Joint Parametric Modeling and Estimation of Time to Cancer Recurrence and Disease Stage at Recurrence

by

Olga V. Marchenko

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2012

Doctoral Committee:

Professor Robert W. Keener, Chair
Professor George Michailidis
Professor Alexander Tsodikov
Associate Professor Edward L. Ionides

To my Parents, Sister, and Children

ACKNOWLEDGEMENTS

I owe a debt of gratitude to many people who contributed to my success in completing this dissertation. First of all, I have been fortunate to work with my advisor, Professor Robert Keener, who provided the guidance and support during my graduate school years and, in particular, during the research and the dissertation preparation process. Professor Alex Tsodikov made invaluable contributions to the development of the ideas and methods for my research. I am extremely thankful to Alex for his time and wiliness to help me through my journey. I am very grateful to Professor George Michailidis for his support during my graduate school years and during the research process. I would like to thank Professor Ed Ionides for his helpful discussions and papers on the topic. I am grateful to the Department of Statistics that gave me an opportunity to be a part of their graduate students program, and to faculty members who made my experience productive and enjoyable. I am thankful to the graduate program assistant Lu Ann Custer for her help with all kinds of administrative work. Additionally, I would like to thank PhotoCure ASA, in particular, Yngvil Kloster Thomas, for allowing to use the data from their trials to demonstrate the models and methods of our research using the real-life example.

I owe much gratitude to my parents, Valentina Kotenko and Vladimir Kotenko who encouraged me to return to the graduate school, believed in me even when I had doubts, and helped with everything they could to give me that extra time to work on the research. I am extremely grateful to my younger sister, Natallia Katenka. Her love, support, and help brought me much further in life than I ever dreamed

of. Natallia has been always there for me and never failed to do everything she could to further my progress. I am very thankful to my children, Victoria Marchenko and Catherine Marchenko, who loved me and did not want me to quit pursuing my education and degree even though it meant less time spent with them. My children always inspired me to be a better person, work hard, and get things done despite of how hard it can get at times. I am grateful to my husband who has been working hard and has been providing mental support to me during these challenging years of my education.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Joint Parametric Modeling and Estimation of Time to Cancer Recurrence and Disease Stage at Recurrence

by

Olga V. Marchenko

Chair: Robert W. Keener

A clinical trial with bladder cancer patients who went through surgery and were followed up for tumor recurrence was used as motivation for this research. The surgery was conducted on patients with an early bladder cancer stage. During the follow-up, patients were evaluated for cancer recurrence at 3 months, 6 months, 9 months, and at about 5 year visits unless they had cancer recurrence in between visits or died prior to a scheduled visit. One of the main objectives of the study was to evaluate the time to cancer recurrence. At the time of cancer recurrence, the disease stage was also evaluated. The stage of the cancer at recurrence significantly impacts future treatment and quality of life. Therefore, modeling and analyzing the time to cancer recurrence and the stage at recurrence jointly makes more sense than an analysis based on the time to recurrence alone.

In our research, we describe a model for the joint distribution of time to recurrence and cancer stage at recurrence that accounts for the recurrence caused by the cancer cells surviving treatment or surgery, and for the recurrence caused by spontaneous carcinogenesis. First, we proceeded with a continuous follow-up assumption using

stochastic models of cancer recurrence. Then we extend the approach to allow for a discrete follow-up process. We provide methods for full maximum likelihood estimation based on the EM algorithm. The methods are illustrated through modeling and estimation of data from a clinical trial in patients with bladder cancer described above. Simulations are used to assess the sensitivity of the methods. An added benefit of such modeling is that it permits using the cancer stage at recurrence to provide adjusted estimates for the time to recurrence distribution and allows for more powerful inference.

CHAPTER I

Introduction

The pharmaceutical industry is highly regulated in the United States and around the world. Before any drug or device becomes available to people, extensive work is done to evaluate the efficacy and safety of investigational drug or device in pre-clinical and clinical trials. Even after the drug or device is approved by regulatory agencies and released to the market, post marketing trials are conducted to detect any rare or long-term adverse effects over a much larger patient population and longer time period than was possible during the Phase I-III clinical trials.

Clinical trials are studies that are conducted in humans to allow safety and efficacy data to be collected and evaluated for new drugs or devices. Clinical trials are commonly classified into four phases. Each phase has its own objectives. While Phase I and most Phase II trials are considered to be exploratory, Phase III studies are aimed at being the definitive assessment of how effective the drug is, in comparison with current gold standard treatment. Because of the large size and comparatively long duration, Phase III trials are the most expensive, time-consuming and difficult trials to design and run. The increasing pressure on pharmaceutical manufacturers to deliver critically important therapies to patients, together with limited funding, has spawned increased efforts to design, analyze, and report clinical trials in a more efficient manner.

Oncology clinical trials face additional ethical issues. Therefore, statistical designs must be sensitive to the associated ethical issues and the choice of the endpoints should appropriately address the research questions.

The National Cancer Institute (NCI) has a Data Modeling Branch whose mission is to support research on statistical and mathematical models in order to understand the impact of cancer control interventions and economic, health care delivery, and utilization factors on the cancer burden. They use mathematical modeling to develop, evaluate and improve estimates of cancer progress measures and develop software for integration of modeling into data systems. The Cancer Intervention and Surveillance Modeling Network (CISNET) is a consortium of NCI-sponsored investigators that use statistical modeling to improve our understanding of cancer control interventions in prevention, screening, and treatment and their effects on population trends in incidence and mortality. More information can be found on www.cancer.gov.

In our research, we describe a model for the joint distribution of time to cancer recurrence and cancer stage at recurrence. Our model accounts for recurrence caused by the cancer cells surviving treatment or surgery and for recurrence caused by spontaneous carcinogenesis. Parametric distributions are used for inference. We describe methods for full maximum likelihood estimation based on the EM algorithm. The methods are illustrated through modeling and estimation of data from a clinical trial in patients with bladder cancer. An added benefit of such modeling is that it permits using the cancer stage at recurrence to provide adjusted estimates for the time to recurrence distribution and allows for more powerful inference.

1.1 Motivation

Oncology clinical trials are conducted mainly in advanced stage cancer patients with high mortality rate. But not all cancers have a high mortality rate although the treatment cost and the burden of these cancers are fairly high. One such cancer

is bladder cancer. The American Cancer Society estimated about 70,530 new cases (about 52,760 men and 17,770 women) and 14,680 deaths (about 10,410 men and 4,270) from bladder cancer in the US in 2010. In spite of increased incidence, the rate of people dying of this cancer has decreased over the past 20 years. More than 500,000 people in the United States are survivors of this cancer[1]. Bladder cancer is one of the most expensive cancers for society because patients live longer and have multiple recurrences.

Photocure ASA, a pharmaceutical company from Norway, conducted a clinical program in patients with bladder cancer which showed that compared to standard white-light cystoscopy, fluorescence cystoscopy using a combination of the photosensitzer hexaminlolevulinate (Hexvix) and blue light improve the visualization of bladder tumors. Results of the Hexvix clinical program conducted in Europe and in USA demonstrated that local instillation of Hexvix significantly increased the number of tumors detected during cystoscopy, which leads to improved patient management in a significant number of patients. The pivotal study 305 also demonstrated for the first time that improved detection of bladder tumors, enables a more complete tumor mapping, and more complete resection, resulting in a significant reduction of recurrence rates at 9 months. More interestingly, of the patients with documented recurrence during the 9 month follow up period, more patients in the group treated with white light only experienced recurrence of higher stage tumors compared to patients in the group treated with both white light and Hexvix fluorescence cystoscopy (*Stenzl et al.*, 2010).

After completion of the pivotal study, Photocure ASA decided to initiate an extension of the pivotal phase III of 305 study to compare the time to recurrence, disease stage at recurrence, and the number of recurrences between two groups. This study investigated whether this improved initial detection and resection of bladder cancer

[1]American Cancer Society, ACS detailed guide: bladder cancer. What are the key statistics for bladder cancer? Available at: www.cancer.org

lesions in patients with non-muscle invasive bladder cancer with experimental fluorescence cystoscopy would also lead to a long-term reduction in recurrence compared to standard white light cystoscopy.

Based on the current knowledge of the bladder cancer recurrence process and the technology available to perform surgeries, there are reasons to believe that a significant percentage of supposed tumor recurrences result from residual tumor left behind at resection or growth of previously undetected lesions (*Sylvester et al.*, 2006). Given the extent of the data that is available from PhotoCure ASA bladder cancer 305 trial, one interesting question to address was how we can model the data assuming the recurrence of cancer can be caused by two reasons: by cancer cells surviving the treatment or surgery, or by spontaneous carcinogenesis. Additionally, given the 305 study conduct and results, it was evident that modeling and analyzing the time to cancer recurrence and the stage at recurrence jointly would make more sense than an analysis based on the time to recurrence alone.

1.2 Bladder Cancer Overview

Bladder cancer is the fourth most common malignant cancer disease in men and the eighth most common malignant cancer disease in women. The disease affects primarily older people; nearly 90% of people with bladder cancer are over the age of 55 years. Men are 3 times as likely to be affected as women. Whites are diagnosed with bladder cancer almost twice as often as blacks. Black patients have generally more advanced cancer at diagnosis. In almost 75% of the cases, patients are first diagnosed with the cancer stage confined to the bladder. In most remaining cases, the cancer has spread to nearby tissues outside the bladder. In about 3% of cases, the cancer has spread to distant sites.

1.2.1 Risk Factors of Bladder Cancer

There are several known and potential risk factors for bladder cancer. Cigarette smoking and occupational exposure to aromatic amines are the most well-established among them. It is estimated that smokers have a twofold to fourfold greater risk of having bladder cancer than nonsmokers, and smoking is believed to contribute to up to 50% of all bladder cancers that are diagnosed. Other risk factors include chronic bladder irritation (e.g., stones or long-term catheter use), occupational exposures (e.g., polychromatic hydrocarbons), family history, and infection with certain parasites (*Lee et al.*, 2006).

1.2.2 Types and Stages of Bladder Cancer

Of urothelial bladder tumors, about 90% are transitional cell carcinoma. Transitional cell carcinoma (Figure 1.1) can be either non-muscle-invasive (pathologic Stages T_a, T_1, and carcinoma in situ (CIS)) or muscle-invasive (pathologic Stages T_2 to T_4).

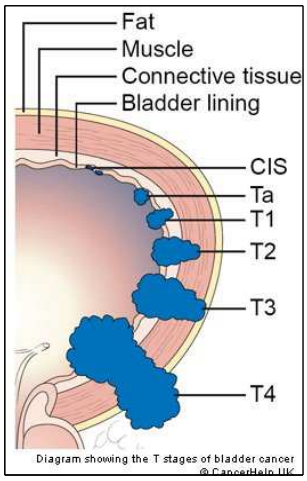

Figure 1.1: Bladder Cancer Pathologic Stages.

In patients with the diagnosis of bladder cancer, about 70% present initially as non-muscle-invasive bladder tumors, with the remainder presenting as invasive can-

cer[2].

1.2.3 Clinical Presentation and Course of Disease

Important endpoints in the natural history of bladder cancer include recurrence, stage at recurrence, progression and survival. Recurrence is defined as appearance of tumors of the same or smaller stage and grade as the primary tumor. The cause of early bladder cancer recurrence may be due to residual tumor after incomplete resection, microsatellites missed during initial transurethral resection of the bladder (TURB) or true recurrence. Recurrence is common; depending on a patient's characteristics after TURB the probability of recurrence at one year ranges from about 15% to 61% and from 31% to 78% at 5 years (*Lee et al.*, 2009). Progression is defined as the development of higher grade tumors with muscle invasion or metastatic disease, and is associated with an increased risk of death. The major prognostic factors for recurrence and progression are tumor multiplicity, size, previous recurrence rates, baseline tumor (T) stage, presence of CIS and tumor grade (*Kurth et al.*, 1995; *Sylvester et al.*, 2006). In the current research, we refer to the cancer recurrence in a general sense; since right after the surgery the patient is considered to be cancer-free, the recurrence is either the true cancer recurrence to the baseline stage (the identified stage prior to the surgery) or any cancer stage at which a patient had the first recurrence diagnosed after the surgery.

1.2.4 Diagnosis and Treatment

The type and severity of clinical signs and symptoms of bladder tumors depend on the extent and location of the tumor. The most common first symptom of bladder cancer is gross or microscopic hematuria, which occurs in over 80% of bladder cancers. Other presenting symptoms include dysuria and urinary frequency or urgency, and

[2]John Hopkins Pathology, Types of Bladder Cancers. Available at: www.pathology2.jhu.edu.

less commonly, flank pain secondary to obstruction, and pain from pelvic invasion or bone metastases (*Lee et al.*, 2009).

The current standard of care for diagnosing bladder cancer is a combination of urine cytology, a visual inspection of the bladder with an cystoscope and white-light illumination (WL cystoscopy) and biopsies for histological verification. WL cystoscopy is used conventionally to detect lesions in the bladder for patients with known or suspected bladder cancer. However, tumors such as flat carcinomas (particularly CIS), dysplasia, multifocal growth and microscopic lesions are often overlooked by conventional WL cystoscopy. Urinary cytology is most accurate in detecting high grade malignancy or CIS, but offers poor sensitivity in detecting low grade carcinomas. A positive cytology may indicate tumor anywhere in the urinary tract, whereas a negative cytology does not necessarily exclude the presence of a low grade malignancy. TURB removes the tumor and allows for pathologic analysis of the resected or biopsied specimen, which establishes the diagnosis and provides important information about the tumor grade and depth of bladder invasion. For patients with low-grade (Ta) tumors, TURB without intravesical therapy is the standard treatment. Immunotherapy with intravesical Bacillus Calmette-Guerin (BCG) or chemotherapy following TURB is the preferred option for patients with high-grade Ta and T1 tumors, as well as for patients with carcinoma in situ (CIS). For invasive disease, total urethrectomy along with cystectomy is performed with adjuvant chemotherapy or radiation (*Lee et al.*, 2006, 2009).

Regular follow-up is required, generally every 3 months for the first 1 to 2 years, then at increasing intervals over the next 2 years, and annually from then on (*Lee et al.*, 2009).

The high rate of early recurrences (up to 60% within 3 months) reported in the literature suggests that a significant percentage of supposed tumor recurrences result from residual tumor left behind at resection or growth of previously undetected

microscopic lesions *Sylvester et al.* (2006).

1.3 Short Overview of PhotoCure Hexvix Trials (NDA 22-555)

The NDA submission of Hexvix consist of 5 Phase III studies (301, 302, 303, 304, and 305) with only one study (305 study) defined as the pivotal/ confirmatory trial. The primary objective of single arm studies 301, 302, and 303 was improved detection, while the study 304 evaluated only the recurrence rate. The pivotal Study 305 is different from other trials in study design, randomization structure, primary and secondary endpoints, and in patient population considered for the primary objectives.

