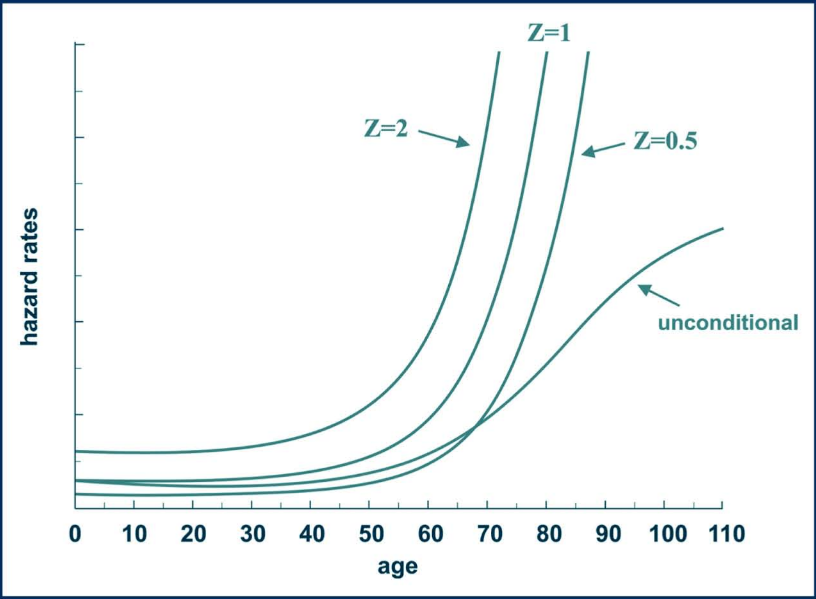
Chapman & Hall/CRC Biostatistics Series

Frailty Models in Survival Analysis



Andreas Wienke

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Frailty Models in Survival Analysis

Chapman & Hall/CRC Biostatistics Series

Chapman & Hall/CRC Biostatistics Series

Published Titles

1. *Design and Analysis of Animal Studies in Pharmaceutical Development,* Shein-Chung Chow and Jen-pei Liu
2. *Basic Statistics and Pharmaceutical Statistical Applications,* James E. De Muth
3. *Design and Analysis of Bioavailability and Bioequivalence Studies, Second Edition, Revised and Expanded,* Shein-Chung Chow and Jen-pei Liu
4. *Meta-Analysis in Medicine and Health Policy,* Dalene K. Stangl and Donald A. Berry
5. *Generalized Linear Models: A Bayesian Perspective,* Dipak K. Dey, Sujit K. Ghosh, and Bani K. Mallick
6. *Difference Equations with Public Health Applications,* Lemuel A. Moyé and Asha Seth Kapadia
7. *Medical Biostatistics,* Abhaya Indrayan and Sanjeev B. Sarmukaddam
8. *Statistical Methods for Clinical Trials,* Mark X. Norleans
9. *Causal Analysis in Biomedicine and Epidemiology: Based on Minimal Sufficient Causation,* Mikel Aickin
10. *Statistics in Drug Research: Methodologies and Recent Developments,* Shein-Chung Chow and Jun Shao
11. *Sample Size Calculations in Clinical Research,* Shein-Chung Chow, Jun Shao, and Hansheng Wang
12. *Applied Statistical Design for the Researcher,* Daryl S. Paulson
13. *Advances in Clinical Trial Biostatistics,* Nancy L. Geller
14. *Statistics in the Pharmaceutical Industry, Third Edition,* Ralph Buncher and Jia-Yeong Tsay
15. *DNA Microarrays and Related Genomics Techniques: Design, Analysis, and Interpretation of Experiments,* David B. Allsion, Grier P. Page, T. Mark Beasley, and Jode W. Edwards
16. *Basic Statistics and Pharmaceutical Statistical Applications, Second Edition,* James E. De Muth
17. *Adaptive Design Methods in Clinical Trials,* Shein-Chung Chow and Mark Chang
18. *Handbook of Regression and Modeling: Applications for the Clinical and Pharmaceutical Industries,* Daryl S. Paulson
19. *Statistical Design and Analysis of Stability Studies,* Shein-Chung Chow
20. *Sample Size Calculations in Clinical Research, Second Edition,* Shein-Chung Chow, Jun Shao, and Hansheng Wang
21. *Elementary Bayesian Biostatistics,* Lemuel A. Moyé
22. *Adaptive Design Theory and Implementation Using SAS and R,* Mark Chang
23. *Computational Pharmacokinetics,* Anders Källén
24. *Computational Methods in Biomedical Research,* Ravindra Khattree and Dayanand N. Naik
25. *Medical Biostatistics, Second Edition,* A. Indrayan
26. *DNA Methylation Microarrays: Experimental Design and Statistical Analysis,* Sun-Chong Wang and Arturas Petronis
27. *Design and Analysis of Bioavailability and Bioequivalence Studies, Third Edition,* Shein-Chung Chow and Jen-pei Liu
28. *Translational Medicine: Strategies and Statistical Methods,* Dennis Cosmatos and Shein-Chung Chow
29. *Bayesian Methods for Measures of Agreement,* Lyle D. Broemeling
30. *Data and Safety Monitoring Committees in Clinical Trials,* Jay Herson
31. *Design and Analysis of Clinical Trials with Time-to-Event Endpoints,* Karl E. Peace
32. *Bayesian Missing Data Problems: EM, Data Augmentation and Noniterative Computation,* Ming T. Tan, Guo-Liang Tian, and Kai Wang Ng
33. *Multiple Testing Problems in Pharmaceutical Statistics,* Alex Dmitrienko, Ajit C. Tamhane, and Frank Bretz
34. *Bayesian Modeling in Bioinformatics,* Dipak K. Dey, Samiran Ghosh, and Bani K. Mallick
35. *Clinical Trial Methodology,* Karl E. Peace and Ding-Geng (Din) Chen
36. *Monte Carlo Simulation for the Pharmaceutical Industry: Concepts, Algorithms, and Case Studies,* Mark Chang
37. *Frailty Models in Survival Analysis,* Andreas Wienke

