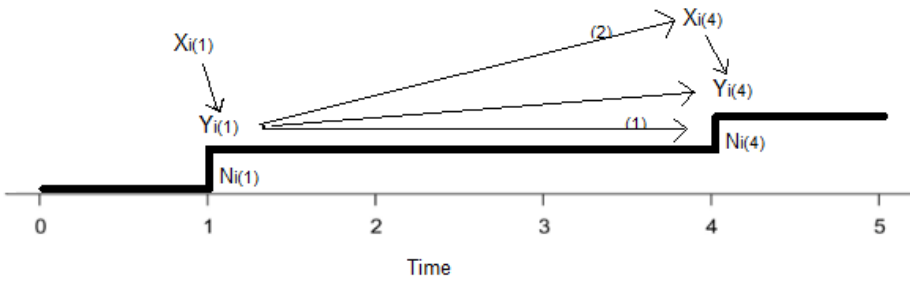
Assumption 2′.

$$Y_i(t) \perp X_i(t') \mid \triangle N_i(t) = 1,\ \mathcal{X}_i(t),\ \mathcal{N}_i(t).$$

Assumption (1′) implies independence between previous patient-outcome and future number of events. It provides an indirect way to test for *informative cluster size* (see Section 6.4). Assumption (2′) implies independence between previous patient-outcome and current patient's exposure. If one of these assumptions is not met, then IEE is the only estimating equation option that provides an unbiased estimator for β. This is a consequence of the diagonal working covariance matrix V_i, involved in independence estimating equations. More technical details can be found in [10]. The robustness of GEEs against misspecification of the correlation structure no longer holds when one of the assumptions (1′) and (2′) is violated; an independence working covariance matrix is then required.

Figure 5.5 describes the recurrent marked point process for an hypothetical hospital i. It can be observed that an outcome is only observed when an event occurs, that is, when the criteria $\triangle N_i(t) = 1$ is satisfied. The arrows between patient-level exposure $X_i(1)$ and patient outcome $Y_i(1)$, and between $X_i(4)$ and $Y_i(4)$ represent the cross-sectional associations of interest. The remaining arrows in Figure 5.5 represents relations that may lead to a bias for the estimation of interest. The association between $Y_i(1)$ and $Y_i(4)$ describes the possible correlation between observations within the same hospital. Arrows (1) and (2) represent violation of assumption (1′) and (2′) respectively [10].

Figure 5.5: Underlying framework for a recurrent marked point process for an hypothetical hospital.



5.3 Results

Table 2 shows estimated associations between different measures for hospital volume and the odds-ratio of dying within 6 months since surgery for ten-patient increase. Associations are obtained using a non-aggregate measure for cumulative total volume, an yearly aggregate measure for present hospital volume and two different yearly aggregate measures for cumulative hospital volume.

Non-aggregate cumulative volume. Results obtained with IEE indicate 1,56% significant decrease in the odds of 6-month patient mortality for a ten-patient increase in non-aggregate cumulative hospital volume. Results obtained with GEE and GLMM-RI show absence and a very weak association respectively. Estimates obtained with GLMM-RS indicate the strongest association, which is equal to a 4,19% significant decrease in the odds of 6-month patient mortality for a ten-patient increase in non-aggregate cumulative hospital volume.

Yearly total. Results obtained with IEE indicate a 15,35% significant decrease in the odds of 6-month patient mortality for a ten-patient increase in yearly total volume. Estimates obtained with IEE show the strongest volume-outcome association, followed by GLMM-RS. All estimated volume-outcome associations for yearly total volume are significant at the 5% level.

Cumulative yearly total. Results obtained with GLMM-RS indicate the strongest association, which is equal to a 4,34% decrease in the odds of 6-month patient mortality for

a ten-patient increase in cumulative yearly total volume. All estimated volume-outcome associations for cumulative yearly total volume are significant at the 5% level.

Running average. The estimated volume-outcome associations obtained using a running average are comparable, but slightly weaker, to those obtained using cumulative yearly total volume. Estimated volume-outcome associations for running average volume obtained with IEE and the GLMM-RS are significant at the 5% level.

Figures 5.6, 5.7 and 5.8 show estimated odds-ratios and their corresponding 95% CI intervals using different estimation methods.

Note that results based on IEE and GEE are population-averaged parameters, quantifying the average volume-outcome associations among the whole population of patients. Results based on GLMM-RI and GLMM-RI are hospital-specific, quantifying the average volume-outcome associations among a population of hospitals.

For each estimation method and measure of hospital volume, the association between hospital volume and the odds of 6-month patient mortality is negative. Many associations are significant, indicating that hospital volume may indeed have a relevant impact on post-treatment outcome. However, estimated associations are different for the four estimation methods used. One explanation might be that the parameters obtained with IEE and GEE, are population-averaged, whereas parameters obtained with the GLMM-RI and GLMM-RS are not. However, differences in estimations are also observed between IEE and GEE. These discrepancies in results might be the consequence of bias, induced in the estimation process because of violation of one or both of the assumptions stated in Section 5.2.2.
Results obtained for non-aggregate cumulative volume and yearly aggregate measures for cumulative volume and running average volume, are quite comparable. Estimated associations for yearly total volume show a much stronger association since they quantify a ten-patient increase in the number of surgeries per year, instead of a ten-patient increase in the number of surgeries performed over all years of study.

Table 2: Estimated associations between hospital volume and odds of dying within 6 months since oesophageal cancer surgery. Results correspond to ten-patient increase in different measures of hospital volume.

Specification of hospital volume	Model	OR	95% CI (based on robust S.E.)
Non-aggregate			
Cumulative total volume			
	IEE	0.9844	(0.9796-0.9893)
	GEE	0.9969	(0.9936-1.0002)
	GLMM-RI	0.9935	(0.9872-0.9999)
	GLMM-RS	0.9580	(0.9452-0.9710)
Aggregate			
Yearly total volume			
	IEE	0.8465	(0.8262-0.8673)
	GEE	0.8564	(0.8290-0.8847)
	GLMM-RI	0.8676	(0.8042-0.9358)
	GLMM-RS	0.8562	(0.8182-0.8959)
Cumulative yearly total volume			
	IEE	0.9849	(0.9803-0.9895)
	GEE	0.9964	(0.9931-0.9996)
	GLMM-RI	0.9935	(0.9873-0.9996)
	GLMM-RS	0.9566	(0.9427-0.9708)
Running average volume			
	IEE	0.9845	(0.9796-0.9894)
	GEE	0.9971	(0.9938-1.0004)
	GLMM-RI	0.9936	(0.9873-1.0000)
	GLMM-RS	0.9544	(0.9401-0.9689)

Figure 5.6: Odds-ratios and their corresponding 95% CI intervals for different estimation methods using a non-aggregate measure of cumulative hospital volume.

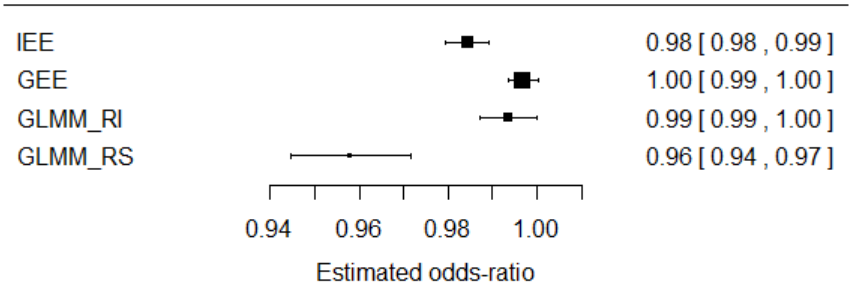


Figure 5.7: Odds-ratios and their corresponding 95% CI intervals for different estimation methods using an aggregate measure of present hospital volume: yearly total.

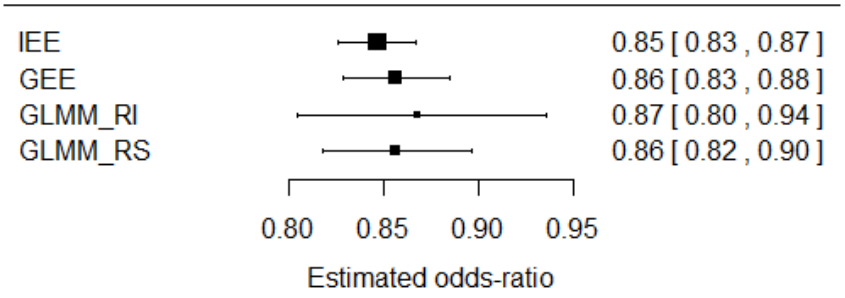
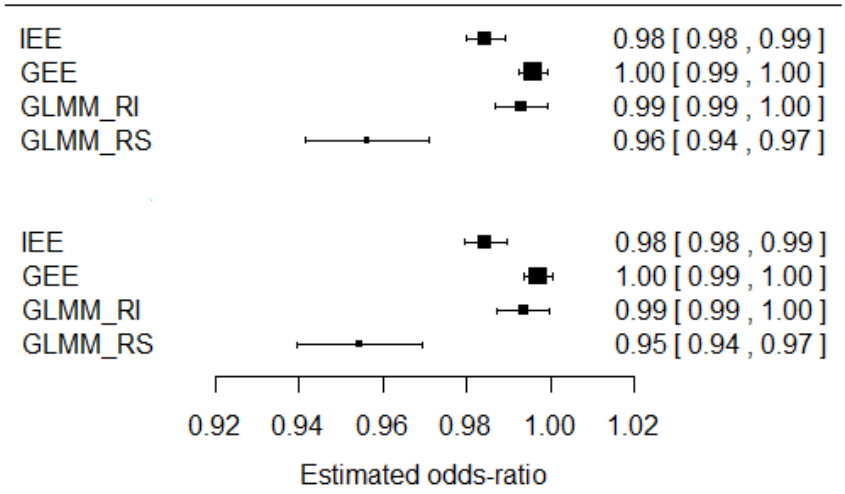


Figure 5.8: Odds-ratios and their corresponding 95% CI intervals for different estimation methods using aggregate measures of cumulative hospital volume. The first four rows in the upper part of the table denote results for cumulative yearly total volume, the last four rows denote results for running average volume.



6 Informative cluster size

As stated before, patients are clustered within hospitals and the dependency among the observations within the same hospital must be taken into account. This chapter introduces an additional issue sometimes associated with clustered data, which is generally referred to as *informative cluster size*. This chapter is organised as follows. In Section 6.1, a definition of informative cluster size is given. Section 6.2 illustrates current methods for marginal inference under informative cluster size. Section 6.3 describes problems erasing when cluster sizes are informative. In Section 6.4 it is examined whether informative cluster size is present in the dataset under study.

6.1 Definition

Informative cluster size has been defined to arise when the outcome of interest is related to the expected number of observations within a cluster. For example, the amount of visits to a general practitioner may give information about how sick a patient feels.
Also, informative cluster sizes may be found for example in a study investigating the relation between maternal cigarette smoking and spontaneous abortion [15]. The cluster of interest is then the set of pregnancy outcomes of a woman. The cluster size will be related to risk, because women at high risk of spontaneous abortion need to have more pregnancies on average to achieve their desired family size.
Another example concerns a study investigating school achievements in different school classes. It may be reasonable to assume that smaller school classes are associated with higher academic achievements. In that case, cluster size will be related to the outcome of interest.
Different but closely related definitions of informative cluster size can be found in literature. [4, Benhin *et al.*], [15, Hoffman *et al.*] and [28, Williamson *et al.*] consider marginal models and use the following definition.

Definition Cluster size is said to be informative if

$$E[Y|N, X] \neq E[Y|X]$$

where Y, X and N respectively denote the outcome, covariates and cluster size.

Cluster size is informative if the conditional expected value of the outcome given the covariates and the cluster size depends on the cluster size.

6.2 Marginal methods for informative cluster size: current methodology

GLMMs and GEE are commonly used to analyse clustered data; GLMMs for cluster-specific inference and GEE for marginal, population-averaged inference. In general, these

methods assume that cluster size is uninformative
This section gives an overview of the current methods concerning marginal inference under informative cluster size: within cluster resampling (WCR) [15, Hoffman *et al.*] and cluster weighted generalized estimating equations (CWGEE) [28, Williamson *et al.*]. Next to marginal inference, [8, Dunson *et al.*] developed a Bayesian approach based on joint modelling the cluster size and the outcome of interest to provide cluster-specific inference. In this thesis only estimating methods for marginal inference under informative cluster size are considered.
In this chapter a slightly different notation is used. Subscripts i ($i = 1, \ldots, I$) and t ($t = 1, \ldots, T_i$) denote cluster and cluster member (i.e. time point of measurements within each cluster), respectively.

6.2.1 Marginal inference: within cluster resampling

Hoffman *et al.* [15] proposed within cluster resampling (WCR), which is a Monte Carlo approach for fitting models with clustered data. WCR produces cluster-based parameters by equally weighting each cluster. This is in contrast to GEE, which equally weights each observation (see Section 6.3).
As the name indicates, within cluster resampling, randomly samples (with replacement) one observation $Y_i(t)$ from each cluster i ($i = 1, \ldots, I$ and $t = 1, \ldots, T_i$). The sampling procedure is repeated, a large number Q, times. The resulting Q resampled datasets contain I independent observations (one observation from each cluster).
Each resampled dataset can be analysed by using any marginal analysis (e.g. a GLM) since the I observations are independent, yielding to the resampling parameters $\widehat{\beta_Q}$.
The WCR regression parameter is obtained by taking the average of the resampling parameters and may be defined as

$$\widehat{\beta_{WCR}} = \frac{1}{Q} \sum_{q=1}^{Q} \widehat{\beta}_q. \tag{12}$$

A consistent estimate of the WCR variance is

$$\widehat{\Sigma} = \widehat{Var}(\widehat{\beta_{WCR}}) = \frac{\sum_{q=1}^{Q} \widehat{\Sigma}(q)}{Q} - \left(\frac{Q-1}{Q}\right) S_{\beta}^2, \tag{13}$$

where $\widehat{\Sigma}(q)$ is the (sandwich) variance estimator of $\widehat{\beta}_q$ (or estimated covariance matrix from the the q^{th} analysis) and S_{β}^2 is the covariance matrix among the Q resample based estimates $\widehat{\beta_Q}'s$ given as

$$S^2_\beta = \frac{\sum_{q=1}^{Q}(\widehat{\beta}_q - \widehat{\beta_{WCR}})(\widehat{\beta}_q - \widehat{\beta_{WCR}})'}{Q-1}. \tag{14}$$

Hoffman *et al.* [15] proved that as $I \to \inf$, $I^{\frac{1}{2}}(\widehat{\beta_Q} - \beta) \to N(0, \Sigma)$ in distribution, where Σ is finite and positive-definite. Figure 6.1 represents the WCR resampling scheme.
Due to the one-per-cluster sampling scheme of the procedure, interpretation of the WCR parameter is cluster-based (not cluster-specific). The parameter reflects the population-averaged difference associated with a unit change in a covariate corresponding to a randomly selected observation from a randomly selected cluster. The interpretation of the WCR parameter in this thesis is based on logistic models; it describes the marginal difference in the log of the odds of dying within 6 months since surgery between a *randomly selected patient* with hospital volume x from a *random hospital* versus a *randomly selected patient* with hospital volume x+1 from a *random hospital*.
WCR is computationally intensive, but it guarantees the consistency and asymptotic normality of the cluster-based marginal parameters. An appealing feature of the method is that it accounts for the within cluster correlation without specifying a working correlation matrix. However, problems may erase when the number of clusters I is small. Individual estimators $\widehat{\beta}_q$ are based on a I observations (i.e. one from each cluster) and might be unstable when $I < p$ is small, with p denoting the number of variables to be estimated. In this specific situation it may happen that the resulting WCR variance $\widehat{\Sigma}$ is not positive-definite. As a consequence variance estimates can be negative. Such a scenario suggests that the number of clusters I does not guarantee the asymptotic approximation to hold [15].