Clinical study 305 was a prospective, randomized multicenter phase III study designed to be comparative and is both within-patient and between-patient controlled. The flowchart (Figure 1.2) illustrates the design of the study:

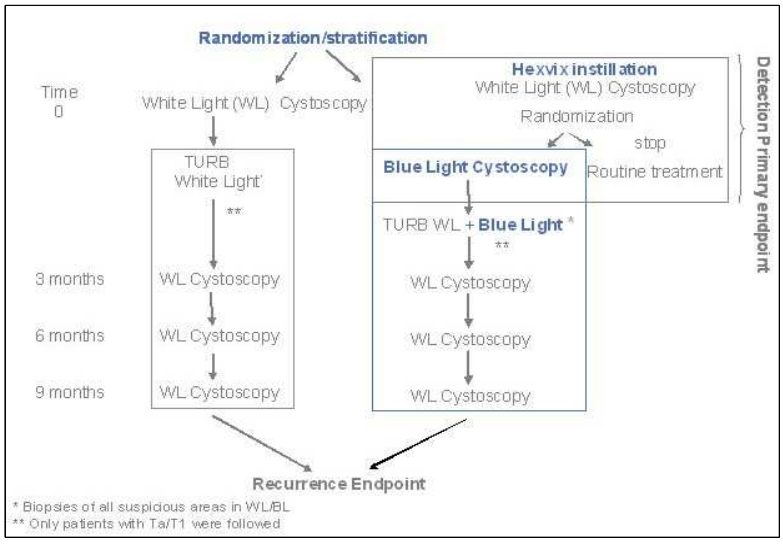

Figure 1.2: Study Design.

The study had two co-primary objectives:

- To compare Hexvix cystoscopy with white light cystoscopy in the detection of histology confirmed papillary bladder cancer in patients with papillary bladder cancer.

- To compare early recurrence rate after Hexvix and white light transurethral resection of bladder (TURB) with white light TURB in patients with superficial bladder cancer (stages Ta and T1).

After completion of the pivotal study, Photocure ASA decided to initiate the extension of the 305 study. This study investigated whether this improved initial detection and resection of bladder cancer lesions in patients with non-muscle invasive bladder cancer with experimental fluorescence cystoscopy will also lead to a long-term reduction in recurrence compared to standard white light cystoscopy. The primary endpoint of the extension study was the recurrence-free survival time. The main secondary endpoints are the time to recurrence, the tumor stage at recurrence and the number of recurrences per patient. All patients who completed the main part of the study were eligible to participate in the extension part of the study.

Table 1.1 provides a sample of the trial data. There are three derived variables: censor status (Censor Status), time to recurrence (Recur Time), and stage at recurrence (Recur Stage). The collected variables include: patient identification number (Patient Num), center number (Center Num), age (Age), gender (Sex), race (Race), procedure group (Group Num), country (Cnt), scheduled visit (Visit Time), baseline cancer stage (Baseline Stage), whether a patient had an initial or recurrent cancer (Cancer History), baseline cancer grade (Baseline Grade), and whether a patient had CIS lesion at baseline or not (CIS).

About 560 patients with histologically confirmed non-muscle invasive bladder cancer (T_a and T_1) by a local pathologist were included in the main pivotal study. The recurrence analysis included 551 patients. At inclusion the patients were randomized to have their cystoscopy including TURB by white light only or by white light plus

Patient Num	Center Num	Age/Sex /Race	Group Num	Cnt	Censor Status	Recur Time	Visit Time	Recur Stage	Baseline Stage	Cancer History	Baseline Grade	CIS
001	001	65/M/ White	1	USA	0	61	5y	0	T_a	1	2	0
002	001	37/M/ White	2	USA	1	2.7	3m	1	T_a	1	2	0
003	001	68/F/ White	2	USA	1	3.1	3m	1	T_a	2	2	0
004	201	71/F/ White	1	CAN	1	10.4	9m	1	T_a	1	2	0
005	011	78/M/ White	1	USA	1	9.2	9m	6	T_a	2	1	0
006	005	61/M/ White	2	USA	1	4.2	6m	7	T_1	1	3	1

Table 1.1: Data Sample.

blue light with experimental drug. The randomization was stratified by cancer history (initial and recurrent bladder cancer). Patients were followed-up by cystoscopy in white light after the resection procedure at 3, 6, and 9 months. The results from the local pathologist at baseline were used to determine if the patient was to be followed up at 3, 6 and 9. Recurrence was to be verified by histology assessment of the local pathologist during visits. Recurrence was defined as presence of either CIS, T_a, T_1, or $T_2 - T_4$ tumor. Patients having a recurrence at three months (based on local pathology) did not continue to the six or nine months follow up. If the patient had recurrence at six months, the patient did not continue to the nine months follow up. Suspected areas seen during cystoscopy at baseline and at follow-up visits were biopsied or resected. If there were multiple pathology results for a single lesion or multiple cancer lesions were identified for a patient, the worst lesion type was used in the analysis. However, if there was a CIS in addition to a papillary lesion reported, both results were included. Prognostic factors for recurrence such as the baseline number of lesions, baseline tumor stage, baseline tumor grade, presence of CIS, and previous recurrences were collected for all patients. The results of this pivotal phase III study (305 study) are in line with the previously published studies and showed that Hexvix fluorescence cystoscopy improves detection of non-muscle invasive papillary bladder tumors compared to white light cystoscopy and TURB alone. In 16.4%

of the patients with Ta or T1 tumors at least one additional tumor was detected with Hexvix fluorescence cystoscopy only ($p = 0.001$) *Sylvester et al.* (2006); *Stenzl et al.* (2010).

The study also demonstrated for the first time that improved detection of bladder tumors, enables a more complete tumor mapping, and more complete resection resulting in a significant reduction of recurrence rates at 9 months. During the surveillance period, for the ITT (intent-to-treat) analysis 128/271 patients (47%) in the Hexvix group and 157/280 patients (56%) in the white light group had tumor recurrence ($p = 0.026$ using CMH test stratified by study center) (*Stenzl et al.*, 2010). The difference of time to recurrence curves was tested by log-rank test at 5% of statistical significance. The analysis of the cancer stage at recurrence did not achieve a statistical significance, although the marginal difference toward the improvement in an experimental group was noted.

551 patients with non-invasive papillary bladder cancer (271 in the fluorescence group, and 280 in the white light group, respectively) enrolled in the previously completed pivotal phase III study who were followed for recurrence were included in the extension phase of the study. The extension part of the study showed that the improved initial detection and resection of bladder cancer lesions in patients with non-muscle invasive bladder cancer with experimental drug fluorescence cystoscopy would also lead to a long-term reduction in recurrence compared to standard white light cystoscopy. Overall time to recurrence difference was tested by Wilcoxon test since the hazard ratio was higher at early survival times than a late ones. The analysis of the cancer stage at recurrence showed again the marginal improvement in an experimental group, but it did not achieve a statistical significance at 5% level.

1.4 Research Objectives

Recurrence-free survival is the suggested endpoint to measure a clinical benefit of bladder cancer patients in long confirmatory trials. There are no requirements on the trial duration in patients with bladder cancer, but the typical trial of 1-5 years in duration will not give an accurate estimation of overall survival or recurrence-free survival due to much longer expected survival time. Such trials mainly use either the proportion of patients without a recurrence at a particular time cut-off (e.g., 1 year or 5 years) or the time to recurrence/progression as the primary endpoint for the comparison. The proportion of patients without a recurrence at a particular point was not a reliable endpoint as was seen in Photocure ASA 305 study: the drop-out rate at 9 month was about 30%. The typical drop-out rate is fairly large ($> 25\%$). Time to recurrence is the recommended primary endpoint in such trials. While one of the primary objectives of the 305 study was to evaluate and compare the time to tumor recurrence, the disease stage was also evaluated at the time of tumor recurrence. A patient recurrence stage was defined as the worst stage among all lesion stages if a patient had multiple lesions at the recurrence diagnosis. The majority of these stages were less advanced, while some patients progressed to more aggressive stages. The stage of the disease at recurrence significantly impacts future treatment and quality of life. Therefore, analyzing and comparing the time to tumor recurrence and the stage at recurrence jointly makes more sense than an analysis based primarily on the time to recurrence.

The main objectives of this research were the following:

- To build and evaluate a joint model of time to cancer recurrence and disease stage at recurrence,
- To provide the appropriate estimates for the time to recurrence distribution adjusting for the cancer stage at recurrence.

Since the current knowledge of the bladder cancer recurrence process and the technology available to perform surgeries suggest that there are reasons to believe that a significant percentage of supposed tumor recurrences result from residual tumor left behind at resection or growth of previously undetected lesions *Sylvester et al.* (2006), we established additional model requirements:

- The model describing the recurrence process should have a biological meaning,
- The model should accommodate two causes of cancer recurrence: recurrence caused by cancer cells surviving a treatment or a surgery, and recurrence caused by spontaneous carcinogenesis.

In our research we describe a model for the joint distribution of time to recurrence and cancer stage at recurrence that accounts for recurrence caused by the cancer cells surviving treatment or surgery, and for recurrence caused by spontaneous carcinogenesis. First, we proceeded with a continuous follow-up assumption using stochastic models of cancer recurrence for the right-censored data (Chapter II). Then we extend the approach to allow for a discrete follow-up process. The interval-censored data model is described in Chapter III. We provide methods for full maximum likelihood estimation based on the EM algorithm. We introduce the random variable U which is not observed, but it is used to differentiate the cause of cancer recurrence. The methods are illustrated through modeling and estimation of data from a clinical trial in patients with bladder cancer described below. We also discuss in Chapter III how the proposed models and methods can be extended to cancer post-surgery surveillance which is represented by discrete process with a non-zero false-negative rate of a diagnostic test.

Before we move further, let us describe the bladder cancer stages and their ordering. The typical bladder cancer stage ordering is the following: $CIS < T_a < T_1 < T_2 < T_3 < T_4$. In clinical trial which was used in our research, most patients had

more than one cancer lesion. Therefore, patient recurrence stage was defined as the worst stage among all lesion stages if a patient had multiple cancer lesions. Since the presence of CIS is considered to be one of the major prognostic factors for recurrence, the suggested order of the stages includes the presence of CIS with another stage as a separate category. In this trial, the clinical stage order from less advanced to more advanced stages was proposed to be the following: $T_a < CIS < (T_a + CIS) < T_1 < (T_1 + CIS) < T_2 < (T_2 + CIS) < T_3 < (T_3 + CIS) < T_4 < (T_4 + CIS)$. In our research, the cancer stage was modeled using a multinomial logit model to account for more general case.

1.5 Literature Review

1.5.1 Models Describing Tumor Latency

Many scientists have investigated the mathematical modeling of carcinogenesis. The majority of models use elements of birth-and-death stochastic processes theory. Tan gives a comprehensive analysis of this class of models in *Tan* (1991). A number of multistage models of the Markov type have been introduced starting from work of *Armitage and Doll* (1954). Recent biological findings provide enough evidence to consider carcinogenesis as a multistage process. Moolgavkar with his colleagues researched two-stage models extensively in *Moolgavkar and Venzon* (1979), *Moolgavkar et al.* (1988), *Moolgavkar et al.* (1990), *Luebeck and Moolgavkar* (1991). A common weak point in many Markovian models of carcinogenesis is that the description of tumor progression is not sufficiently advanced. The time to observing a tumor is not equal to the time at which the first malignant cell is generated. Additionally, the estimation procedure is quite tedious even in computationally feasible cases (*Tan and Chen*, 1993). Therefore, the search for new ways of modeling carcinogenesis seems to be important. The mathematical description of tumor latency with regard to

the tumor recurrence and regression analysis of tumor recurrence data is described in work by *Yakovlev and Tsodikov* (1996). The authors discuss different stochastic models with parameters that have clear biological meaning. Their proposed parametric models describe the process of cancer recurrence. They suggest that there are several causes of local cancer recurrence including recurrence caused by the cancer cells surviving a treatment or a surgery, and recurrence caused by spontaneous carcinogenesis which have different mathematical representation and biological interpretation. In our research, we used these models to build the joint model of the time to bladder cancer recurrence and the cancer stage at recurrence.

1.5.2 Joint Distribution Models and Estimation

Numerous papers were published on joint modeling and analysis of time to event outcome and repeated measurements on a continuous response. The motivation for such modeling and analysis arose from medical studies. The most popular motivating example given in literature is an HIV study with the progression of CD4 cell counts over time and the time to patients' death. A mixed-effects model with normal random effects is used to model the repeated measurements and a proportional intensity model is used to model the hazard function of survival time. Random effects are used to account for the dependence between repeated measurements and survival time due to unobserved heterogeneity. In the literature, such a joint model is described as either a selection model if the conditional distribution of survival time given repeated measurements is modeled, or as a pattern-mixture model if the conditional distribution of repeated measurements given survival time is modeled. In most of the joint analysis literature, nonparametric maximum likelihood estimation has been proposed. The EM algorithm has often been used to calculate the maximum likelihood (ML) estimates.

Selection models have been studied by many scientists in difference contexts, for

example, by *Tsiatis and Davidian* (2001), and by *Henderson et al.* (2000). *Zeng and Cai* (2005) proved the consistency of the maximum likelihood estimators in the selection model and derived their asymptotic distributions.

Hogan and Laird (1997) described a mixture model for the joint distribution which accommodates incomplete repeated measures and right-censored event times. The EM algorithm was used to calculate the ML estimators. The parameter estimates from the model were used to make a treatment comparison after adjusting for the effects of dropout.

Faucett and Thomas (1996) proposed a joint model for censored survival data and repeated measured covariates. They used the Markov chain Monte Carlo technique of Gibbs sampling to estimate the joint posterior distribution of the unknown parameters of the model.

Ankerst and Finkelstein (2006) used a shared parameter selection model, to model the prostate cancer biomarker PSA level following radiotherapy and disease recurrence. A Markov chain Monte Carlo method comprised of a series of Gibbs and Metropolis-Hastings steps was used to estimate the joint posterior distribution of the unknown parameters and to assess sensitivity of the estimators using different priors.

Law et al. (2002) considered the cure model which is a special case of the mixture model. The longitudinal disease progression marker (PSA) and the failure time process were modeled jointly, in the joint-cure model setting. The EM algorithm was used to obtain the ML parameter estimators.

Tsodikov and Chefo (2009) modeled the prostate cancer data using the complex joint survival-multinomial mixed model. Observed outcomes represented the age at diagnosis and stage which was a combination of the actual cancer stage and grade. Chefo and Tsodikov developed a stable and structured MLE approach obtaining the model estimates iteratively. The approach was based on generalized self-consistency and the quasi-EM algorithm was used to handle the mixed multinomial part of the

response through a Poisson likelihood.

1.5.3 Survival Analysis

Fleming and Lin (2000) gave a nice overview of survival analysis methods, techniques, and areas of the research. The developments in this field that had the most impact on clinical trials were the Kaplan-Meier method for estimating the survival function, the log-rank statistic for comparing two survival distributions, and the Cox proportional hazards model for quantifying the effects of covariates on the survival time. Significant progress has been achieved and further developments are expected in many areas including the accelerated failure time model, multivariate failure time data, depending censoring, joint modeling of failure time and longitudinal data, Bayesian survival methods, etc.

Cox and Oakes (1984), *Klein and Moeschberger* (1999), and *Hosmer and Lemeshow* (1999) provide a detailed explanation, including examples of the standard survival data analysis and techniques for censored and truncated data.