Chapman & Hall/CRC Biostatistics Series

Frailty Models in Survival Analysis

Andreas Wienke
Institute of Medical Epidemiology, Biostatistics, and Informatics
Martin-Luther-University Halle-Wittenberg, Germany

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20140514

International Standard Book Number-13: 978-1-4200-7391-1 (eBook - PDF)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

All models are wrong, some are useful. (Box 1976)

Contents

List of Tables

List of Figures

Symbol Description

ACE	genetic model
AFT	accelerated failure time
AIC	Akaike information criterion
AR	autoregressive process
BMI	body mass index
BLUB	best linear unbiased predictor
CAD	coronary artery disease
CSRF	corrected scale reduction factor
CHD	coronary heart disease
cdf	cumulative density function
DIC	Bayesian information criterion
DZ	dizygotic
E	expectation
EBCT	electron-beam computed tomography
ECOG	Eastern Cooperative Oncology Group
EM	expectation-maximization
f	generic symbol for pdf
F	generic symbol for cdf
Γ	gamma function
H_0	null hypothesis
H_A	alternative hypothesis
ICD	International Classification of Diseases
iid	independent and identically distributed
L	Laplace transform
L	likelihood function
μ	generic symbol for a hazard
μ_0	generic symbol for a baseline hazard
M	generic symbol for a cumulative hazard
M_0	generic symbol for a cumulative baseline hazard
$\log N(m, s^2)$	log-normal distribution with parameters m, s^2
MCEM	Markov Chain EM
MCMC	Markov Chain Monte Carlo
ML	maximum likelihood
MZ	monozygotic
NSCLC	non-small cell lung carcinoma
P(A)	probability of event A
pdf	probability density function
PH	proportional hazards
PPL	penalized partial likelihood
PVF	power variance function
S	generic symbol for a survival function
t^+	truncation time
TNM	classification of malignant tumours
V	variance
$U(a, b)$	uniform distribution in the interval $[a, b]$
$N(\mu, \sigma^2)$	normal distribution with parameters μ, σ^2
$Exp(\lambda)$	exponential distribution with parameter λ
$W(\lambda, \nu)$	Weibull distribution with parameters λ, ν
$G(\lambda, \varphi)$	Gompertz distribution with parameters λ, φ
$\Gamma(k, \lambda)$	gamma distribution with parameters k, λ
$\log L(\nu, \kappa)$	log-logistic distribution with parameters ν, κ
$Ps(\alpha)$	positive stable distribution with parameter α
$cP(\gamma, k, \lambda)$	compound Poisson distribution with parameters γ, k, λ

Preface

The analysis of lifetime data (or more exactly, time-to-event, event-history, or duration data) plays an important role in medicine, epidemiology, biology, demography, economics, engineering, actuarial science, and other fields. It has expanded rapidly in the last three decades, with works having been published in various disciplines in addition to statistics. But what distinguishes survival analysis from other fields of statistics? Why does survival data need a special statistical theory? The main problem is censoring, which means that, for some individuals in the study population, the researcher only has the information that the event of interest did not occur before a particular time point. To put it plainly, a censored observation contains only partial information about the random variable of interest. This kind of incomplete observation needs special methods. As a consequence of censoring, survival times are usually a mixture of discrete (censoring indicator) and continuous (event/censoring time) data that lend themselves to a different type of analysis from that used in the traditional discrete or continuous case. The mixture is the result of censoring and has an important effect on data analysis. The Kaplan–Meier estimator (Kaplan and Meier 1958) of the survival function is a major step in the development of suitable models for such kind of data. Furthermore, most evaluations are made conditionally on what is known at the time of the analysis, and this changes over time. Usually, as the population under study is changing, we only consider the individual risk to die for those who are still alive, but this means that many standard statistical approaches cannot be applied.