6.2.2 Marginal inference: cluster weighted generalized estimating equations

WCR is computationally intensive. Cluster weighted generalized estimating equations (CWGEE), proposed in [28], or the same concept under the name mean estimating equations proposed in [4], provide an estimator that is asymptotically equivalent to WCR as $Q \to \infty$ but avoids the Monte Carlo element of WCR. The CWGEE are given as

$$\sum_{i=1}^{I} \frac{1}{T_i} \sum_{t=1}^{T_i} \mu_{it}(\beta) = 0. \tag{15}$$

Thus, instead of solving Q estimating equations separately, CWGEE combines them into a single equation in which each cluster i is weighted by the inverse of its cluster size T_i. Hence the name cluster weighted generalized estimating equations. The CWGEE parameters are iteratively estimated.

Figure 6.1: WCR resampling scheme

Consider a hypothetical data set containing I clusters...

$$Y_1 = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \end{bmatrix} \quad Y_2 = \begin{bmatrix} Y_{21} \\ Y_{22} \end{bmatrix} \quad Y_3 = \begin{bmatrix} Y_{31} \end{bmatrix} \quad Y_4 = \begin{bmatrix} Y_{41} \\ Y_{42} \\ Y_{43} \\ Y_{44} \\ Y_{45} \end{bmatrix} \quad Y_I = \begin{bmatrix} Y_{I1} \\ Y_{I2} \end{bmatrix}$$

$$R(1) = \begin{bmatrix} Y_{11}, x_{11} \\ Y_{22}, x_{22} \\ Y_{31}, x_{31} \\ Y_{44}, x_{44} \\ \dots \\ Y_{I1}, x_{I1} \end{bmatrix} \; GLM -> \hat{\beta}_1 \; \dots \; R(Q) = \begin{bmatrix} Y_{12}, x_{12} \\ Y_{22}, x_{22} \\ Y_{31}, x_{31} \\ Y_{42}, x_{42} \\ \dots \\ Y_{I2}, x_{I2} \end{bmatrix} \; GLM -> \hat{\beta}_q$$

The WCR regression parameter $\hat{\beta_{WCR}}$ is obtained by averaging the resample-based estimators accross the resamples...

$$\widehat{\beta}_{WCR} = \frac{1}{Q} \sum_{q=1}^{Q} \hat{\beta}_q$$

6.3 Problems associated with informative cluster size

Generalized estimating equations (GEE) assume that cluster size is unrelated to the outcome of interest and provide biased estimates in case this assumption is violated [15], [28]. This is because the working covariance structure chosen within GEE may change the weights given to each cluster. Among the GEE-based methods there is one exception; IEE still provide unbiased marginal estimates when cluster size is informative [28] due to the independence working correlation structure specified within IEE.

In volume-outcome studies, the concept of informative cluster size is more complicated since cluster size is or is closely linked to the covariate of primary interest, whereas generally, cluster size is regarded as a nuisance variable. It should be noted that a clear and concise definition of informative cluster size in volume-outcomes studies is lacking.

In this thesis hospital volume is closely linked but not equivalent to cluster size. Hospital volume $N(t)$ changes over the coarse of the study, whereas cluster size N is fixed throughout the study period. Cluster sizes denotes the number of patients (i.e surgeries) within

each hospital accumulated over all years of study and therefore is only equal to hospital volume at the end of the study period. As a consequence, including hospital volume in the model might not be enough to capture the relationship between outcome and cluster size. If there is any residual relation between outcome and cluster size that is not captured by the model by the inclusion of hospital volume, cluster size is informative.
Although the IEE approach might still provide consistent estimates when informative cluster size is present, depending on the situation, there may be some limitations to its use. It may be inefficient relative to a covariance-weighted method. Furthermore, IEE estimates are observation-based, whereas cluster-based parameters might better address the scientific question of interest.
IEE and GEE are marginal methods, providing population-averaged parameters. Their parameters should be interpreted as the population-averaged differences in the log of the odds of 6-month patient mortality corresponding to a unit increase in the covariate. In this thesis, this reflects the association between hospital volume and post treatment outcome for a *randomly selected patient* among the whole population of patients. This means that the observation-based parameter is averaged across clusters and across observations.
Cluster-based marginal parameters on the other hand, would reflect the volume-outcome associations of interest for a *randomly selected patient* from a *randomly selected hospital.*
Observation-based and cluster-based parameters use different schemes of weighting. The former equally weighs observations, whereas the latter equally weighs clusters. When cluster size is unrelated to outcome, both parameters coincide. However, when cluster size is informative the estimated parameters are different since an observation-based parameter gives greater weight to the outcome associated with large cluster size since large clusters consist of more observations.
Consider an hypothetical dataset, reporting the number of children within a school class that have to repeat the school year in rural versus non-rural areas (see Table 3). In Table 3 the first row represents the number of students who have to repeat the school year, the second row denotes the size of the school class. The observation-based parameter obtained from IEE represents the risk for a *randomly selected student* to repeat the school year. For the *rural* group from the hypothetical data in Table 3 this probability is calculated as

$$\begin{aligned}
\widehat{P}_{(O-B)}(\text{repeating year}) &= \frac{\text{total number of students repeating a class}}{\text{total number of students}} \\
&= \frac{0+1+0+1+3+0+0+2+1+0}{18+25+23+29+31+21+22+30+27+25} \\
&= 0.032
\end{aligned}$$

The cluster-based, marginal risk denotes the probability that a *randomly selected student* from a *randomly selected class* has to repeat the school year. This risk is estimated by averaging the class-specific probabilities of repeating the school year. For the *rural* group from the data in Table 3 the cluster-based, marginal risk, is calculated as

$$\begin{aligned}\widehat{P}_{(C-B)}(\text{repeating year}) &= \frac{\text{sum of class-specific risk}}{\text{number of classes}} \\ &= \frac{\frac{0}{18}+\frac{1}{25}+\frac{0}{23}+\frac{1}{29}+\frac{3}{31}+\frac{0}{22}+\frac{0}{21}+\frac{2}{30}+\frac{1}{27}+\frac{0}{25}}{10} \\ &= 0.027\end{aligned}$$

Cluster sizes are informative in this toy dataset; the risk of repeating the school year is higher for students within large classes. The observation-based parameter gives larger weight to large classes and therefore produces a higher estimate for the risk of repeating a class.
In case the target of inference is the population of all cluster members, an observation-based IEE will provide the desired parameter estimates. However, when the target of inference denotes a population of randomly selected cluster members and informative cluster sizes are present, a cluster-based parameter must be used in order to provide valid estimates.
In the next section, the presence of informative cluster size in the data under study will be investigated.

Table 3: Hypothetical dataset regarding school classes in rural versus non-rural areas and the number of students that have to repeat a class.

Rural	0	1	0	1	3	0	0	2	1	0
	18	25	23	29	31	21	22	30	27	25
Non-rural	0	2	1	1	3	1	0	0	2	1
	25	30	29	30	32	29	23	28	34	28

6.4 Validating assumption (1′)

Recall that assumption (1′) implies independence between patient-outcome and future number of events,

$$Y_i(t) \perp N_i(t') \mid \triangle N_i(t) = 1,\ \mathcal{X}_i(t),\ \mathcal{N}_i(t).$$

The estimates provided in Table 2 show quite some discrepancies. While both IEE and GEE estimates should provide the population-averaged decrease in the odds of 6-month patient mortality for a ten-patient increase in hospital volume, their parameter estimates do not coincide. These differences between estimates may exist because one or both assumptions might not be satisfied.
In this thesis only assumption (1′) is investigated, which implies independence between patient outcome and future patient covariates. Validating assumption (2′) is difficult in

this context since none of the covariates is constant over the coarse of the study. This is because covariates $X_i(t)$ are specified at patient-level, belonging to a patient corresponding to time point t at hospital i.
As stated in the previous section, including hospital volume in the model might not be enough to capture the relationship between outcome and cluster size. Because assumption $(1')$ implies independence between patient-outcome and occurrence of a subsequent event given hospital volume, it provides an indirect mechanism to test whether there is still some informative cluster size that is not captured by the model through the inclusion of hospital volume.
A Cox regression model is used for the validation of assumption $(1')$. A Cox model regresses the hazard rate on several explanatory variables. The hazard rate function is the rate that an individual, who has not experienced the event of interest (e.g. death) until time x, will experience the event in the next interval of time. The hazard rate function is given by

$$h(x) = \lim_{x \to 0} \frac{P[x \leq X < x + \triangle x \mid X \geq x]}{\triangle x}, \tag{16}$$

with $X \geq 0$ a random variable, representing the time to the event of interest.
In order to validate assumption $(1')$ a Cox model is fitted by regressing time until a next surgery on patient outcome and adjusting for hospital volume and patients' characteristics (see Table 1). Hospitals are defined as clusters. This gives an estimated hazard rate equal to 0.8798, 95% CI (0.8755, 0.8840) indicating that the hazard of a subsequent surgery among hospitals with a previous death are 12% lower than among hospitals without a previous death. This difference is statistically significant (p=0.00). There is therefore evidence that assumption $(1')$ is violated, implying that there is still some informative cluster size that is not captured by the inclusion of hospital volume in the model. In the R code below it is shown how the dataset is prepared to estimate a Cox model. A new column concerning time between successive surgeries has to be added to the dataset. This information should be set equal to NA (non-available) for the last surgery at each hospital.

```
#########################
# Evaluating assumption (1)
#########################

# Preparing new dataset
# myind.hospital indicates when we move to different hospital
myind.hospital <- (c(0, data.order$surgHospital[2:nrow(data.order)] -
                        data.order$surgHospital[1:(nrow(data.order) - 1)]
                        ))
```

```
data.order$myind.hospital <- myind.hospital

dat.dif <-  c(data.order$surgDate[2:nrow(data.order)] -
              data.order$surgDate[1:(nrow(data.order) - 1)], NA)
data.order$dat.dif <- dat.dif

# dat.dif should be NA for the last surgery at each hospital
data.order$dat.dif <- ifelse(data.order$dat.dif < 0, NA,
                             data.order$dat.dif)

data.order[1:5, c(33,35,74,80,81)]
surgHospital surgDate y.num myind.hospital dat.dif
        1    7087.679     0              0    169
        1    7256.679     0              0    33
        1    7289.679     0              0    63
        1    7352.679     1              0    180
        1    7532.679     0              0    91

# Cox regression model:
# Regressing time between successive surgeries on previous patient
# outcome, adjusting for patient characteristics.
# Previous patient outcome defined as outcome in previous year in
# same hospital

coxph.model.a1 <- coxph(Surv(dat.dif, as.numeric(surgery)) ~
                  y.num + hv.na + ses + sex + ageCat + diagYearCat +
                  pathMorph + pathStage + preOpTherapy + postOpTherapy,
                   cluster(surgHospital), data = data.order)

# Coefficient
round(coef(summary(coxph.model.a1))[1, 2], 4)
# 95% CI
cc <- coef(summary(coxph.model.a1))
citab.coxph.model.a1 <- round(cbind(lwr = (cc[1, 2]) - 1.96 * cc[1, 3],
```

In the next Chapter a new method for volume-outcome studies that account for informative cluster size is proposed.

7 Thesis contribution: Within cluster resampling in combination with recurrent marked point process

In this thesis the recurrent marked point process is used as an approach to model volume-outcome associations of interest. Recall that independence estimating equations (IEE) are the only estimating equations that provides unbiased estimates under the recurrent marked point process model when informative cluster size is present. In Chapter 6 it was shown that cluster sizes are informative in the dataset employed in this thesis. As discussed before, depending on the target of inference, IEE might not always gives the desired parameter estimates.
As described in Section 6.2.1, within cluster resampling (WCR) is a proper method to analyse informative cluster size data due to one-per-cluster sampling scheme. The novelty of this thesis is to use WCR in the framework of a recurrent marked point process to study a longitudinal volume-outcome association.
This chapter is set out as follows. Section 7.1 highlights a formal notation of the proposed method. Section 7.2 discusses the application of the current method and in Section 7.3 results based on the proposed method are presented.

7.1 New approach

In this thesis it is proposed to use WCR in the framework of a recurrent marked point process to study a longitudinal volume-outcome association. To the best of our knowledge WCR has never been used in combination with a recurrent marked point process before.
The proposed method consists of the following steps.

Step 1: Repeatedly (Q times) sample one observation from each cluster I, forming Q resampled datasets.

Step 2: Fit the marginal model

$$\mu_i(t) = \quad E[Y_i(t) \mid \triangle N_i(t) = 1,\ \mathcal{X}_i(t),\ \mathcal{N}_i(t)] = x_{it}\beta, \tag{17}$$

to each of the Q resampled datasets, quantifying the volume-outcome associations between hospital volume $N_i(t)$ and patient outcome $Y_i(t)$. The estimated parameters $\widehat{\beta_q}$ represent the average volume-outcome association within the Q resampled dataset. The equation $\triangle N_i(t) = 1$ is required for the mark $Y_i(t)$ to exist.

Step 3: Obtain the WCR parameter by taking the average.

$$\widehat{\beta_{WCR}} = \frac{1}{Q}\sum_{q=1}^{Q}\widehat{\beta}_q \tag{18}$$

7.2 Application

In this thesis patient outcome (e.g. dead or alive) is only observed for patients undergoing surgery. Since in a recurrent marked point process the outcome or mark only exists if an event occurs, it is a realistic approach to model the volume-outcome associations of interest by employing this methodology.
In Section 6.4 it was shown that cluster size is informative. IEE provide unbiased estimates under the recurrent marked point process even when cluster sizes are informative. However, estimates obtained with this methodology are observation-based. Suppose we are interested in the volume-outcome associations for a randomly selected patient of a randomly selected hospital. In this case, IEE does not provide the desired parameter estimates. As described in Sections 6.3 and 6.2.1, WCR provides cluster-based parameters.

The proposed method comes down to fitting the mean model

$$\begin{aligned} E[Y_i(t) \mid \triangle N_i(t) = 1,\ X_i(t),\ \mathcal{N}_i(t)] \\ = \beta_0 + \beta_1 X_i(t) + \beta_2 N_i(t), \end{aligned} \tag{19}$$

describing the association between hospital volume and the average outcome among patients satisfying $\triangle N_i(t) = 1$. In Chapter 5 this model was fitted by using IEE and GEE assuming an exchangeable correlation structure, taking into account the dependent structure of the data. Under the proposed methodology, WCR is used to estimate the parameters in model.
Using WCR, one observation $Y_i(t)$ is repeatedly (Q times) sampled from each cluster. Note that clusters in this thesis denote hospitals i ($i = 1, \ldots, n$). Since 148 hospitals are included in the dataset, the resulting Q resampled datasets contain 148 independent observations. On each of the Q resampled datasets, model (19) fitted by using GLM.

