A Practical Guide to Survival Analysis: Understanding Absolute Risk, Competing Risk, Multistate Models and Causal Inference

---

Terry M. Therneau, Elizabeth J. Atkinson and Cynthia S. Crowson

Contents

Preface

This book is designed as an aid for statistical practioners who analyze time-to-event data, also known as survival data. It covers the basics of Kaplan-Meier estimates and Cox proportional hazards models, as well as more complex methods, such as competing risks and multistate models. The authors all work in medical research at a tertiary care institution and "time until" endpoints form a major component of our work. Time-to-event data apply to many fields, and although the examples are medical, they should be easily understood by those outside medicine. This book is written in the same conversational style as Therneau and Grambsch (2000), but with less theory and even more examples from our work. In addition, a separate resource is available online with detailed examples corresponding to each chapter. All of the code shown in the book and extended examples was written using R, but other software packages can also complete many of these same tasks.

Chapter 1

Introduction

"Prediction is hard, especially about the future." Yogi Berra

1.1 Classical survival analysis

Survival data, also known as time-to-event data, is very common in medical research. Survival analysis methods can be used to analyze data for outcomes other than just the alive/dead outcome implied by the "survival" nomenclature. Examples include time until cancer recurrence, or time until liver transplant for subjects on the organ waiting list.

A basic property of time-to-event data is that *it takes time to observe time*. This has direct consequences:

1. Delayed analysis. Suppose, for instance, we wish to describe the 5 year survival probability of a particular treatment for pancreatic cancer. In order to do so, a longitudinal study is undertaken to enroll a set of treated patients and follow their outcome for 5 years. At the time of analysis, however, the study will be describing a treatment plan that is 5 years old; it might well be out of date before the first report! This leads to an unavoidable tension between the desire for more complete information and the timely dissemination of results.
2. Not all subjects will reach the endpoint. One consequence of the time constraint is that the analysis of a dataset will occur before all the subjects have reached an observed endpoint. For a subject who accrued 243 days before the analysis and who is still alive at the time of analysis, we only know that their actual survival time is > 243 days. In most computer packages this is encoded as a pair of variables `(time, status)`, with the status variable indicating whether the endpoint had or had not yet been observed for the subject at the ending time. The latter subjects are said to be *censored*.
3. Prediction gets more difficult over time: it is simply harder to predict long term outcomes than short term — for life in general as well as science. As a consequence, the predictive ability of a covariate often appears to weaken over time. This is universally true for lab measurements in studies with long observation periods. One consequence is that when fitting models for the purpose of short or medium term prediction, it is not always advantageous

to include endpoints that occur after that time frame, leading to a tension between a decrease in the standard deviation, due to more total events, and a decrease in effect size.

In some datasets the total number of observed events is only a small fraction of the observations. For instance, Figure 1.1 shows data on the time until failure of diesel generator fans taken from Nelson [76]. Seventy generators were studied, and the goal of the study was to decide whether or not to replace working fans with a higher quality fan to prevent future failures. For each fan, the number of hours of running time from its first being put into service until fan failure or until the end of the study (whichever came first) was recorded. Only 12 of the 70 fans failed during the observation window.

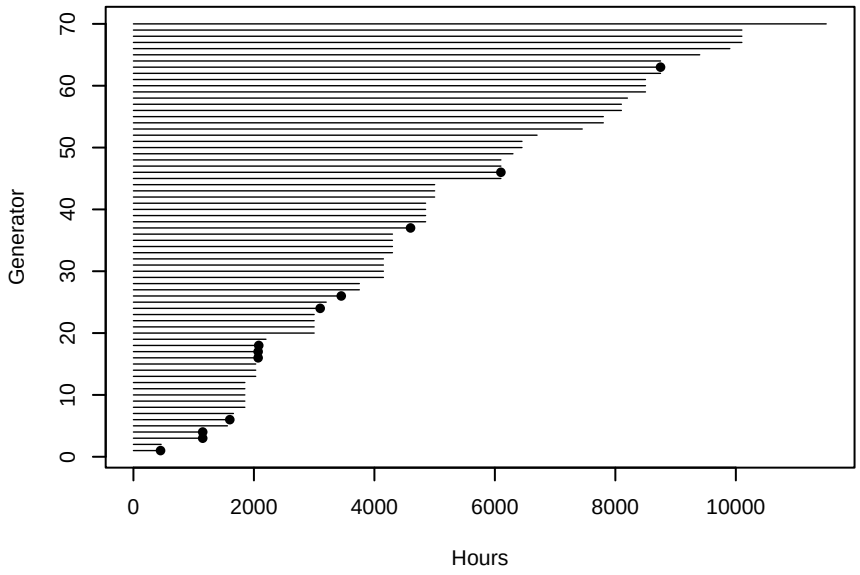

Figure 1.1 *Time until failure or last follow-up of the fans from 70 diesel generators. Fan failure times are indicated with a closed circle at the time of the fan failure. Generators without a closed circle still had working fans at the end of the observation time (i.e., last follow-up).*

1.2 Multistate models

Another way to think about survival data is in terms of multistate transition models, as shown in Figure 1.2. In the diagrams, boxes represent the set of possible states for a subject and arrows represent the possible transitions. We divide multistate models into 4 classes with respect to whether any given subject can have one vs. more than one event (see the columns of Figure 1.2),

and whether there is a single or multiple endpoints (see the rows of Figure 1.2).

1. The two state model in the upper left panel is classical survival data. For instance, using the generator fan data, the labels would be "in service" and "failed".
2. The upper right panel shows an example of repeated outcomes of the same type. An example of this is children with chronic granulomatous disease (CGD) who can experience multiple serious infections. We will demonstrate this using a dataset from a placebo-controlled trial of gamma interferon in children with CGD where patients may have multiple serious infections during the observation time of the study.
3. The lower left panel is an example of competing risks and shows an entry state with two different outcome states. These are mutually exclusive outcomes, such as death from cancer or death from cardiovascular disease. We will illustrate this useding a study of subjects with monoclonal gammopathy of undetermined significance (MGUS) where the endpoint is time until plasma cell malignancy (PCM) or death without PCM.
4. The lower right panel is a general multistate model. The diagram shows the classic illness-death model schema. This representation can be expanded to include additional states and transitions. The MGUS data can also be examined using a multistate model, if we add a transition (arrow) from PCM to death. Another example we will use is a study of non-alcoholic fatty liver disease (NAFLD) and its impact on incident metabolic comorbidities, cardiovascular events, and mortality.

The analysis of multiple events requires more thought, both in the creation of a valid dataset and in the analysis plan, than the simple survival scenario in the upper left panel of the figure.

1.3 Counting processes

Over the last 30 years, the theoretical basis for survival data and models for these data has been solidified by connecting it to the study of counting processes and martingale theory, which in turn arose out of the study of games of chance. The key difference between the classical survival theory and the counting process approach is that the counting process approach considers each subject as a process over time. These processes can also be represented using state space diagrams, such as those in Figure 1.2.

Discussion of the formal mathematics which underlies the counting process has been well covered by other authors, e.g., Cook and Lawless [62], and the details will not be repeated here. Our purpose in introducing this concept is, rather, to show how thinking about time-to-event data in terms of states and transitions can be a useful tool for understanding time-to-event data. It is also a critical step in extending the models to more complex situations. In terms of notation, the counting process describes the data using three quantities.

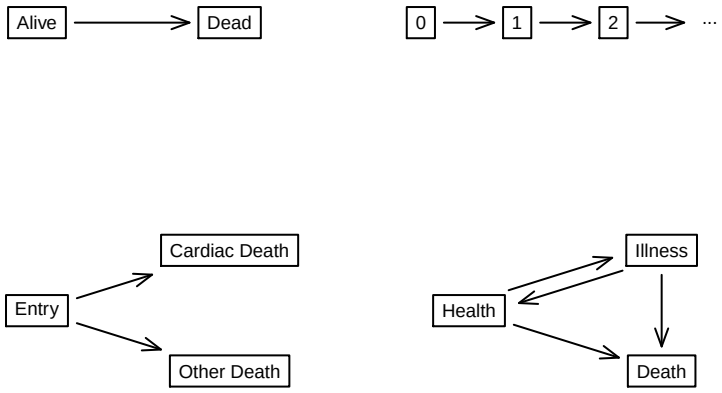

Figure 1.2 *State space diagrams for four multistate models. Upper left: simple survival, upper right: repeated events of the same type, lower left: competing risks, and lower right: a general multistate case (i.e., the illness-death model).*

- $Y_{ij}(t)$ is a 0/1 variable, which is 1 if subject i is currently under observation in state j and consequently at risk for an observed transition out of state j.
- $N_{ijk}(t)$ is the number of transitions for subject i, to date, from state j to state k
- $X_i(t)$ is the set of covariates for subject i, some of which may be time dependent

For the simple alive/dead model in the upper left panel of Figure 1.2, we can dispense with the j and k subscripts: in this case N is known to have a maximum value of 1, and $Y = 0$ for the death state (there is nowhere else to go). The shorthand $dN(t)$ will often be used for an event exactly at time t.

1.4 Important concepts

There are a few concepts that are inherent in every type of time-to-event data, which are often overlooked in practice. These concepts include immortal time bias and censoring. We thought these concepts were important enough to include them both here in the introduction and also later in the book.

1.4.1 Immortal time bias

A pervasive problem in time-to-event (survival) analysis is an issue known as immortal time bias. This bias refers to situations when some subjects cannot have events observed during a portion of their follow-up, which make them look "immortal". This general issue has long been recognized, though the *immortal time bias* label is newer, coming from an overview article by Suissa [101] which provides several examples.

A basic tenet of valid analysis in survival modeling is that you must not look into the future. A very common form of the mistake is a comparison of survival between those who respond or do not respond to treatment, e.g., display some shrinkage of a tumor mass at the 6 week follow-up visit after randomization. Inclusion of a 0/1 variable in the analysis, at time 0, leads to incorrect results: since response is not ascertained before 6 weeks, those in the response group have a guaranteed survival of 6 weeks, an immortal time bias.