Models based on the hazard function have dominated survival analysis since the construction of the proportional hazards model by Cox (1972). One of the reasons this model is so popular is the ease with which technical difficulties such as censoring and truncation are handled. This is due to the appealing interpretation of the hazard as a risk that changes over time. Naturally, the concept allows for the entering of covariates in order to describe their influence and to model different levels of risk for different subgroups.

This book focuses on frailty models, a specific area in survival analysis. The concept of frailty provides a convenient way of introducing unobserved heterogeneity and associations into models for survival data. In its simplest form, frailty is an unobserved random proportionality factor that modifies the hazard function of an individual or related individuals. In essence, the concept goes back to the work of Greenwood and Yule (1920) on "accident proneness" with binary data. The first univariate frailty model was suggested by Beard (1959), considering different mortality models. The term *frailty* itself was introduced by Vaupel et al. (1979) in the univariate context. Its first applications to problems in multivariate survival analysis date from a seminal paper by Clayton (1978).

Ordinary methods in survival analysis implicitly assume that populations are homogenous, meaning that all individuals have the same risk of death. However, in general, it is impossible to include all relevant risk factors, perhaps because we have no information on individual values, which is often the case in demography. Furthermore, we may not know all relevant risk factors, or it is impossible to measure them without great financial costs, something that is common in medical and biological studies. The neglect of covariates leads to unobserved heterogeneity. That is, the population consists of subjects with different risks. As a consequence, it is important to consider the population as heterogeneous, i.e., as a mixture of individuals with different hazards. A frailty model is a random effects model for time-to-event data, where the random effect (frailty) has a multiplicative effect on the baseline hazard. It can be used for univariate (independent) lifetimes, i.e., to adjust for unobserved risk factors in a proportional hazards model (heterogeneity). The variability of event-time data is split into one part that depends on covariates and is thus theoretically predictable, and one part that is initially unpredictable, even knowing all relevant information at that time. There are advantages in separating these sources of variability: unobserved heterogeneity can explain some unexpected results or give an alternative interpretation, for example, crossing-over or leveling-off effects of hazards.

However, considering multivariate (dependent) duration times is especially interesting. The introduction of a common random effect – frailty – is a natural way of modeling the dependence of event times. The random effect explains the dependence in the sense that had we known the frailty, the event times would have been independent. In other words, because we do not know the frailty, the lifetimes are independent conditionally on the frailty. This approach can be used for survival times of related individuals such as twins or family members, where independence cannot be assumed, or for recurrent events in the same individual or for times to several events for the same individual, such as onset of different diseases, relapse, or death (competing risks). Different extensions of univariate frailty models to multivariate models are possible and will be considered in this book.

The standard assumption is to use a gamma distribution for frailty, but other distributions are also possible. The relationships between individual and observed survival characteristics play a key role in the statistical analysis of duration data in heterogeneous populations.

Various frailty models have been developed in the past. However, compared with standard mixed models, frailty models pose additional difficulties in developing inferential methods, caused by incomplete data due to censoring and truncation. Thus, inferential methods have been less developed here than in other mixed models.

To keep the book to a reasonable length, some topics are discussed only briefly, and references are given for further reading. Because the literature on frailty models is extensive (especially in the last few years), the choice of subject matter is difficult. The material discussed in detail is to some extent a reflection of the author's interest in this research field. However, my attempt has been to present a relatively comprehensive and complete overview of the fundamental approaches in the field of frailty models.

The present monograph is primarily aimed at the biostatistical community with applications from biomedicine, (genetic) epidemiology, and demography. Some efforts were also undertaken to include literature from other fields like econometrics if interesting methodological problems are raised. The practical use of models is a key issue in biostatistics, where the data at hand often are motivating for the development of new models. The language of this book is nontechnical and therefore it can be understood by nonspecialists. Nevertheless, some experience with survival analysis is an advantage.