The theory for the analysis of interval-censored data has been developed over the past three decades and several good reviews have been written. However, it is still a common practice in clinical trials to simplify the interval censoring structure of the data into a more standard right censoring case. Reviews written by *Huang and Wellner* (1997) and *Lindsey and Ryan* (1998) have been a keystone, but are outdated by many of the newer interval-censored methods. The more recent book by *Sun* (2006) addresses statistical issues and describes statistical methods for the analysis of singly and doubly interval-censored survival data arising from AIDS, cancer and other disease studies. Parametric survival models for interval-censored data with time-dependent covariates are described in work by *Sparling et al.* (2006). The most recent review by *Gomez et al.* (2009) includes methodology on non-parametric, parametric, and semi-parametric estimating approaches, and reviews software for an-

alyzing interval-censored data.

1.5.4 Cancer Post-treatment Surveillance

Post-treatment cancer surveillance represents a discrete observational process yielding incomplete information on the time to cancer recurrence. Instead of the actual time of recurrence only the time of examination is available, which usually follows the specific discrete schedule. Additionally, false-positive and false-negative rates of the diagnostic test may be present. There exists a broad range of literature on parametric and non-parametric estimation of the disease natural history from discrete observations including *Albert et al.* (1978a,b), *Flehinger and Kimmel* (1991), *Klebanov et al.* (1993), *Ivankov et al.* (1993), and *Yakovlev et al.* (1993). If surveillance is error free, the corresponding sample can be considered as the interval-censored.

1.5.5 Multivariate Survival Techniques

Another way to model the recurrence data, time to recurrence to a particular stage or to a grouped stage, is by use of dependent competing risks. A parametric model with two dependent competing risks and the estimation of parameters are briefly discussed in *Yakovlev and Tsodikov* (1996). More information on the multivariate survival data and analysis including the multivariate parametric and non-parametric estimation can be found in *Hougaard* (2000).

Thall et al. (2000) proposed an approximate Bayesian method for comparing two treatments based on multivariate patient outcomes. They partitioned the parameter space into four sets: a set where the experimental treatment is superior to the control treatment, a set where two treatments are equivalent, a set where the control treatment is superior to the experimental one, and a set where the treatment effects are discordant. Then they computed posterior probabilities of the parameter sets by treating an estimator of the parameter vector as a random variable in the Bayesian

paradigm.

1.5.6 Stochastic Processes

Stochastic processes theory provides another way to look at and model the recurrence data. A general review of stochastic processes, including the Poisson process, Markov chains, martingales, and Brownian motion theory, is provided by *Ross* (1996).

The counting-process martingale theory pioneered by Aalen provided a unified framework for studying the small- and large-properties of survival analysis statistics, see *Fleming and Lin* (2000). *Fleming and Harrington* (2005) give the detailed description and provide applications of counting processes and martingales to survival analysis.

Yakovlev and Tsodikov (1996) consider threshold models of tumor latency. The simple model describing the dynamics of tumor growth is a linear birth-and-death process with two absorbing states. More general model, a semistochastic threshold model of tumor recurrence, is introduced and evaluated by the authors.

King et al. (2008) model the process of cholera (an infectious disease) using an iterated filtering algorithm. The models were formulated as stochastic differential equations which were integrated using the Euler-Maruyama algorithm. *Breto et al.* (2009) continued working with cholera data and developed a framework for constructing nonlinear mechanistic models. This work builds on recently developed plug-and-play inference methodology for partially observed Markov models. *He et al.* (2010) model the measles data using the plug-and-play approach.

In Chapter II we proceeded with a continuous follow-up assumption using stochastic models of cancer recurrence and describe the models and methods for the right censored data. Then we extend the approach to allow for a discrete follow-up process. Chapter III describes the models and methods for the interval-censored data. Our conclusions are given in Chapter IV.

CHAPTER II

Joint Modeling of Time to Recurrence and Cancer Stage at Recurrence in Oncology Trials - When Event Times Are Right Censored (Continuous Follow-up Observation Process)

2.1 Introduction

2.1.1 Motivation

This research was motivated by a clinical trial with bladder cancer patients who went through surgery and were followed up for tumor recurrence. The surgery was conducted in patients with an early cancer stage (T_a and T_1) and either with first or recurrent cancer at diagnosis. There was a control group using the standard procedure, and an experimental group with a drug designed to enhance a detection of suspected cancer lesions. One of the study objectives was to evaluate and compare the time to tumor recurrence of patients in control and experimental groups. At the time of tumor recurrence, the disease stage was also evaluated. The stage of disease at recurrence significantly impacts future treatment and quality of life of a patient. Therefore, modeling and analyzing the time to tumor recurrence and the stage at recurrence jointly makes a lot of sense and gives more powerful inference.

Oncology clinical trials are conducted mainly in advanced stage cancer patients with high mortality rate. But not all cancers have high mortality rate while the treatment cost and the burden of cancers are fairly high. One of such cancers is bladder cancer. The American Cancer Society estimated about 70,530 new cases (about 52,760 men and 17,770 women) and 14,680 deaths (about 10,410 men and 4,270) from bladder cancer in the US in 2010 (www.cancer.org). In spite of the increased incidence, the rate of people dying of this cancer has decreased over the past 20 years. More than 500,000 people in the United States are survivors of this cancer. Bladder cancer is one of the most expensive cancers for society because patients live longer and have multiple recurrences. Depending on a patient's characteristics after TURB (transurethral resection of the bladder) the probability of recurrence at one year ranges from about 15% to 61% and from 31% to 78% at 5 years (*Lee et al.*, 2006). The major prognostic factors for recurrence and progression are tumor multiplicity, size, previous recurrence rates, baseline tumor (T) stage, presence of CIS and tumor grade (*Kurth et al.*, 1995). The high rate of early recurrences (up to 60% within 3 months) reported in the literature suggests that a significant percentage of tumor recurrences result from residual tumor left behind at resection or growth of previously undetected microscopic lesions. See *Sylvester et al.* (2006) for more details and references therein.

This problem in not unique to the bladder cancer trials, cancer trials in other indications evaluating patients at the early stage when the surgery or the treatment with expectation of complete recovery is possible, anticipate the cancer recurrence in some patients for whom the time to recurrence and the cancer stage at recurrence would make a difference with respect to subsequent treatment and a quality of the life.

2.1.2 Brief Overview

Many scientists investigated mathematical modeling of carcinogenesis problem. The majority of models used the elements of the birth-and-death stochastic processes theory. Tan gives a comprehensive analysis of this class of models in his work *Tan* (1991). A common weak point in many Markovian models of carcinogenesis is that the description of tumor progression is not sufficiently advanced. The time to observing a tumor is not equal to the time at which the first malignant cell is generated. Additionally, the estimation procedure is quite tedious even in computationally feasible cases *Tan and Chen* (1993). Therefore, the search for new ways of modeling the carcinogenesis seems to be very reasonable. The mathematical description of tumor latency with regard to the tumor recurrence and regression analysis of tumor recurrence data described in work by *Yakovlev and Tsodikov* (1996). Authors discuss different stochastic models with parameters that have clear biological meaning. The proposed parametric models describe the process of cancer recurrence. Authors suggest that there are several causes of local cancer recurrence including the recurrence caused by the cancer cells surviving a treatment or a surgery and the recurrence caused by spontaneous carcinogenesis which have different mathematical representation and biological interpretation. In our research, we used the proposed models to build the joint model of the time to bladder cancer recurrence and the cancer stage at recurrence.

Multinomial-Poisson (MP) transformation has been a popular technique to simplify maximum likelihood estimation and has been researched by many scientists including *Baker* (1994). The approach works by substituting a Poisson likelihood for the multinomial likelihood at the cost of augmenting the model parameters by axillary ones. The MP transformation is justified through the method of Lagrange multipliers by *Lang* (1996). *Tsodikov and Chefo* (2009, 2008) proposed an alternative approach based on generalized self-consistency methodology that allows to use Pois-

son likelihood with arbitrary covariate structure. The authors modeled the prostate cancer data using the complex joint survival-multinomial mixed model. Observed outcomes represented the age at diagnosis and stage which was a combination of the actual cancer stage and grade. Chefo and Tsodikov developed a stable and structured MLE approach obtaining the model estimates iteratively. The approach was based on generalized self-consistency and the quasi-EM algorithm used to handle the mixed multinomial part of the response through Poisson likelihood. This work was extended from the work of *Tsodikov* (2003). Tsodikov developed a generalized self-consistency approach to MLE estimation and model building in a survival analysis setting.

In this chapter, we describe a model for the joint distribution of time to recurrence and cancer stage at recurrence that accounts for the recurrence caused by the cancer cells surviving a treatment or a surgery and for the recurrence caused by spontaneous carcinogenesis when event times are right censored. We provide methods for full maximum likelihood estimation based on the EM algorithm. The methods are described in Section 2.2 of this chapter. Section 2.3 outlines the real-life example based on the clinical trial in patients with bladder cancer. The methods are illustrated through modeling and estimation of data from this trial. The simulations used to assess the sensitivity of the methods are presented in Section 2.4. Section 2.5 summarizes the results and gives conclusions.

2.2 Models and Methods

2.2.1 The Model of Cancer Recurrence

In this section, we outline a parametric model that will be used to model the cancer recurrence in patients with bladder cancer who went through the surgery. *Yakovlev* (1993) proposed a simple stochastic model for cancer recurrence incorporating parameters that have clear biological meaning. At the end of surgery, the

cancer cells that were not resected possess the capacity of giving rise to an overt tumor. These cells, clonogens, will propagate into a newly detectable tumor. The initial number of clonogens is modeled as a Poisson random variable with expectation θ_1. Let X_i be a random time for the ith clonogen to produce a detectable tumor. The non-negative random variables X_i are independent and identically distributed with a common cumulative distribution function $F_1(t)$. This assumption is natural if the surviving tumor clonogens are in small proportion and wide apart from each other which is likely to occur in a treated tumor. The time to tumor recurrence (latent period) can be defined as the random variable V such that

$$V = \min_{\{i:0\leq i\leq \text{ num. of surv. clonogens}\}} X_i, \tag{2.1}$$

where $X_0 = +\infty$ with probability one. Then the survival function for the random variable V is the following:

$$S_1(t) = \sum_{k=0}^{\infty} \left\{ \frac{\theta_1^k \exp(-\theta_1)}{k!} (1 - F_1(t))k \right\} = \exp\left(-\theta_1 F_1(t)\right). \tag{2.2}$$

The key advantage of expression (2.2) is to show the contribution of the two distinct characteristics of tumor growth: the expected number of surviving clonogens θ_1 and the rate of their progression described by the c.d.f. $F_1(t)$. Estimation of both characteristics is feasible and provides additional information on the biology of tumor recurrence. Another advantage is due to the fact that survival function (2.2) corresponds to an improper (substochastic) distribution and its limiting value $S_1(+\infty) = \exp(-\theta_1)$ represents the probability of tumor cure (no recurrence) or the surviving fraction. The difficulties associated with the estimation of surviving fraction from censored observations within the non-parametric framework are known and described in works by *Pepe and Fleming* (1989) and *Cantor and Shuster* (1992). Most parametric survival models implicitly assume a zero limiting survival probability as in

Kalbfleisch and Prentice (2002). The importance of allowing for surviving fractions in failure-time models has been recognized by many scientists. In parametric analyses, this concept leads to the necessity of improper distributions in the analysis of failure time data. These distributions do not need to be of the mixture type as deliberated by *Yakovlev* (1994). The model specified by (2.2) allows for the surviving fraction in a natural way.

When considering the cause of tumor recurrence, it is important to consider the possibility of tumor appearance due to an enhanced transformation rate and depression of the immune system in the subject. The model proposed by *Klebanov et al.* (1993) includes the description of spontaneous carcinogenesis as a special case. A non-stationary generalization of the model was given by Yakovlev, Tsodikov, and Bass in *Yakovlev* (1993). Once a malignant cell comes into existence, its growth is irreversible and the progression begins resulting in a detectable tumor after some time. The primary event in the process of carcinogenesis is the formation of an intracellular lesion which is potentially carcinogenic. These precancerous lesions located in different target cells possess the capacity for producing a detectable tumor in the long run. The primary events occur at random times and their sequence in time represents a point stochastic process. This process will be considered a Poisson process with intensity $\theta_2(t)$, so that the number of lesions accumulated by time T is a Poisson random variable with expectation $\int_0^T \theta_2(t)dt$. Let a random variable Y_i be the time from the ith lesion formation to the observable overt tumor caused by this lesion. The nonnegative random variables Y_i, $i = 1, 2, \ldots$, are assumed to be independent and identically distributed with the common c.d.f. $F_2(y)$. Let $\nu(t)$ be the number of misrepaired lesions (cells are endowed with a capacity to repair lesions, but some lesions remain unrecognized by the repair system and some lesions happen to be misrepaired) accumulated in an organism by time t. We assume that $\nu(t)$ is independent of random variables $Y_1, Y_2, \ldots$. The latent period is defined as a random variable W

such that:

$$W = \min_{i:0\leq i\leq \nu(t)} (E_i + Y_i), \tag{2.3}$$

where E_i is the time of the ith lesion formation given that this time is less than T, and E_i and Y_i are mutually independent with $E_0 + Y_0 = +\infty$ (no lesion) with probability one. Tumor recurrence remains latent until either it is detected or a censoring event occurs. The survival function of the random variable W is:

$$S_2(t) = \exp\left\{ -\int_0^t \theta_2(x) F_2(t-x)dx \right\}, \tag{2.4}$$

where $\theta_2(t)$ is the rate of formation of intracellular lesions, and the rate of their progression described by the function F_2. When $\theta_2(t)$ is constant over time, $S_2(t)$ becomes:

$$S_2(t) = \exp\left\{ -\int_0^t \theta_2 F_2(x)dx \right\}, \tag{2.5}$$

which is best matched to model $S_1(t)$ with regards to estimation problems.

There are no pathological or clinical criteria for discrimination between possible causes of cancer recurrence. By studying the temporal characteristics of tumor latency, an appropriate solution to this problem might be found. The results by *Hoang et al.* (1996) show that discrimination between true recurrence and spontaneous carcinogenesis is feasible. In our real-life example we are evaluating patients with bladder cancer who went through the surgery. Depending on a patient's characteristics after TURB (transurethral resection of the bladder) the probability of recurrence at one year ranges from about 15% to 61% and from 31% to 78% at 5 years (*Lee et al.* (2006)). The high rate of early recurrences (up to 60% within 3 months) reported in the literature suggests that a significant percentage of tumor recurrences result from residual tumor left behind at resection or growth of previously undetected microscopic lesions.

As discussed above, a recurrence of cancer can be caused by cancer cells surviving the treatment or surgery or by spontaneous carcinogenesis. Assuming that these two reasons are the cause of a cancer recurrence, the survival function of time to recurrence can be written as a product of two survival functions:

$$S(t) = \exp\left(-\theta_1 F_1(t) - \int_0^t \theta_2 F_2(x)dx\right) = S_1(t)S_2(t), \tag{2.6}$$

where $S_1(t) = \exp\left(-\theta_1 F_1(t)\right)$ describes the time to tumor recurrence from cancer cells that survive treatment, and $S_2(t) = \exp\left(-\int_0^t \theta_2 F_2(x)dx\right)$ describes the time to tumor recurrence by spontaneous carcinogenesis. Here θ_1 is the mean number of cancer cells surviving a treatment or surgery and $F_1(t)$ is a c.d.f. describing the rate of their progression; θ_2 is the rate of formation of intracellular lesions and $F_2(t)$ is a c.d.f. describing the rate of their progression.