Here it is shown how the proposed method is implemented in R in order to obtain an estimate for the association between a non-aggregate measure for cumulative hospital volume and 6-month patient mortality. In a similar way, estimates for the aggregate measures of hospital volume may be obtained (see Appendix B).

```
#########################
# Application of WCR
#########################

# Data preparation
temp7 <- count(data.order, "surgHospital")[,                                    2]
temp8 <- cumsum(temp7)
temp9 <- c(1, temp8 + 1)
x <- NULL

# Clusters
I <- length(unique(data.order$surgHospital)) # I = 148 clusters
Q <- 5000
# Generate Q datasets of size I = 148

#########################
# Non-aggregate cumulative hospital volume
#########################

rownumber <- numeric(148)
vec.coef <- matrix(0, Q, 1)
vec.st.error <- matrix(0, Q, 1)

set.seed(123)

for (q in 1:Q) {

  for (i in 1:I){
    x <- temp9[i]:temp8[i] # row number to choose from per cluster
    rownumber[i] <- sample(x, size = 1)
  }

  # Analyse resampled dataset by a GLM
  dat.set <- data.order[rownumber, ]
  glm.6M <- glm(y.num ~ hv.na + ses + sex + ageCat + diagYearCat
                + pathMorph + pathStage + preOpTherapy
                , data = dat.set, family = binomial(link = logit))

  vec.coef[q, ] <- exp(coef(summary(glm.6M))[2, 1])
```

```
  vec.st.error[q, ] <- coeftest(glm.6M,  vcov = sandwich)[2, 2]

}

B.WCR.0 <- mean(vec.coef)

# WCR variance estimator

S.B.sig <- (sum((vec.coef - B.WCR.0) * (vec.coef - B.WCR.0)))/(Q - 1)
sig.Q <- sum(vec.st.error^2) #sumcovs

Var.0 <- sig.Q / Q - ((Q - 1)/ Q) * S.B.sig

# Results
round(B.WCR.0, 4)
round(Var.0, 4)
```

7.3 Results

Table 4 provides estimates under the proposed methodology for the odds-ratio of 6-month patient mortality, using non-aggregate cumulative total, yearly total, cumulative total and a running average as measures for hospital volume. For non-aggregate cumulative total volume, a ten-patient increase is estimated to be associated with a 5.34% decrease in the odds of 6-month patient mortality.
For yearly total volume, cumulative total volume and running average volume, a ten-patient increase is estimated to be associated with respectively a 2.91%, 5.08% and a 5.42% decrease in the odds of 6-month patient mortality. The estimate for yearly total is associated with a substantial higher variance than the other measures for hospital volume. As stated before, the variance estimators can take negative values when number of clusters (i.e. hospitals) is too small for the asymptotic variance approximation to hold. This is a drawback of using WCR. Another problem was encountered when implementing the proposed method. Some covariates included in the model are categorical (see Table 1). It is possible that observations for certain levels of categorical variables are lacking in one or more of the Q resampled datasets, making it impossible to obtain estimates for these covariate levels. As a result, the estimated covariance matrices, obtained from each q^{th} analysis, may not be equally large and in that case it is impossible to obtain variance estimates for all covariates.

Table 4: Estimated associations between several specifications of hospital volume and the odds of 6-month mortality after oesophageal cancer surgery, obtained with the proposed method. Estimates are based on Q=5000 resampled datasets.

Specification of hospital volume	Model	OR	Variance estimator, using (14)
Non-aggregate			
Cumulative total volume			
	WCR	0.9466	0.0017
Aggregate			
Yearly total volume			
	WCR	0.9709	-0.6044
Cumulative total volume			
	WCR	0.9492	0.0015
Running average volume			
	WCR	0.9458	0.0018

8 Simulation Study

When outcome is dependent on cluster size, cluster size is said to be informative. In this situation, GEE and the GLMM might give biased parameter estimates. IEE might give unbiased but undesired parameter estimates when the target of inference is a randomly selected cluster member. WCR is a method suitable to analyse informative cluster size data due to one-per-cluster sampling scheme. In the previous chapter it was proposed to use within cluster resampling (WCR) within the framework of a recurrent marked point process to study a longitudinal volume-outcome association.
In this chapter the performance of the proposed method is evaluated by means of a simulation study. In Section 8.1 the simulation algorithm used in this thesis is illustrated. Performance of the proposed method is assessed under different scenarios. In Section 8.2 it is described how these different scenarios are created by varying design factors. Simulation results are presented in Section 8.3.

8.1 The Gauge

The goal of the simulation study is to asses the performance of the proposed method in estimating the volume-outcome associations of interest when:

- Assumption $(1')$ is violated. As stated before, this assumption $(1')$ implies independence between patient outcome and future number of events. Violating assumption $(1')$ is a mechanism to generate data in which informative cluster size is present.
- An aggregate measure for hospital volume is used.

The data are generated in the framework of a recurrent marked point process. For each of the total 1000 iterations, a population of 10.000 hospitals is simulated at time $t = 1, \ldots, 100$ discrete time points. For each hospital i, at each time point t, it is simulated whether or not an event (i.e. surgery) occurs. In case a surgery takes place a patient-level covariate $X_i(t)$, denoting patient's risk and a patient's outcome $Y_i(t)$ are generated. These generated quantities refer to a patient corresponding to hospital i at time point t. The following algorithm is used to generate data.

For $t = 1$:
Step 1: Generate a continuous patient's risk $X_i(t)$ for each hospital i at $t = 1$ by

$$X_i(1) \sim N(0, \tau^2(1 - \rho^2).$$

Step 2: Generate $\triangle N_i(t)$ indicating whether or not a surgery occurs at hospital i at $t = 1$

$$\triangle N_i(1) \mid X_i(1) \sim B(expit[\ \eta_0 + \eta_2 X_i(1)\ ]),$$

where $expit(\cdot) = exp(\cdot)/[1 + exp(\cdot)]$.

Step 3: Generate hospital specific random intercepts, exposure and volume effects $\boldsymbol{\gamma_i} = \{\gamma_{i0}, \gamma_{i1}, \gamma_{i2}\}$, serial correlation $W_i(t)$ and measurement error $\epsilon_i(t)$ at $t = 1$

$$\begin{gathered}\boldsymbol{\gamma_i}(1) \sim N_3(0, D)\\ W_i(1) \sim N(0, \tau^2(1-\rho^2)\\ \epsilon_i(1) \sim N(0, \sigma^2).\end{gathered}$$

Step 4: Center each of the hospital specific random effects $\gamma_{i0}(t), \gamma_{i1}(t), \gamma_{i2}(t)$ and serial correlation $W_i(t)$ by subtracting their conditional expectation of the subset $\triangle N_i(1) = 1$ at $t = 1$

$$\begin{gathered}\widetilde{\gamma_{i0}}(1) = \gamma_{i0}(1) - \bar{\gamma_{i0}}(1)\\ \widetilde{\gamma_{i1}}(1) = \gamma_{i1}(1) - \bar{\gamma_{i1}}(1)\\ \widetilde{\gamma_{i2}}(1) = \gamma_{i2}(1) - \bar{\gamma_{i2}}(1)\\ \widetilde{W_i}(1) = W_i(1) - \bar{W}_i(1).\end{gathered}$$

Centering must be done in order to have a correct marginal expectation for $Y_i(t)$.

Step 5: Generate an outcome at $t = 1$ if an event occurs, that is if $\triangle N_i(1) = 1$, by

$$Y_i(1) = \beta_0 + \beta_1 X_i(1) + \beta_2 N_i(1) + \widetilde{\gamma_{i0}}(1) + \widetilde{\gamma_{i1}}(1) X_i(1) + \widetilde{\gamma_{i2}}(1) N_i(1) + \widetilde{W_i}(1) + \epsilon_i(1),$$

where $\{\beta_0, \beta_1, \beta_2\}$={1, -1 , 0.05}.

If $\triangle N_i(t) = 0$, go to **Step 6**.

Step 5: Calculate the centered residual $R_i(t)$, which is defined as the difference between observed and expected outcome given $X_i(t)$, at $t = 1$

$$R_i(1) = \widehat{Y}_i(1) - Y_i(1),$$

where $Y_i(1) = \widehat{a} + \widehat{b}X_i(1) + \epsilon_i(1)$.

for $t = 2, \ldots, 100$:
Step 6: Generate $X_i(t)$ by employing a normal distribution with mean dependent on previous patient's risk

$$\begin{aligned} X_i(t) \mid \mathcal{X}_i(t),\, \mathcal{N}_i(t),\, \mathcal{Y}_i(t) &= X_i(t) \mid X_i(t-1) \\ &\sim N(\rho X_i(t-1), \tau^2(1-\rho^2). \end{aligned} \tag{20}$$

Step 7: Generate the event-time process $\triangle N_i(t)$, both dependent on $X_i(t)$ and previous patient outcome $Y_i(t-1)$ via $R_i(t-1)$ by

$$\begin{aligned} dN_i(t) \mid \mathcal{X}_i(t),\, \mathcal{N}_i(t),\, \mathcal{Y}_i(t) &= dN_i(t) \mid X_i(t),\, Y_i(t-1) \\ &\sim B(expit[\, \eta_0 + \eta_1 R_i(t-1) + \eta_2 X_i(t)\,]). \end{aligned}$$

Therefore, the chance of a surgery to occur increases with the difference between observed and expected previous outcome. Using $R_i(t-1)$ instead of simply using previous outcome $Y_i(t-1)$ is done in order to prevent the probabilities to become very large. Note that for $R_i(t-1)$ the last existing centered residual is used.

Step 8: Generate hospital specific random intercepts, exposure and volume effects $\boldsymbol{\gamma_i} = \{\gamma_{i0}, \gamma_{i1}, \gamma_{i2}\}$, serial correlation $W_i(t)$ and measurement error $\epsilon_i(t)$

$$\begin{gathered} \boldsymbol{\gamma_i}(t) \sim N_3(0, D) \\ W_i(t) \sim N(\rho W_i(t-1), \tau^2(1-\rho^2) \\ \epsilon_i(t) \sim N(0, \sigma^2). \end{gathered}$$

Generate $W_i(t)$ by employing a normal distribution with mean dependent on the previous value for serial correlation $W_i(t-1)$.

Step 9:
Center each of the hospital specific random effects $\gamma_{i0}(t), \gamma_{i1}(t), \gamma_{i2}(t)$ and serial correlation $W_i(t)$ by subtracting their conditional expectation of the subset $\triangle N_i(t) = 1$

$$\begin{gathered} \widetilde{\gamma_{i0}}(t) = \gamma_{i0}(t) - \bar{\gamma_{i0}}(t) \\ \widetilde{\gamma_{i1}}(t) = \gamma_{i1}(t) - \bar{\gamma_{i1}}(t) \\ \widetilde{\gamma_{i2}}(t) = \gamma_{i2}(t) - \bar{\gamma_{i2}}(t) \\ \widetilde{W_i}(t) = W_i(t) - \bar{W}_i(t). \end{gathered}$$

Step 10: If $\triangle N_i(t) = 1$, generate an outcome $Y_i(t)$

$$\begin{aligned} & Y_i(t) \mid \triangle N_i(t) = 1,\ \mathcal{X}_i(t),\ \mathcal{X}_i(t) \\ = \ & \beta_0 + \beta_1 X_i(t) + \beta_2 N_i(t) + \widetilde{\gamma_{i0}}(t) + \widetilde{\gamma_{i1}}(t) X_i(t) + \widetilde{\gamma_{i2}}(t) N_i(t) + \widetilde{W_i}(t) + \epsilon_i(t), \end{aligned}$$

where $\{\beta_0, \beta_1, \beta_2\}$={1, -1 , 0.05}.

If $\triangle N_i(t) = 0$ go to **Step 6**.

Step 11: Calculate the centered residual $R_i(t)$

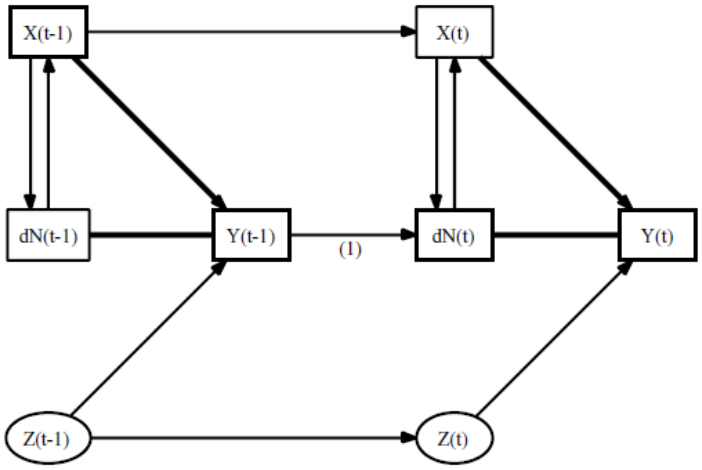
$$R_i(t) = \widehat{Y}_i(t) - Y_i(t),$$

where $Y_i(t) = \widehat{a} + \widehat{b} X_i(t) + \epsilon_i(t)$ and go back to **Step 6**.

Step 12: Repeat steps 6 through 11 until $t = 100$.

Figure 8.1 shows conditional relations between patient's risk, the event-time process, patient outcome and unmeasured error $Z_i(t)$ in the performed simulation study [10]. Unmeasured error includes serial correlation $W_i(t)$, hospital specific random effects $\gamma_i(t)$ and measurement error $\epsilon_i(t)$.
For each iteration, 1000 hospitals, together with their corresponding patients' risk and patients' outcomes at time $t = 1, \ldots, 100$, are randomly selected and volume-outcome analysis is performed. Special focus is on the parameter β_2 since this parameter in the mark process quantifies the effect of hospital volume on patient outcome.
WCR, based on 5000 resamplings, is used to obtain mean point estimates for the regression parameter β_2, along with the corresponding mean squared error (MSE), mean standard error, empirical standard error and estimated coverage of 95% confidence intervals. Additionally, IEE and GEE by assuming an exchangeable correlation structure are applied to obtain estimates for the parameter β_2. IEE assumes that pairwise observations within a cluster are independent. The exchangeable correlation structure assumes that all pairwise correlations within a cluster are equal, which might be more realistic. Also, a linear mixed model with random intercepts (LMM-RI) and a linear mixed model with random hospital specific intercepts, exposure and volume effects (LMM-RS) are fitted.
As stated before, a negative variance for the WCR regression parameter is possible. If a negative variance estimate occurs, standard error estimate is deleted and the simulation is included in the coverage calculation as it did not cover. In our simulations occurrence of a negative variance estimate appeared to be very rare for most scenarios because the

Figure 8.1: Underlying framework for the simulation study.



number of resamples Q and the number of clusters I are chosen sufficiently large. Simulation results presented are based on the following values for the parameters: $\rho = 0.9$, denoting the amount of autocorrelation, $\tau^2 = 1.5$, used in the specification of the Gaussian variance of the covariate process, $\sigma^2 = 1$, denoting measurement variance, $\eta_0 = 0$, denoting the intercept in the event-time process. For the generation of the hospital specific random effects, a Gaussian variance with parameters $D_{00} = 0.5^2, D_{11} = 0.2^2, D_{22} = 0.5^2, D_{33} = 0.1^2$ is used. For the parameter β_2 a value of 0.05 is chosen in order to incorporate a moderate effect of hospital volume on patient-outcome. In Appendix A, it is shown how the simulation is implemented in R.