The bias will arise when any of the three key aspects of a survival model depend on the future: covariate values, inclusion in the risk set, or outcomes: $X(t)$, $Y(t)$ and $N(t)$ in the counting process notation. This is dealt with more fully in Section 4.2, which gives a large number of examples. The error is surprisingly easy to make, and distressingly common in the scientific literature.

In summary, there are many ways to introduce immortal time bias in study data, and it is important to be aware of how this bias can be accidentally introduced into your work. Understanding how the study was established and when information is known are important parts of the analysis.

1.4.2 Censoring and truncation

Censoring occurs when there is incomplete information available about the event time for some of the subjects. In basic time-to-event data, this generally means that the observation time has ended, but the event has not yet occurred, and this is referred to as *right censoring*. For example, in the analysis of the simple alive/dead outcome, those who are still alive at last follow-up are censored.

A key assumption of all survival analysis methods is that censoring is not informative (also referred to as independent censoring). All survival methods assume that when someone is censored, from that point forward those who remain under observation can "stand in" for the censored subject. The image is that the departing subject distributes their assets (case weight) evenly among those remaining, and that the collective behavior of those remaining is a proxy for what would have transpired. If we were able at some future date to extend the follow-up of those who were censored, and then compare this new information to what was available before, we expect that a) the event rate over this added follow-up will be the same as it was in those originally followed for a longer time, b) the fraction still alive will be the same as in those originally followed and c) the distribution of time-remaining-until-event will be the same as well.

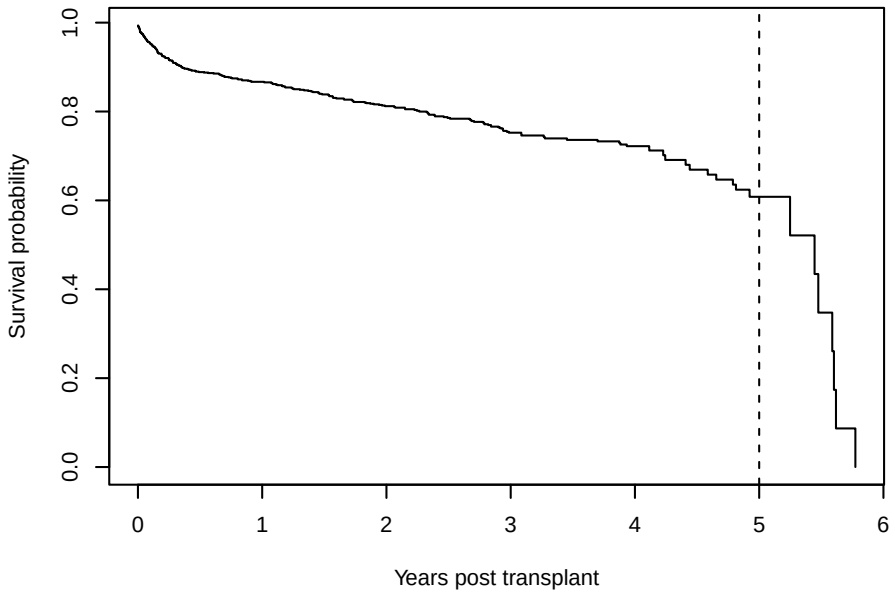

Figure 1.3 *Survival after liver transplant for a particular study. An example of an informative censoring problem.*

Informative censoring

When there is informative censoring, the survival estimates can be biased. The bias can go in either direction, resulting in either longer or shorter survival estimates compared to the true survival experience of the cohort. For instance, suppose that you have followed subjects via correspondence, and those who enter a nursing home stop answering your letters. "No answer" is in this case a predictor of death; those who stop answering have a higher death rate than patients who continue to correspond. This is an example of *informative censoring*. The opposite effect may be seen in communities with major medical centers. When residents retire, those who are healthy are free to relocate, while the more fragile may decide to remain local.

Figure 1.3 shows the Kaplan-Meier survival curve after liver transplant. (Actual data from an initial analysis.) The tail of the curve is clearly wrong. In this study there was nearly complete follow-up for 5 years, at which time funding ceased, and the study assistant transferred to another project. However, entries received from a government death index were still added to the file. Between years 5 and 6 there are only 7 data points, all from the government index and thus by definition all deaths. The usual Kaplan-Meier estimate assumes uninformative censoring, i.e., that this 100% death rate between years 5 and 6 will hold true for for all study subjects. Another description for this

case would be informative *non*-censoring. This type of endpoint bias will be an increasing issue as more use is made of automated systems of follow-up.

Another example of informative censoring is differential enrollment. Verhuel et al. [112] examined the survival of 634 consecutive patients over the age of 20 years who had an aortic valve replacement at Amsterdam Medical Center from 1966 to 1986, with analysis of the data in 1991. Over the first 10 years of the study, 1% of the patients were over 70 years of age, rising to 27% in the second 10 years. Informative censoring in this study is a by product of the enrollment: those censored before 10 years are also systematically older, and as a consequence the right tail of the survival curve is too optimistic, with too low an estimated death rate. One obvious solution is to stratify any analysis or reporting by age group, but that is unlikely to be completely sufficient for this study since age is simply one marker of an underlying expansion of the surgical practice to higher risk subjects as the procedure became more familiar.

Administrative and random censoring

Another classification of censoring is based on how it arises.

- Random censoring. Each observation is censored at a random time. This is rare outside of simulation studies.
- Administrative censoring.
  - Subjects were enrolled over a range of dates, and all are followed until some fixed time. This leads to different censoring times for different subjects. If the enrollment process is random, then this is a variant of random censoring and will be uninformative.
  - Note that administrative censoring dates may be different for different outcomes. For example, you may only have information on a medical outcome, such as fracture, cardiovascular disease or cancer recurrence until each patient's last medical visit. However, you may have additional information on mortality from other sources that extend beyond the last medical visit. These different administrative censoring dates can lead to challenges when analyzing competing risks.
- Type I and II, which are found in industrial experiments but less often in medical studies.
  - type I: fixed follow up for all units
  - type II: follow until a fixed number of events
- Lost to follow-up. You are no longer able to contact someone and don't know why. This is the most dangerous, as the censoring is likely to be informative.

Left, right, and interval censoring

There are also different directions with censoring. More generally, we might know that an event occurs in some interval (s, t). Then, (s, ∞), $(-\infty, t)$ and

(s, t) represent right, left, and interval censoring, respectively. By far, the most common is right-censored data such as the generator fans example above; almost all the examples in this book will be right censored. Right censored data also fits naturally into the multistate framework.

Interval censored data occurs when the event time is only known to be in a certain range. The most common source of such data are conditions that are detected only at a patient visit. An example is data from patients with chronic liver disease who are at risk for esophageal varices (varicose veins in the esophagus due to elevated pressure in the portal vein). These are found by endoscopic examination, which is done at most once a year. When varices are found, one only knows that the onset date was sometime between the prior and current exams. Interval censoring is also known as *panel data*, from the situation where subjects undergo a panel of tests at each visit.

Left censoring occurs when all that is known is that the event time occurred prior to entry in the study. For instance, in a cohort of subjects with a rare liver disease, it may only be known that the liver transplant occurred prior to being recruited for the study. One of the better known examples of left censoring is Tobit regression [108] which is popular in econometrics; y is assumed to follow a Gaussian distribution left censored at zero. Another example are laboratory assays which have a lower limit of detection, so that for the smallest values we only know that $y < c$, where c is the detection threshold. Left censoring of the time scale is uncommon and will not play any further role in the book.

Censoring versus Truncation

There is often confusion between *censoring* and *truncation*. As an example of left truncation consider a review of 3914 multiple myeloma cases seen at Mayo Clinic [59]. The question of interest was the duration of patient survival after a diagnosis of multiple myeloma. However, nearly half of the patients in the study had originally been diagnosed by their local provider, and later referred to Mayo Clinic for further assessment and treatment. A subject diagnosed with MM on 1983-10-21 with a first Mayo appointment on 1984-03-14 and a last follow-up after enrollment on 1992-05-16 would be left truncated on day 145 and right censored on day 3130, both expressed as time since myeloma diagnosis since that is the time scale of interest. A subject who was known to have died sometime before day 145 but with unknown date of death would be left *censored* at day 145.

For left truncated data, the analysis methods assume non-informative *enrollment*. That is, assume that those who enter the study at diagnosis are not different than those who join after a delay. As with informative censoring, there is no statistical "test" for this. This concept and the specifics of this example will be further discussed in Chapter 2.

Understanding the concepts of censoring and truncation, and how they apply to your study data is very important and often overlooked. Knowing enough about a data source to determine whether informative censoring may be present is critical to avoiding bias in the results of your analysis.

1.5 Example datasets

There are many publicly available survival datasets; a small number of them, available in the survival package, have been selected to help illustrate concepts described in the book. Additional datasets are used in the online examples. In a few situations, the examples used in the book do not use publicly available datasets, however similar examples will be demonstrated in the online material using publicly available datasets.

1.5.1 Single endpoints

Listed below are three datasets with a single endpoint. These datasets, along with others of a similar nature, will be used in the first part of the book.

Advanced lung cancer

The `lung` data is derived from a study of overall survival in advanced cancer patients, conducted by the North Central Cancer Treatment Group [69]. The study was developed to determine whether descriptive information from a patient-completed questionnaire could provide prognostic information that was independent from that already obtained by the patient's physician. The data in the survival package is a subset of the full dataset and contains 228 lung cancer patients.