To allow functional dependence on covariate information, the rates θ_1 and θ_2 will be modeled parametrically as:

$$\theta_1(Z) = \exp\left(\beta_{01} + \beta_1^T Z\right),\ \theta_2(Z) = \exp\left(\beta_{02} + \beta_2^T Z\right), \tag{2.7}$$

where Z is a vector of values of explanatory variables and β_{ij} are regression coefficients.

Let introduce random variable U which will take the following values:

$$U = \begin{cases} 1, \text{ if recurrence is caused by spontaneous carcinogenesis,} \\ 0, \text{ if recurrence is caused by surviving a treatment cancer cells.} \end{cases} \tag{2.8}$$

Note that U is a random variable which is not observed, but is used to differentiate the cause of cancer recurrence.

2.2.2 Multinomial Logit Model

Let $X_i \in \{1, 2, \cdots, M\}$ be the ith subject's multinomial response (cancer stage) in one of the M possible categories. On the complete-data level, multinomial probabilities are modeled using log-linear predictors $\pi_m(z_i, t_i, u_i)$ specific to categories m and conditional on a vector of covariates Z_i, time T_i, and indicator U_i:

$$\Pr\{X_i = m | Z_i, T_i, U_i\} = \frac{\pi_m(z_i, t_i, u_i)}{1 + \sum_{c=2}^{M} \pi_c(z_i, t_i, u_i)}, \tag{2.9}$$

where for identifiability, regression coefficients corresponding to the first category are set to zero. We will use the following parameterization of function π_m using regression coefficients α_m:

$$\pi_m(z_i, t_i, u_i) = \exp\left(\alpha_m \cdot z_i + \alpha_{t,m} \cdot t_i + \alpha_{u,m} \cdot u_i\right). \tag{2.10}$$

2.2.3 Joint Distribution and Likelihood

In survival analysis we observe the time to recurrence or the time at which a subject did not have a cancer recurrence which was confirmed by an objective medical evaluation and then the information was not collected after a certain period or it was missing. In this case, the event is considered to be right censored at the time of an objective evaluation confirming no recurrence. The event is right censored when follow-up is curtailed with observing the event. Let

$$\delta_i = \begin{cases} 1, \text{ if } i^{th} \text{ subject had cancer recurrence,} \\ 0, \text{ otherwise.} \end{cases} \tag{2.11}$$

It is convenient to represent time to recurrence data subjected to random censoring by the n pairs of the form (t_i, δ_i), where t_i are observed times, and δ_i is a censoring index, $i = 1, \ldots, k$. If the censoring is non-informative, then the likelihood for right

censored data is:

$$\mathbb{L} \propto \prod_{i=1}^{k} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}. \tag{2.12}$$

Let's build the joint distribution for $k = 1$. Given $\delta = 0$, $S(t) = S_1(t)S_2(t)$. Then given $\delta = 1$, the joint density for a time to recurrence and a cancer stage at recurrence can be described by the following pdf:

$$\begin{aligned} (T, X) \sim f(t, m) &= s_1(t)S_2(t)\frac{\pi_m(t, Z, u = 0)}{1 + \sum_{c=2}^{M} \pi_c(t, Z, u = 0)} \\ &+ S_1(t)s_2(t)\frac{\pi_m(t, Z, u = 1)}{1 + \sum_{c=2}^{M} \pi_c(t, Z, u = 1)}, \end{aligned} \tag{2.13}$$

where $s_1(t)$ and $s_2(t)$ are pdf given by $1-S_1(t)$ and $1-S_2(t)$ distributions, respectively.

Denote

$$\rho_{0,m}(t, Z) = \frac{\pi_m(t, Z, u = 0)}{1 + \sum_{c=2}^{M} \pi_c(t, Z, u = 0)} \quad \text{and} \quad \rho_{1,m}(t, Z) = \frac{\pi_m(t, Z, u = 1)}{1 + \sum_{c=2}^{M} \pi_c(t, Z, u = 1)}.$$

Then since $s_1(t) = \theta_1(Z)f_1(t)S_1(t)$ and $s_2(t) = \theta_2(Z)F_2(t)S_2(t)$, the joint pdf $f(t, m)$ can be expressed in the following way:

$$\begin{aligned} f(t, m) &= \theta_1(Z)f_1(t)S_1(t)S_2(t)\rho_{0,m}(t, Z) + \theta_2(Z)F_2(t)S_1(t)S_2(t)\rho_{1,m}(t, Z) \\ &= S(t)\left(\theta_1(Z)f_1(t)\rho_{0,m}(t, Z) + \theta_2(Z)F_2(t)\rho_{1,m}(t, Z) \right), \end{aligned} \tag{2.14}$$

where $f_1(t)$ is the pdf corresponding to a distribution given by $F_1(t)$. Therefore,

$$f(t, m, u) = \begin{cases} S(t)\theta_1(Z)f_1(t)\rho_{0,m}(t, Z), & \text{if } u = 0, \\ S(t)\theta_2(Z)F_2(t)\rho_{1,m}(t, Z), & \text{if } u = 1. \end{cases} \tag{2.15}$$

The conditional pdf of $U = u$ given $T = t$ and $X = m$ is

$$f(u|t, m) = \frac{f(t, m, u)}{f(t, m)},$$

so that if $u = 0$, then

$$f(u = 0|t, m) = \frac{\theta_1(Z) f_1(t) \rho_{0,m}(t, Z)}{\theta_1(Z) f_1(t) \rho_{0,m}(t, Z) + \theta_2(Z) F_2(t) \rho_{1,m}(t, Z)}, \tag{2.16}$$

and if $u = 1$

$$f(u = 1|t, m) = \frac{\theta_2(Z) F_2(t) \rho_{1,m}(t, Z)}{\theta_1(Z) f_1(t) \rho_{0,m}(t, Z) + \theta_2(Z) F_2(t) \rho_{1,m}(t, Z)}. \tag{2.17}$$

The full likelihood is proportional to the likelihood associated with the event time distribution and cancer stage $L(\beta)$, where β is a vector of parameters need to be estimated from the model. The observed data log-likelihood is $\log L(\beta)$ calculated as following:

$$\begin{aligned} l &= \log L(\beta) = \sum_{i \in \text{ non-recurrences}} \left(-\theta_1(Z_i) F_1(t_i) - \int_0^{t_i} \theta_2(Z_i) F_2(x) dx \right) \\ &+ \sum_{i \in \text{ recurrences}} \log \left(S(t_i) \theta_1(Z_i) f_1(t_i) \rho_{0,x_i}(t_i, Z_i) + S(t_i) \theta_2(Z_i) F_2(t_i) \rho_{1,x_i}(t_i, Z_i) \right). \end{aligned} \tag{2.18}$$

The complete data log-likelihood is:

$$\begin{aligned} l_{cd} &= \sum_{i \in \text{ non-recurrences}} \left(-\theta_1(Z_i) F_1(t_i) - \int_0^{t_i} \theta_2(Z_i) F_2(x) dx \right) \\ &+ \sum_{i \in \text{ recurrences}} \{ u_i \cdot \log \left(S(t_i) \theta_2(Z_i) F_2(t_i) \rho_{1,x_i}(t_i, Z_i) \right) \\ &+ (1 - u_i) \cdot \log \left(S(t_i) \theta_1(Z_i) f_1(t_i) \rho_{0,x_i}(t_i, Z_i) \right) \}. \end{aligned} \tag{2.19}$$

Our approach will be to use EM algorithm, with the E-step solving the problem of imputation of U and the M-step maximizing a log-likelihood obtained from the complete data model.

2.2.4 The EM Algorithm

The EM algorithm is formulated as follows.

Step 1: Set initial values of regression coefficients and distribution parameters

$$\beta^{(0)} = (\beta_1, \beta_2, \alpha, \text{ parameters from } F_1(t) \text{ and } F_2(t) \text{ distributions}) .$$

Step 2: E-step. Calculate a vector

$$\begin{aligned} \hat{U}(\beta^{(k)}) &= E(U|\text{Observed data} = (t, m), \delta = 1) \\ &= \frac{\theta_2(Z_i)F_2(t_i)\rho_{1,x_i}(t_i, Z_i)}{\theta_1(Z_i)f_1(t_i)\rho_{0,x_i}(t_i, Z_i) + \theta_2(Z_i)F_2(t_i)\rho_{1,x_i}(t_i, Z_i)}. \end{aligned} \tag{2.20}$$

Step 3: M-step. Maximize the log-likelihood obtained from the complete data model at $\hat{U}(\beta^{(k)})$ which can be achieved by maximizing separately l^{ρ}_{cd} and $l^1_{cd} + l^2_{cd}$, where

$$l^{\rho}_{cd} = \sum_{i\in \text{ recurrences}} \left\{ \hat{u}_i(\beta^{(k)}) \cdot \log \rho_{1,x_i}(t_i, Z_i) + (1 - \hat{u}_i(\beta^{(k)})) \cdot \log \rho_{0,x_i}(t_i, Z_i) \right\},$$

$$\begin{aligned} l^1_{cd} + l^2_{cd} &= \sum_{i\in \text{ non-recurrences}} \left(-\theta_1(Z_i)F_1(t_i) - \int_0^{t_i} \theta_2(Z_i)F_2(x)dx \right) + \\ &+ \sum_{i\in \text{ recurrences}} \{\hat{u}_i(\beta^{(k)}) \cdot \log \left(S(t_i)\theta_2(Z_i)F_2(t_i)\right) \\ &+ (1 - \hat{u}_i(\beta^{(k)})) \cdot \log \left(S(t_i)\theta_1(Z_i)f_1(t_i)\right)\} \\ &= \sum_{i\in \text{ all events}} -\theta_1(Z_i)F_1(t_i) + \sum_{i\in \text{ recurrences}} (1 - \hat{u}_i(\beta^{(k)})) \log(\theta_1(Z_i)f_1(t_i)) \\ &+ \sum_{i\in \text{ all events}} -\int_0^{t_i} \theta_2(Z_i)F_2(x)dx + \sum_{i\in \text{ recurrences}} \hat{u}_i(\beta^{(k)}) \log(\theta_2(Z_i)F_2(t_i)). \end{aligned} \tag{2.21}$$

Denote the solution by $\beta^{(k+1)}$.

Step 4: Set $k = k + 1$. Continue with Step 2 and Step 3 iterations until convergence.

Standard error estimates are based on the inverse of the observed information matrix:

$$I = -\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T}, \tag{2.22}$$

where β is the vector of model parameters and $l(\beta) = \log E\{L(\beta|U)\}$ is the model log-likelihood maximized as a result of EM algorithm. The observed information matrix is derived by an application of the missing information principle representing the observed information as the difference between expected complete-data information and the missing information, given observed data, see *McLachlan and Krishnan* (2008). Alternatively, a bootstrap estimate of standard errors could be done using Efron's approach, see *Efron* (1994).

2.3 Real-life Example: Bladder Cancer Trial

2.3.1 Background

The clinical 305 trial described in Section 2.2.2 was a trial conducted in patients with bladder cancer. The trial had two parts: a pivotal 9 month study and an extension part of the study which captured patient information for approximately 5 years after the completion of the pivotal study. About 560 patients with histologically superficial bladder cancer (T_a and T_1) confirmed by a local pathologist were included in the study. The recurrence analysis included 551 patients. At inclusion the patients were randomized to have their cystoscopy including TURB by white light only or by white light plus blue light with the experimental drug. In addition, the patients were stratified by cancer history (initial and recurrent bladder cancer).

Patients were followed-up by cystoscopy in white light after the resection proce-

dure at three, six and nine months. The results from the local pathologist at baseline were used to determine if the patient was to be followed up at 3, 6 and 9 months. Recurrence was to be verified by histology assessment of the local pathologist during visits. Recurrence was defined as presence of either a CIS, T_a, T_1 or $T_2 - T_4$ tumor. Suspected areas seen during cystoscopy at baseline and at follow-up visits were biopsied or resected. The urologist recorded the bladder sector in which the lesion or suspected area was found, whether the lesion appeared visually to be flat or papillary, and whether the lesion/suspected area was visible in white light (or blue light at the baseline). The results were recorded as normal, flat lesions (classified as dysplasia, hyperplasia, carcinoma in situ (CIS)), or papillary lesions (classified as T_a, T_1, T_2, T_3, T_4 according to the TNM staging). In addition, the WHO grade was recorded if applicable as 1, 2, 3 for the papillary tumors. If there were multiple pathology results for a single lesion or multiple cancer lesions were identified for a patient, the worst lesion type was used in the analysis. However, if there was a CIS in addition to a papillary lesion reported, both results were included. In the analysis of the trial data, the cancer stage order from less advanced to more advanced stages was determined clinically as the following: $T_a < CIS < (T_a + CIS) < T_1 < (T_1 + CIS) < T_2 < (T_2 + CIS) < T_3 < (T_3 + CIS) < T_4 < (T_4 + CIS)$. Risk factors such as smoking, occupational exposure to aromatic amines, history of kidney stones, and family history were not collected in this study while prognostic factors for recurrence/ progression (e.g., number of lesions, tumor stage, tumor grade, presence of CIS, and previous recurrences) were collected for all patients.

2.3.2 Data Summary

Both therapy groups had similar patients with respect to the baseline characteristics such as gender, race, age, and the bladder history *Stenzl et al.* (2010).

The scheduled visits for the cystoscopy were at 3, 6, and 9 months. The extension

part of the study collected the follow up data for 5-6 years after the original therapy unless patient died prior to the follow-up period. The recurrence data used in the study was treated slightly different for the analysis in this paper from what was reported after the study was completed. Only patients who had the cancer stage at recurrence available and confirmed by histology were considered as ones with recurrence. If the cancer stage was missing or not confirmed, the time to recurrence was censored.

The visual summary of the data is provided by Figure 2.1. It gives a histogram of the observed data (top left panel), a density function of observed data (top right panel), a histogram of recurrence data (bottom left panel), and a histogram of censored data (bottom right panel). One can notice that the majority of patients with cancer recurrences were diagnosed during the first year after the surgery.

The Kaplan-Meier (KM) estimate of the median recurrence time in the standard group was 9.5 months with the number of events of 142. The Kaplan-Meier estimate of the median recurrence time in the experimental group was 16.4 months with the number of events of 125. The p-value from the Wilcoxon test was 0.043. From the KM Figure 2.2 (left panel), it is noticeable that the separation between the survival curves started after 6 months and continued until the end of the follow-up period suggesting the better outcome in the experimental group. Figure 2.2 (right panel) shows the cumulative hazard by therapy group.

The Kaplan-Meier curve of the overall time to cancer recurrence with the 95% confidence intervals is presented on Figure 2.3.

The proportional hazard model (Cox regression) of the recurrence time revealed the significant effect of therapy, country, baseline cancer stage, and cancer history. The significant effect of baseline cancer stage and cancer history was expected as these variables are considered to be the prognostic factors for time to recurrence/ progression. The effect of the country can be explained by the following fact: this

study was the multi-country study ran in Europe, Canada, and USA; while the sites in Canada and USA used the technology the first time, the European sites had some experience with it already (this experimental therapy was approved by the European Medicine Agency in 2004).