8.2 Design factors

In the simulation study the following quantities are varied:

- η_1, quantifying the extent to which assumption $(1')$ is violated, having two levels $\eta_1 = \{\log(1), \log(2), \log(4)\}$. Log(1), log(2) and log(4) correspond to no violation, a moderate violation and a strong violation of assumption $(1')$ respectively. For all simulations $\eta_2 = -\eta_1$.
- Non-aggregate and aggregate measures for hospital volume. For aggregate measures, hospital volume is specified by using three measures, where $N_i(t)$ is aggregated into ten years of 10 time points:
  - Yearly total volume

$$N_{1i}(j) = N_i(T_j) - N_i(T_{j-1}), \tag{21}$$

– Cumulative total volume

$$N_{2i}(T_j) = \sum_{s=1}^{Tj} dN_i(s), \tag{22}$$

– Running average volume

$$N_{3i}(j) = \frac{N_{2i}(T_{j-1}) + N_{2i}(T_j)}{2}, \tag{23}$$

all for hospital i at year j and with T_j denoting the end of each year. Non-aggregate cumulative volume $N_i(t)$ is calculated as

$$N_{0i}(t) = \sum_{s=1}^{t} \triangle N_i(s), \tag{24}$$

denoting the total number of surgeries performed at hospital i through time t.

8.3 Simulation results

Non-aggregate cumulative volume. Table 5 provides simulation results using a non-aggregate measure for cumulative hospital volume. When assumption $(1')$ is not violated, all estimation methods provide unbiased parameter estimates with good coverage. Only LMM-RI shows poor coverage because the standard error is underestimated.
When assumption $(1')$ is moderately violated, parameter estimates are slightly or substantially underestimated. For both IEE and WCR bias is relatively small with estimated parameter estimates equal to 0.045 and 0.046 respectively for β_2 and an observed coverage of 912 and 907 respectively. In this case, covariance weighted methods provide biased parameters estimates with poor coverage.
When assumption $(1')$ is strongly violated, parameter estimates are underestimated for all estimation methods. However, performance of IEE and WCR is a great deal better than for the other estimation methods.
Recall that hospital specific intercepts, exposure and volume effects are specified to generate the data. Only LMM-RS specifies within hospital correlation structure correctly; it

is the only method that includes hospital specific random intercepts, volume and exposure effects. LMM-RI does not specify within correlation structure correctly, explaining why it gives standard errors that are substantially underestimated. Also IEE and GEE do not specify within correlation structure correctly, however obtain standard errors are still consistent because the robust (sandwich) variance estimator is used, which is robust to misspecification of the correlation structure (see Section 4.2). WCR accounts for the within cluster correlation structure, without the need of specifying it.

Aggregate measures for cumulative volume. Table 6 provides simulation results when assumption $(1')$ is not violated, using yearly aggregate measures for cumulative hospital volume. All methods provide (approximately) unbiased parameter estimates with good coverage. Only LMM-RI shows poor coverage because it substantially underestimates the standard error.

Table 5: Simulation results using a non-aggregate measure for cumulative hospital volume: mean point estimates of regression coefficients, mean squared error (MSE) of β_2 (x10), mean of standard error estimates for β_2 (mean SE), empirical standard error of estimated regression coefficients for β_2 (ESE) and estimated coverage for β_2 for η_1={log(1), log(2), log(4)}. Results are based on 1000 iterations.

η_1	Method	Mean $\widehat{\beta_2}$	MSE $\widehat{\beta_2}$ x 10	Mean SE $\widehat{\beta_2}$	ESE $\widehat{\beta_2}$	Coverage $\widehat{\beta_2}$
log 1	IEE	0.050	0.0001	0.0038	0.0036	971
	GEE	0.050	0.0001	0.0035	0.0032	971
	LMM-RI	0.050	0.0001	0.0006	0.0032	287
	LMM-RS	0.050	0.0001	0.0034	0.0031	976
	WCR	0.050	0.0001	0.0037	0.0036	950
log 2	IEE	0.045	0.0010	0.0086	0.0088	912
	GEE	0.042	0.0011	0.0070	0.0067	798
	LMM-RI	0.042	0.0011	0.0007	0.0087	100
	LMM-RS	0.037	0.0020	0.0045	0.0052	156
	WCR	0.046	0.0010	0.0083	0.0088	907
log 4	IEE	0.040	0.0022	0.0099	0.0105	818
	GEE	0.037	0.0024	0.0085	0.0086	672
	LMM-RI	0.038	0.0023	0.0008	0.0125	54
	LMM-RS	0.029	0.0050	0.0056	0.0065	35
	WCR	0.040	0.0022	0.0095	0.0106	795

Table 6: Simulation results using aggregate measure for cumulative hospital volume: mean point estimates of regression coefficients, mean squared error (MSE) of β_2 (x10), mean of standard error estimates for β_2 (mean SE), empirical standard error of estimated regression coefficients for β_2 (ESE) and estimated coverage for β_2 for η_1=(log1, log2). Results are based on 1000 iterations.

η_1	Method	Mean $\widehat{\beta_2}$	MSE $\widehat{\beta_2}$ x 10	Mean SE $\widehat{\beta_2}$	ESE $\widehat{\beta_2}$	Coverage $\widehat{\beta_2}$
Cumulative total						
log 1	IEE	0.050	0.0001	0.0039	0.0036	969
	GEE	0.050	0.0001	0.0035	0.0032	968
	LMM-RI	0.050	0.0001	0.0007	0.0239	284
	LMM-RS	0.050	0.0001	0.0034	0.0031	966
	WCR	0.050	0.0001	0.0037	0.0036	949
Running average						
log 1	IEE	0.051	0.0002	0.0040	0.0037	957
	GEE	0.052	0.0001	0.0036	0.0033	964
	LMM-RI	0.052	0.0001	0.0007	0.0517	250
	LMM-RS	0.052	0.0001	0.0035	0.0032	947
	WCR	0.051	0.0002	0.0038	0.0037	939

Yearly total. Table 7 provides simulation results using an aggregate measure for present hospital volume when assumption $(1')$ is not violated. Using yearly total, all estimation methods substantially underestimate the regression parameter. The point estimate for both IEE and WCR is half the true value. Point estimates for covariance weighted estimates are all zero. None of the estimation methods shows good coverage, however coverage for IEE and WCR is a great deal better than for the other estimation methods. Coverage for WCR is not as good because in this scenario the variance estimator frequently takes a negative value.

Simulations are performed to evaluate the performance of different methods in various scenarios. Recall that the parameter β_2 quantifies the volume-outcome association of interest. Simulations show that in case assumption $(1')$ is violated, performance of IEE and WCR is still acceptable whereas covariance weighted methods provide substantially biased parameter estimates with poor coverage.
Recall that when cluster size is uninformative, the IEE and WCR parameter coincide. For each measure of hospital volume, it is indeed shown that parameters estimated by employing IEE and WCR are equal when assumption $(1')$ is not violated. When cluster size is informative however, it is difficult to compare WCR and GEE since they are not

estimating the same parameters. However, their similarities in parameter estimates are noticeable. Likewise, IEE and GEE by assuming exchangeable correlation structures are not estimating the same parameters.
Simulations also show that parameter estimates for β_2 are unbiased when an aggregate measure for cumulative hospital volume is used, however that substantial bias may be introduced in the estimation process when an aggregate measure for present hospital volume is used.

Table 7: Simulation results using an aggregate measure for present hospital volume: mean point estimates of β_2, mean squared error (MSE) of β_2 (x10), mean of standard error estimates for β_2 (mean SE), empirical standard error of estimated regression coefficients for β_2 (ESE) and estimated coverage for β_2 for η_1=(log1, log2). Results are based on 1000 iterations.

η_1	Method	Mean $\widehat{\beta_2}$	MSE $\widehat{\beta_2}$ x 10	Mean SE $\widehat{\beta_2}$	ESE $\widehat{\beta_2}$	Coverage
Yearly total						
log1	IEE	0.025	0.0113	0.0235	0.0225	823
	GEE	0.000	0.0268	0.0149	0.0145	86
	LMM-RI	0.000	0.0268	0.0071	0.0145	5
	LMM-RS	0.000	0.0273	0.0150	0.0145	70
	WCR	0.025	0.0113	0.0209	0.0225	728

Figure 8.2: Simulation results: point estimates of regression parameters and their corresponding 95% CI intervals using different estimation methods when cluster size is non-informative.

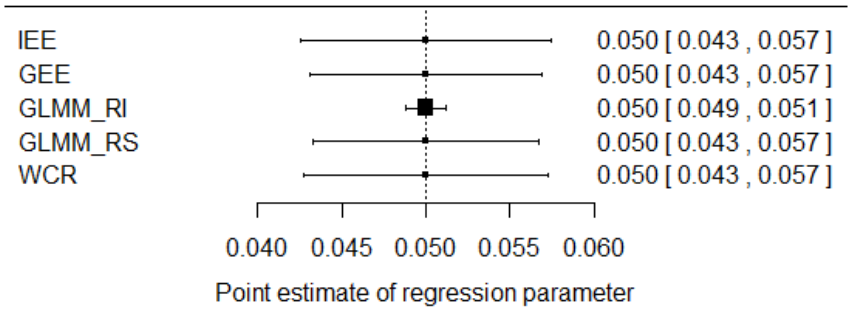
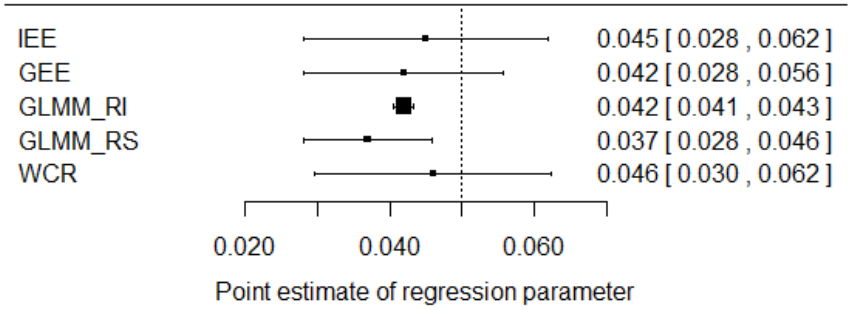


Figure 8.3: Simulation results: point estimates of regression parameters and their corresponding 95% CI intervals using different estimation methods when cluster size is informative.



9 Critical appraisal

During the past decades, the focus of much research has been on investigating the association between hospital volume and health outcome of patients after a surgical procedure. Many studies have reported an inverse relationship between hospital volume and postoperative mortality. As stated before, there are several features of volume-outcome studies that propose statistical challenges. However, many existing volume-outcome studies use naive estimation methods, not taking into account these difficulties. This chapter presents a critical appraisal on statistical methodology used in existing volume-outcome studies.

Cross-sectional analysis. Patients treated at the same hospital are more likely to experience similar results than patients treated at a different hospital. As a consequence, outcomes of patients treated at the same hospital tend to be correlated. Correlation might still exist after adjusting for hospital volume since hospitals may show variations in quality that are unrelated to the number of surgeries performed, for example due to different surgical techniques or processes of care. Therefore, it is necessary to consider in the analysis the correlated nature of the data. Panageas *et al.* [21] reviewed statistical methodology of volume-outcome studies that were published between 1995 and 2003 and found that in a great majority of these studies (73%) cross-sectional analyses were performed. In a cross-sectional study comparisons are made at a single time point, thereby not taking into account correlation between observations within the same hospital over time. For example [3, Begg *et al.*] and [26, Sollano *et al.*] examined volume-outcome relations by using standard logistic regression.
As described in Chapter 4, two longitudinal methods that account for the effect of correlation are generalized estimating equations (GEE) and the generalized linear mixed model (GLMM). The regression parameter estimated by the GLMM may be different of the regression parameter estimated by standard logistic regression due to correction for clustering. Using GEE compared to standard logistic regression, does not change the estimated regression parameter, only their estimated standard error. In general, when outcomes are positively correlated within clusters, using methods that do not adjust for correlation will lead to estimated standard errors that are underestimated [22, Panageas *et al.*]. Analysis based on standard logistic regression might therefore blow up the statistical significance.
In Table 8 it is shown how the estimated odds-ratio change analysing the data employed in this thesis by standard logistic regression. It can be seen that the estimated odds-ratio is not different, however the CI is narrowed, showing that statistical significance is indeed exaggerated.

Categorical variable for hospital volume. As described in Chapter 5, hospital volume is not a fixed quantity but rather a quantity that changes over time. Many volume-outcome studies analyse hospital volume as a categorical variable, thereby neglecting that hospital volume at the beginning of the study period may be different from its value at the end of

the study period. In addition, the selection of cut-off points for the different categories may highly influence the significance of the obtained volume-outcome associations. [27, Wen *et al.*], for example, defined three hospital groups: *low*, *medium* and *high volume* and compared in-hospital mortality between the different categories. [14, Hannan *et al.*] defined three hospital volume thresholds and calculated the odds-ratio of in-hospital mortality relative to the other side of the threshold. This means that three analyses were performed, each time defining two different categories *low* and *high*.
Table 8 provides estimated volume-outcome associations in the dataset employed in this thesis when a categorical variable for hospital volume is used, both by defining two and four levels. Cut-off points are chosen such that each category contains the same amount of observations. Analysing hospital volume as a categorical variable with two levels, for example, indicates a 36.57% significant decrease in the odds of 6-month patient mortality for high hospital volume versus low hospital volume. However, it is difficult to compare results using a continuous and categorical variable since they quantify different associations.

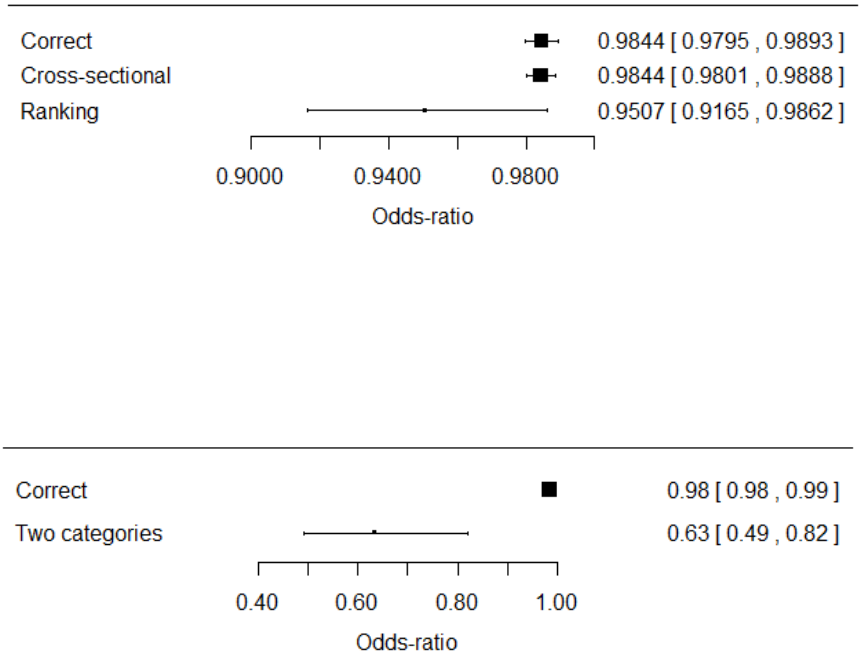
Ranking hospital volume. Volume-outcome studies such as [25, Schrag *et al.*] and [3], rank hospitals according to the number of surgeries performed over the study period and use this rank number as a variable for hospital volume. Ranking hospitals in a volume-outcome study is inappropriate for the same reason as using a categorical variable; in this way it is not taken into account that hospital volume may change over the coarse of the study. A specific hospital might be assigned a rank number one because it has the largest amount of surgeries accumulated over the study period. In case surgeries are concentrated in the last part of the study period, this rank number is not appropriate for the first time period. Table 8 shows volume-outcome associations analysing hospital volume as a ranked variable. Analysing hospital volume in this manner, indicates a significant of 4.93% decrease in the odds of 6-month patient mortality for a ten point increase in rank number. As mentioned before, it is difficult to compare results using a continuous variable or rank number for hospital volume.