Veteran's Administration Lung cancer

The `veteran` data was originally published by Kalbfleisch and Prentice [53]. In this study, males with advanced inoperable lung cancer were randomized to either a standard or test chemotherapy and were followed until death or last follow-up; death was observed in 128 of the 137 subjects. This data is useful for illustrating how a variable (e.g., Karnofsky score) can have a large effect early in the follow-up period, but a much weaker effect as time progresses.

Free light chain

The `flchain` data is a stratified random sample containing half (7,874) of the subjects from a population-based study that focused on the relationship between free light chain (FLC) levels and mortality. In this population-based study, a blood sample was solicited from all subjects over the age of 50 years in Olmsted County, Minnesota. To minimize patient burden, for any subject who had a visit to the Mayo Clinic during the enrollment period (1995-2003), and who had excess blood remaining from laboratory testing, the excess was frozen and a letter sent asking for consent to use the sample in the research study; assays were then performed on the samples for which consent was given. (It is standard practice to draw an extra 1-2mm from patients on whom blood work has been ordered and save it for 3 days, so that if one of the requested assays fails it can be repeated without subjecting the patient to a new blood draw.)

Figure 1.4 *Two possible approaches for viewing a model with multiple events of the same type. In the left panel, the events are seen as sequential, so the states are 0 infections, 1st infection, 2nd infection, etc. In the right panel, all the infections are considered to be independent, so the states are 0 infections and infection, where infections are counted as return visits to the same state.*

In the original analysis [28], the continuous variable FLC was divided into three groups: those below the 70th percentile of FLC (low), 70–90th percentile (medium) and above the 90th percentile (high). Division into three groups is convenient to illustrate various methods, but we do not make any claim that such a categorization is optimal or even a sensible statistical practice. This data is used to illustrate different adjustment approaches when comparing overall survival in the three FLC groups, and to illustrate different time scales.

1.5.2 Multiple events of the same type

The following datasets include events of the same type that occur multiple times per patient. These can be modeled two different ways, as illustrated in Figure 1.4.

Infections and chronic granulomatous disease

The `cgd` dataset comes from a randomized clinical trial of interferon gamma in children with chronic granulomatous disease (CGD). CGD is a heterogeneous group of uncommon inherited disorders characterized by recurrent pyogenic infections that usually begin early in life and may lead to death in childhood.

It was hypothesized that treatment with interferon gamma might reduce the frequency of serious infections in patients with CGD. In 1986, Genentech, Inc. conducted a randomized, double-blind, placebo-controlled trial in 128 CGD patients who received Genentech's humanized interferon gamma (rIFN-g) or placebo three times daily for a year. The primary endpoint of the study was the time to the first serious infection. However, data were collected on all serious infections until the end of the follow-up, which occurred before day 400 for most patients. Thirty of the 65 patients in the placebo group and 14 of the 63 patients in the rIFN-g group had at least one serious infection; one patient in the placebo group had 7 infections. The total number of infections was 56 and 20 in the placebo and treatment groups, respectively. Follow-up for 1 patient ends on the day of his last infection; all others had some follow-up after their last episode. The data is found in the appendix of Fleming and Harrington [34].

The `dnase` data comes from a study of cystic fibrosis patients who are prone to chronic respiratory infections. In 1992, Genentech conducted a randomized double-blind trial comparing a treatment called rhDNase to placebo. Patients were monitored for pulmonary exacerbations, along with measures of lung volume and flow. The primary endpoint was the time until first pulmonary exacerbation; however, data on all exacerbations were collected for 169 days. The definition of an exacerbation was an infection that required the use of intravenous (IV) antibiotics. Subjects who had an event were not considered to be at risk for another event during the course of antibiotics, nor for an additional 6 days after they end. (If the symptoms reappeared immediately after cessation then from a medical standpoint this would not be a new infection.) Additionally, a few subjects were infected at the time of enrollment, such as subject 173 who had a first infection interval of -21 to 7.

1.5.3 Competing risks

As shown in Figure 1.5, the subjects in this scenario are at risk for more than one outcome, but only one outcome or the other can occur. The outcomes are mutually exclusive: a subject cannot have both outcomes. This phenomenon is often referred to as competing risks. While both outcomes may be of interest, often the competing risk of death is treated as a nuisance that needs to be properly accounted for, but it is not an event of interest.

Primary biliary cirrhosis

This `pbc` data is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984 [103]. PBC is a progressive disease thought to be of an autoimmune origin; the subsequent inflammation process eventually leads to cirrhosis and destruction of the liver's bile ducts and death of the patient. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 subjects in

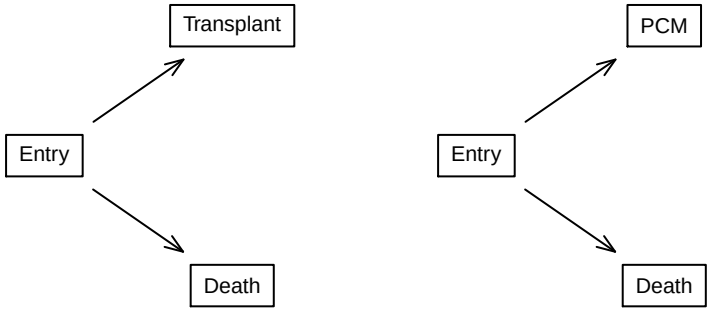

Figure 1.5 *This figure illustrates time-to-event data where subjects are at risk for more than one outcome, but only one outcome or the other can occur. A subject cannot have both outcomes. This phenomenon is often referred to as competing risks. The two examples shown here correspond to the primary biliary cirrhosis (PBC) data (left panel) and the monoclonal gammopathy of undetermined significance (MGUS) data (right panel). PCM is plasma cell malignancy.*

the dataset participated in the randomized trial and contain largely complete data. The additional 112 subjects were not in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so the `pbc` dataset includes data for the 106 additional subjects, as well as the 312 randomized participants. The `pbcseq` data includes multiple laboratory results per subject for the original 312 randomized subjects. The status variable indicates whether the patient was alive at last contact, had a liver transplant, or died.

Monoclonal gammopathy of undetermined significance (MGUS)

The `mgus` data is from an observational study of monoclonal gammopathy of undetermined significance (MGUS). The plasma cell lineage comprises only a small portion of human blood cells ($< 3\%$) but is responsible for the production of immunoglobulins, which is an important part of the body's immune defense. In the case of a plasma cell malignancy (PCM), the immunoglobulin assay will often reveal a sharp spike. The presence of a monoclonal spike in persons without evidence of overt disease is termed "monoclonal gammopathy of undetermined significance" (MGUS); it may often be discovered inadver-

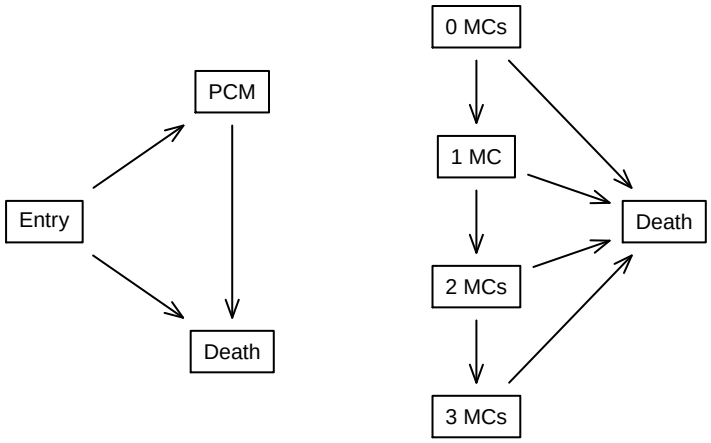

Figure 1.6 *Two examples of multistate models. The left panel shows that the competing risk model for the monoclonal gammopathy of undetermined significance (MGUS) study can become a multistate model similar to the illness-death model by adding the arrow between plasma cell malignancy (PCM) and death. The right panel shows the multistate model for analyzing the non-alcoholic fatty liver disease (NAFLD) data. In this scenario, the states indicate the number of metabolic comorbidities [MC] (i.e., diabetes, hypertension, and dyslipidemia) plus a death state.*

tently when protein electrophoresis is performed for other diagnostic reasons. Given its connection to serious plasma cell disease, the potential prognostic importance of MGUS is of interest: is it a precursor state to malignancy, an incidental finding of no prognostic importance, or something in between? Dr. Kyle studied all 241 cases of MGUS identified at Mayo Clinic before January 1, 1971, with between 20 and 35 years of total follow-up on each patient [58]. The response variable is time to the development of PCM or death. Important covariates are the age at diagnosis, the size of the monoclonal spike, hemoglobin, and creatine levels. A second data set `mgus2` expands this to 1384 subjects.

1.5.4 Multistate

The most complex data focuses on multiple states/endpoints and is represented in Figure 1.6.

Acute myeloid leukemia

The `myeloid` data originated from a clinical trial of subjects with acute myeloid leukemia. The data in the `survival` package were modified to protect patient confidentiality, but the study results are essentially unchanged [63]. However, these data are intended for illustration purposes only and are not meant to provide recommendations for clinical practice or to replace the results presented in the primary trial manuscript.

In this comparison of two conditioning regimens, the canonical path for a subject is that the conditioning regimen will induce a removal of tumor cells from the circulation (below the level of detection), termed a complete response (CR). This sets the patient up for a more aggressive regimen of hematologic stem cell transplant (SCT) which targets any remaining, occult tumor cells, then a sustained remission, followed eventually by relapse or death. Thus a common path of patient states is: initial therapy $\rightarrow$ CR $\rightarrow$ SCT, then either relapse or death.