The number of patients with recurrence in the standard therapy group was 142 out of 280 patients, and the number of patients with recurrence in the experimental group was 125 out of 271 patients. The distribution of the stages by group is described in Table 2.1. The χ-squared test did not show a significant difference between groups

Cancer Stage at Recurrence	Experimental Therapy (Number of Patients)	Standard Therapy (Number of Patients)
Missing	16	19
0 (None or not confirmed)	130	119
1 (T_a)	102	109
2 (CIS)	3	5
3 ($T_a + CIS$)	4	5
4 (T_1)	7	11
5 ($T_1 + CIS$)	3	3
6 ($T_2 - T_4$)	5	7
7 ($T_2 - T_4 + CIS$)	1	2

Table 2.1: Study Cancer Stage at Recurrence.

using the ordered outcomes of stages although the marginal difference toward the better outcomes (e.g., less recurrences and recurrences at less aggressive stages) in the experimental group needs to be noted here.

The cancer recurrence was observed as early as at 3 months after the surgery. While only 7 patients with recurrences were observed in the experimental group, 21 patients with recurrences were observed in the standard group. It was noted that out of 21 patients 7 progressed to more advanced stages. The distribution of the stages is presented in the table below:

Cancer Stage at Recurrence during first 3 months	Experimental Therapy (Number of Patients)	Standard Therapy (Number of Patients)
1 (T_a)	6	12
2 (CIS)	1	2
3 ($T_a + CIS$)	0	0
4 (T_1)	0	3
5 ($T_1 + CIS$)	0	1
6 ($T_2 - T_4$)	0	1
7 ($T_2 - T_4 + CIS$)	0	2

Table 2.2: Recurrence Cancer Stage at 3 months.

2.3.3 Data Modeling

The time to cancer recurrence and the cancer stage at recurrence data were modeled using joint modeling methods described in Section 2.2. Covariates that showed significant effect during the preliminary analysis such as procedure group, cancer history (previous recurrences), and baseline tumor stage were included in the model. Therefore, the mean number of cancer cells surviving the surgery θ_1 and the rate of formation of intracellular lesions θ_2 were based on the following parametrization of the predictors:

$$\theta_i(z) = \exp\left(\beta_{i0} + \beta_{i1} \cdot z1 + \beta_{i2} \cdot z2 + \beta_{i3} \cdot z3\right), \tag{2.23}$$

where $z1$ = procedure group, $z2$ =cancer history, and $z3$ =baseline tumor stage.

Seven cancer stages (T_a, CIS, T_a+CIS, T_1, T_1+CIS, T_2-T_4, and T_2-T_4+CIS) were evaluated in the model.

$$\Pr\left\{X_i = m | Z_{ij}, T_i, U_i\right\} = \frac{\pi_m(z_{ij}, t_i, u_i)}{1 + \sum_{c=2}^{7} \pi_c(z_{ij}, t_i, u_i)}, \tag{2.24}$$

$$\pi_m(z_{ij}, t_i, u_i) = \exp\left(\alpha_{m1} \cdot z_{i1} + \alpha_{m2} \cdot z_{i2} + \alpha_{m3} \cdot z_{i3} + \alpha_{t,m} \cdot t_i + \alpha_{u,m} \cdot u_i\right),$$

where $i \in \{1, \ldots, 551\}$; and $j \in \{1, 2, 3\}$.

The modified cancer stages such as less aggressive (T_1), moderately aggressive

Parameter	MLE	SE	Wald Statistics	P-value
β_{10}	-1.7513	0.36608	-4.6678	<0.001
β_{11}	0.3039	0.13554	2.2023	0.0281
β_{12}	0.5122	0.14704	3.3707	<0.001
β_{13}	0.0289	0.06227	0.3925	0.6949
γ_1	8.2432	0.37193	22.1194	<0.001
γ_2	1.7529	0.08579	20.4572	<0.001
β_{20}	-5.3287	0.20151	-26.4439	<0.001
β_{21}	-0.1296	0.01583	-8.1869	<0.001
β_{22}	0.1811	0.08112	2.2325	0.0201
β_{23}	0.0674	0.00342	19.7076	<0.001
λ_1	0.1254	0.02231	5.6208	<0.001

Table 2.3: Parameter Estimations for Survival Function of Time to Recurrence.

(CIS, $T_a + CIS$, T_1, and $T_1 + CIS$), and more aggressive ($T_2 - T_4$ and $T_2 - T_4 + CIS$) were evaluated in the model as well, but these stage combinations did not result in a better fit. Several combinations of distributions were fit the recurrence time data. Gamma, Weibull, log-normal, Makeham, Compertz and exponential distributions were evaluated as possible candidates for the survival function describing the rate of progression of cancer cells surviving the surgery. Gamma, Weibull, and exponential distributions were evaluated as potential candidates for the survival function describing the rate of spontaneous carcinogenesis. Weibull distribution (γ_1, γ_2) for time to tumor recurrence from cancer cells that survive surgery and exponential distribution (λ_1) for time to tumor recurrence by spontaneous carcinogenesis showed the most appropriate fit graphically and analytically. Models were fit using the proposed EM algorithm which was written in R software. The results of the parameter estimations for the survival function of time to recurrence are presented in Table 2.3.

Figure 2.4 (left panel) gives the graphical estimation of the survival curve of time to cancer recurrence: Kaplan-Meier non-parametric estimation and a parametric estimation adjusted for the recurrence stage. Stepwise curve is the Kaplan-Meier estimate, and the smooth line is the MLE. Figure 2.4 (right panel) gives a similar graphical estimates by procedure group.

Fifty four parameters for the model-predicted marginal probabilities of the cancer stage associated with the cancer recurrence caused by cancer cells surviving the surgery and caused by spontaneous carcinogenesis were estimated from the model, they are presented in Table 2.4, Maximum Likelihood Estimators (MLE) of coefficients from the stage model.

Categories	$\alpha_{m,1}$; $u=0$	$\alpha_{m,2}$; $u=0$	$\alpha_{m,3}$; $u=0$	$\alpha_{t,m}$; $u=0$	$\alpha_{m,1}$; $u=1$	$\alpha_{m,2}$; $u=1$	$\alpha_{m,3}$; $u=1$	$\alpha_{t,m}$; $u=1$	$\alpha_{u,m}$; $u=1$
Stage 2	-0.3150	-0.6763	0.9910	-1.2168	2.0395	1.6534	0.8097	-0.7246	0.6936
Stage 3	-1.2113	0.4469	0.5294	-0.9591	1.2060	-0.7715	0.8260	-0.1446	0.6357
Stage 4	-0.1667	-0.8487	0.4121	-0.3154	0.3737	-0.2968	0.1679	-0.0850	0.9847
Stage 5	-0.3015	-1.5345	0.6671	-0.7112	0.6429	-0.1053	2.4061	-0.4861	0.6499
Stage 6	-0.4717	-1.2408	0.5968	-0.2206	1.0309	-1.2957	0.2788	-0.1798	2.2075
Stage 7	0.3107	-1.9972	0.8325	-0.9091	4.5126	-3.9208	0.0490	-0.9377	3.4217

Table 2.4: MLE.

2.4 Simulations

Simulations were used to assess the sensitivity of the method. Similar to the real-life example, three covariates were simulated and included in the model: procedure group, cancer history (previous recurrences), and baseline tumor stage. It was assumed that only three stages: less aggressive (1), moderately aggressive (2), and more aggressive (3) were the possible cancer stage outcomes.

Parameter	MLE	SE	P-value
β_{10}	0.4497	0.4234	0.2911
β_{11}	0.3632	0.0314	<0.001
β_{12}	6.6384	0.4563	<0.001
β_{13}	-0.6936	0.5811	0.2357
γ_1	0.0179	0.0010	<0.001
γ_2	1.4253	0.1342	<0.001
β_{20}	-1.7796	0.3203	<0.001
β_{21}	-0.1226	0.2752	0.6564
β_{22}	0.3539	0.2784	0.2068
β_{23}	0.0683	0.2738	0.8035
λ_1	5.7801	4.8189	0.2333

Table 2.5: Table X: F1=Weibull(γ_1, γ_2), F2=exponential(λ_1).

Parameter	MLE	SE	P-value
β_{10}	-9.6819	7.7231	0.2131
β_{11}	8.9729	0.7832	<0.001
β_{12}	-3.7458	3.2341	0.2497
β_{13}	-0.7269	0.6342	0.2547
μ_1	56.0891	5.4562	<0.001
μ_2	10.5673	9.9721	0.2920
β_{20}	-1.0888	0.2143	<0.001
β_{21}	0.0812	0.1734	0.6404
β_{22}	-0.1919	0.3245	0.5557
β_{23}	0.7505	0.6931	0.2817
λ_1	1.9570	0.8231	0.0195

Table 2.6: Table Y: F1=Gamma(μ_1, μ_2), F2=exponential(λ_1).

Parameter	MLE	SE	P-value
β_{10}	2.4316	0.4254	<0.001
β_{11}	0.3077	0.0312	<0.001
β_{12}	-0.2567	0.2316	0.2705
β_{13}	-0.3264	0.2811	0.2485
λ_2	0.0179	0.0011	<0.001
β_{20}	-1.016	0.2805	<0.001
β_{21}	0.3499	0.2734	0.2038
β_{22}	-0.2875	0.2788	0.3051
β_{23}	-0.3883	0.2813	0.1707
λ_1	4.3423	3.0959	0.1640

Table 2.7: Table Z: F1=exponential(λ_2), F2=exponential(λ_1).

The Weibull distribution (γ_1, γ_2), gamma distribution (μ_1, μ_2), and exponential distribution (λ_2) were evaluated as suitable candidates for the time to tumor recurrence from cancer cells that survive surgery, and the exponential distribution (λ_1) was evaluated as a suitable candidate for the time to tumor recurrence by spontaneous carcinogenesis. The results of the parameter estimations for the survival function of time to recurrence are presented in Tables 2.5-2.7. These results were obtained from samples of size 100. Given the results in these tables, one can build confidence intervals for model parameters.

2.5 Discussion

We have described a framework for modeling the joint distribution of time to cancer recurrence and cancer stage at recurrence. Our approach accommodates two different causes of the cancer recurrence: recurrence caused by cancer cells surviving a treatment or a surgery and recurrence caused by spontaneous carcinogenesis. The case considered in this chapter based on the continuous follow-up observation process with right censored event times. ML estimation with the EM algorithm was used to estimate the necessary parameters in the model. One could apply the approach proposed by Tsodikov and Chefo to simplify the likelihood maximization of the cancer stage and use the quasi-EM algorithm instead of the EM algorithm.

Modeling the time to cancer recurrence and cancer stage at recurrence jointly allows for more powerful inference. Real-life data and simulations are used to assess the sensitivity and provide robustness of the method. In the real life example, we had to model and estimate 65 parameters. Eleven parameters were used to estimate the time to recurrence distribution. An added benefit of such modeling is that it permits using the cancer stage at recurrence to provide adjusted estimates for the time to recurrence distribution and use them in tests. The cancer stage at recurrence significantly impacts patient quality of life and further treatment. Therefore, it should be accounted in the estimation and analysis of time to cancer recurrence.

A potential limitation of the proposed approach is the use of the parametric functions. There are many different parametric survival distributions that one might need to evaluate in order to find the most appropriate fit for the data.

In this chapter, we assumed the continuous follow-up observation process with right censored event times. In the real-life example, during the first year of the follow-up period patients had scheduled visits at 3, 6, and 9 months; after one year, the follow-up process was based on the regular visits set by investigators and their patients participating in the clinical trial. The more typical clinical trials have scheduled visits

throughout the follow-up period defined by a protocol. In the next chapter we extend our approach to allow for a discrete follow-up process. The interval censored data model will be described in the next chapter.

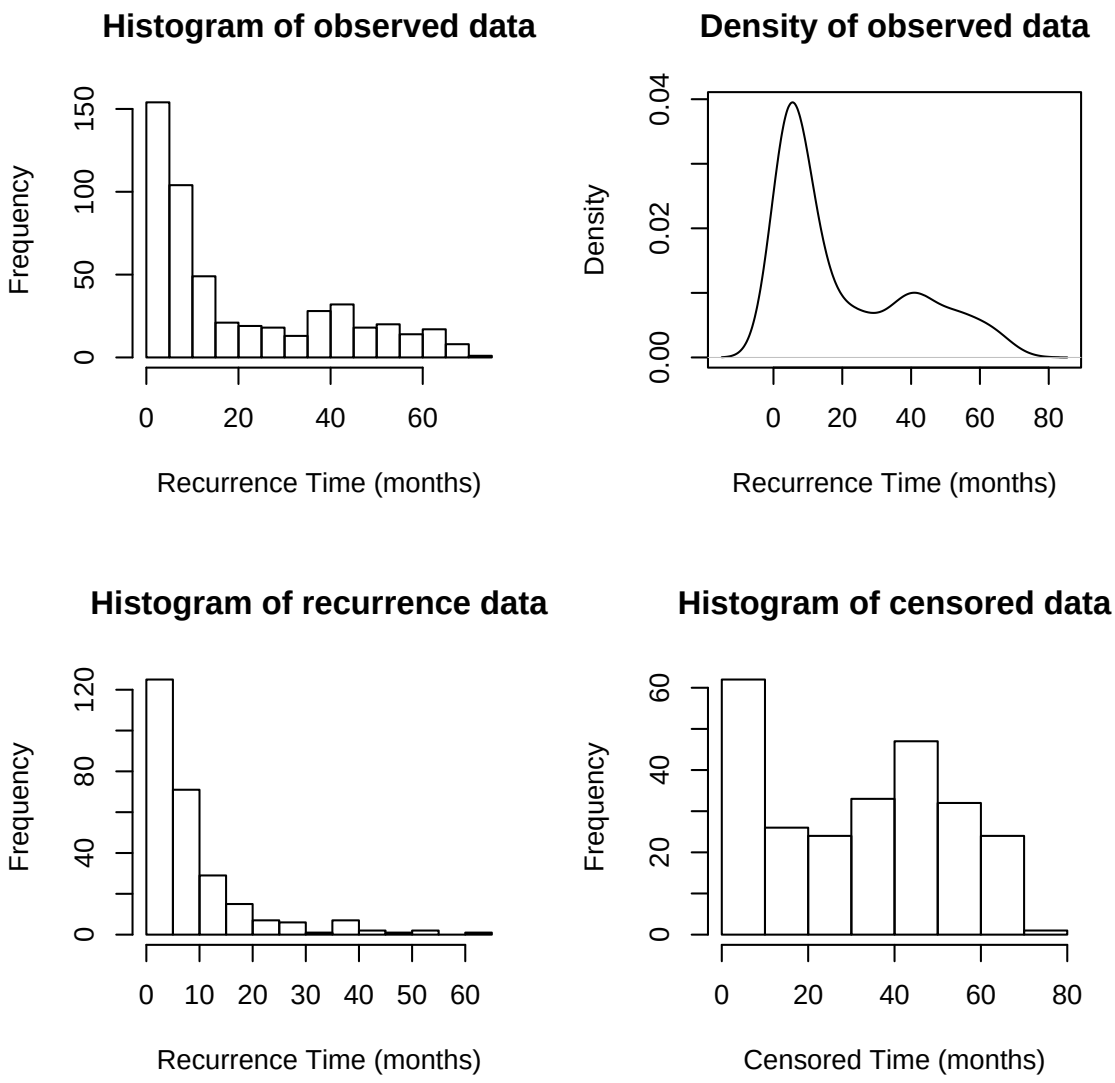

Figure 2.1: Histograms of Data.