Acknowledgments

I would like to express my sincere thanks to everyone who supported directly or indirectly this book; my former Ph.D. advisor Friedrich Liese (Rostock) for awakening my interest in statistical research, Anatoli Yashin and Konstantin Arbeev (Duke), Paul Janssen and Niel Hens (Hasselt), Catherine Legrand (Louvain-la-Neuve), Alexander Begun (Hamburg), Nicole Giard (Gütersloh), Kaare Christensen and Ivan Iachine (Odense), Isabella Locatelli (Lausanne), Slobodan Zdravkovic (Kopenhagen), Samuli Ripatti (Helsinki), Oliver Kuß (Halle), and Juni Palmgren (Stockholm) for many helpful discussions and fruitful collaboration. I take the opportunity to thank all my colleagues at the Institute for Medical Epidemiology, Biostatistics, and Informatics at the Medical Faculty of the Martin Luther University Halle-Wittenberg, especially Johannes Haerting as head of the institute, for providing an excellent research environment. I would like to thank the Danish and Swedish Twin Registries for making the unique twin data available. I extend my sincere thanks to the Max Planck Institute for Demographic Research (Rostock) and its founding director, James Vaupel, where I began to work on frailty models and where I returned several times as a guest researcher during the last years. Special thanks go to Luc Duchateau (Ghent) for careful reading an earlier version of the manuscript and giving many very helpful suggestions. Furthermore, I would like to thank my Ph.D. students Katharina Hirsch and Diana Pietzner (Halle) for many fruitful discussions and help in preparing the manuscript.

Finally, acknowledgments go to my family. To my parents Margit and Kurt Wienke for all their love and support. Also, to my sons Moritz, Jakob, and Finn, showing that there is a life beyond research and, finally, to my loving wife Kati Moeller, who always strongly encouraged me to finish this book.

Halle, June 2010 Andreas Wienke

Chapter 1

Introduction

1.1 Goals and Outline

Survival analysis is one of the core research methods used in many fields such as medicine, biology, epidemiology, demography, and engineering. Notion survival analysis reflects the origin of the methods in medical and demographic studies of mortality. Especially since the end of the 1970s, the empirical analysis of event history data has become widespread by the development of the proportional hazards model in the seminal paper by Cox (1972) and several extensions during the last three decades. The present monograph deals with one important direction of extensions in this field, namely frailty models. A frailty model is a multiplicative hazard model consisting of three components: a frailty (random effect), a baseline hazard function (parametric or nonparametric), and a term modeling the influence of observed covariates (fixed effects).

Only a few books exist on this subject, which contain short chapters devoted to frailty models. Ibrahim et al. (2001) consider parametric as well as semiparametric shared frailty models based on a Bayesian approach. Klein and Moeschberger (2003) consider the application of the EM algorithm in semiparametric shared frailty models. Aalen et al. (2008) take a process point of view dealing with different frailty models. The present book extends in two main directions the presentation of frailty models made in the seminal monographs by Hougaard (2000), Therneau and Grambsch (2000), and Duchateau and Janssen (2008). First, univariate frailty models with their focus on unobserved heterogeneity are covered in more detail compared to previous books. In univariate models all durations describe the time to the same type of event, and event times are considered as independent. Second, the main emphasis is placed on correlated frailty models as natural extensions of shared frailty models. Here, different strengths of association between clustered lifetimes are of special interest.

One of the main problems in the application of frailty models to real data is the limited availability of standard software in this area. Consequently, one aim of this monograph is to show which of the models considered can be applied to real data by using standard statistical packages such as R, SAS, and STATA. Here, the link to generalized linear mixed models will be exploited. Both parametric as well as semiparametric models are considered. Furthermore, models are fitted by the frequentist and Bayesian approaches. Most of the Bayesian analyses are performed with WinBUGS, but the new PROC MCMC in SAS opens new possibilities for the future.

In this first chapter we will introduce different data sets used throughout the book to illustrate modeling techniques and practical interpretations of the results. In Chapter 2 an introduction to basic and general concepts in survival analysis and a definition of common terminology are given. The topic is also covered by other books, for example, Miller (1981), Cox and Oakes (1984), Andersen et al. (1993), Lawless (2002), Kalbfleisch and Prentice (2002), Klein and Moeschberger (2003), Collett (2003), and Machin, Cheung, and Parmar (2006). The recent book by Finkelstein (2008) has a special focus on reliability but also covers a wide range of exiting topics in biostatistics and demography. Though, it is not necessary for people acquainted with this field to read it, it does contain notations and key results and lays the basis for the more advanced frailty models treated in the following chapters. After this preparatory chapter we deal with univariate frailty models (single spell data) in Chapter 3, discussing the broad range of possible frailty distributions with their specific features. Gamma distribution is the most often applied frailty distribution because frailties appearing in conditional likelihood can be integrated out, giving simple expressions of unconditional likelihood. Then, maximization of unconditional likelihood can be used for estimation. Here, the interpretation of frailty is as unobserved heterogeneity due to nonobserved covariates. The focus of Chapter 4 is on the shared frailty model, which has already been discussed in detail by other authors (Hougaard 2000, Therneau and Grambsch 2000, Duchateau and Janssen 2008). Shared frailty models are an important tool for analyzing multivariate (clustered) survival data. Hence, this chapter forms the basis of the correlated frailty model and its extensions considered in detail in Chapter 5. Advantages and limitations of the proposed models are discussed, and simulations show the properties of the parameter estimates for finite sample sizes. Different approaches and applications are presented to demonstrate the flexibility of the correlated frailty approach in modeling associations in clustered event times. Chapter 6 deals with copula models and analyzes similarities and dissimilarities between frailty and copula models. Chapter 7 gives an overview of different problems related to frailty models such as tests for homogeneity, identifiability aspects, and available software. The Appendix provides a series of technical mathematical results and background about genetic models used throughout the book.