Ignoring informative cluster size. The situation in which cluster size is related to outcome, conditional on the covariates, is called informative cluster size. As stated before, including hospital volume in the model might not be enough to capture the relation between outcome and cluster size in a volume-outcome study. If there is any residual relation, not explained through the inclusion of hospital volume in the model, then informative cluster size is present. Assumptions present in the traditional methods which correct for clustering, such as generalized estimating equations, are violated in this setting.
The concept of informative cluster size in volume-outcome studies has not been made very clear yet. Most existing volume-outcome studies use other methods than IEE to quantify the association between hospital volume and patient outcome, without testing for informative cluster size. In Chapter 5 it was shown how other methods than IEE resulted in different estimated odds-ratios.

Table 8: Odds-ratios and their corresponding 95% CI intervals for different naive estimation methods.

Method	OR	95% CI
Appropriate statistical method	0.9844	(0.9796-0.9893)
Cross-sectional analysis	0.9844	(0.9801-0.9888)
Categorical variable for hospital volume Defining 2 levels		
low	1.00	
high	0.6343	(0.4915-0.8168)
Defining 4 levels		
very low	1.00	
low	0.9505	(0.7774-1.1620)
medium	0.8066	(0.6645-0.9791)
high	0.4445	(0.3568-0.5539)
Ranking hospital volume	0.9507	(0.9165-9861)

Figure 9.1: Odds-ratios and their corresponding 95% CI intervals for different naive estimation methods.



10 Discussion

In this thesis two different issues were investigated. The first one was to employ appropriate statistical methodology to investigate volume-outcome associations concerning patients with oesophageal cancer surgery. A recurrent marked point process was used to address the first problem. Statistical issues in the specification of both aggregate and non-aggregate measures were considered. Different estimation methods were used to provide volume-outcome estimates under the recurrent marked point process. Challenging issues that may be involved in volume-outcome studies were described: possible correlation between patients within the same hospital, specification of an appropriate measure for hospital volume that takes into account its time-dependent character and bias that may occur in the estimation process when certain assumptions are violated. Critical notes concerning naive statistical methodology applied in existing volume-outcome literature were presented.
The second goal of this thesis was to test for the presence of informative cluster size in the data used in this thesis and to propose a new method suitable for volume-outcome studies in which informative cluster size is present. In the new methodology proposed in this thesis WCR is used within the framework of the recurrent marked point process, providing cluster-based parameters. Simulations showed that when informative cluster size is present, the proposed method estimates the parameter for volume with relatively small bias. Simulations also indicate that bias might be introduced in the estimation process when an aggregate measure for present hospital volume is taken.

In the first part of this thesis, the association between hospital volume and patient outcome was evaluated. Different estimation methods showed an inverse relation between hospital volume and patient mortality, many of them being significant. In daily practice, not only hospitals but also surgeons are associated with different case volumes. This thesis has been limited to the consideration of the effect of surgeon volume. Therefore, the question whether the performance of large hospitals is better due to expertise of the hospital or the experience of the individual surgeon, remains unsolved and will need to be explored in future research.

In the second part of this thesis, assumption $(1')$ was used as a mechanism to test for the presence of informative cluster size. Using this approach, it was tested whether hospitals with a previous death (within 6 months since surgery) were associated with more time until a next surgery. It should be mentioned that this method is somewhat delicate because time between surgeries might be less than 6 months. Therefore, patient outcome after six months since surgery might not be observed yet.
However, using assumption $(1')$ provides an excellent starting point to test for informative cluster size. Future research is needed to provide alternative methods to test for the presence of informative cluster size. An example might be to test whether the occurrence of a subsequent event depends on the percentage of successful surgeries up to the current time

point. However, focus of this thesis was on proposing a new method suitable when cluster sizes are informative, rather than on testing for it.

A limitation of using WCR, in general and in the proposed method, is the possibility of a negative variance estimate to occur. According to Hoffman *et al.* [15], which proposed WCR, the probability of a negative variance estimator is exceedingly small and if it is still observed such an aberration should suggests that the number of resamples or number of clusters is insufficient. We do not fully agree with the last statement since information on 148 hospitals was available and 10000 resamples were used, which we do not think is exceptionally low.
Next to the association between hospital volume and patient outcome, the association between surgeon volume and patient outcome is often investigated in volume-outcome studies. Surgeons are nested within hospitals. Therefore, for volume-outcome studies focusing on the impact of surgeon volume, typically more clusters (i.e. surgeons) are available. This makes the proposed methodology more suitable for volume-outcome studies assessing the association between surgeon volume and patient outcome.

Under informative cluster size, WCR and GEE are not estimating the same parameters. Simulation results under violation of assumption $(1')$, however, showed a rather suspicious amount of resemblance between their parameter estimates. To investigate this issue further, simulations were also performed in which only very small and very large hospital were generated, making the differences in cluster size larger. However, parameter estimates obtained with IEE and WCR were still very similar in this scenario. An explanation for this phenomenon might be that the presence of informative cluster size is not strong enough when introduced through the violation of assumption $(1')$.

The novelty of this thesis is to use WCR in combination with a recurrent marked point process to account for informative cluster size. Simulations suggest that the proposed method provides estimates for the volume parameter with relatively small bias. In the application for this thesis, a binary outcome variable was modelled, indicating whether or not the patient is still alive after three months since surgery. For future research it may also be of interest to extend the proposed method to survival analyses (e.g. modelling time to event).

11 Appendix A: R code for simulation study

```
#########################
# Load necessary packages
#########################

library(plyr)
library(caTools)
library(gee)
library(lme4)
library(mvtnorm)
library(hydroGOF)
library(sandwich)
library(lmtest)

#########################
# Specify parameters
#########################

p         <- 0.9
v.squared <- 1.5
t.squared <- v.squared
eta0      <- 0
eta1      <- log(1)
eta2      <- -eta1
beta0     <- 1
beta1     <- -1
beta2     <- 0.05 # moderate effect of hospital volume on patient outcome
sigma     <- 1
D         <- matrix(c(0.25, 0, 0, 0, 0.04, 0, 0, 0, 0.01), 3, 3, byrow = TRUE)

Q         <- 5000
nclus     <- 10000
nit       <- 1000
t.p       <- 100

#########################
# Prepare dataframes
#########################
```

```
# Dataframes for X, N, Y, HV
X.gen         <- matrix(0, nclus, t.p + 1)
N             <- matrix(0, nclus, t.p + 1)
Y             <- matrix(NA, nclus, t.p)
HV            <- matrix(NA, nclus, t.p)

# Dataframes for centered residuals, serial correlation and hospital specific effects
res.cent      <- matrix(NA, nclus, t.p + 1)
res.cent[, 1] <- rep(0, nclus)
W             <- numeric(nclus)
gamma.tilde   <- matrix(NA, nclus, 3)

# Dataframes for WCR
mat.coef      <- matrix(0, Q, 3)
vec.var       <- numeric(Q)

# Dataframes for coefficients
coefs.iee     <- matrix(0, nit, 3)
coefs.gee     <- matrix(0, nit, 3)
coefs.lmm.ri  <- matrix(0, nit, 3)
coefs.lmm.rs  <- matrix(0, nit, 3)
coefs.wcr     <- matrix(0, nit, 3)

# Vectors for std.errors
std.iee       <- numeric(nit)
std.gee       <- numeric(nit)
std.lmm.ri    <- numeric(nit)
std.lmm.rs    <- numeric(nit)
std.WCR       <- numeric(nit)


lastValue <- function(x) {
# Takes last value for each row of x
  tail(x[!is.na(x)], 1)
  }

#########################
# Function to simulate results
#########################

SimResults <- function(nit, nclus, tp, Q, eta1) {
```

```
# Performs simulation algorithm described in Section 8.2 and uses different
# estimation methods to obtain estimates within the recurrent marked point
# process framework
#
# Args:
# nit:   number of iterations
# nclus: number of clusters
# tp:    number of time points within hospital
# eta1:  quantifies the extent to which assumption 1 is violated
#
# Returns for each estimation method:
# regression parameter estimates
# mean of standard errors
# mean of empirical standard errors
# means squared errors
# coverage
#
# For each of 1000 iterations..
  for(i in 1:nit){
    cat(i)
    # Prepare new matrix of outcomes Y
    Y <- matrix(NA, nclus, t.p)


    # Generating exposure process
    for (t in 1:(t.p)) {
      std <- sqrt(v.squared * (1 - p ^ 2))
      X.gen[, t+1] <-  rnorm(nclus, p * X.gen[, t], std)
    }
    X <- X.gen[, -1]


    # Generate hospital specific random effects and measurement error
    gamma <- rmvnorm(nclus, rep(0, nrow(D)), D)
    e     <- rnorm(t.p * nclus, 0, sd = sigma)


    # Generating event-time process and mark process simultaneously

    for (t in 1:t.p) {
```

```
  # Use previous centered residual
  # Event is likely to occur if difference between expected and
  # observed outcome is large
  last.res.cent <- apply(res.cent, 1, lastValue)
  temp          <- eta0 + eta1 * last.res.cent+ eta2 * X[, t]
  prob          <- exp(temp) / (1 + exp(temp))

  N[, t + 1]    <- rbinom(nclus, 1, prob) # moved with one column in order to prepare H
  HV[, t]       <- apply(N[, 1:(t + 1)], 1, sum, na.rm = TRUE)


  # Center gamma
  gamma.mat        <- N[, t + 1] * gamma
  gamma.exp        <- apply(gamma.mat, 2, sum) / sum(N[, t + 1])
  gamma.tilde[, 1] <- N[, t + 1] * (gamma[, 1] - gamma.exp[1])
  gamma.tilde[, 2] <- N[, t + 1] * (gamma[, 2] - gamma.exp[2])
  gamma.tilde[, 3] <- N[, t + 1] * (gamma[, 3] - gamma.exp[3])

  # Center serial correlation W
  W       <- rnorm(nclus, p * W, sqrt(t.squared * (1 - p ^ 2)))
  W.vec   <- N[, t + 1] * W
  W.exp   <- sum(W.vec) / sum(N[, t + 1])
  W.tilde <- N[, t + 1] * (W - W.exp)

  # Generate outcome
  (Y[,t] <- beta0 + beta1 * X[, t]+ beta2 * HV[, t] + gamma.tilde[, 1]+ gamma.tilde[, 2
   + gamma.tilde[, 3] * HV[, t] + W.tilde + e[((t * nclus) - (nclus- 1 )):(t * nclus)])

  Y[, t] <- N[, t + 1] * Y[, t]
  Y[, t] <- ifelse(N[, t + 1]==0, NA, Y[, t])

  # Calculate conditional expectation and centered residual
  # Linear model on observations with delta N = 1
  lmo               <- lm(Y[, t] ~ X[, t], na.action = na.exclude)
  cond.exp          <- fitted.values(lmo)
  res.cent[, t + 1] <- Y[, t] -  cond.exp

}

# Generate data matrix
```

```
Yrow  <- as.vector(t(Y))
Xrow  <- as.vector(t(X))
HVrow <- as.vector(t(HV))
crow  <- rep(1:nclus, each = t.p)
year  <- rep(rep(1:(t.p / 10),  each = (t.p / 10), nclus))

# df.sim <- as.data.frame(cbind(Yrow, Xrow, HVrow, crow))
df.sim = data.frame(Y    = Yrow,
                    X    = Xrow,
                    HV   = HVrow,
                    clus = crow,
                    year = year)

# Add a column with aggregated hospital volume measures
temp <- df.sim[seq(10, nclus * t.p, 10), 3]
df.sim$cum.yearly.tot <- rep(temp, each = 10)

temp2 <- temp[2:length(temp)] - temp[1:(length(temp) - 1)]
temp2[seq(10, (((t.p * nclus) / 10) - 10),
          by = 10)] <- temp[seq(11, ((t.p * nclus) / 10) - 9), by = 10)]
temp2 <- c(temp[1], temp2)
df.sim$yearly.tot <- rep(temp2, each = 10)

temp3 <- runmean(temp, k = 2)
temp3[seq(11, (((t.p * nclus) / 10) - 9),
          by = 10)] <- temp2[seq(11, (((t.p * nclus) / 10) - 9), by = 10)]
df.sim$run.ave <- rep(temp3, each = 10)


# Sample 1000 clusters (hospitals)
clussample <- sample(1:nclus, 1000)
# Select these 1000 clusters
df.sim <- df.sim[df.sim$clus %in% clussample, ]

print("data frame ready")


# Within Cluster Resampling

for (q in 1:Q) {
  # Select one row per cluster
```

```
  rownumber <- round(runif(length(clussample),
        min = seq(1, ((t.p * nclus) - (t.p - 1)), by = t.p), max = seq(t.p, (t.p * nclu

  dat.set <- df.sim[rownumber, ]
  lmo     <- lm(Y ~ X + HV, data = dat.set)

  mat.coef[q, ] <- summary(lmo)$coef[1:3, 1] # results for each iteration
  vec.var[q]    <- (coeftest(lmo,  vcov = sandwich)[3, 2]) ^ 2
  }

print("WCR finished")


# Results WCR
B.WCR             <- colMeans(mat.coef, na.rm = TRUE)[3]
std.error.WCR     <- sqrt((sum(vec.var) / Q) - ((sum((mat.coef[, 3] - B.WCR) *
                                                  (mat.coef[, 3] - B.WCR))) / Q))
coefs.wcr[i, 1:3] <- colMeans(mat.coef, na.rm = TRUE)
std.WCR[i]        <- ifelse(std.error.WCR >= 0, std.error.WCR, NA)
# In case of negative variance estimator std.WCR is NA


# Fitting models and save results

# Fitting IEE
iee.sim   <- gee(Y ~ X + HV, data = df.sim, id = clus, corstr = "independence")
coefs.iee[i, 1:3]    <- summary(iee.sim)$coef[1:3, 1]
std.iee[i]           <- summary(iee.sim)$coef[3,4]
print("iee ready")

# Fitting GEE
gee.sim    <- gee(Y ~ X + HV, data = df.sim, id = clus, corstr = "exchangeable")
coefs.gee[i, 1:3]    <- summary(gee.sim)$coef[1:3, 1]
std.gee[i]           <- summary(gee.sim)$coef[3,4]
print("gee ready")

# Fitting LMM_RI
lmm.ri.sim <- lmer(Y ~ X + HV + (1 | clus), data = df.sim)
coefs.lmm.ri[i, 1:3] <- summary(lmm.ri.sim)$coef[1:3, 1]
std.lmm.ri[i]        <- summary(lmm.ri.sim)$coef[3, 2]
print("lmm.ri ready")
```

```
  # Fitting LMM_RS
  lmm.rs.sim <- lmer(Y ~ X + HV + (1 + X + HV | clus), data = df.sim)
  coefs.lmm.rs[i, 1:3] <- summary(lmm.rs.sim)$coef[1:3, 1]
  std.lmm.rs[i]        <- summary(lmm.rs.sim)$coef[3, 2]
  print("lmm.rs ready")

}

object <- list(
               # Regression parameter estimates

               means.iee    <- colMeans(coefs.iee),
               means.gee    <- colMeans(coefs.gee),
               means.lmm.ri <- colMeans(coefs.lmm.ri),
               means.lmm.rs <- colMeans(coefs.lmm.rs),
               means.wcr    <- colMeans(coefs.wcr, na.rm = TRUE),


               # Mean of standard errors

               mean.std.iee    <- mean(std.iee),
               mean.std.gee    <- mean(std.gee),
               mean.std.lmm.ri <- mean(std.lmm.ri),
               mean.std.lmm.rs <- mean(std.lmm.rs),
               mean.std.wcr    <- mean(std.WCR , na.rm = TRUE),


               # Mean of empirical standard errors

               emp.se.iee   <-  sqrt(sum((coefs.iee[, 3] -
               mean(coefs.iee[, 3])) ^ 2) / (nit - 1)),
               emp.se.gee   <- sqrt(sum((coefs.gee[, 3] -
               mean(coefs.gee[, 3])) ^ 2) / (nit - 1)),
               emp.se.lmm.ri <- sqrt(sum((coefs.lmm.ri[, 3] -
               mean(coefs.lmm.rs[, 3])) ^ 2)/(nit - 1)),
               emp.se.lmm.rs <- sqrt(sum((coefs.lmm.rs[, 3] -
               mean(coefs.lmm.rs[, 3])) ^ 2)/(nit - 1)),
               emp.se.wcr    <- sqrt(sum(((coefs.wcr[, 3] -
               mean(coefs.wcr[, 3])) ^ 2), na.rm = TRUE) / (nit - 1)),
```

```
                # Mean squared errors

                mse.iee    <- mean((beta2 - coefs.iee[, 3]) ^ 2),
                mse.gee    <- mean((beta2 - coefs.gee[, 3]) ^ 2),
                mse.lmm.ri <- mean((beta2 - coefs.lmm.ri[, 3]) ^ 2),
                mse.lmm.rs <- mean((beta2 - coefs.lmm.rs[, 3]) ^ 2),
                mse.wcr    <- mean((beta2 - coefs.wcr[, 3]) ^ 2, na.rm = TRUE),


                # Coverage

                cov.iee    <- sum((beta2 > coefs.iee[, 3] - 1.96 * std.iee) &
                                  (beta2 < coefs.iee[, 3] + 1.96 * std.iee)),
                cov.gee    <- sum((beta2 > coefs.gee[, 3] - 1.96 * std.gee) &
                                  (beta2 < coefs.gee[, 3] + 1.96 * std.gee)),
                cov.lmm.ri <- sum((beta2 > coefs.lmm.ri[, 3] - 1.96 * std.lmm.ri) &
                                  (beta2 < coefs.lmm.ri[, 3] + 1.96 * std.lmm.ri)),
                cov.lmm.rs <- sum((beta2 > coefs.lmm.rs[, 3] - 1.96 * std.lmm.rs) &
                                  (beta2 < coefs.lmm.rs[, 3] + 1.96 *std.lmm.rs)),
                cov.wcr    <-  sum((beta2 > coefs.wcr[, 3] - 1.96 * std.WCR) &
                                   (beta2 < coefs.wcr[, 3] + 1.96 * std.WCR),
                                        na.rm = TRUE)
  )
  return(object)

} # end of function

set.seed(123)
simresults(nit, nclus, t.p, Q, eta1)

# Simulations are performed for different measures for hospital volume:
# HV (Non-aggregate measure for cumulative hospital volume)
# cum.yearly.tot and run.ave (Aggregate measures for cumulative total)
# yearly.tot (Aggregate measures for present measures for hospital volume)
```