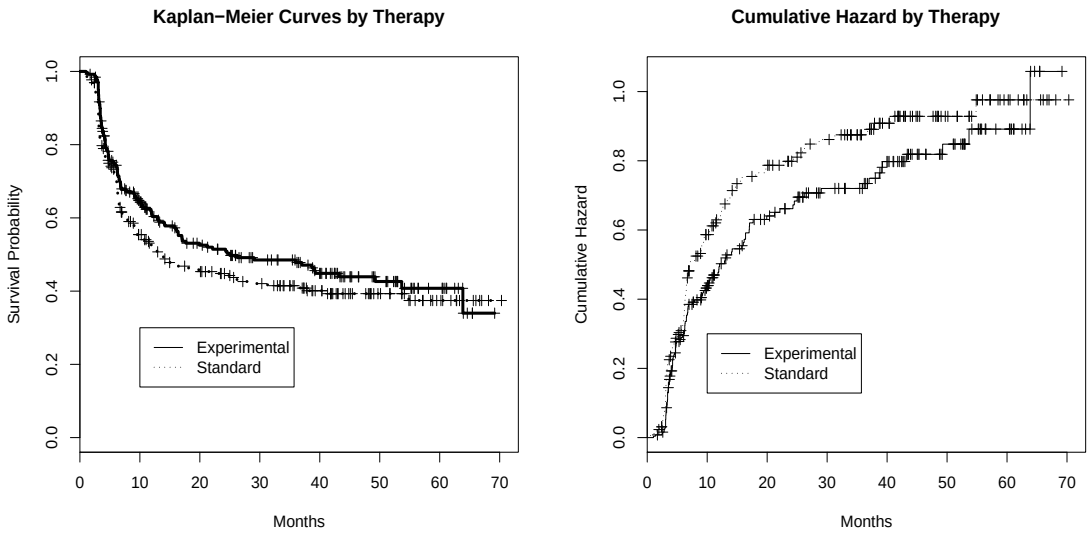

Figure 2.2: Kaplan-Meier and Cumulative Hazard Curves by Therapy.

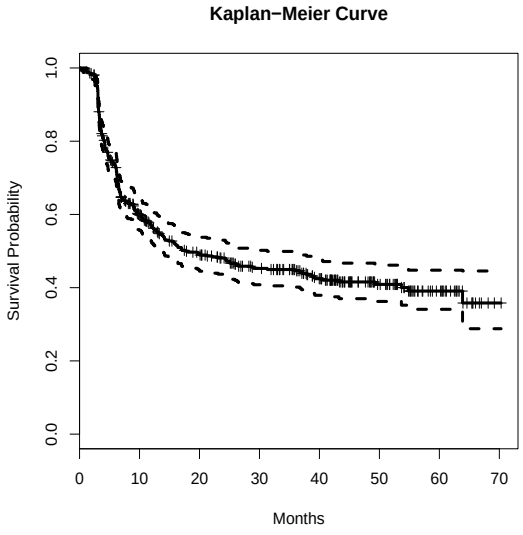

Figure 2.3: Kaplan-Meyer Curves with 95% CI.

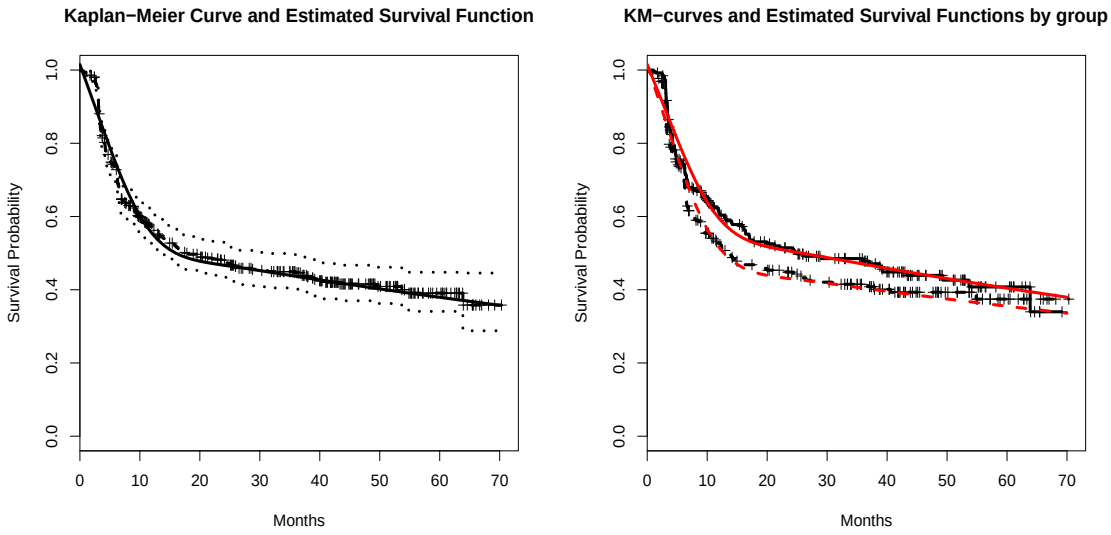

Figure 2.4: Kaplan-Meyer Curves and Estimated Survival Function Overall (left panel) and Sorted by Group (right panel).

CHAPTER III

Joint Modeling of Time to Recurrence and Cancer Stage at Recurrence in Oncology Trials - When Event Times Are Interval - Censored (Discrete Follow-up Observation Process)

3.1 Introduction

3.1.1 Interval-Censored Data: Brief Overview

In the real-life example described in Chapter I and Section 2.3, during the first year of the follow-up period patients had scheduled visits at 3, 6, and 9 months; after one year, the follow-up process was based on the regular visits set by investigators and their patients participating in the clinical trial. The extension part of the 305 study collected a 5-year follow up information based on a single patient visit. The more typical way to collect the extension trial information would be to have scheduled visits throughout the follow-up period defined by a protocol (the same scheduled visits for all patients in a study).

In many situations, the event of interest cannot be observed and it is only known to have occurred within two times. In this set-up, we say that the time is interval-censored. Interval-censored data are quite usual in longitudinal studies where subjects

are not monitored continuously but scheduled to be inspected at particular visits.

The theory for the analysis of interval-censored data has been developed over the past three decades and several good reviews have been written. However, it is still a common practice in clinical trials to simplify the interval censoring structure of the data into a more standard right censoring case. Reviews written by *Huang and Wellner* (1997) and *Lindsey and Ryan* (1998) have been a keystone but are outdated by many of the interval-censored methods. The more recent book by *Sun* (2006) addresses statistical issues and describes statistical methods for the analysis of singly and doubly interval-censored survival data arising from AIDS, cancer and other disease studies. Parametric survival models for interval-censored data with time-dependent covariates are described in work by *Sparling et al.* (2006). The most recent review provided by *Gomez et al.* (2009) includes the methodology on non-parametric, parametric, and semi-parametric estimating approaches, and the review of software for analyzing interval-censored data.

There are several types of interval-censored data.

Case I interval-censored data or current status data: T is only known to be larger or smaller than an observed monitoring time, L. In this case, the study subject is observed only once producing either a left- or a right-censored observation.

Case II interval-censored data: In experiments with two monitoring times, L and R with $L < R$, the survival time of interest T is only known to be before the first monitoring time ($T \leq L$), between the two monitoring times ($L < T \leq R$), or after the second monitoring time ($T > R$).

Case K interval-censored data: In longitudinal studies with periodic follow-up and K monitoring times $M_1, M_2, \ldots, M_K$, the event of interest is only observed between two consecutive inspecting times M_l, M_{l+1} and the observed data reduce to the interval $(M_l,\ M_{l+1}]$. This censoring scheme corresponds to a natural extension of case I and case II mechanisms and is discussed and extended in *Schick and Yu* (2000).

Authors generalized the model assuming that the number of monitoring times K is random.

The main assumption for many interval-censored techniques is that censoring occurs non-informatively, that is, the only information provided by censoring interval about the survival time t is that the interval contains t. The non-informative assumption is relevant and not always fulfilled. More information on it can be found in *Oller et al.* (2004).

In the next section, we will describe a framework for modeling the joint distribution of time to cancer recurrence and cancer stage at recurrence accommodating two causes of the cancer recurrence: recurrence caused by cancer cells surviving a treatment or a surgery and recurrence caused by spontaneous carcinogenesis using interval-censored techniques.

3.2 Models and Methods

3.2.1 Survival Function

Survival function in this case will be the same as in Section 2.2.1 and will be written as

$$S(t) = \exp\left(-\theta_1 F_1(t) - \int_0^t \theta_2 F_2(x)dx\right) = S_1(t)S_2(t), \tag{3.1}$$

where $S_1(t) = \exp\left(-\theta_1 F_1(t)\right)$ describes the time to tumor recurrence from cancer cells that survive treatment, and $S_2(t) = \exp\left(-\int_0^t \theta_2 F_2(x)dx\right)$ describes the time to tumor recurrence by spontaneous carcinogenesis. Here θ_1 is the mean number of cancer cells surviving a treatment or surgery and $F_1(t)$ is a c.d.f. describing the rate of their progression; θ_2 is the rate of formation of intracellular lesions and $F_2(t)$ is a c.d.f. describing the rate of their progression.

To allow functionally dependence on covariate information, the rates θ_1 and θ_2

will be modeled parametrically as:

$$\theta_1(Z) = \exp\left(\beta_{01} + \beta_1^T Z\right),\ \theta_2(Z) = \exp\left(\beta_{02} + \beta_2^T Z\right), \tag{3.2}$$

where Z is a vector of values of explanatory variables and β_{ij} are regression coefficients.

Let introduce random variable U which will take the following values:

$$U = \begin{cases} 1, \text{ if recurrence is caused by spontaneous carcinogenesis,} \\ 0, \text{ if recurrence is caused by surviving a treatment cancer cells.} \end{cases} \tag{3.3}$$

Note that U is a random variable which is not observed, but is used to differentiate the cause of cancer recurrence.

3.2.2 Multinomial Logit Model

Again, let $X_i \in \{1, 2, \cdots, M\}$ be the ith subject's multinomial response (cancer stage) in one of the M possible categories. On the complete-data level, multinomial probabilities are modeled using log-linear predictors $\pi_m(z_i, t_i, u_i)$ specific to categories m and conditional on a vector of covariates Z_i, time T_i, and indicator U_i:

$$\Pr\{X_i = m | Z_i, T_i, U_i\} = \frac{\pi_m(z_i, t_i, u_i)}{1 + \sum_{c=2}^{M} \pi_c(z_i, t_i, u_i)}, \tag{3.4}$$

where for identifiability, regression coefficients corresponding to the first category are set to zero. We will use the following parameterization of function π_m using regression coefficients α_m:

$$\pi_m(z_i, t_i, u_i) = \exp\left(\alpha_m \cdot z_i + \alpha_{t,m} \cdot t_i + \alpha_{u,m} \cdot u_i\right). \tag{3.5}$$

3.2.3 Joint Distribution and Likelihood

In many situations, the event of interest cannot be observed and it is only known to have occurred within two times, say L and R. In this set-up, we say that the time is interval-censored. Interval-censored data are quite usual in clinical trials where subjects are inspected at particular visits. Let

$$\delta_i = \begin{cases} 1, \text{ if } i^{th} \text{ subject had cancer recurrence detected,} \\ 0, \text{ otherwise.} \end{cases} \tag{3.6}$$

Let T_i denote the time to recurrence for subject i, $i = 1, \ldots, k$ and suppose the T_i's follow a parametric model with survival function $S(t, \beta)$ where vector β denotes unknown parameters. Also suppose that only interval-censored data are available and have the form:

$$\{(L_i,\, R_i], Z_i;\, i = 1, \ldots, k\}, \tag{3.7}$$

where $(L_i,\, R_i]$ denotes the interval to which an observed T_i belongs, and Z_i is the covariate vector associated with subject i, $i = 1, \ldots, k$. Then the likelihood from the interval-censored times is proportional to:

$$\mathbb{L}_1(\beta) = \prod_{i=1}^{k} \mathrm{L}_i(\beta) \propto \prod_{i=1}^{k} [S(L_i, \beta) - S(R_i, \beta)], \tag{3.8}$$

assuming that $L_i < R_i$ for all $i = 1, \ldots, k$.

If we assume that the recurrence is interval-censored while the non-recurrence data is right censored, the likelihood function will be:

$$\mathbb{L}_2 = \prod_{i=1}^{k} \left\{ (S(L_i))^{1-\delta_i} \cdot (S(L_i) - S(R_i))^{\delta_i} \right\}. \tag{3.9}$$

The computations and model fitting procedure are simplified if only a few values are possible for L and R. In our case, it is easier to refer to the interval-censored values by

intervals on the time scale common to all subjects. Assume that we have J intervals denoted by $(\tau_{j-1},\, \tau_j]$ for $j = 1, 2, \ldots, n$ with $\tau_0 = 0$ and $\tau_J = \infty$, and these intervals are the same for all subjects. In the real-life example we used in the previous chapter the intervals would be the following:

$$\{(0, 3], (3, 6], (6, 9], (9, 60], (60, \infty]\}.$$

Let denote I_j the j^{th} interval $(\tau_{j-1}, \tau_j]$. The binary variable indicating the specific time interval observed for the i^{th} subject is defined as:

$$a_{ij} = \begin{cases} 1, \text{ if } (L_i,\, R_i] = I_j, \\ 0, \text{ otherwise.} \end{cases} \tag{3.10}$$

Now for $k = 1$ and $\delta = 0$, the likelihood contribution is $\prod_{j=1}^{n} [S(\tau_{j-1})]^{a_{ij}}$,where:

$$S(\tau_{j-1}) = S_1(\tau_{j-1}) S_2(\tau_{j-1}).$$

For $k = 1$ and $\delta = 1$, the joint density for a time to recurrence and a cancer stage at recurrence can be described by the following pdf:

$$\begin{aligned} (T, X) \sim f(t, m) &= \Pr(T \in (\tau_{j-1}, \tau_j], x = m) \\ &= \int_{\tau_{j-1}}^{\tau_j} S_2(t) s_1(t) \rho_{0,m}(t, Z) dt + \int_{\tau_{j-1}}^{\tau_j} S_1(t) s_2(t) \rho_{1,m}(t, Z) dt, \end{aligned} \tag{3.11}$$

where $s_1(t)$ and $s_2(t)$ are pdf's given by $1 - S_1(t)$ and $1 - S_2(t)$ distributions, respectively.