Non-alcoholic fatty liver disease

The `nafld` datasets come from an observational study of the incidence of non-alcoholic fatty liver disease (NAFLD). It is defined by three criteria: presence of greater than 5% fat in the liver (steatosis), absence of other indications for the steatosis such as excessive alcohol consumption or certain medications, and absence of other liver disease [80]. It is essentially the presence of excess fat in the liver, and parallels the ongoing obesity epidemic. NAFLD is currently responsible for almost 1/3 of liver transplants and its impact is growing. It is expected to be a major component of hepatology practice in the coming decade [102], Approximately 20-25% of NAFLD patients will develop the inflammatory state of non-alcoholic steatohepatitis (NASH), leading to fibrosis and eventual end-stage liver disease. NAFLD can be accurately diagnosed by magnetic resonance imaging (MRI) methods, but NASH diagnosis currently requires a biopsy. The available data is based on a 90% random sample of a population-based cohort including all adult NAFLD subjects from 1997 to 2014 along with 4 controls for each case [2]. The goal of the study was to analyze the impact of NAFLD on development of incident metabolic comorbidities (MC), such as diabetes, hypertension, or dyslipidemia, and death. In this analysis, the number of MCs (0-3) plus death were each treated as separate states.

The NAFLD data is represented as 3 datasets, `nafld1` has one observation per subject containing baseline information (age, sex, etc.), `nafld2` has information on repeated laboratory tests, e.g. blood pressure, and `nafld3` has information on yes/no endpoints. To protect patient confidentiality, all time intervals are in days since the index date; none of the dates from the original data were retained. Subject age is their integer age at the index date, and the subject identifier is an arbitrary integer. As a final protection, a 10% random

sample of subjects was excluded. As a consequence analyses results will not exactly match the original paper.

1.6 Overview of this book

The book is partitioned into two major parts: the first deals with models that have a single endpoint, and the second deals with models that have multiple endpoints. Common themes used to approach every example throughout the book include these steps.

1. Determine the type of outcome or outcomes and draw the state space diagram
2. Organize the data into the format needed for analysis
3. Analyze the data using models and figures
4. Interpret the results

Chapters 2-4 describe ways to summarize the data descriptively and visually, various types of models used for single endpoints, Cox model basics and checking assumptions. Chapter 5 discusses issues with time-dependent covariates and various approaches for creating an appropriate analysis dataset. Chapter 6 provides details on how to adjust survival curves to represent a meaningful population of interest. Chapter 7 discusses risk prediction models and how to assess their performance. Chapter 8 introduces pseudovalues. The second part of the book starts with an overview of multistate models, then expands to competing risk analysis, more complex multistate models and multiple events of the same type.

Throughout the book there are sections marked with the  symbol indicating that this material is not necessary to understand the general concepts, but may be of interest in certain situations. Feel free to skip over these sections when initially reading the book.

Part I

Single endpoint

Chapter 2

Survival models

"A model is a lie that shows us the truth." Howard Skipper

The favorite model in statistics is the simple linear model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

and the second is the generalized linear model (GLM), which is very much like it,

$$\begin{aligned} E(y) &= g\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots\right) \\ &= g(\eta) \end{aligned}$$

where g is a monotone transformation function. Logistic regression, where y is binary and $g(\eta) = \exp(\eta)/[1 + \exp(\eta)]$, is the most common GLM.

What makes these two so very popular? Assume for the moment that there is one variable in which we are particularly interested, placed first in the equation as x_1, e.g., the treatment arm in a disease trial. If the model holds, in particular the feature that x_1 appears only once in the equation as an additive term, then β_1 is a wonderfully condensed summary of x_1: β_1 is *THE* effect of

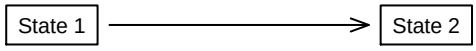

Figure 2.1 *The simplest multistate model.*

treatment regardless of age, sex, cholesterol, or whatever other factors appear as x_2, x_3, etc. Statisticians (and their customers) dearly love this simplicity, because it makes the underlying science much easier to think about, to say nothing of how much easier it is to describe and summarize the results for others. Generalized additive models (GAM) replace one or more of the linear terms $\beta_k x_k$ with a smooth function $s(x_k)$, but retain the separability of terms.

We would argue that successful statistical models, successful in the sense of wide use, tend to have 3 attributes.

1. Simplicity: in the sense described above, leading to simple explanations for the effect of key predictors.
2. Statistical validity: the model must describe the data adequately.
3. Numerical stability: the code to fit a model does not require hand-holding or fiddling with tuning parameters: it just runs.

Using logistic regression instead of linear regression to model a binomial response y is a good example of the importance of attributes 2 and 3. Modeling a binomial response y with an ordinary linear model is actually simpler to understand: a linear term $0.2x_1$ for a 0/1 treatment covariate x_1 predicts that treatment increases the probability of the outcome y by $.2 = 20$ percent, making it straightforward to reason about the cost-benefit of such a treatment. However, if there are multiple important covariates, the overall linear predictor $\eta = \beta_0 + \beta_1 x_1 + \ldots$ will nearly always contain values that are < 0 or > 1 for at least a few observations. Probabilities outside 0-1 are statistically unacceptable (violating attribute 2); and such values lead to the logarithm of a negative value in the underlying formula, which in turn leads to computational failure (attribute 3) as well. The imposition of a sinusoidal function g that maps the entire real interval onto 0–1 prevents both failures, and so has become the standard.

We will look at 3 common approaches for modeling (t, δ) survival data; each using the 3 attributes above. We will give a short example of each below, with a more full exploration to follow in subsequent sections.

1. Focus on the alive/dead outcome δ, accounting for the variable follow-up t as a nuisance.
2. Directly model the time to failure t, accounting for the censoring δ as a nuisance.
3. Model the hazard rate λ, i.e., the arrows of Figure 2.1.

Here is simple R code for all 3:

```
> # Focus on a binomial endpoint at a specific timepoint
> fit1 <- glm(I(time <730) ~ age + sex, data=lung, family=binomial)
> # Focus on time with Weibull model
> fit2 <- survreg(Surv(time, status) ~ age + sex, data=lung)
> # Focus on the hazard rate with Cox model
> fit3 <- coxph(Surv(time, status) ~ age + sex, data=lung)
```

```
          (Intercept)         age        sex
binomial    1.396785  0.02830842 -0.2298300
Weibull     6.274853 -0.01225703  0.3820851
Cox                  0.01704533 -0.5132185
```

- Details
  - The coefficients for the Cox model of the hazard and the Weibull model of the failure time have opposite signs for age and sex. This is because a higher hazard is bad and and a longer survival is good. The positive coefficient for age in the Cox model indicates the hazard increases as age increases. The negative coefficient for age in the Weibull model indicated the failure time decreases as age increases.
  - The binomial model above treats all subjects with a time of less than 2 years as a failure, completely ignoring censoring, which is wrong. Even so, the coefficients are not too dissimilar.
  - The Cox model does not have an intercept coefficient.
- General principles
  - Hazard models are the most prevalent in practice, and the majority of our presentation will be focused on that approach. Both the expected time to event and expected probability of event can be obtained from a hazard model; so these important predictions are not lost.
  - Accelerated failure time (AFT) models are nearly as easy to fit as hazard models, at least for single event data.
  - Binomial models are the most difficult to use, in terms of a correct data and model setup that will deliver valid estimates. However, they play an important secondary role in summaries.

Why are the hazard models, and in particular the Cox model, so popular? One reason, of course, is that they are set up to be additive and so satisfy axiom 1 (simplicity). Another is that additivity of log hazards often holds, at least approximately, in real data. Figure 2.2 shows the 2010 Minnesota death rates from age 40 to 90 years, the age range of a majority of medical studies, which fits remarkably well to the simple exponential model $\lambda(t) = \exp(\beta_0 + \beta_1 \text{age} + \beta_2 \text{sex})$. This log-additivity holds for the total US mortality as well, and, somewhat surprisingly, for many of the individual causes that make up this overall death rate, e.g., stroke, heart disease, etc. This does not mean that all outcomes and predictors will fall into this pattern, but it is a reasonable starting point. More surprisingly, the same log additivity often holds for acute disease processes such as advanced cancer, at least in the short term.

In the remaining sections we will cover the three most common hazard models, followed by accelerated failure time models and binomial models, and will attempt to justify each of the summary statements above.

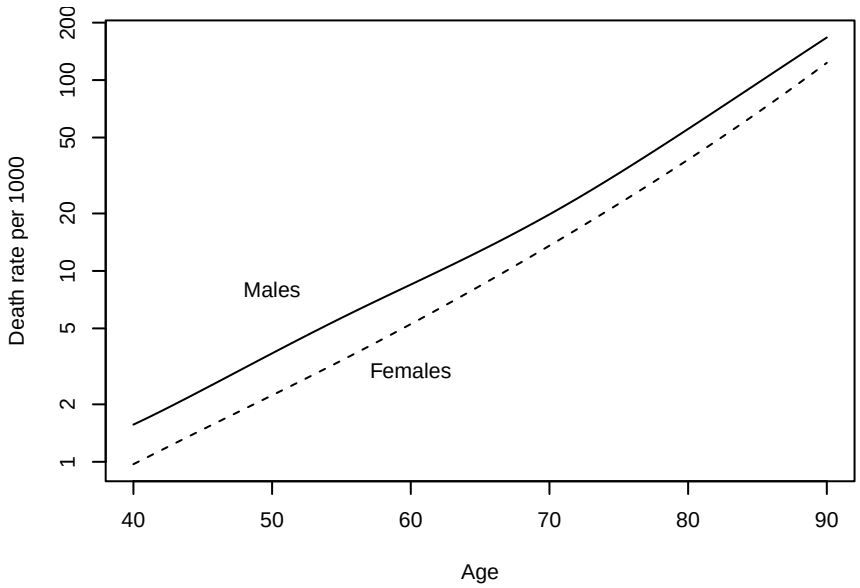

Figure 2.2 *Minnesota 2010 death rates for males (solid) and females (dashed) from age 40 to 90 years.*

2.1 Cox model

The most commonly used approach for time to event data is to model the hazards, and the Cox proportional hazards model is far and away the most popular hazard model. The three most common hazard models are shown below, equation (2.1) is Poisson regression, (2.2) is the Cox model and (2.4) is the Aalen additive hazards model. In each case X is the matrix of covariates, without an intercept term, one row per observation and one column per predictor, X_i the ith row of that matrix and β a column vector of coefficients.

$$\begin{aligned}
\lambda_i(t) &= \exp(\beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j) && (2.1)\\
&= \exp(\beta_0 + X_i\beta) \\
\lambda_i(t) &= \exp(\beta_0(t) + X_i\beta) && (2.2)\\
&= \lambda_0(t)\exp(X_i\beta) && (2.3)\\
\lambda_i(t) &= \beta_0(t) + X_i\beta(t) && (2.4)\\
&= \lambda_0(t) + X_i\beta(t)
\end{aligned}$$

The obvious difference between the Poisson and Cox models is in the intercept term β_0, which for the Poisson model is constant while for the Cox proportional hazards model it is an arbitrary function of time. The Aalen

model allows all of the coefficients to depend on time, and also removes the exponential function. Both the Poisson and Aalen models will be explored in later sections.