The present monograph does not attempt to give a complete overview of the fast growing literature on frailty models; this would not be possible. The treatment of the topics covered are restricted to explaining the basic ideas in frailty modeling and statistical techniques, with focus on real data application and interpretation of the results. In many cases different models are applied to the same data to compare and discuss their advantages and limitations under varying model assumptions.

1.2 Examples

Different survival models are considered in this book. Most of them will be applied to real data, mainly using examples from research fields such as medicine, epidemiology, and demography. Survival analysis deals with the analysis of times until the occurrence of a well defined event. The occurrence of this event describes the transition from one state to another, for example, occurrence of a disease is the transition from the state of being healthy to the state of being sick. Sometimes the transition is of special interest (incidence of the disease), and in other cases the state (prevalence of the disease) is the target of the analysis. For such kind of analysis it is necessary to define the time scale and a starting time point zero. In many cases the time scale is the age of the individual. In clinical trials the starting point is often beginning of treatment. If the focus is on the development of a disease, the time of diagnosis is usually the starting point. In occupational cohort studies the starting point is often the beginning of employment or unemployment.

We first consider the univariate event times, which means data with no clustering. Such data set is given in Example 1.1, based on a prognostic study analyzing the value of electron-beam computed tomography (EBCT) derived calcium scores for risk stratification in symptomatic patients. Example 1.2 presents the malignant melanoma data. The models fitted to these data sets are parametric and semiparametric proportional hazard models. The most important goal of Chapter 3 is to analyze the effect of including unobserved heterogeneity on regression parameter estimates.

However, the main focus of this book is on multivariate frailty models, where event times are clustered. Example 1.3 will serve as an example of univariate as well as multivariate data. In the last situation the cancer-diagnosing units were considered as clusters. The cluster size differs from cluster to cluster, which is common in multicenter clinical trials. Here, frailty can describe center-to-center variations not explained by observed covariates. In addition to the problem of analyzing the effect of observed covariates, an important research problem is evaluating the dependence between event times in clusters. In genetic studies, correlations between family members are the basis of the analysis of heritability of specific traits, for example, the times of onset of breast cancer or cause of death specific lifetimes. We use Danish and Swedish twin data provided by the Danish Twin Registry at the University of Southern Denmark in Odense and the Swedish Twin Registry at the Karolinska Institute in Stockholm to emphasize the practical purpose of the frailty models with fixed and small cluster sizes. In Example 1.4, cause-specific lifetimes of Danish twins are considered. A subsample with additional covariate information is presented in Example 1.5. Example 1.6 provides data on the age of onset of breast cancer in Swedish twins, whereas Example 1.7 deals with current status data. The next section provides a brief description of these data.

Example 1.1 **Prognostic Study of the EBCT Calcium Score**

EBCT-derived calcium score is a measure of coronary arteriosclerotic plaque that can be used for more precise risk stratification in symptomatic patients (Schmermund et al. 2004). In the study presented here, it was investigated whether EBCT-derived calcium score can add prognostic information compared with clinical information derived from risk-factor assessment, exercise stress testing, and coronary angiography. Patients with recent (<3 months) onset of symptoms were retrospectively identified and examined for possible coronary artery disease (CAD) and underwent EBCT. Complete follow-up after 42 months was available for 255 patients with mean age at baseline of 58 years, who were finally included into the study.

Table 1.1: Five patients in the EBCT study

id	time	status	risk group	calcium score	age
1	42	0	2	0	70
2	42	0	1	0	59
3	42	0	1	1	74
4	14	1	4	1	70
5	42	0	1	0	50

Four clinical risk groups with increasing evidence of CAD were constructed based on risk factor assessment, exercise stress testing, coronary angiographic anatomy, and revascularization at baseline. The main interest was in the occurrence of a combined event consisting of major adverse cardiac events such as myocardial infarction, cardiac death, and revascularization. The event was observed in 40 (16%) patients during the follow-up, the observations of the other patients are mainly censored after 42 months at the end of the study. The data for five patients are presented in Table 1.1.