As previously,

$$\rho_{0,m}(t, Z) = \frac{\pi_m(t, Z, u = 0)}{1 + \sum_{c=2}^{M} \pi_c(t, Z, u = 0)} \quad \text{and} \quad \rho_{1,m}(t, Z) = \frac{\pi_m(t, Z, u = 1)}{1 + \sum_{c=2}^{M} \pi_c(t, Z, u = 1)}.$$

Since $s_1(t) = \theta_1(Z)f_1(t)S_1(t)$ and $s_2(t) = \theta_2(Z)F_2(t)S_2(t)$, the joint pdf $f(t,m)$ can be expressed as follows:

$$\begin{aligned} f(t,m) &= \int_{\tau_{j-1}}^{\tau_j} S_2(t)\theta_1(Z)f_1(t)S_1(t)\rho_{0,m}(t,Z)dt \\ &+ \int_{\tau_{j-1}}^{\tau_j} S_1(t)\theta_2(Z)F_2(t)S_2(t)\rho_{1,m}(t,Z)dt \\ &= \int_{\tau_{j-1}}^{\tau_j} \{S(t)\,(\ \theta_1(Z)f_1(t)\rho_{0,m}(t,Z) + \theta_2(Z)F_2(t)\rho_{1,m}(t,Z)\ )\}\,dt, \end{aligned} \tag{3.12}$$

where $f_1(t)$ is the pdf corresponding to a distribution given by $F_1(t)$.

Therefore,

$$f(t,m,u) = \begin{cases} \int_{\tau_{j-1}}^{\tau_j} S(t)\theta_1(Z)f_1(t)\rho_{0,m}(t,Z)dt, \text{ if } u=0, \\ \int_{\tau_{j-1}}^{\tau_j} S(t)\theta_2(Z)F_2(t)\rho_{1,m}(t,Z)dt, \text{ if } u=1. \end{cases} \tag{3.13}$$

The conditional pdf of $U = u$ given $T = t$ and $X = m$ is

$$f(u|t,m) = \frac{f(t,m,u)}{f(t,m)}.$$

So, if $u = 0$, then

$$f(u=0|t,m) = \frac{\int_{\tau_{j-1}}^{\tau_j} S(t)\theta_1(Z)f_1(t)\rho_{0,m}(t,Z)dt}{\int_{\tau_{j-1}}^{\tau_j} S(t)\theta_1(Z)f_1(t)\rho_{0,m}(t,Z)dt + \int_{\tau_{j-1}}^{\tau_j} S(t)\theta_2(Z)F_2(t)\rho_{1,m}(t,Z)dt},$$

and if $u = 1$

$$f(u=1|t,m) = \frac{\int_{\tau_{j-1}}^{\tau_j} S(t)\theta_2(Z)F_2(t)\rho_{1,m}(t,Z)dt}{\int_{\tau_{j-1}}^{\tau_j} S(t)\theta_1(Z)f_1(t)\rho_{0,m}(t,Z)dt + \int_{\tau_{j-1}}^{\tau_j} S(t)\theta_2(Z)F_2(t)\rho_{1,m}(t,Z)dt}.$$

The full likelihood is proportional to the likelihood associated with the event time distribution and cancer stage $L(\beta)$, where β is a vector of parameters need to be

estimated from the model. The observed data log-likelihood is $\log L(\beta)$ calculated as following:

$$
\begin{aligned}
l_{obs} &= \log \mathbb{L}(\beta) = \sum_{i\in \text{ non-recurrences}} \sum_{j=1}^{n} a_{ij} \log S(\tau_{i,j-1}) \\
&+ \sum_{i\in \text{ recurrences}} \sum_{j=1}^{n} a_{ij} \log \Bigg(\int_{\tau_{i,j-1}}^{\tau_{i,j}} S(t)\theta_1(Z_i) f_1(t) \rho_{0,x_i}(t, Z_i) dt \\
&+ \int_{\tau_{i,j-1}}^{\tau_{i,j}} S(t)\theta_2(Z_i) F_2(t) \rho_{1,x_i}(t, Z_i) dt \Bigg).
\end{aligned} \tag{3.14}
$$

The complete data log-likelihood is:

$$
\begin{aligned}
l_{cd} &= \sum_{i\in \text{ non-recurrences}} \sum_{j=1}^{n} a_{ij} \log S(\tau_{i,j-1}) \\
&+ \sum_{i\in \text{ recurrences}} \sum_{j=1}^{n} a_{ij} \Big\{ (1-u_i) \cdot \log \int_{\tau_{i,j-1}}^{\tau_{i,j}} S(t)\theta_1(Z_i) f_1(t) \rho_{0,x_i}(t, Z_i) dt \\
&+ u_i \cdot \log \int_{\tau_{i,j-1}}^{\tau_{i,j}} S(t)\theta_2(Z_i) F_2(t) \rho_{1,x_i}(t, Z_i) dt \Big\}.
\end{aligned} \tag{3.15}
$$

The EM algorithm, with the E-step solving the problem of U imputation and the M-step maximizing a log-likelihood build from the complete data model, can be used to estimate the necessary parameters. These complex computations can be simplified if we use a single point imputation approach for the multinomial probabilities of the cancer stage model. When a patient has a cancer recurrence in between visits, a cancer is in its earliest stage. Since a cancer lesion is growing over time, the cancer stage is increasing until detected. When a patient comes to the hospital for a visit and a cancer recurrence detected, the cancer stage is determined at that time point. Therefore, using the right time point of the interval to model and estimate the multinomial probabilities of the cancer stage seems to be reasonable. In this case, the

joint density for a time to cancer recurrence and a cancer stage at recurrence can be described by the following pdf:

$$\begin{aligned}(T,X)\sim f(t,m) &= \Pr(T\in(\tau_{j-1},\tau_j],x=m) \\ &= \rho_{0,m}(\tau_j,Z)\int_{\tau_{j-1}}^{\tau_j} S_2(t)s_1(t)dt+\rho_{1,m}(\tau_j,Z)\int_{\tau_{j-1}}^{\tau_j} S_1(t)s_2(t)dt \\ &= \rho_{0,m}(\tau_j,Z)\int_{\tau_{j-1}}^{\tau_j} S_2(t)\theta_1(Z)f_1(t)S_1(t)dt \\ &+ \rho_{1,m}(\tau_j,Z)\int_{\tau_{j-1}}^{\tau_j} S_1(t)\theta_2(Z)F_2(t)S_2(t)dt.\end{aligned} \tag{3.16}$$

Therefore,

$$f(t,m,u)=\begin{cases}\rho_{0,m}(\tau_j,Z)\int_{\tau_{j-1}}^{\tau_j} S(t)\theta_1(Z)f_1(t)dt, \text{ if } u=0,\\ \rho_{1,m}(\tau_j,Z)\int_{\tau_{j-1}}^{\tau_j} S(t)\theta_2(Z)F_2(t)dt, \text{ if } u=1.\end{cases} \tag{3.17}$$

If $u=0$, then

$$f(u=0|t,m)=\frac{\rho_{0,m}(\tau_j,Z)\int_{\tau_{j-1}}^{\tau_j} S(t)\theta_1(Z)f_1(t)dt}{\rho_{0,m}(\tau_j,Z)\int_{\tau_{j-1}}^{\tau_j} S(t)\theta_1(Z)f_1(t)dt+\rho_{1,m}(\tau_j,Z)\int_{\tau_{j-1}}^{\tau_j} S(t)\theta_2(Z)F_2(t)dt},$$

and if $u=1$

$$f(u=1|t,m)=\frac{\rho_{1,m}(\tau_j,Z)\int_{\tau_{j-1}}^{\tau_j} S(t)\theta_2(Z)F_2(t)dt}{\rho_{0,m}(\tau_j,Z)\int_{\tau_{j-1}}^{\tau_j} S(t)\theta_1(Z)f_1(t)dt+\rho_{1,m}(\tau_j,Z)\int_{\tau_{j-1}}^{\tau_j} S(t)\theta_2(Z)F_2(t)dt}.$$

The full likelihood is proportional to the likelihood associated with the event time distribution and cancer stage $L(\beta)$, where β is a vector of parameters need to be estimated from the model. The observed data log-likelihood is $\log L(\beta)$ calculated as

following:

$$
\begin{aligned}
l_{obs} &= \log \mathbb{L}(\beta) = \sum_{i\in \text{ non-recurrences}} \sum_{j=1}^{n} a_{ij} \log S(\tau_{i,j-1}) \qquad (3.18)\\
&+ \sum_{i\in \text{ recurrences}} \sum_{j=1}^{n} a_{ij} \log\Big(\rho_{0,x_i}(\tau_{i,j}, Z_i) \int_{\tau_{i,j-1}}^{\tau_{i,j}} S(t)\theta_1(Z_i) f_1(t) dt \\
&+ \rho_{1,x_i}(\tau_{i,j}, Z_i) \int_{\tau_{i,j-1}}^{\tau_{i,j}} S(t)\theta_2(Z_i) F_2(t) dt \Big).
\end{aligned}
$$

The complete data log-likelihood is:

$$
\begin{aligned}
l_{cd} &= \sum_{i\in \text{ non-recurrences}} \sum_{j=1}^{n} a_{ij} \log S(\tau_{i,j-1}) \qquad (3.19)\\
&+ \sum_{i\in \text{ recurrences}} \sum_{j=1}^{n} a_{ij} \Big\{ (1-u_i) \cdot \log \rho_{0,x_i}(\tau_{i,j}, Z_i) \int_{\tau_{i,j-1}}^{\tau_{i,j}} S(t)\theta_1(Z_i) f_1(t) dt \\
&+ u_i \cdot \log \rho_{1,x_i}(\tau_{i,j}, Z_i) \int_{\tau_{i,j-1}}^{\tau_{i,j}} S(t)\theta_2(Z_i) F_2(t) dt \Big\} \\
&= \sum_{i\in \text{ non-recurrences}} \sum_{j=1}^{n} a_{ij} \log S(\tau_{i,j-1}) \qquad (3.20)\\
&+ \sum_{i\in \text{ recurrences}} \sum_{j=1}^{n} a_{ij} \Big\{ (1-u_i) \cdot \log \rho_{0,x_i}(\tau_{i,j}, Z_i) \\
&+ (1-u_i) \cdot \log \int_{\tau_{i,j-1}}^{\tau_{i,j}} S(t)\theta_1(Z_i) f_1(t) dt \\
&+ u_i \cdot \log \rho_{1,x_i}(\tau_{i,j}, Z_i) + u_i \cdot \log \int_{\tau_{i,j-1}}^{\tau_{i,j}} S(t)\theta_2(Z_i) F_2(t) dt \Big\}.
\end{aligned}
$$

Our approach will be to use EM algorithm, with the E-step solving the problem of imputation U and the M-step maximizing a log-likelihood obtained from the complete data model.

3.2.4 The EM Algorithm

The EM algorithm is formulated as follows.

Step 1: Set initial values of regression coefficients and distribution parameters

$$\beta^{(0)} = (\beta_1, \beta_2, \alpha, \text{ parameters from } F_1(t) \text{ and } F_2(t) \text{ distributions}) .$$

Step 2: E-step. Calculate a vector

$$\begin{aligned} & \hat{U}(\beta^{(k)}) = E(U|\text{Observed data} = (t, m), \delta = 1) \qquad (3.21) \\ = \; & \frac{\rho_{1,x_i}(\tau_{i,j}, Z_i) \int_{\tau_{i,j-1}}^{\tau_{i,j}} S(t)\theta_2(Z_i)F_2(t)dt}{\rho_{0,x_i}(\tau_{i,j}, Z_i) \int_{\tau_{i,j-1}}^{\tau_{i,j}} S(t)\theta_1(Z_i)f_1(t)dt + \rho_{1,x_i}(\tau_{i,j}, Z_i) \int_{\tau_{i,j-1}}^{\tau_{i,j}} S(t)\theta_2(Z_i)F_2(t)df}. \end{aligned}$$

Step 3: M-step. Maximize the log-likelihood obtained from the complete data model at $\hat{U}(\beta^{(k)})$, which can be achieved by maximizing separately l_{cd}^{ρ} and l_{cd}^{t}:

$$\begin{aligned} l_{cd}^{\rho} \; = \; & \sum_{i \in \text{ recurrences}} \sum_{j=1}^{n} a_{ij} \Big\{ \hat{u}_i(\beta^{(k)}) \cdot \log \rho_{1,x_i}(\tau_{i,j}, Z_i) \Big\} \\ + \; & \sum_{i \in \text{ recurrences}} \sum_{j=1}^{n} a_{ij} \Big\{ (1 - \hat{u}_i(\beta^{(k)})) \cdot \log \rho_{0,x_i}(\tau_{i,j}, Z_i) \Big\}, \end{aligned}$$

$$\begin{aligned} l_{cd}^{t} \; = \; & \sum_{i \in \text{ non-recurrences}} \sum_{j=1}^{n} a_{ij} \log S(\tau_{i,j-1}) \qquad (3.22) \\ + \; & \sum_{i \in \text{ recurrences}} \sum_{j=1}^{n} a_{ij} \Big\{ (1 - \hat{u}_i(\beta^{(k)})) \cdot \log \int_{\tau_{i,j-1}}^{\tau_{i,j}} S(t)\theta_1(Z_i)f_1(t)dt \\ + \; & \hat{u}_i(\beta^{(k)}) \cdot \log \int_{\tau_{i,j-1}}^{\tau_{i,j}} S(t)\theta_2(Z_i)F_2(t)dt \Big\}. \end{aligned}$$

Denote the solution by $\beta^{(k+1)}$.

Step 4: Set $k = k + 1$. Continue with Step 2 and Step 3 iterations until conver-

gence.

Standard error estimates are based on the inverse of the observed information matrix:

$$I = -\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T}, \tag{3.23}$$

where β is the vector of model parameters and $l(\beta) = \log E\{L(\beta|U)\}$ is the model log-likelihood maximized as a result of EM algorithm. The observed information matrix is derived by an application of the missing information principle representing the observed information as the difference between expected complete-data information and the missing information, given observed data, see *McLachlan and Krishnan* (2008). Alternatively, a bootstrap estimate of standard errors could be done using Efron's approach, see *Efron* (1994).

3.3 Discussion

In this chapter, we described a framework for modeling the joint distribution of time to cancer recurrence and cancer stage at recurrence using interval-censored data techniques considering two causes of the cancer recurrence: recurrence caused by cancer cells surviving a treatment or a surgery and recurrence caused by spontaneous carcinogenesis. The proposed EM algorithm can be used to estimate the necessary parameters in the model.

The availability of software for the right censoring techniques in survival analysis made it easy for the scientists in the pharmaceutical industry to apply the right censored data techniques to the time to event outcomes. It is one of the main reasons why it is still a common practice in clinical trials to simplify the interval censoring structure of the data into a more standard right censoring case. The commercial software S-PLUS and the free software R from the R Development Core Team *Team*

(2008) are the most complete sources for survival analysis with interval-censored data.

In general, the parametric approach for analyzing interval-censored data is computationally easier than non-parametric. A variety of parametric models can be used, for example see *Lindsey and Ryan* (1998) to obtain smooth representations of both the hazard and the survival functions. Maximum likelihood methods can be applied to provide useful and meaningful parameter estimation. Under the non-informative censoring assumption, standard likelihood inference and usual large sample properties apply. The parametric approach is appealing because of its simplicity but its disadvantage is that all the inferences depend upon the assumption of a model which is difficult to assess based on an interval-censored sample, with the risk of deriving inconsistent estimators for the parameters of interest leading to inaccurate conclusions. *Ren* (2003) proposed a goodness-of-fit method, called the leveraged bootstrap, and *Calle and Gomez* (2008) proposed a sampling-based chi-squared test.