The Cox model is almost invariably written in the second form (2.3), with the intercept separated out as the *baseline hazard* $\lambda_0(t)$; we have included the first form to further emphasize the parallelism with Poisson regression. The concept that $\beta_0(t)$ is "just an intercept" will reappear, however, since it can often help clarify certain aspects of the model. When the Cox model holds, the ratio of hazards for two subjects i and j will be

$$\begin{aligned} \frac{\lambda_i(t)}{\lambda_j(t)} &= \frac{e^{\beta_0(t)+X_i\beta}}{e^{\beta_0(t)+X_j\beta}} \\ &= e^{(X_i-X_j)\beta} \end{aligned} \tag{2.5}$$

a value which is constant in time, hence the label of "proportional hazards".

The Cox partial likelihood (PL) is defined as

$$PL(\beta) = \prod_{i\in\text{deaths}}^{n} \frac{r_i}{\sum_{k=1}^{n} Y_k(t_i) r_k} \tag{2.6}$$

$$LPL(\beta) = \sum_{i=1}^{n} \int \left[\log(r_i(s)) - \log\left(\sum_{k=1}^{n} Y_k(s) r_k(s) \right) \right] dN_i(s) \tag{2.7}$$

There is a term in the PL for each observed event; the term compares the risk score $r_i = \exp(X_i\beta)$ of the subject who experienced the event to the sum of scores for all those who were at risk at that time. Looking at the individual terms in (2.6), we can think of each of them as a lottery model, assessing the probability that the subject who experienced the event should have done so. In the analogy, each observation's risk score r_i corresponds to the number of tickets they hold. Given that an event (lottery drawing) occurred, the probability for the subject who "won" is $r_i / \sum r_k$, where the denominator sum is over all those who were at risk, i.e., had opportunity to win the drawing by virtue of being alive and under observation at that timepoint. The analogy may be a bit macabre when the prize is death, but turns out to be very useful when thinking through some of the more subtle aspects of risk sets as discussed in Section 4.2.

Equation (2.7) rewrites this in terms of the log partial likelihood (LPL). The formula sums over all subjects via the artifice of using $dN_i(s)$ as the variable of integration, which reveals two important facets. The first is that the PL is well defined when a single subject can have multiple events (multiple jumps in dN), and so extends directly to data such as repeated infections. More important is that since each term is distinct in time, we can use time-dependent risk scores $r_i(s) = \exp(X_i(s)\beta)$. This allows the fitted hazard at any timepoint to depend on up-to-date information about a patient, in addition to any information gathered at a baseline study visit. Time-dependent covariates

have a very powerful attraction in medical studies, since in caring for a patient, the physician will also make use of the most recent information available. Time-dependent covariates mimic clinical practice.

One computationally important aspect of the Cox model is that it is *separable*: computation of the baseline hazard $\lambda_0(t) = \exp(\beta_0(t))$ can be deferred until after a solution has been found for the remaining coefficients. In fact, many users will postpone that step ad infinitum, which in turn leads to a major issue: namely that the baseline hazard is rarely if ever reported in published work, which in turn means that prediction of absolute risk is not possible from published results. (Absolute prediction is stymied for any model — linear, logistic, or hazard — when the intercept coefficient has been hidden.) This in turn has led to a somewhat widespread fiction that absolute prediction from a Cox model is not possible. In truth absolute risk estimates are relatively easy to calculate using any modern software package.

Cox model summary measures

Consider our simple Cox model for the advanced lung cancer dataset shown earlier, where age, sex, and performance score are included as predictors. Relative hazards are given by the standard printout, shown below, one which is essentially the same in all packages.

The death rate increases by 1.6 fold for each 1 point increase on the 4 point ECOG performance scale, and the rate is reduced by a factor of 0.6 for females relative to males, but there is only a modest increase of 1.1 for each additional decade of age. "Hazard ratio", "relative hazard", "risk ratio" and "relative risk" are all commonly used labels for these values, leading to significant gnashing of teeth among our more theoretical brethren, since only the first of these is technically correct.

```
Call:
coxph(formula = Surv(time, status) ~ age10 + sex + ph.ecog, data = lung)

            coef exp(coef) se(coef)      z        p
age10    0.11067   1.11702  0.09267  1.194 0.232416
sex     -0.55261   0.57544  0.16774 -3.294 0.000986
ph.ecog  0.46373   1.58999  0.11358  4.083 4.45e-05

Likelihood ratio test=30.5  on 3 df, p=1.083e-06
n= 227, number of events= 164
   (1 observation deleted due to missingness)
```

The `coef` column shows the estimated coefficients. The exponentiated coefficients provide the multiplicative change in risk for each covariate and are the hazard ratios.

- Age was divided by 10 in the Cox model fit, so the age coefficient represents a 10 unit change in age. In this fit, each 10 year increase in age is

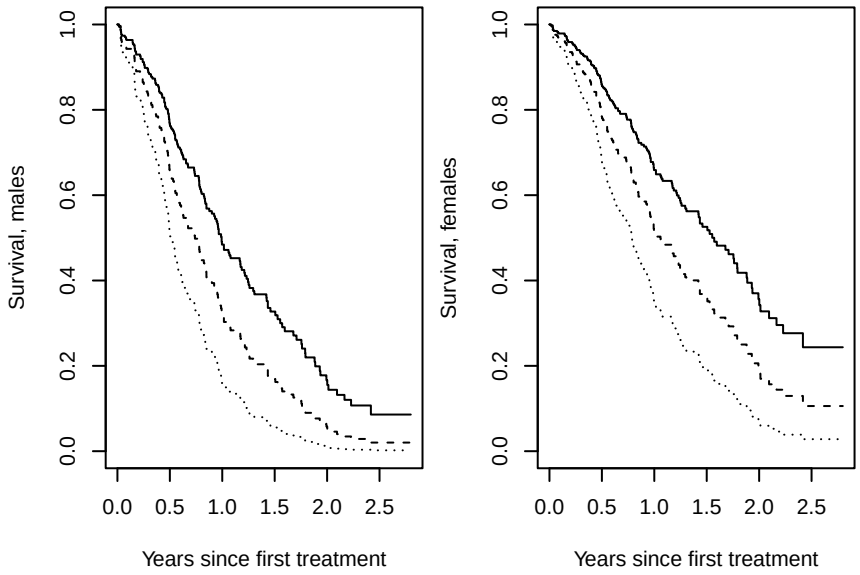

Figure 2.3 *Predicted survival for the lung cancer study, for subjects with ECOG performance score of 0 (solid), 1 (dashed) or 2 (dotted). age 65, males (left) or females (right).*

associated with a 12% increase in a patient's risk based on the hazard ratio of 1.12. Ages in the study range from 39 to 82 years, with the first and third quartiles at 56 and 69 years. The difference in risk between the first and third quartiles is exp(.11* 13/10) = 1.15.

- Sex is coded as 1=male and 2=female, so a 1 unit increase in sex compares females to males. The negative sex coefficient and the hazard ratio < 1 indicate female's death rate is lower than males, .58 fold as large. (Sometimes reported as a 42% lower risk of death, i.e., 1-.58=.42). Many researchers prefer to re-code associations with negative coefficients to report only increased risks for consistency and ease of interpretability, e.g., males have a risk that is exp(.55261)= 1.74 fold greater than females. This is equivalent taking the reciprocal of the hazard ratio (i.e., 1/0.575 = 1.74). Thus, males have a 78% increased risk of death compared to females.
- The physician's estimate of the ECOG performance score can range from 0-4, and values of 0-3 were present in this dataset. A 1 unit increase in physician's ECOG score is associated with an 1.59 fold increase in the mortality rate.

Absolute prediction can be done in terms of survival curves, as shown in Figure 2.3, the predicted survival at a particular time, or in terms of the restricted mean survival time (RMST) as shown in Table 2.1. All of these are easily obtained from the fitted `coxph` model in R. The risk estimates can be

			Pr(death)			
Sex	Age	ECOG	1	1.5	2	Mean survival
Male	55	0	0.48	0.63	0.80	1.23 (0.07)
Male	55	1	0.64	0.80	0.92	0.93 (0.04)
Male	55	2	0.81	0.92	0.98	0.68 (0.02)
Male	71	0	0.54	0.70	0.85	1.11 (0.06)
Male	71	1	0.71	0.85	0.95	0.82 (0.03)
Male	71	2	0.86	0.95	0.99	0.60 (0.02)
Female	55	0	0.31	0.44	0.60	1.59 (0.12)
Female	55	1	0.45	0.60	0.77	1.29 (0.08)
Female	55	2	0.61	0.77	0.90	0.98 (0.05)
Female	71	0	0.36	0.50	0.67	1.48 (0.10)
Female	71	1	0.51	0.66	0.83	1.17 (0.07)
Female	71	2	0.68	0.82	0.94	0.88 (0.04)

Table 2.1 *Predicted absolute risk at 1, 1.5, and 2 years, and the restricted mean survival time (RMST) at 2.5 years (with standard error), for selected predictor combinations.*

read off of the survival curve(s), and the mean is the restricted mean time, also known as the sojourn time in the initial state. Because of finite follow-up, about 2.5 years in the lung study, it is not possible to estimate the overall mean survival and instead we estimate a conditional mean $E[\min(t, \tau)]$ for a cutoff of $\tau = 2.5$ years. This result can be interpreted as the portion of the first 2.5 years that the average patient will be alive.