Table 1.2: Description of the EBCT study population

covariate	category	absolute frequency	relative frequency
risk group	1	79	31.0%
	2	78	30.6%
	3	42	16.5%
	4	56	21.9%
calcium score	< 100	150	58.8%
	≥ 100	105	42.2%
age (years)	23 – 54	85	33.3%
	55 – 62	79	31.0%
	63 – 84	91	35.7%

The first column gives the patient specific identification number, the observed event or censoring time in the second column is measured in months. The covariate of main interest in this study is the CAD risk divided into four prognostic groups (group 1 – no evidence of ischemia, ≤ 1 conventional risk factor; group 2 – evidence of ischemia and/or ≥ 2 conventional risk factors, no angiographic stenoses; group 3 – angiographic stenoses, no revascularization at baseline; group 4 – early revascularization). The dichotomous covariate calcium indicates an EBCT-derived calcium score larger than 100. There is no clustering in this data set. One of the research questions was to examine whether the EBCT-derived calcium score can add prognostic information compared with the clinical information summarized in the risk groups. Age (in years) was categorized into three groups with the youngest age group as the reference. The covariate frequencies are given in Table 1.2. □

Example 1.2 **Malignant Melanoma Data**

The data set contains observations of 205 patients with radical surgery for malignant carcinoma (skin cancer) at the University Hospital of Odense in Denmark during 1962–1977. Radical surgery means that the tumor was completely removed including the skin within a distance of about 2.5 cm around it. Patients were followed up until 1977 and 57 deaths from malignant melanoma, and 14 deaths due to other causes (coded as censored event times) were observed. The data of five patients are given in Table 1.3.

Table 1.3: Data of five patients of the malignant melanoma study

id	time	status	gender	age	thickness
1	10	0	1	76	676
2	30	0	1	56	65
3	35	0	1	41	134
4	99	0	0	71	290
5	185	1	1	52	1208

The first column provides the unique patient identification number. Variable time measures time since surgery in months and variable status indicates the occurrence of death caused by malignant melanoma. There are several covariates available in the data set; for ease of presentation we will restrict them in this application to the following three covariates: gender (0 = female, 1 = male), age at surgery (years), and tumor thickness (in 1/100 mm). This is a univariate data set without clustering. The data was first analyzed by Drzewiecki et al. (1980a,b) and later published and reanalyzed in more detail by Andersen et al. (1993). □

Example 1.3 **Halluca Study**

The Halle Lung Cancer (Halluca) study was a study investigating provision of medical care to lung cancer patients in the region of Halle and Dessau in the eastern part of Germany (Bollmann et al. 2004, Kuß et al. 2008). The study region covers about 1.5 million inhabitants and belongs to the State of Saxony-Anhalt. In cooperation with the regional clinical tumor registries, all lung cancer patients in the study region were recorded from April 1996 to September 1999, and follow-up was done until September 2000. A total of 1696 lung cancer patients were observed, and survival was defined as time from clinical, histological or cytological diagnosis to death. 1349 patients (79.5%) died until the end of follow-up; median survival in the study population was 9.3 months. To validate and complement survival information, the data from the Clinical Cancer Registry were compared to death certificates collected by the local health institutions and linked to the data from the Common Cancer Registry of Eastern Germany. Minimal follow-up time was 12 months, and the median follow-up time 33 months. To judge the influence of prognostic and risk factors on overall survival, five fixed-effects covariates, known to be important in the prognosis of survival regarding lung cancer, were included. The information for five patients is given in Table 1.4. The first column gives the patient-specific id number, and the second column the survival time (in months). The third column contains the survival status (1 = death, 0 = alive) and the fourth column the cluster variable diagnosing unit. Lung cancer was diagnosed in 56 different diagnosing units with numbers of patients ranging from 1 to 392 (mean 30.3). The Halluca data is analyzed using univariate approaches in Chapters 2 and 3. In Chapter 4 the data is treated like from a multicenter study with the diagnosing unit as cluster variable. A cluster effect by diagnosing unit is indicated by Figure 1.1. In multicenter studies (with treatment center as cluster), despite the tight study protocols, often center-to-center variation occurs, which cannot be explained by covariates. Frailty models can be used to investigate this variation. The other columns represent variables gender (0 = female, 1 = male), age (years), histologic type (1 = small-cell lung cancer, 2 = non-small-cell lung cancer), ECOG status (range 0 to 4), and UICC stage (1 = I, 2 = II, 3 = IIIa, 4 = IIIb, 5 = IV).