12 Appendix B: Source of R Code used in this thesis

```
#########################
# Load necessary packages
#########################

library(foreign)
library(plyr)
library(caTools)
library(gee)
library(lme4)
library(KMsurv)
library(survival)
library(spatstat)
library(mvtnorm)
library(hydroGOF)
library(lmtest)
library(sandwich)
library(chron)
library(metafor)

#########################
# Read SPSS datafile
#########################

mydata <- read.spss("C:\\SAS\\NKR Chir Eso VolumeCat 1989-2009.sav",
                    to.data.frame = TRUE)

#########################
# Recode variables and add measures for hospital volume
#########################

first.letter  <- tolower(substring(names(mydata), 1, 1))
other.letters <- substring(names(mydata), 2)
newnames      <- paste(first.letter, other.letters, sep = "")

names(mydata) <- newnames
mydata$ses <- mydata$sES

# Recoding variables surgDate and surgYear
```

```
mydata$surgDate <- as.numeric(as.chron(ISOdate(1582, 10, 14) + mydata$surgDate))
mydata$diagDate <- as.numeric(as.chron(ISOdate(1582, 10, 14) + mydata$diagDate))

mean.d <- mean((mydata$surgDate - mydata$diagDate), na.rm=TRUE)
mydata$surgDate <- ifelse(is.na(mydata$surgDate),
                          (mydat$diagDat + rnorm(1, mean.d, 0.01)), mydat$surgDate)
mydata$surgYear <- ifelse(is.na(mydata$surgYear), mydata$diagYear, mydata$surgYear)

# Order data on hospital id, surgery year and surgery date respectively
data.order <- mydata[order(mydata$surgHospital, mydata$surgYear, mydata$surgDate), ]

# Non-aggregate measure for cumulative hospital volume
temp0 <- count(data.order, c("surgHospital"))[, 2]
hv.na <- NULL
for (i in 1:length(temp0)) {
  temp.na0 <- 1:temp0[i]
  hv.na <- c(hv.na, temp.na0)
}

data.order$hv.na <- hv.na

# Aggregate measure1: Yearly total
temp1 <- count(data.order, c("surgHospital", "surgYear"))
yearly.tot <- rep(temp1$freq, temp1$freq)

data.order$yearly.tot <- yearly.tot

# Aggregate measure2: Cumulative yearly total
temp2 <- tapply(temp1$freq, temp1$surgHospital, cumsum)
temp3 <- unname(unlist(temp2))
cum.yearly.tot <- rep(temp3, temp1$freq)

data.order$cum.yearly.tot <- cum.yearly.tot

# Aggregate measure3: Running average
temp4 <- count(data.order, c("surgHospital", "surgYear", "cum.yearly.tot"))
temp5 <- tapply(temp4$cum.yearly.tot, temp4$surgHospital, runmean, k=2)
temp6 <- unname(unlist(temp5))
run.ave <- rep(temp6, temp4$freq)

data.order$run.ave <- run.ave
```

```
# Add a column with patient status (numeric)
data.order$y.num <- as.numeric(data.order$survivalStatus6M) - 1
```

12.1 R code Chapter 2

```
#########################
# Make some explanatory plots
#########################

## Figure 2.1

hist(x    = data.order$age,
     xlab = "Age in years",
     main = "",
     col  = "lightblue")

## Figure 2.2

x <- temp0
names(x) <- 1:148
barplot(x     = temp0,
        horiz = T,
        col   = "gold",
        ylab  = "Hospital id",
        xlab  = "Total number of surgeries")

## Figure 2.3

#Make categories for cumulative yearly total
summary(sort(data.order$cum.yearly.tot))
data.order$cum.yearly.tot.cat[data.order$cum.yearly.tot > 159] <- "Large"
data.order$cum.yearly.tot.cat[data.order$cum.yearly.tot > 65
                              & data.order$cum.yearly.tot < 159] <- "Med.Large"
data.order$cum.yearly.tot.cat[data.order$cum.yearly.tot > 24
                              & data.order$cum.yearly.tot < 65] <- "Med.Small"
data.order$cum.yearly.tot.cat[data.order$cum.yearly.tot < 24] <- "Small"

counts <- table(data.order$survivalStatus6M, data.order$cum.yearly.tot.cat)
```

```
barheights <- rbind(counts,0)
dim(barheights) <- NULL
plot(c(0,12), c(0,2500))
barplot(height    = barheights,
        col       = rep(c("lightgreen", "lightpink", "white"), 3),
        names.arg = c("Very Large", "", "", "Large", "", "", "Medium", "", "",
                      "Small", "", ""),
        main      = "Survival Status 6M",
        ylab      = "Number of Patients")

legend(locator(1),
       legend = c("Alive", "Death"),
       fill   = c("lightgreen", "lightpink"),
       bty    = "n",
       cex    = 1)

## Figure 2.4

plot(survfit(Surv(fUPDaysDiagnosis6M, y.num) ~ cum.yearly.tot.cat, data = data.order),
     col       = c("lightgreen","khaki1","lightpink","lightcoral"),
     lwd       = 5,
     mark.time = F,
     ylim      = c(0.80,1),
     xlab      = "Days",
     ylab      = "Proportion of deaths")

legend(locator(1),
       legend = c("VeryLarge","Large","Medium","Small"),
       col    = c("lightgreen","khaki1","lightpink","lightcoral"),
       lwd    = rep(5,4),
       lty    = rep(1,4),
       bty    = "n")
```

12.2 R code Chapter 3

```
#########################
# Marked point processes
#########################
```

```
## Figures 3.6, 3.7, 3.8, 3.9

par(mfrow=c(1,1))
v <- plot(amacrine)
X <- rpoispp(25)
plot(X)

X <- rpoispp(function(x, y) {
  exp( 2 + 5 * x)
  }
  )

plot(X)
X <- rpoispp(100)
M <- sample(1:3, X$n, replace = TRUE)
plot(X %mark% M)
```