Relative prediction in the sense of equation (2.5) is always possible, leading to the standard table of β, se(β) and p which forms the statistical backbone, and sometimes the only substance, of the final results in most scientific papers. Lukewarm support for this practice of ignoring $\beta_0(t)$ can be based on the idea that absolute event rates may differ from one population to another while the relative effects of one or more covariates on that rate will stay approximately constant, and that therefore the relative rates are "portable" to other studies. The same argument underlies the common practice of only reporting relative rates (odds ratios) from binomial models, rather than absolute probabilities of the endpoint.

However, we will argue here and throughout the examples that absolute rates are a critical portion of the research report. Relative rates are useful but only tell part of the story. One of the more well known examples is the effect of smoking. It leads to a 10-30 fold risk increase in small cell lung cancer (SCLC) and a 2-4 fold increase in heart disease. The first of these gets all of the attention. However, SCLC is a very rare tumor while heart disease is common; the absolute number of excess deaths from heart disease, for smokers, turns out to be 5x the excess due to cancer. Reference for this?

The disadvantage of absolute prediction is that the predictions are for a particular combination of covariates: we no longer have a single number

summary of *the* effect for a covariate. If we compare the sojourn times from row 1 to row 7, row 2 to row 8, etc in Table 2.1, for example, the estimated gain in RMST for females is .36 (1.59 - 1.23), .36, .31, .37, .34, and .28 years for the 6 combinations of age and sex in the table; all slightly different. Additivity on the log hazard scale does not translate to additivity on another scale. This makes presentation of the results more difficult. We will discuss strategies to combat this using marginal estimates in Chapter 5 and using pseudovalues in Chapter 7.

Risk sets

A key insight into the Cox model is that at each event time, the partial likelihood of a Cox model is based on comparison of the subject who had the event to all others who were at risk at that time. This is demonstrated in the form of the Cox model score statistic

$$\sum_{d \in \text{deaths}} x_d - \overline{x}(t_d) \tag{2.8}$$

which is a sum of differences between the covariate vector of each death and a weighted average covariate vector for those at risk at the time of the death. The subjects under observation at a particular time are referred to as a risk set.

Understanding the concept of risk sets provides valuable insights into the Cox model results. Several extensions to the Cox model (e.g., strata, matching, and time scales) arise from this.

Strata

An important extension of the basic Cox model is stratified models, where the intercept $\beta_{0g}(t)$ is allowed to differ for different subsets of the data, here indexed by a group label g. An example of this would be to stratify by each enrollment center in a multi-center clinical trial. Stratification allows each centers' baseline hazard to have a different *shape* — they no longer are required to be proportional.

As a simple example of this consider the Chronic Granulomatous Disease (CGD) trial. There was concern, and some a priori evidence, that the 13 enrolling centers would not all attract the same patient population, and based on this the centers were divided into 4 hospital categories represented by a variable `hos.cat` in the data. Figure 2.4 shows cumulative hazard functions for the 4 hospital category groups. What is important with respect to stratification is not the absolute size of the hazard functions but their *shape*. In the graph the 2 groups of US:NIH and US:other are essentially identical in this metric but the 2 European groups differ, both from the US groups and from each other.

The stratified model gives the most general adjustment for different subsets

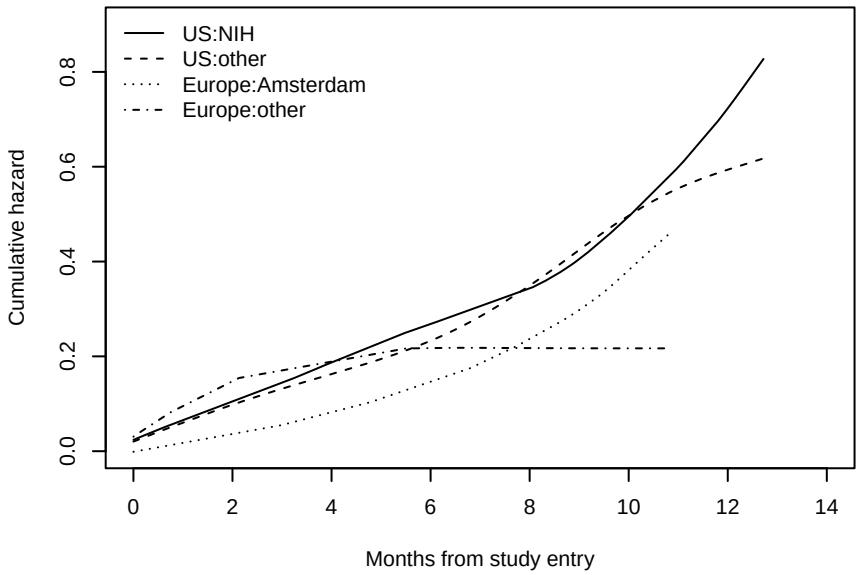

Figure 2.4 *Estimated cumulative hazards for hospital categories from the CGD dataset. The hazard functions have been smoothed to aid visualization.*

of data (e.g., institution). In particular, the effect included in as a strata need not assume proportional hazards. The trade-off is that no *estimate* of the covariate effect is provided by the stratified model, nor is there a test statistic for comparing the stratified and un-stratified approach. A general rule for strata is to use it for covariates that are expected to have an important effect on the outcome, but the estimated size of the effect is not of direct interest. Common choices are enrolling institution or covariates whose effect is already well established.

The change to the program call to stratify by hospital category is trivial as shown below. In this case the effect of stratification on the coefficients is minor, results are in Table 2.2. The estimated treatment effect changes slightly, inheritance type remains minor, and the estimated effect of pre-study steroid treatment increases from 2 to 3, all with a very modest increase in standard errors. This is not an atypical result: coefficient changes will often be small.

```
> # Only use first infection
> cgd1 <- subset(cgd, tstart==0)
> # standard
> coxph(Surv(tstop, status) ~ treat + inherit + steroids, data=cgd1)
> # stratified
> coxph(Surv(tstop, status) ~ treat + inherit + steroids +
            strata(hos.cat), data=cgd1)
```

	treat:IFN-g	inherit:autosomal	steroids
standard	0.34 (0.17-0.65)	1.02 (0.55-1.91)	2.25 (0.54-9.36)
stratified	0.32 (0.16-0.61)	1.04 (0.54-2.00)	3.14 (0.68-14.56)

Table 2.2 *Hazard ratios (confidence intervals) for treat, inherit, and steroids using the CGD data. Models were fit using a standard Cox model and a Cox model stratified by hospital category.*

The model now has 4 baseline hazards, one for each hospital group.

Figure 2.5 compares fits of the lung cancer data, treating institution as a stratum or as a covariate. In this case we see that the coefficients of interest — age (decades), male sex, and ECOG performance score — hardly change with the choice of method.

A second, independent way to view stratification is based on risk sets: it allows the model to compare like with like. For a stratified model, the risk set is limited to those who are both at risk and members of the same stratum. For example, if a model is stratified by enrolling institution, then each event is compared to other observations from the same institution.

Relationship to log-rank test

When there is a single categorical variable in the Cox model, the score test from the Cox model is identical to the log-rank test. As pointed out by Schoenfeld [93], statistical tests such as the log-rank that ignore covariates other than treatment are less efficient than tests that use the covariates, often substantially less when the omitted covariates have greater impact on the hazard than the treatment. In fact, when Schoenfeld considered a study with two years of accrual and two years of follow-up, he showed that the log-rank test had an efficiency of 61% compared to a test using the covariate.

Power considerations

The risk set insight helps us understand a limitation of stratification, which is a potential loss of power. A devastating effect on power can come from over-stratification, where some or many of the events have 0 other subjects that share their strata. Such observations are effectively ignored in the partial likelihood.

A general epidemiology rule of thumb for case-control studies is that 2 controls per case is better than 1, 3 better than 2, etc., but that the gains become very minor beyond 4. The argument is similar to the variance of a 2 group t-test with n and kn observations in the two groups, for $k = 1, 2, 3, 4$ the variance will be $(\sigma^2/n)(1+1)$, $(\sigma^2/n)(1+1/2)$, $(\sigma^2/n)(1+1/3)$, $(\sigma^2/n)(1+1/4)$, giving multipliers of 2, 1.5, 1.33, 1.25; clearly a case of diminishing returns.

A stratified Cox model that has > 10 subjects in each stratum will thus have essentially no loss of power.