Table 1.4: Data of five patients in the Halluca study

id	time	status	unit	gender	age	type	ECOG	stage
1	54.74	1	1	1	75.02	1	0	4
2	6.68	1	1	1	63.60	1	0	5
3	0.33	1	1	1	52.68	2	.	.
4	24.28	1	2	1	55.14	.	0	.
5	15.46	0	2	0	79.28	2	0	1

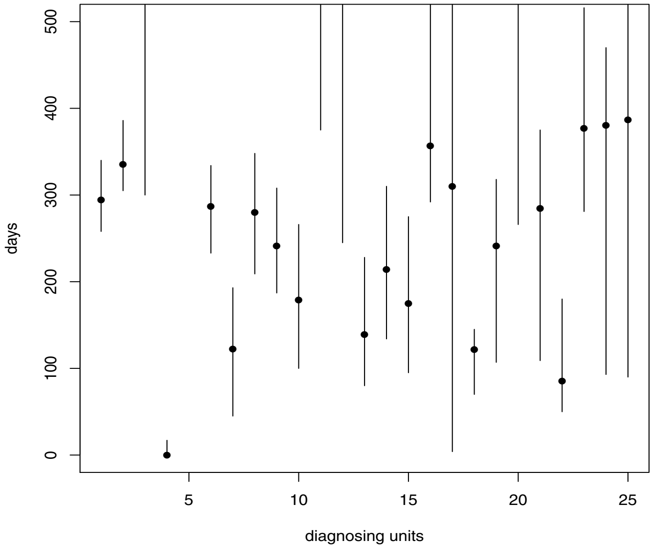


Figure 1.1: Median survival (95% CI) of 25 diagnosing units in Halluca

Dots indicate missing covariate values. It was initially decided to model these as separate categories despite the dangers of this procedure. Throughout the book, in the tables presenting the results, these missing categories are omitted.

Table 1.5: Description of the Halluca study population

covariate	**category**	**absolute frequency**	**relative frequency**
gender	male	1374	81.0%
	female	322	19.0%
histologic type	NSCLC	1183	69.8%
	SCLC	366	21.5%
	missing	147	8.7%
ECOG	ECOG 0-2	1366	68.7%
	ECOG 3-4	123	7.3%
	missing	407	24.0%
stage	I	185	10.9%
	IIa	79	4.7%
	IIb	195	11.5%
	III	280	16.5%
	IV	621	36.6%
	missing	336	19.8%

We further explicitly omitted the primary treatment as a covariate. This was to prevent unjustified treatment recommendations, which should only be derived from randomized trials and not from observational studies. The study population is described in Table 1.5. Mean age at diagnosis was 65 years. □

***Example 1.4* Danish Twins Cause-Specific Mortality Data**

The Danish Twin Registry was the world's first nation–wide twin registry, established in 1954 by Bent Harvald and Mogens Hauge. The older part of this population based registry includes all twins born in Denmark during the period 1870–1910 and all like-sex pairs born between 1911 and 1930 in which both partners survived to the age of six years. Pairs with deaths before the age of six were excluded because it turns out to be very difficult to obtain detailed information especially about the zygosity of such twin pairs. The birth registers from all 2200 parishes of Denmark during the relevant calendar years were manually scrutinized to identify multiple births. After such births were identified, a search was then carried out for the twins or, whenever needed, for their closest relatives in regional population registers (in operation since 1924) or other public sources, especially the archives of probate courts and censuses. As soon as a twin was traced, a questionnaire was sent to the twin (if she or he was alive) or to the closest relatives (if she or he was not alive). Questions about phenotypic similarities were included in the questionnaires to assess the zygosity by self-reported similarities. The reliability of this method was validated by comparison with the results of laboratory methods based on blood serum enzyme group determination in a subgroup of twins. The misclassification rate of this method was found to be less than 5% in the Danish twin data (Holm 1983). Similar results are known from the Swedish Twin Registry (Cederlöf et al. 1961) The follow-up procedure traced nearly all twins who did not die or emigrate before the age of six years. For further, more detailed information about the construction and the composition of the Danish Twin Registry see Hauge (1981).

The data provided by the Danish twin registry contain 8201 monozygotic (MZ) and dizygotic (DZ) twin pairs who were born between 1 January 1870 and 31 December 1930, and who were both still alive on 1 January 1943. As a consequence of this restriction, around two-thirds of the twin pairs born were excluded because of the high infant mortality of this period (1870–1930). Furthermore, twins have a higher infant mortality than singletons because of their lower birth weight. It was necessary to exclude early deaths because it was nearly impossible to obtain zygosity information when one or both twin partners died at young ages. Zygosity information is crucial to the application of the methods in twin research.