12.3 R code Chapter 5

```
## Figure 5.1

g <- c(16,17)
c <- c("green" ,"red")
dat.plot <- data.order[data.order$surgHospital==5, c(35,52)]
dat.plot <- dat.plot[order(dat.plot[,1]), ]
plot(x    = dat.plot$surgDate,
     y    = dat.plot$survivalStatusSurgery6M,
     yaxt = "n",
     xaxt = "n",
     xlab = "Surgery time",
     ylab = "Patient Status",
     pch  = g[dat.plot$survivalStatusSurgery6M],
     col  = c[dat.plot$survivalStatusSurgery6M])

axis(side   = 2,
     at     = c(1,2),
     labels = c("Alive", "Dead"))
# Same for hospital 43 and hospital 71
```

```
## Figure 5.2

plot(data.order$hv.na[data.order$surgHospital==1],
     lwd  = 2,
     xaxt = "n",
     xlab = "Years",
     ylab = "Hospital volume",
     type = "l",
     col  = "black")
lines(data.order$cum.yearly.tot[data.order$surgHospital==1],
      lwd = 2,
      col = "cornflowerblue")
lines(data.order$run.ave[data.order$surgHospital==1],
      lwd = 2,
      col = "orange")
lines(data.order$yearly.tot[data.order$surgHospital==1],
      lwd = 2,
      col = "hotpink")

axis(side   = 1,
     at     = c(1,152),
     labels = c("1989", "2010"))

legend(locator(1),
       legend = c("Non-aggregate", "Cumulative yearly total", "Running average",
                  "Yearly total"),
       fill   = c("black", "cornflowerblue", "orange", "hotpink"),
       bty    = "n",
       cex    = 1)

## Figure 5.3

deduped.data <- unique(data.order[,c("surgHospital", "surgYear", "yearly.tot")])

plot(deduped.data$yearly.tot[deduped.data$surgHospital==1],
     xaxt = "n",
     xlab = "Years",
     ylab = "Yearly total volume",
     type = "l",
     col  = "cornflowerblue")
lines(deduped.data$yearly.tot[deduped.data$surgHospital==6],
```

```
      col="orange")
lines(deduped.data$yearly.tot[deduped.data$surgHospital==18],
      col="hotpink")

axis(side   = 1,
     at     = c(1,22),
     labels = c("1989","2010"))

legend(locator(1),
       legend = c("Hospital 1", "Hospital 2", "Hospital 3"),
       fill   = c("cornflowerblue", "orange", "hotpink"),
       bty    = "n",
       cex    = 1)


#########################
# Fitting the models
#########################

# Results associated with ten-patient increase
data.order$hv.na          <- data.order$hv.na / 10
data.order$yearly.tot     <- data.order$yearly.tot / 10
data.order$cum.yearly.tot <- data.order$cum.yearly.tot / 10
data.order$run.ave        <- data.order$run.ave / 10

#########################
# Non-aggregate cumulative hospital volume
#########################

## Fitting IEE (GEE with independence correlation structure)
gee.in.0 <- gee(y.num ~ hv.na + ses + sex + ageCat + diagYearCat +
                pathMorph + pathStage + preOpTherapy + postOpTherapy,
                data = data.order, id = surgHospital,
                family = binomial(link = logit), corstr = "independence")

# Coefficient
iee.hv <- round(exp(coef(summary(gee.in.0))[2, 1]), 4)
# Std. error and 95% CI
cc.iee.hv <- coef(summary(gee.in.0))[2, 4]
cc <- coef(summary(gee.in.0))
citab.iee.hv <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 4]),
```

```
                            upr = exp(cc[2, 1] + 1.96 * cc[2, 4])), 4)

## Fitting GEE
gee.ex.0 <- gee(y.num ~ hv.na + ses + sex + ageCat + diagYearCat +
                pathMorph + pathStage + preOpTherapy + postOpTherapy,
                data = data.order, id = surgHospital,
                family = binomial(link = logit), corstr = "exchangeable")

# Coefficient
gee.hv <- round(exp(coef(summary(gee.ex.0))[2, 1]), 4)
# Std.error and 95% CI
cc.gee.hv <- coef(summary(gee.ex.0))[2, 4]
cc <- coef(summary(gee.ex.0))
citab.gee.hv <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 4]),
                            upr = exp(cc[2, 1] + 1.96 * cc[2, 4])), 4)

## Fitting GLMM-RI
glmm.ri.0 <- glmer(y.num ~ hv.na + ses + sex + ageCat + diagYearCat +
                   pathMorph + pathStage + preOpTherapy + postOpTherapy +
                   (1|surgHospital), data = data.order,
                   family = binomial(link = "logit"))

# Coefficient
glmm.ri.hv <- round(exp(coef(summary(glmm.ri.0))[2, 1]), 4)
# Std.error and 95% CI
cc.glmm.ri.hv <- coef(summary(glmm.ri.0))[2, 2]
cc <- coef(summary(glmm.ri.0))
citab.glmm.ri.hv <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 2]),
                                upr = exp(cc[2, 1] + 1.96 * cc[2, 2])), 4)

## Fitting GLMM-RS
glmm.rs.0 <- glmer(y.num ~  hv.na + ses + sex + ageCat + diagYearCat +
                   pathMorph + pathStage + preOpTherapy + postOpTherapy +
                   (1 + hv.na | surgHospital), data = data.order,
                   family = binomial(link = "logit"))

# Coefficient
glmm.rs.hv <- round(exp(coef(summary(glmm.rs.0))[2, 1]), 4)
# Std.error and 95% CI
cc.glmm.rs.hv <- coef(summary(glmm.rs.0))[2, 2]
cc <- coef(summary(glmm.rs.0))
```

```
citab.glmm.rs.hv <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 2]),
                                upr = exp(cc[2, 1] + 1.96 * cc[2, 2])), 4)

#########################
# Yearly total
#########################

## Fitting IEE (GEE with independence correlation structure)
gee.in.1 <- gee(y.num ~ yearly.tot + ses + sex + ageCat + diagYearCat +
                pathMorph + pathStage + preOpTherapy + postOpTherapy,
                data = data.order, id = surgHospital,
                family = binomial(link = logit), corstr = "independence")

# Coefficient
iee.yt <- round(exp(coef(summary(gee.in.1))[2, 1]), 4)
# Std.error and 95% CI
cc.iee.yt <- coef(summary(gee.in.1))[2, 4]
cc <- coef(summary(gee.in.1))
citab.iee.yt <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 4]),
                            upr = exp(cc[2, 1] + 1.96 * cc[2, 4])), 4)

## Fitting GEE
gee.ex.1 <- gee(y.num ~ yearly.tot + ses + sex + ageCat + diagYearCat +
                pathMorph + pathStage + preOpTherapy + postOpTherapy,
                data = data.order, id = surgHospital,
                family = binomial(link = logit), corstr = "exchangeable")

#Coefficient
gee.yt <- round(exp(coef(summary(gee.ex.1))[2, 1]), 4)
#Std. error and 95% CI
cc.gee.yt <- coef(summary(gee.ex.1))[2, 4]
cc <- coef(summary(gee.ex.1))
citab.gee.yt <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 4]),
                            upr = exp(cc[2, 1] + 1.96 * cc[2, 4])), 4)

# Fitting GLMM-RI
glmm.ri.1 <- glmer(y.num ~ yearly.tot + ses + sex + ageCat + diagYearCat +
                   pathMorph + pathStage + preOpTherapy + postOpTherapy +
                   (1 | surgHospital), data = data.order,
                   family = binomial(link = "logit"))
```

```
# Coefficient
glmm.ri.yt <- round(exp(coef(summary(glmm.ri.1))[2, 1]), 4)
# Std.error and 95% CI
cc.glmm.ri.yt <- coef(summary(glmm.ri.1))[2, 2]
cc <- coef(summary(glmm.ri.1))
citab.glmm.ri.yt <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 2]),
                                upr = exp(cc[2, 1] + 1.96 * cc[2, 2])), 4)

## Fitting GLMM-RS
glmm.rs.1 <- glmer(y.num ~  yearly.tot + ses + sex + ageCat + diagYearCat +
                   pathMorph + pathStage + preOpTherapy + postOpTherapy +
                   (1 + yearly.tot | surgHospital), data = data.order,
                   family = binomial(link = "logit"))

# Coefficient
glmm.rs.yt <- round(exp(coef(summary(glmm.rs.1))[2, 1]), 4)
# Std.error and 95% CI
cc.glmm.rs.yt <- coef(summary(glmm.rs.1))[2, 2]
cc <- coef(summary(glmm.rs.1))
citab.glmm.rs.yt <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 2]),
                                upr = exp(cc[2, 1] + 1.96 * cc[2, 2])), 4)

#########################
# Cumulative yearly total
#########################

## Fitting IEE (GEE with independence correlation structure)
gee.in.2 <- gee(y.num ~ cum.yearly.tot + ses + sex + ageCat + diagYearCat +
                pathMorph + pathStage + preOpTherapy + postOpTherapy,
                data = data.order, id = surgHospital,
                family = binomial(link=logit), corstr = "independence")

# Coefficient
iee.cyt <- round(exp(coef(summary(gee.in.2))[2, 1]), 4)
# Std.error and 95% CI
cc.iee.cyt <- coef(summary(gee.in.2))[2, 4]
cc <- coef(summary(gee.in.2))
citab.iee.cyt <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 4]),
                             upr = exp(cc[2, 1] + 1.96 * cc[2, 4])), 4)

## Fitting GEE
```

```
gee.ex.2 <- gee(y.num ~ cum.yearly.tot + ses + sex + ageCat + diagYearCat +
                pathMorph + pathStage + preOpTherapy + postOpTherapy,
                data = data.order, id = surgHospital,
                family = binomial(link = logit), corstr = "exchangeable")

# Coefficient
gee.cyt <- round(exp(coef(summary(gee.ex.2))[2, 1]), 4)
# Std.error and 95% CI
cc.gee.cyt <- coef(summary(gee.ex.2))[2, 4]
cc <- coef(summary(gee.ex.2))
citab.gee.cyt <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 4]),
                             upr = exp(cc[2, 1] + 1.96 * cc[2, 4])), 4)

## Fitting GLMM-RI
glmm.ri.2 <- glmer(y.num ~ cum.yearly.tot + ses + sex + ageCat + diagYearCat +
                   pathMorph + pathStage + preOpTherapy + postOpTherapy +
                   (1 | surgHospital), data = data.order,
                   family = binomial(link = "logit"))

# Coefficient
glmm.ri.cyt <- round(exp(coef(summary(glmm.ri.2))[2, 1]), 4)
# Std.error and 95% CI
cc.glmm.ri.cyt <- coef(summary(glmm.ri.2))[2, 2]
cc <- coef(summary(glmm.ri.2))
citab.glmm.ri.cyt <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 2]),
                                 upr = exp(cc[2, 1] + 1.96 * cc[2, 2])), 4)

## Fitting GLMM-RS
glmm.rs.2 <- glmer(y.num ~  cum.yearly.tot + ses + sex + ageCat + diagYearCat +
                   pathMorph + pathStage + preOpTherapy + postOpTherapy +
                   (1 + cum.yearly.tot | surgHospital), data = data.order,
                   family = binomial(link = "logit"))

# Coefficient
glmm.rs.cyt <- round(exp(coef(summary(glmm.rs.2))[2, 1]), 4)
# Std.error and 95% CI
cc.glmm.rs.cyt <- coef(summary(glmm.rs.2))[2, 2]
cc <- coef(summary(glmm.rs.2))
citab.glmm.rs.cyt <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 2]),
                                 upr = exp(cc[2, 1] + 1.96 * cc[2, 2])), 4)
```

```
########################
# Running average
########################

## Fitting IEE (GEE with independence correlation structure)
gee.in.3 <- gee(y.num ~ run.ave + ses + sex + ageCat + diagYearCat +
                pathMorph + pathStage + preOpTherapy + postOpTherapy,
                data = data.order, id = surgHospital,
                family = binomial(link = logit), corstr = "independence")

# Coefficient
iee.ra <- round(exp(coef(summary(gee.in.3))[2, 1]), 4)
# Std.error and 95% CI
cc.iee.ra <- coef(summary(gee.in.3))[2, 4]
cc <- coef(summary(gee.in.3))
citab.iee.ra <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 4]),
                            upr = exp(cc[2, 1] + 1.96 * cc[2, 4])), 4)

## Fitting GEE
gee.ex.3 <- gee(y.num ~ run.ave + ses + sex + ageCat + diagYearCat +
                pathMorph + pathStage + preOpTherapy + postOpTherapy,
                data = data.order, id = surgHospital,
                family = binomial(link = logit), corstr = "exchangeable")

# Coefficient
gee.ra <- round(exp(coef(summary(gee.ex.3))[2, 1]), 4)
# Std.error and 95% CI
cc.gee.ra <- coef(summary(gee.ex.3))[2, 4]
cc <- coef(summary(gee.ex.3))
citab.gee.ra <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 4]),
                            upr = exp(cc[2, 1] + 1.96 * cc[2, 4])), 4)

## Fitting GLMM-RI
glmm.ri.3 <- glmer(y.num ~ run.ave + ses + sex + ageCat + diagYearCat +
                   pathMorph + pathStage + preOpTherapy + postOpTherapy +
                   (1 | surgHospital), data = data.order,
                   family = binomial(link = "logit"))

# Coefficient
glmm.ri.ra <- round(exp(coef(summary(glmm.ri.3))[2, 1]), 4)
# Std.error and 95% CI
```

```
cc.glmm.ri.ra <- coef(summary(glmm.ri.3))[2, 2]
cc <- coef(summary(glmm.ri.3))
citab.glmm.ri.ra <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 2]),
                                upr = exp(cc[2, 1] + 1.96 * cc[2, 2])), 4)

##Fitting GLMM-RS
glmm.rs.3 <- glmer(y.num ~  run.ave + ses + sex + ageCat + diagYearCat +
                   pathMorph + pathStage + preOpTherapy + postOpTherapy +
                   (1 + run.ave | surgHospital), data = data.order,
                   family = binomial(link = "logit"))

# Coefficient
glmm.rs.ra <- round(exp(coef(summary(glmm.rs.3))[2, 1]), 4)
# Std.error and 95% CI
cc.glmm.rs.ra <- coef(summary(glmm.rs.3))[2, 2]
cc <- coef(summary(glmm.rs.3))
citab.glmm.rs.ra <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 2]),
                                upr = exp(cc[2, 1] + 1.96 * cc[2, 2])), 4)

## Figure 5.6, 5.7, 5.8

y1   <- cbind(iee.hv,  gee.hv,  glmm.ri.hv,  glmm.rs.hv)
err1 <- cbind(cc.iee.hv,  cc.gee.hv,  cc.glmm.ri.hv,  cc.glmm.rs.hv)

y2   <- cbind(iee.yt,  gee.yt,  glmm.ri.yt,  glmm.rs.yt)
err2 <- cbind(cc.iee.yt,  cc.gee.yt,  cc.glmm.ri.yt,  cc.glmm.rs.yt)

y3   <- cbind(iee.cyt, gee.cyt, glmm.ri.cyt, glmm.rs.cyt,
            iee.ra,  gee.ra,  glmm.ri.ra,  glmm.rs.ra)
err3 <- cbind(cc.iee.cyt, cc.gee.cyt, cc.glmm.ri.cyt, cc.glmm.rs.cyt,
              cc.iee.ra,  cc.gee.ra,  cc.glmm.ri.ra,  cc.glmm.rs.ra)

forest(x      = log(y1),
       sei    = err1,
       digits = 2,
       transf = exp,
       xlab   = "Estimated odds-ratio",
       slab   = c("IEE", "GEE", "GLMM_RI", "GLMM_RS"))
forest(x      = log(y2),
       sei    = err2,
       digits = 2,
```

```
       transf = exp,
       xlab   = "Estimated odds-ratio",
       slab   = c("IEE", "GEE", "GLMM_RI", "GLMM_RS"))
forest(x      = log(y3),
       sei    = err3,
       digits = 2,
       transf = exp,
       xlab   = "Estimated odds-ratio",
       slab   = c("IEE", "GEE", "GLMM_RI", "GLMM_RS", "IEE", "GEE", "GLMM_RI", "GLMM_RS"))
```

12.4 R code Chapter 6

```
#########################
# Evaluating assumption (1)
#########################

# Preparing new dataset

data.order2 <- data.order
# Indicate when moving to a new hospital
myind.hospital <- (c(0, data.order2$surgHospital[2:nrow(data.order2)] -
                     data.order2$surgHospital[1:(nrow(data.order2) - 1)]))
data.order2 <- cbind(data.order2, myind.hospital)

dat.dif <- c(data.order2$surgDate[2:nrow(data.order2)] -
               data.order2$surgDate[1:(nrow(data.order2) - 1)], NA)
data.order2 <- cbind(data.order2, dat.dif)

data.order2$dat.dif <- ifelse(data.order2$dat.dif < 0, NA, data.order2$dat.dif)

# Evaluating assumption (1)
# Cox regression model:
# regressing time between successive surgeries on previous patient outcome
# adjusting for patient characteristics
# previous patient outcome defined as outcome in previous year in same hospital
coxph.model.a1 <- coxph(Surv(dat.dif, as.numeric(surgery)) ~
                          y.num + hv.na + ses + sex + ageCat + diagYearCat +
                          pathMorph + pathStage + preOpTherapy +
                          postOpTherapy, cluster(surgHospital),
```

```
                          data = data.order2)

# Coefficient
round(coef(summary(coxph.model.a1))[1, 2], 4)
# 95% CI
cc <- coef(summary(coxph.model.a1))
citab.coxph.model.a1 <- round(cbind(lwr = (cc[1, 2] - 1.96 * cc[1, 3]),
                                    upr = (cc[1, 2] + 1.96 * cc[1, 3])), 4)

# The hazard of a subsequent surgery among hospitals with a previous death is
# 1 - 0.8798 = 0.1202 -> 12 % lower. Thus suggesting that assumption 1 is violated.
```