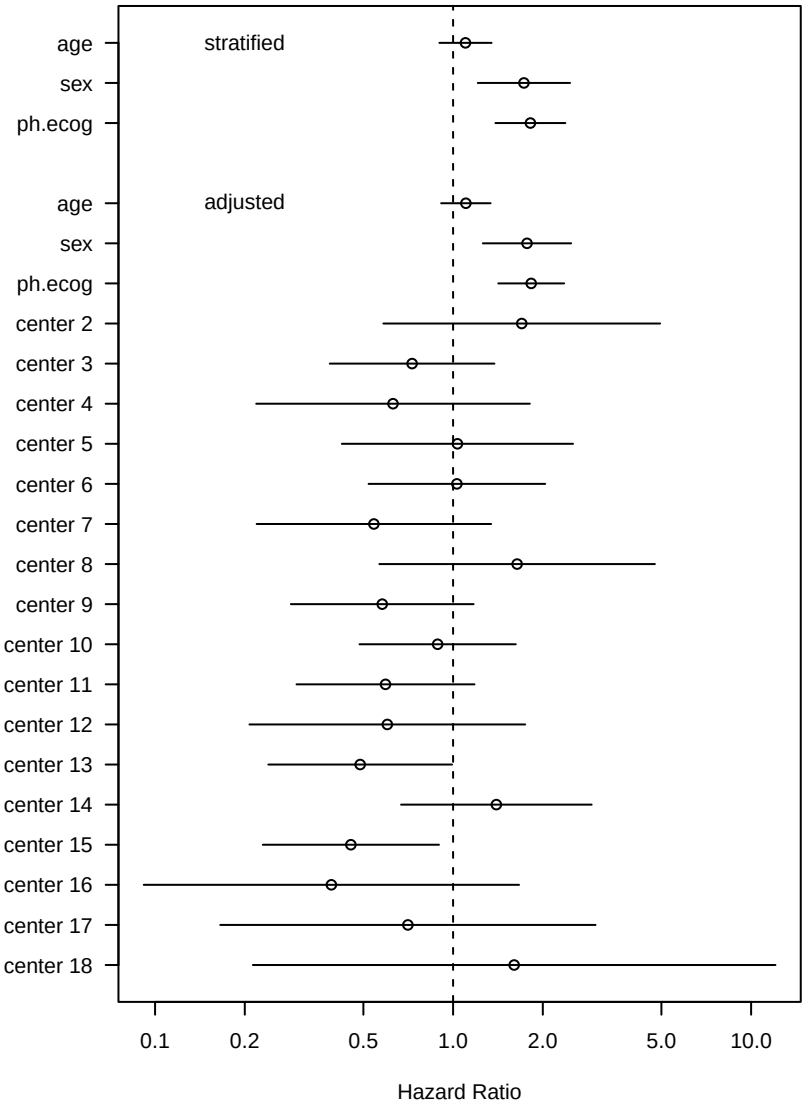

Figure 2.5 *Hazard ratios and 95% confidence intervals from a model stratified by institution (top 3) and a model that adjusts for institution (bottom).*

As the number of subjects per stratum approaches 2 the loss of power will be noticeable.

Matching

There is a natural tension between matching and covariate adjustment, closely related to bias/variance trade off. Matching gives the most general adjustment for factors, at the expense of potentially higher variance and loss of a coefficient estimate for the matched covariate. In the extreme, matching on many factors can lead to each event being paired with only a handful of non-event observations. Thus, counter to the usual increase in power associated with matched methods in other analyses, matched analysis of case-comparator pairs in the Cox model setting can lead to a loss of power. When a matched pair is used as a strata in a Cox model, a comparison between the subjects in the pair occurs at the first event for the pair. Following that first event, only 1 subject is under observation, so no comparison can be made at the latter subject's event time. This is why a matched analysis in a Cox model setting is often thought of as truncating the follow-up to the minimum follow-up of the pair. This analysis has the advantage of ensuring the case-comparator groups remain balanced on the matching factors throughout follow-up, but this advantage is offset by a loss of power.

Sampled risk sets

In the other extreme of big data, the statistical gain from a very large risk set will be minimal in an ordinary Cox model. Sampling can be used to reduce computational burden or resource use when analyzing large datasets.

This concept is used in the nested case-control study design, where the subject with an event is compared to a *random sample* of subjects from the risk set, instead of to all subjects in the risk set. The nested case-control approach is commonly used when one or more covariates requires significant resources (e.g., an expensive laboratory test or the need to manually abstract information from the medical history). It may not be feasible to collect such information on all subjects in such a situation. Sampling a small number of controls can result in a minimal loss of power with a potentially large decrease in the study's cost.

Even better gains are possible with targeted re-sampling. A key insight is based on the form of the Cox model score statistic in equation (2.8) Any sub-sampling approach which preserves the expectation of $\overline{x}$, will preserve the expected value of the score statistic, leading to the same asymptotic solution. Since this is just a mean, an entire literature of sampling estimates for a mean is available to us [61]. There are now a large number of approaches that utilize this idea, under the general heading of risk set sampling. They include:

- Case-cohort design. A single random sample is drawn at the start of the follow-up, which will be used as the comparison group henceforth.

- Nested case-control. A random selection is made from the risk set at each death.
- Counter-matching. Ancillary data is used to select risk sets that will be most informative. Roughly, power is increased when var(X) is large within each risk set, similar to how a larger spread in X values minimizes the variance of the slope in a linear model.

Two important principles apply to all of these approaches. The first is that subject selection can never depend on the future, e.g., selecting controls who will be event free for the next year, say. It is ok to remember the past, however, e.g., to not reuse a subject who was chosen as a control at a prior event time. The second is that the proper survey sampling weights and variance must be determined. In selected applications the increase in either study efficiency and/or precision of the final estimates can be substantial.

Given a set of events and a defined risk set for each of them, a common way to arrange the data is by blocks. Each block has a row for the 'event' observation plus rows for the selected comparator observations from the risk set at that event time, covariates X for those subjects, a grouping variable which contains a unique value for each risk set, and a status variable which is 1 for the event and 0 for comparators. A dummy time variable is then added to the data, with a constant value of 1 (or any positive value); it is only necessary that within each risk set the dummy time value for the case is less than or equal to that for the control. Then fit a Cox model as `Surv(dummy, status)` $\sim$ `strata(group) + x1 + ...`, where `x1, x2`, etc. are the covariates of interest. Essentially, the appropriate risk sets are enforced by the stratification variable.

This approach can also be used to analyze a binary outcome in a matched case-control study using logistic regression models. Quite surprisingly, the Cox partial likelihood (CPL) for the stratified dataset described above is exactly identical to the log-likelihood for a matched logistic regression, that is, whenever each case is matched to a defined set of controls. As a consequence of this, and of the reliability and speed of Cox model algorithms that has resulted from their heavy usage in statistics, many programs for matched case-control binomial data are set up as a simple front end to a proportional hazards model "engine", simply recoding the data per the prior paragraph. The `status` variable encodes the case status of 1 or 0, and the stratification variable identifies the matched sets.

In the more general (but quite rare) case of m:n matching, where there are m controls matched to n cases in each group, the equivalence between the Cox partial likelihood and the logistic still holds, but only if the Cox model computation uses the exact partial likelihood approximation for ties, see Section A.3.2 for more details on the approximations. For the standard case of $m = 1$, all of the Cox model approaches for ties yield the same result.

A very interesting case of risk set sampling is provided by the analysis approach used in the Nurse's Health Study. The study has over 120 thousand enrollees, each of whom has up to 16 biennial follow-up visits. Because of time-dependent covariates, the analysis dataset has up to 16 (age1, age2)

rows for each subject, each containing the participant's most recently known values over the interval from age1 to age2; in all almost 2 million total rows. Because of the long follow-up and low death rate, the terms in an ordinary Cox partial likelihood have on average over 60 thousand participants in the risk sets for each death. This is clearly overkill — half a dozen would suffice from a statistical point of view. Leaving any subjects totally out of the analysis, however, is unacceptable, given the effort that the participants themselves have contributed to the research project. A solution is to use the age at prior visit, in months, as a stratification variable. This creates about 700 bins; essentially `floor(age1* 12/365.25)`, when starting with `age1` in days. The dataset is large enough that there are multiple deaths in every month of age, each death will be compared to other at risk subjects of the same age. (If two people were the same age at a prior visit 2 years ago they are, obviously, still the same age at the current visit.) Two observations for the same subject will never be in the same stratum, so the (time1, time2) form of the response is not needed; an important bonus since this form of input runs much slower in some software implementations. The model is then `Surv(age2, death)` $\sim$ `strata(group) + x1 + ...`. As a final bonus, a new dataset did not need to be created.

Time scales

As stated above, the Cox partial likelihood has a term for each event, which compares the observation with the event to others who were at risk for said event, with "at risk" defined with respect to the time scale t in the hazard $\lambda_0(t)\exp(X\beta)$. The most common time scale is t = time from study entry, in which every subject enters the analysis at time zero by definition; other possible choices are t = age or t = calendar time. In the first case, the fitted coefficients estimate the increase in hazard of each subject relative to others with the same time since enrollment; when using age scale the coefficients estimate the increase relative to others of the same age, and etc.

Figure 2.6 illustrates an example dataset shown on a time scale and an age scale. The vertical dashed line shows who is included in the risk set for a subject who has an event. On the time scale there are multiple subjects in the comparison group whereas on the age scale there are fewer available subjects. Also of note, for the age time scale the time intervals do not start at time 0. To accommodate this in the analysis, instead of just specifying one time variable, two time variables are needed: a starting age at the onset of the study and a stopping age at the last follow-up. If only the stopping age is used in the analysis, this would introduce immortal time bias, as the patients were identified and enrolled in the study at ages that were considerably later than their birthdates. It is easy to prevent this bias by properly indicating the starting age when each patient enters the study.