Table 1.6: Data of three Danish twin pairs

id	time	status	pair	gender	zygosity	cause	birth
1	76.09	1	1	1	1	2	1889
2	76.02	1	1	1	1	2	1889
3	64.73	1	2	0	2	4	1908
4	94.75	1	2	0	2	1	1908
5	85.62	0	3	0	1	0	1881
6	68.61	1	3	0	1	2	1881

Observed covariates are gender, zygosity and year of birth. A total of 246 twin pairs with incomplete information about the cause of death were excluded, leaving a study population of 7955 twin pairs. Individuals were followed up through 31 December 1993, and those identified as deceased after that date have been classified here as living. Altogether, we have 1344 male MZ twin pairs and 2411 DZ twin pairs, and 1470 female MZ twin pairs and 2730 DZ twin pairs. In addition to the lifetimes, there is information about cause of death for all noncensored lifetimes, that is, for all individuals in the study population who died before 31 December 1993. For the present analysis, only the underlying cause of death was considered.

The data for the first six twins are given in Table 1.6. The first column gives the identification number of the individual, and the second one the survival or censoring time (in years). The third column contains the censoring indicator (1 = death, 0 = alive), the fourth column is the identification number of the twin pair (cluster), and the four other columns represent the covariates gender (0 = female, 1 = male), zygosity (1 = monozygotic, 2 = dizygotic), cause of death (0 = alive, 1 = cancer, 2 = coronary heart disease, 3 = stroke, 4 = respiratory diseases, 5 = other), and year of birth. For more detailed information about cause of death, gender, and zygosity of the study population see Table 1.7.

Table 1.7: Causes of death in the Danish twin population (number of individuals)

	males		females	
cause of death	MZ twins	DZ twins	MZ twins	DZ twins
cancer	440	809	423	823
coronary heart disease	666	1180	548	999
stroke	161	278	186	335
respiratory diseases	143	203	89	205
other causes	336	661	330	555
all causes together	1746	3131	1576	2917
alive	942	1691	1364	2543

Information regarding death status, age at death, and cause of death was obtained from the Central Person Register, the Danish Cancer Register, the Danish Cause–of–Death Register, and other public registries in Denmark. The main source for obtaining information on cause of death was the Death Register at the National Institute of Public Health. Information about cause of death is available from this register for individuals who died after 1942 (Juel and Helweg-Larsen 1999). Consequently, cause of death is included in the twin register only for twins who died after this year.

Table 1.8: Cause of death groups by ICD number

cause of death	**ICD revision 6 & 7**	**ICD revision 8**
cancer	140 – 205	140 – 209
coronary heart disease	420	410 – 414
stroke	330 – 334	430 – 439
respiratory diseases	470 – 527	460 – 519

The validity of the twin register was checked on the basis of a comparison of information about year of death with the nationwide Danish Cancer Register. There was around 99% agreement, although both registries were independent. Further data corrections increased this level of agreement to almost 100%. Cause of death was coded following the sixth, seventh, and eighth edition of the International Classification of Diseases (ICD). Four different groups of main causes of death are considered in the present example: cancer, coronary heart disease (CHD), stroke, and diseases of the respiratory system. ICD codes in three revisions of the ICD for these broad cause-of-death groups are given in Table 1.8. Causes of death are a common example for competing risks. □

Example 1.5 **Danish twins CHD mortality data with covariates**

The data in this example is a subset of the cause-specific mortality data in the foregoing example. Here the main focus is on age at death with death caused by coronary heart disease and the influence of BMI and smoking on this outcome.

In 1966, a questionnaire including questions about smoking, height, and weight was mailed by the Danish Twin Registry to all Danish twins born in the period 1890–1920 who were alive and traceable on 1 January 1966. 3709 individuals answered the questionnaire (response rate 65%). Excluded from the study were 813 twins with nonresponding partners, four pairs with unknown zygosity, and 212 pairs with incomplete or uncertain information on height and weight. A total of 23 pairs were excluded because of incomplete information about the cause of death, resulting in a total study population of 1209 complete twin pairs.

Individuals were followed from 1 January 1966 to 31 December 1993. Those persons identified as deceased after the follow-up period are classified for our purposes as censored. At the end of follow-up period, approximately 40% of the twins were still alive, resulting in the right censored data. Altogether, there were 210 male monozygotic twin pairs and 316 dizygotic twin pairs, and 273 female monozygotic twin pairs and 410 dizygotic twin pairs. The data for the first six twins in the study population are given in Table 1.9. The first column provides the unique identification number of the twin, the second column the observation time (in years). The third column contains the