3.4 Remarks

The post-treatment cancer surveillance represents a discrete observational process yielding incomplete information on the time to cancer recurrence. Instead of the actual time of recurrence only the time of examinations is available which usually follows the specific discrete schedule. Additionally, false-positive and false-negative rates of the diagnostic test may be present. There exists a broad range of literature on parametric and non-parametric estimation of the disease natural history from discrete observations including *Albert et al.* (1978a,b), *Flehinger and Kimmel* (1991), *Klebanov et al.* (1993), *Ivankov et al.* (1993), and *Yakovlev et al.* (1993). If surveillance is error free, the corresponding sample can be considered as interval-censored.

Cancer surveillance represents a discrete observation process. During cancer surveillance only the time of disease diagnosis is available while the time of the tumor onset is unknown. The diagnostic time is usually discretized according to the specific

schedule of visits.

Every individual under study is supposed to be initially at the disease free stage. In other words, the disease escapes detection for some time right after the surgery. At some point of time the preclinical stage begins during which a disease is detectable but asymptotic. Having stayed in the preclinical stage without being diagnosed, a patient enters the clinical stage characterized by apparent symptoms. If a preclinical disease is detected by screening at time τ, its natural history is interrupted. The probability, p, that a test detects cancer given the individual under examination is in the preclinical stage is called sensitivity. Sensitivity estimation is a difficult task for many reasons, see *Yakovlev and Tsodikov* (1996). In our case, we can assume that it is known and it is a constant or a simple function.

Let τ_i be the time points at which patients are scheduled to be examined, $i = 1, \ldots, n$:

$$0 \leq \tau_0 \leq \tau_1 \leq \ldots \leq \tau_n \leq T,$$

where T is the planning period of observation. The individual's outcome can be one of the following:

- { an individual is censored in $[\tau_{j-1}, \tau_j)$ },
- { an individual is detected with cancer at τ_j during scheduled visit },
- { an individual is detected with cancer prior to visit τ_j based on clinical symptoms }.

Let p_1 be the detection probability of the cancer caused by the cancer cells surviving a treatment or surgery, and $q_1 = 1 - p_1$ is false negative rate of the corresponding diagnostic test. Let p_2 be the detection probability caused by spontaneous carcinogenesis, and $q_2 = 1 - p_2$ false negative rate of the corresponding diagnostic test.

If a surveillance is error free, the corresponding sample will be interval censored by the points $\tau_1, \tau_2, \ldots$.

Let us assume that a cancer lesion is growing deterministically which depends on the follow-up time. We have two types of cancer diagnosis during the discrete cancer surveillance: clinical diagnosis, when a patient comes to the office prior to the scheduled visit because of the symptoms that he/she experiences; or visit diagnosis, when a patients is diagnosed with the cancer during a scheduled visit. It would be interesting to build and evaluate the models for this case.

CHAPTER IV

Conclusions

We have described a framework for modeling the joint distribution of time to cancer recurrence and cancer stage at recurrence. Our approach accommodates two different causes of the cancer recurrence: recurrence caused by cancer cells surviving a treatment or a surgery and recurrence caused by spontaneous carcinogenesis. We evaluated the model and provided the estimates for different outcomes of the recurrence time. First, we proceeded with a continuous follow-up assumption using stochastic models of cancer recurrence. Then we extended the approach to allow for a discrete follow-up process. ML estimation with the EM algorithm was used to estimate the necessary parameters in the models. We introduced the random variable U which is not observed, but was used to differentiate the cause of cancer recurrence. By using the random variable U, the maximization step in the EM algorithm was simplified by splitting the complete-data likelihood into separate parts for stage and event time.

Modeling the time to cancer recurrence and cancer stage at recurrence jointly allows for more powerful inference. Real-life data from a bladder cancer trial and simulations were used to assess the sensitivity and robustness of the method. An added benefit of such modeling is that it permits using the cancer stage at recurrence to provide adjusted estimates for the time to recurrence distribution and use them in

tests. The cancer stage at recurrence significantly impacts patient quality of life and further treatment. Therefore, it should be accounted in the estimation and analysis of time to cancer recurrence.

The research described in this paper is not unique to the bladder cancer trials. The proposed method can be used in evaluating the time to recurrence or progression jointly with the disease stage at recurrence or progression in other indications, settings (e.g., adjuvant therapies), and even therapeutic areas.

We considered cancer post-surgery surveillance which is represented by a discrete process with non-zero false-negative rate of a given test. The extension of our methods to this case is a natural next step of the research. The pharmaceutical industry is highly regulated industry. Before any drug or device becomes available to people, extensive work is done to evaluate the efficacy and safety of investigational drug or device in pre-clinical and clinical trials. Clinical trials are classified into exploratory and confirmatory studies. While Phase I and most Phase II trials are considered to be exploratory ones, Phase III studies are aimed at being the definitive assessment of how effective the drug is, in comparison with current gold standard treatment. Therefore, testing, sample size and power calculations are always of interest. As a future research, tests that can be applied to compare joint and marginal survival distributions between two or more treatment or procedure groups could be developed and evaluated. The sample size and power calculation of the tests would be a logical next step in a future research. The recurrence data from the extension study 305 have the information on patients' multiple recurrences during the follow-up period. Therefore, as an extension to the current problem, the join distribution approach can be considered for recurrent event data and the corresponding analysis techniques might be proposed.

BIBLIOGRAPHY

BIBLIOGRAPHY

Albert, A., P. M. Gertman, and T. A. Louis (1978a), Screening for the early detection of cancer. 1. the temporal history of a progressive disease state, *Mathematical Biosciences*, *40*.

Albert, A., P. M. Gertman, T. A. Louis, and S. I. Liu (1978b), Screening for the early detection of cancer. 1. the temporal history of the natural history of the disease, *Mathematical Biosciences*, *40*.

Ankerst, D., and D. Finkelstein (2006), Clinical monitoring based on joint models for longitudinal biomarkers and event times, *Handbook of Statistics in Clinical Oncology. Second Edition*, pp. 383 – 394.

Armitage, P., and R. Doll (1954), The age distribution of cancer and a multistage theory of carcinogenesis, *British Journal of Cancer*, *8*.

Baker, S. (1994), The multinomial-poisson transformation, *The Statistician*, *43*, 495–504.

Breto, C., D. He, E. Ionides, and A. King (2009), Time series analysis via mechanistic models, *The Annals of Applied Statistics*, *3*(1), 319 – 348.

Calle, M., and G. Gomez (2008), *A sampling based chi-squared test for interval-censored data in statistical models and methods for biomedical and technical systems*, 303–314 pp., In Vonta F, Nikulin MS, Limnios N and Huber-Carol C eds. Statistics for industry and technology. Springer: Birkhauser.

Cantor, A. B., and J. J. Shuster (1992), Parametric versus nonparametric methods for estimating cure rates based on censored survival data, *Statistics in Medicine*, *11*.

Cox, D., and D. Oakes (1984), *Analysis of Survival Data*, Chapman and Hall, CRC.

Efron, B. (1994), Missing data, imputation and the bootstrap (with discussion), *Journal of the American Statistical Association*, *89*, 463–479.

Faucett, C., and D. Thomas (1996), Simultaneously modeling censored survival data and repeatedly measured covariates: a gibbs sampling approach, *Statistical Medicine*, *15*, 1663–1685.

Flehinger, B., and M. Kimmel (1991), Screening for cancer in relation to the natural history of the disease, in *In Proceedings of 2nd International Conference Mathematical Population Dynamics, Pau, France*.

Fleming, T., and D. Harrington (2005), *Counting Processes and Survival Analysis*, Wiley Series in Probability and Statistics.

Fleming, T., and D. Lin (2000), Survival analysis in clinical trials: Past developments and future directions, *Biometrics*, *56*, 971–983.

Gomez, G., M. Calle, R. Oller, and K. Langohr (2009), Tutorial on methods for interval-censored data and their implementation in r, *Statistical Modeling*, *9*(4), 259–297.

He, D., E. Ionides, and A. King (2010), Plug-in-play inference for disease dynamics: Measles in large and small populations as a case study, *Journal of the Royal Society*, *7*, 271 – 283.

Henderson, R., P. Diggle, and A. Dobson (2000), Joint modeling of longitudinal measurements and event time data, *Biostatistics*, *1*, 465–480.

Hoang, T., A. Tsodikov, and A. Yakovlev (1996), A parametric analysis of tumor recurrence data, *Journal of Biological Systems*, *4*(3), 391–403.

Hogan, J., and N. Laird (1997), Mixture models for the joint distribution of repeated measures and event times, *Statistics in Medicine*, *16*, 239–257.

Hosmer, D., and S. Lemeshow (1999), *Applied Survival Analysis. Regression Modeling of Time to Event Data*, Wiley Series in Probability and Statistics.

Hougaard, P. (2000), *Analysis of multivariate survival data*, Springer.

Huang, J., and J. Wellner (1997), Interval censored survival data: a review of recent progress, in *Proceedings of The First Seattle Symposium in Biostatistics: Survival Analysis*, pp. 123–169, Springer-Verlag, New York.

Ivankov, A. A., B. Asselain, A. Fourquet, T. Hoang, A. D. Tsodikov, T. P. Yakimova, and A. Y. Yakovlev (1993), Estimating the growth potential of a treated tumor from time to recurrence observations, in *In Statistique des Processus en Milieu Medical, B. Bru, C. Huber, B. Prum (eds)*.

Kalbfleisch, J. D., and R. L. Prentice (2002), *The Statistical Analysis of Failure Time Data. Second Edition*, New York: Wiley.

King, A., E. Ionides, M. Pascuel, and M. Bouma (2008), Inapparent infections and cholera dynamics, *Nature*, *454*, 877 – 890.

Klebanov, L. B., S. T. Rachev, and A. Y. Yakovlev (1993), A stochastic model of radiation-induced carcinogenesis. latent time distributions and their properties, *Mathematical Biosciences*, *113*.

Klein, J., and M. Moeschberger (1999), *Techniques for Censored and Truncated Data*, Springer.

Kurth, K., L. Denis, C. Bouffioux, R. Sylvester, F. Debruyne, M. Pavone-Macaluso, and a. et (1995), Factors affecting recurrence and progression in superficial bladder tumors, *European Journal of Cancer*, *31A*(11), 1840–1846.

Lang, J. (1996), On the comparison of multinomial and poisson log-linear models, *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *58*, 253–266.

Law, N., J. Taylor, and H. Sandler (2002), The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure, *Biostatistics*, *3*, 547–563.

Lee, C., B. Hollenbeck, and D. Wood (2006), *Ureter, Bladder, Penis, and Urethra. Oncology an Evidence-Based Approach.*, Springer.

Lee, C., B. Hollenbeck, and D. Wood (2009), *Bladder Cancer*, vol. 2, Practice Guidelines in Oncology.

Lindsey, J., and L. Ryan (1998), Tutorial in biostatistics methods for interval censored data, *Statistics in Medicine*, *17*, 219–238.

Luebeck, E., and S. Moolgavkar (1991), Stochastic analysis of intermediate lesions in carcinogenesis experiments, *Risk Analysis*, *11*.

McLachlan, G., and T. Krishnan (2008), *The EM Algorithm and Extensions. Second Edition*, Wiley Series in Probability and Statistics.

Moolgavkar, S., and D. Venzon (1979), Two-event models for carcinogenesis: incidence curves for childhood and adult tumors, *Mathematical Biosciences*, *47*.

Moolgavkar, S., A. Dewanji, and D. Venzon (1988), A stochastic two-stage model for cancer risk assessment. the hazard function and the probability of tumor, *Risk Analysis*, *8*.

Moolgavkar, S., F. Cross, G. Luebeck, and G. Dagle (1990), A two-mutation model for random-induced lung tumors in rats, *Radiation Research*, *121*.

Oller, R., G. Gomez, and M. Calle (2004), Interval censoring: model characterizations for the validity of the simplified likelihood, *The Canadian Journal of Statistics*, *32*, 315–326.

Pepe, M. S., and T. R. Fleming (1989), Weighted kaplan-meier statistics: A class of distance tests for censored survival data, *Biometrics*, *45*.

Ren, J. (2003), Goodness of fit test with interval censored data, *Scandinavian Journal of Statistics*, *30*, 211–226.

Ross, S. (1996), *Stochastic Processes*, Wiley Series in Probability and Statistics.

Schick, A., and Q. Yu (2000), Consistency of the gmle with mixed case interval-censored data, *Scandinavian Journal of Statistics*, *27*, 45–55.

Sparling, Y., N. Younes, J. Lachin, and O. Bautista (2006), Parametric survival models for interval-censored data with time-dependent covariates, *Biostatistics*, *7*, 599–614.

Stenzl, A., et al. (2010), Hexaminolevulinate guided fluorescence cystoscopy reduces recurrence in patients with nonmuscle invasive bladder cancer, *Journal of Urology*, *184*, 1907–1914.

Sun, J. (2006), *The statistical analysis of interval-censored failure time data*, NewYork: Springer.

Sylvester, R., A. van der Meijden, W. Oosterlinck, J. Witjes, C. Bouffioux, L. Denis, and a. et (2006), Predicting recurrence and progression in individual patients with stage ta t1 bladder cancer using eortc risk tables: a combined analysis of 2596 patients from seven eortc trials, *European Urology*, *49*(3), 466–475.

Tan, W. (1991), *Stochastic Models of Carcinogenesis*, Marcel Dekker, New York.

Tan, W., and C. Chen (1993), A nonhomogeneous stochastic model of carcinogenesis and its applications, in *Proceedings of the III International Conference of Mathematical Population Dynamics, Pau, France*.

Team, R. D. C. (2008), *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing, available at http://www.R-project.org.

Thall, P., R. Simon, and Y. Shen (2000), Approximate bayesian evaluation of multiple treatment effects, *Biometrics*, *56*, 213 – 219.

Tsiatis, A., and M. Davidian (2001), Semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error, *Biometrika*, *88*, 447–458.

Tsodikov, A. (2003), Semiparametric models: a generalized self-consistency approach, *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *65*(3), 759–774.

Tsodikov, A., and S. Chefo (2008), Generalized self-consistency: multinomial logit model and poisson likelihood, *Journal of Statistical Planning and Inference*, *138*, 2380–2397.

Tsodikov, A., and S. Chefo (2009), Stage-specific cancer incidence: An artificially mixed multinomial logit model, *Statistics in Medicine*, *28*, 2054 – 2076.

Yakovlev, A., and A. Tsodikov (1996), *Stochastic Models of Tumor Latency and Their Biostatistical Applications*, World Scientific.

Yakovlev, A. Y. (1993), Comments on the distribution of clonogens in irradiated tumors, *Radiation Research*, *134*.

Yakovlev, A. Y. (1994), Letter to the editor, *Statistics in Medicine*, *13*, 983–986.

Yakovlev, A. Y., B. Asselain, V.-J. Bardou, A. Fourquet, T. Hoang, A. Rochefordiere, and A. D. Tsodikov (1993), A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer, *In Biometrie et Analyse de Donnees Spatio-Temporelles. B. Asselain, M. Boniface, C. Duby, C. Lopez, J. P. Masson, and J. Tranchefort (eds)*, *12*, 66–82.

Zeng, D., and J. Cai (2005), Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time, *The Annals of Statistics*, *33*, 2132 – 2163.