12.5 R code Chapter 7

```
# Randomly sample one observation and it's corresponding
# covariate vector from each cluster with replacement

# Data preparation
temp7 <- count(data.order, "surgHospital")[, 2]
temp8 <- cumsum(temp7)
temp9 <- c(1, temp8 + 1)
x <- NULL

# Clusters
I <- length(unique(data.order$surgHospital)) #I = 148 clusters
Q <- 10000
# Generate Q datasets of size I = 148

##########################
# Non-aggregate cumulative hospital volume
##########################

rownumber <- numeric(148)
vec.coef <- matrix(0, Q, 1)
vec.st.error <- matrix(0, Q, 1)

set.seed(123)

for (q in 1:Q) {
```

```
  for (i in 1:I) {
    x <- temp9[i]:temp8[i] # row number to choose from per cluster
    rownumber[i] <- sample(x, size = 1)
  }

  # Analyse resampled datset by a GLM
  dat.set <- data.order[rownumber, ]
  glm.6M <- glm(y.num ~ hv.na + ses + sex + ageCat + diagYearCat +
                pathMorph + pathStage + preOpTherapy,
                data = dat.set, family = binomial(link=logit))

  vec.coef[q, ] <- exp(coef(summary(glm.6M))[2, 1])
  vec.st.error[q, ] <- coef(summary(glm.6M))[2, 2]

}

# WCR parameter
B.WCR.0 <- mean(vec.coef)

# WCR variance estimator
S.B.sig <- (sum((vec.coef - B.WCR.0) * (vec.coef - B.WCR.0)))/(Q - 1)
sig.Q <- sum(vec.st.error ^ 2)
Var.0 <- sig.Q / Q - ((Q - 1) / Q) * S.B.sig

# Results
round(B.WCR.0, 4)
round(Var.0, 4)

#########################
# Yearly total
#########################

set.seed(123)

for (q in 1:Q) {

  for (i in 1:I) {
    x <- temp9[i]:temp8[i] #row number to choose from per cluster
    rownumber[i] <- sample(x, size = 1)
  }
```

```
  # Analyse resampled datset by a GLM
  dat.set <- data.order[rownumber, ]
  glm.6M <- glm(y.num ~ yearly.tot + ses + sex + ageCat + diagYearCat +
                pathMorph + pathStage + preOpTherapy,
                data = dat.set, family = binomial(link = logit))

  vec.coef[q, ] <- exp(coef(summary(glm.6M))[2, 1])
  vec.st.error[q, ] <- coef(summary(glm.6M))[2, 2]

}

# WCR parameter
B.WCR.1 <- mean(vec.coef)

# WCR variance estimator
S.B.sig <- (sum((vec.coef - B.WCR.1) * (vec.coef - B.WCR.1)))/(Q - 1)
sig.Q <- sum(vec.st.error ^ 2)
Var.1 <- sig.Q / Q - ((Q - 1) / Q) * S.B.sig

# Results
round(B.WCR.1, 4)
round(Var.1, 4)

#########################
# Cumulative yearly total
#########################

set.seed(123)

for (q in 1:Q) {

  for (i in 1:I) {
    x <- temp9[i]:temp8[i] # row number to choose from per cluster
    rownumber[i] <- sample(x, size=1)
  }

  # Analyse resampled datset by a GLM
  dat.set <- data.order[rownumber, ]
  glm.6M <- glm(y.num ~ cum.yearly.tot + ses + sex + ageCat + diagYearCat +
                pathMorph + pathStage + preOpTherapy,
```

```
                data = dat.set, family = binomial(link = logit))

  vec.coef[q,] <- exp(coef(summary(glm.6M))[2,1])
  vec.st.error[q, ] <- coef(summary(glm.6M))[2, 2]

}

# WCR parameter
B.WCR.2 <- mean(vec.coef)

# WCR variance estimator
S.B.sig <- (sum((vec.coef - B.WCR.2) * (vec.coef - B.WCR.2)))/(Q - 1)
sig.Q <- sum(vec.st.error ^ 2)
Var.2 <- sig.Q / Q - ((Q - 1) / Q) * S.B.sig

# Results
round(B.WCR.2, 4)
round(Var.2, 4)

#########################
# Running average
#########################

set.seed(123)

for (q in 1:Q) {

  for (i in 1:I) {
    x <- temp9[i]:temp8[i] #row number to choose from per cluster
    rownumber[i] <- sample(x, size = 1)
  }

  # Analyse resampled dataset by a GLM
  dat.set <- data.order[rownumber, ]
  glm.6M <- glm(y.num ~ run.ave + ses + sex + ageCat + diagYearCat +
                pathMorph + pathStage + preOpTherapy,
                data = dat.set, family = binomial(link = logit))

  vec.coef[q, ] <- exp(coef(summary(glm.6M))[2, 1])
  vec.st.error[q, ] <- coef(summary(glm.6M))[2, 2]
```

```
}

# WCR parameter
B.WCR.3 <- mean(vec.coef)

# WCR variance estimator
S.B.sig <- (sum((vec.coef - B.WCR.3) * (vec.coef - B.WCR.3)))/(Q - 1)
sig.Q <- sum(vec.st.error ^ 2)
Var.3 <- sig.Q / Q - ((Q - 1) / Q) * S.B.sig

# Results
round(B.WCR.3, 4)
round(Var.3, 4)
```

12.6 R code Chapter 8

```
# See Appendix A for simulation algorithm

## Figure 8.2, 8.3
B2.1 <- rep(0.05, 5)
se.1 <- c(0.0038, 0.0035, 0.0006, 0.0034, 0.0037)
forest(x       = B2.1,
       sei     = se.1,
       digits  = 3,
       refline = 0.05,
       xlab    = "Point estimate of regression parameter",
       slab    = c("IEE", "GEE", "GLMM_RI", "GLMM_RS", "WCR"))

B2.2 <- c(0.045, 0.042, 0.042, 0.037, 0.046)
se.2 <- c(0.0086, 0.0070, 0.0007, 0.0045, 0.0083)
forest(x       = B2.2,
       sei     = se.2,
       digits  = 3,
       refline = 0.050,
       xlab    = "Point estimate of regression parameter",
       slab    = c("IEE", "GEE", "GLMM_RI", "GLMM_RS", "WCR"))
```

12.7 R code Chapter 9

```
# Categorical variable
temp.cat <- count(data.order, "surgHospital")
total.hv <- rep(temp.cat[, 2], temp.cat[, 2])
data.order$total.hv <- total.hv
quantile(total.hv)

# Two levels
data.order$hv.cat2 <- ifelse(total.hv > 146, ">146", "1-146")
data.order$hv.cat2 <- factor(data.order$hv.cat2, levels = c("1-146", ">146"))

# Four levels
data.order$hv.cat4 <- ifelse(total.hv > 251, ">251", total.hv)
data.order$hv.cat4 <- ifelse((total.hv > 146 & total.hv <= 251), "146-251",
                             data.order$hv.cat4)
data.order$hv.cat4 <- ifelse((total.hv > 65 & total.hv <= 146), "65-146",
                             data.order$hv.cat4)
data.order$hv.cat4 <- ifelse((total.hv <= 65), "<65", data.order$hv.cat4)

# Variable for ranking
rankings <- rank(temp.cat[, 2]) #ties are averaged
data.order$rank <- rep(rankings,temp.cat[, 2])
data.order$rank.10 <- data.order$rank / 10


#########################
#Fitting models
#########################

gee.in.cat2 <- gee(y.num ~ hv.cat2 + ses + sex + ageCat + diagYearCat +
                   pathMorph + pathStage + preOpTherapy + postOpTherapy,
                   data = data.order, id = surgHospital,
                   family = binomial(link = logit), corstr = "independence")

gee.in.cat4 <- gee(y.num ~ hv.cat4  + ses + sex + ageCat + diagYearCat +
                   pathMorph + pathStage + preOpTherapy + postOpTherapy,
                   data = data.order, id = surgHospital,
                   family = binomial(link = logit), corstr = "independence")
```

```
glm.cross <- glm(y.num ~ hv.na + ses + sex + ageCat + diagYearCat +
                 pathMorph + pathStage + preOpTherapy + postOpTherapy,
                 data = data.order, family = binomial(link = logit))

gee.in.rank <- gee(y.num ~ rank  + ses + sex + ageCat + diagYearCat +
                   pathMorph + pathStage + preOpTherapy + ostOpTherapy,
                   data = data.order, id = surgHospital,
                   family = binomial(link = logit), corstr = "independence")

gee.in.rank.10 <- gee(y.num ~ rank.10 + ses + sex + ageCat + diagYearCat +
                      pathMorph + pathStage + preOpTherapy + postOpTherapy,
                      data = data.order, id = surgHospital,
                      family = binomial(link = logit), corstr = "independence")


#########################
# Coefficients and standard errors for forest plots
#########################

cat2.coef <- exp(coef(summary(gee.in.cat2))[2, 1])
cc.cat2.coef <- coef(summary(gee.in.cat2))[2, 4]

cat4.1 <- round(exp(coef(summary(gee.in.cat4))[2, 1]), 4)
cat4.2 <- round(exp(coef(summary(gee.in.cat4))[3, 1]), 4)
cat4.3 <- round(exp(coef(summary(gee.in.cat4))[4, 1]), 4)

cc.cat4.1 <- coef(summary(gee.in.cat4))[2, 4]
cc.cat4.2 <- coef(summary(gee.in.cat4))[3, 4]
cc.cat4.3 <- coef(summary(gee.in.cat4))[4, 4]

cross <- exp(coef(summary(glm.cross)))[2, 1]
cc.cross <- coef(summary(glm.cross))[2, 2]

rank <- round(exp(coef(summary(gee.in.rank.10))[2, 1]), 4)
cc.rank <- coef(summary(gee.in.rank.10))[2, 4]

## Figure 9.1

coef <- c(iee.hv, cross, rank)
se <- c(cc.iee.hv, cc.cross, cc.rank)
forest(x       = log(coef),
```

```
       sei     = se,
       refline = NA,
       digits  = 4,
       transf  = exp,
       xlab    = "Odds-ratio",
       slab    = c("Correct", "Cross-sectional", "Ranking"))

coef2 <- c(iee.hv, cat2.coef)
se2 <- c(cc.iee.hv, cc.cat2.coef)
forest(x = log(coef2), sei = se2, refline = NA, transf = exp,
       xlab = "Odds-ratio", slab = c("Correct", "Two categories"))

# Std.error and 95% CI
cc.cat2 <- coef(summary(gee.in.cat2))[2, 1]
cc <- coef(summary(gee.in.cat2))
citab.cat2 <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 4]),
                          upr = exp(cc[2, 1] + 1.96 * cc[2, 4])), 4)

# Std.error and 95% CI
cc.cat4 <- coef(summary(gee.in.cat4))[2:4, 1]
cc <- coef(summary(gee.in.cat4))
citab.cat4 <- round(cbind(lwr = exp(cc[2:4, 1] - 1.96 * cc[2:4, 4]),
                          upr = exp(cc[2:4, 1] + 1.96 * cc[2:4, 4])), 4)

# Std.error and 95% CI
cc.rank <- coef(summary(gee.in.rank.10))[2, 1]
cc <- coef(summary(gee.in.rank.10))
citab.rank.10 <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 4]),
                             upr = exp(cc[2, 1] + 1.96 * cc[2, 4])), 4)

# Std.error and 95% CI
cc.glm.cross <- coef(summary(glm.cross))[2, 1]
cc <- coef(summary(glm.cross))
citab.glm.cross <- round(cbind(lwr = exp(cc[2, 1] - 1.96 * cc[2, 2]),
                               upr = exp(cc[2, 1] + 1.96 * cc[2, 2])), 4)
```

References

[1] Baddeley A., Bárány, I., Schneider, R., & Weil, W. (2004). *Stochastic geometry* (Vol. 1892). Berlin Heidelberg: Springer.

[2] Baddeley A., & Turner R. (2005). Spatstat: an R package for analysing spatial point patterns. *Journal of Statistical Software*, 12(6):142. URL: www.jstatsoft.org, ISSN: 1548-7660

[3] Begg, C. B., Cramer, L. D., Hoskins, W. J., & Brennan, M. F. (1998). Impact of hospital volume on operative mortality for major cancer surgery. *Jama*, 280(20), 1747-1751.

[4] Benhin, E., Rao, J. N. K., & Scott, A. J. (2005). Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes. *Biometrika*, 92(2), 435-450.

[5] Daley, D. J., & Vere-Jones, D. (1988). *An introduction to the theory of point processes* (Vol. 2). New York: Springer.

[6] Davoli, M., Amato, L., Minozzi, S., Bargagli, A. M., Vecchi, S., & Perucci, C. A. (2004). Volume and health outcomes: an overview of systematic reviews]. *Epidemiologia e prevenzione*, 29(3-4), 3-63.

[7] Dikken, J. L., Dassen, A. E., Lemmens, V. E., Putter, H., Krijnen, P., van der Geest, L., & Wouters, M. W. (2012). Effect of hospital volume on postoperative mortality and survival after oesophageal and gastric cancer surgery in the Netherlands between 1989 and 2009. *European Journal of Cancer*, 48(7), 1004-1013.

[8] Dunson, D. B., Chen, Z., & Harry, J. (2003). A Bayesian Approach for Joint Modeling of Cluster Size and SubunitSpecific Outcomes. *Biometrics*, 59(3), 521-530.

[9] French, B., Farjah, F., Flum, D. R., & Heagerty, P. J. (2012). A general framework for estimating volume-outcome associations from longitudinal data. *Statistics in medicine*, 31(4), 366-382.

[10] French, B., & Heagerty, P. J. (2009). Marginal mark regression analysis of recurrent marked point process data. *Biometrics*, 65(2), 415-422.

[11] Fritz, A. G. (Ed.). (2000). International classification of diseases for oncology: ICD-O. World Health Organization.

[12] Ghisletta, P., & Spini, D. (2004). An introduction to generalized estimating equations and an application to assess selectivity effects in a longitudinal study on very old individuals. *Journal of Educational and Behavioral Statistics*, 29(4), 421-437.

[13] Halekoh, U., Hjsgaard, S., & Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software*, 15(2), 1-11.

[14] Hannan, E. L., Wu, C., Walford, G., King, S. B., Holmes, D. R., Ambrose, J. A., Sharma, S., Katz, S., Clark, L. T., & Jones, R. H. (2005). Volume-outcome relationships for percutaneous coronary interventions in the stent era. *Circulation*, 112(8), 1171-1179.

[15] Hoffman, E. B., Sen, P. K., & Weinberg, C. R. (2001). Within-cluster resampling. *Biometrika*, 88(4), 1121-1134.

[16] Huang, Y., & Wang, M. C. (2003). Frequency of recurrent events at failure time: modelling and inference. *Journal of the American Statistical Association*, 98(463), 663-670.

[17] Kulkarni, G. S., Laupacis, A., Urbach, D. R., Fleshner, N. E., & Austin, P. C. (2009). Varied definitions of hospital volume did not alter the conclusions of volume-outcome analyses. *Journal of clinical epidemiology*, 62(4), 400-407.

[18] Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13-22.

[19] Livingston, E. H., Elliott, A. C., Hynan, L. S., & Engel, E. (2007). When policy meets statistics: the very real effect that questionable statistical analysis has on limiting health care access for bariatric surgery. *Archives of Surgery*, 142(10), 979-987.

[20] McCulloch, C. E., & Neuhaus, J. M. (2001). *Generalized linear mixed models*. John Wiley & Sons, Ltd.

[21] Panageas, K. S., Schrag, D., Riedel, E., Bach, P. B., & Begg, C. B. (2003). The effect of clustering of outcomes on the association of procedure volume and surgical outcomes. *Annals of internal medicine*, 139(8), 658-665.

[22] Panageas, K. S., Schrag, D., Russell Localio, A., Venkatraman, E. S., & Begg, C. B. (2007). Properties of analysis methods that account for clustering in volume-outcome studies when the primary predictor is cluster size. *Statistics in medicine*, 26(9), 2017-2035.

[23] Pepe, M. S., & Couper, D. (1997). Modelling partly conditional means with longitudinal data. *Journal of the American Statistical Association*, 92(439), 991-998.

[24] Schouten, L. J., Jager, J. J., & Van den Brandt, P. A. (1993). Quality of cancer registry data: a comparison of data provided by clinicians with those of registration personnel. *British journal of cancer*, 68(5), 974.

[25] Schrag, D., Panageas, K. S., Riedel, E., Cramer, L. D., Guillem, J. G., Bach, P. B., & Begg, C. B. (2002). Hospital and surgeon procedure volume as predictors of outcome following rectal cancer resection. *Annals of surgery*, 236(5), 583.

[26] Sollano, J. A., Gelijns, A. C., Moskowitz, A. J., Heitjan, D. F., Cullinane, S., Saha, T., Chen, J. M., Roohan, P. J., Reemstra, K. & Shields, E. P. (1999). Volume-outcome relationships in cardiovascular operations: New York State, 1990-1995. *The Journal of thoracic and cardiovascular surgery*, 117(3), 419-430.

[27] Wen, H. C., Tang, C. H., Lin, H. C., Tsai, C. S., Chen, C. S., & Li, C. Y. (2006). Association between surgeon and hospital volume in coronary artery bypass graft surgery outcomes: a population-based study. *The Annals of thoracic surgery*, 81(3), 835-842.

[28] Williamson, J. M., Datta, S., & Satten, G. A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics*, 59(1), 36-42.

[29] Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 1049-1060.