The time scale can be thought of in similar fashion to stratification or risk set sampling. The concept of "comparing like to like" is a relevant criterion for choosing which scale is best. In a controlled clinical trial, participants will

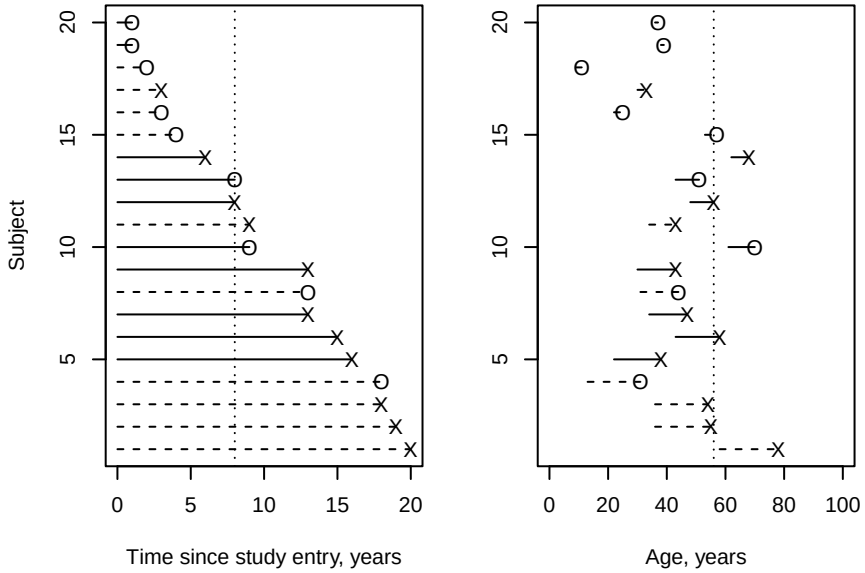

Figure 2.6 *20 hypothetical subjects with follow-up and event status plotted on time-since-study-entry scale and on an age scale. Subjects in group 1 have solid lines and subjects in group 2 have dashed lines, events are marked with an "X" and censored observations are marked with an "O". The vertical dotted line represents the covariate comparisons that occur at one of the event times.*

have often passed through the eye of a needle, in terms of a multi-page list of eligibility criteria, and are consequently closely matched as of the enrollment date. Thus time since enrollment is often the obvious choice.

The free light chain study provides an example where age is the more natural time scale. In this population-based study, a blood sample was solicited from all subjects over the age of 50 years in Olmsted County, Minnesota. The samples were obtained from excess blood drawn during usual clinical care of these patients. In this case "time since blood draw" is a fairly meaningless metric, since the reasons for the patient visit range from an annual checkup to chronic disease to acute illness. Moreover, in a population-based sample, the primary drivers of the underlying mortality rate are age and sex; using an age scale for t in the Cox model along with stratification on sex more naturally fulfills the criteria of comparing like to like. This idea of using an age scale for epidemiologic cohort studies is supported by simulations presented in work by Thiébaut et al [106].

Another way to think of this is to imagine that you were given only one piece of time information: time since enrollment, time since birth (age), time since disease onset, etc., from which to make a prediction of a participant's current risk. Which would you choose to create the most accurate prediction?

Often the choice of time scale will make only a minor difference in the

estimated coefficients. Here is an example using the MGUS dataset. Male sex and low hemoglobin are both markers of an increased risk of death, whether the comparison group is others of the same age or others of the same duration since testing. For the age time scale, the `Surv` function includes both the starting and ending age of each subject's follow-up interval.

```
> fit1 <- coxph(Surv(futime, death) ~ hgb + sex, data=mgus)
> fit2 <- coxph(Surv(age, age + futime/365, death) ~ hgb + sex, data=mgus)

              hgb            sex
1: time scale -0.178 (0.043) 0.496 (0.142)
2: age scale  -0.150 (0.044) 0.354 (0.140)
```

For many datasets multiple time scales may affect the hazard. For instance, subjects enter the MGUS dataset when serum electrophoresis (SEP) is ordered, a somewhat uncommon laboratory test. Subjects ranged from age 34 to 90 years and very few were found to have a plasma cell malignancy, the condition for which the test is definitive. Most subjects died of the ordinary afflictions of old age, with a median survival from the test date of over 13 years. Age will clearly be an important predictor of the individual hazard rate. However, SEP is normally only ordered as part of a work up for more serious indications. A comparison of the death rate for the subjects, as compared to the population expected rate matched by age and sex, shows a 3.3 fold increase over the first 6 months versus 1.3 fold for the interval from 6 months to 5 years. A subset of the subjects, at least, is being enrolled in an interval of very high risk.

Should we use an age time scale in the Cox model and model time since enrollment, or use time since enrollment as the time scale and model age? One of the scales will be addressed via "matching" (i.e., risk sets compare like with like) and one via modeling. Arguments about the relative merits of matching (or stratification) versus covariate modeling are as old as the statistics profession. One of the beauties of the Cox model is that whichever measure is used for the time scale will be adjusted for in the most general way, due to the non-parametric nature of the baseline hazard function. One of the challenges is that you only get to choose one. A sensible approach is to choose the aspect which has the larger effect, age or time since enrollment, as the primary time scale. Another would be to select whichever aspect will be more difficult to model explicitly, and use that as the time scale.

When there may be multiple time scales influencing the results, it is important to adjust for the other scale in the model. If using the time scale, age should be included in the model, and if using the age scale, time should be included in the model.

```
> ## time scale, adjust for age
> fit1 <- coxph(Surv(futime, death) ~ pspline(age,2) + hgb + sex, data=mgus)
>
```

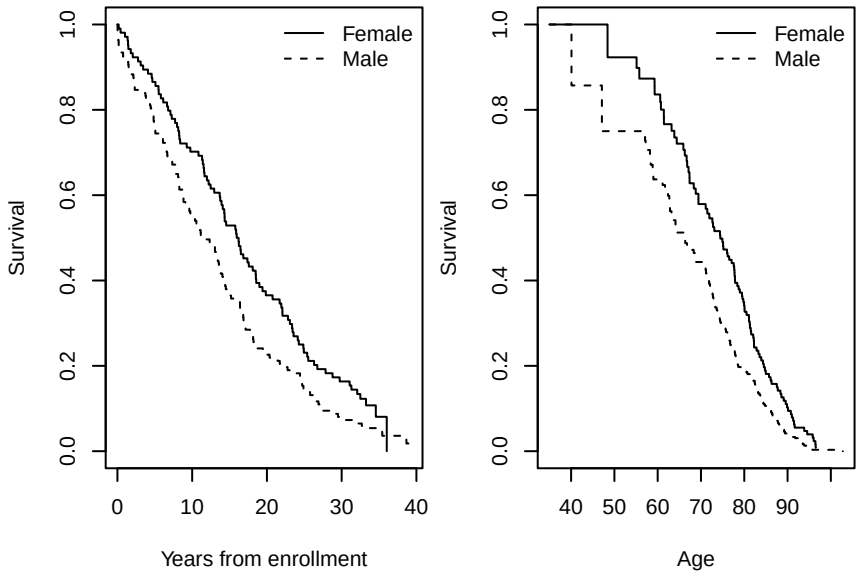

Figure 2.7 *Survival curves for the MGUS data stratified by sex. The first panel is on the time-since-enrollment scale and the second panel is on the age scale, starting at age 35 years.*

```
> ## age scale, adjust for time
> mgus$age0 <- mgus$age
> tmp <- survSplit(Surv(age, age+futime/365,death)~age0+hgb + sex,data=mgus,
                   id="id",cut = c(34:90), episode="time")
> tmp$time <- tmp$age-tmp$age0
> fit2 <- coxph(Surv(age, tstop, death) ~ pspline(time,2) + hgb + sex, data=tmp)

              hgb            sex
1: time scale -0.162 (0.045) 0.341 (0.142)
2: age scale  -0.161 (0.045) 0.382 (0.143)
```

One issue to be sensitive to when doing analysis on age scale, for a study with baseline covariates measured at time of entry, is that many lab tests become less predictive the further one is from the measurement date. The risk set for an event at age 60, for instance, might well include a subject who enrolled 1 year ago and another who enrolled 10 years ago (at age 50). This mixing of "fresh" with "stale" laboratory measurements can lead to unexpected results. I don't see how. This needs more expansion.

We are all aware that KM curves can become unstable at the far right, when the number at risk is very small (<= 5) and the steps become large. For KM curves on the age scale, we need to remember that this same effect can happen at both the left and right edges. Figure 2.7 shows the MGUS data

on the two scales. For age, we specified that the curves start at age 35 years, resulting in an estimate of survival after age 35, given survival to age 35. (The youngest age in both the males and females was 34 years.) However, it is clear from the size of the steps that there are few subjects joined the cohort at less than 50 years of age, whereas the curves on the time-since-enrollment scale start off with small jump sizes. If an event occurs at a young age when only a few subjects are under observation, this can lead to unrealistic event rates. It is best to start the computation at a point with larger sample size, e.g., survival after 50 years of age. This is a form of landmarking, which will be discussed more formally in Chapter 5.

Transformation models

An extension of the Cox model that is sometimes suggested is the family of transformation models

$$\lambda(t) = g(e^{\beta_0(t)+\beta_1 x_1+\beta_2 x_2+\cdots}) \tag{2.9}$$

If g is the identity function, this is an ordinary Cox model; thus the formulation allows the Cox model to be embedded in a larger family. Common choices for the transformation function g are the Box-Cox family or the logarithmic family $g(x) = \log(1 + rx)/r$. For the latter, $r = 0$ corresponds to the identity (an ordinary Cox model) and $r = 1$ to a logistic link. A significant downside to this model is that the computation of $\beta_0(t)$ is no longer separable, leading to a more difficult estimation problem; so much so that the model has never gained significant traction. The approach has definite merits in terms of flexibility, but we will not consider it further.

2.2 Poisson regression

The simplest hazard model is Poisson regression (2.1). It can be fit using standard GLM software by using a simple trick. Let d_i be the 0/1 status variable, t_i the follow-up time for a subject, and x_i the vector of covariates for the subject, including an intercept. Then

$$\begin{aligned} E(d_i) &= \lambda_i t_i & (2.10)\\ &= \left(e^{x_i\beta}\right) t_i \\ &= e^{x_i\beta+\log(t_i)} & (2.11) \end{aligned}$$

Equation (2.10) writes the expected number of events as rate*time, the second line inserts our model for the rate, and the third absorbs the time value into the model as a new variable with a known coefficient of 1, this is known as an offset. Using the `rats` cancer data, the results below show a Poisson model fit with sex and treatment as covariates.

```
glm(formula = status ~ rx + sex + offset(log(time)), family = poisson,
    data = rats)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.11       0.22  -27.69   <0.001
rx              0.73       0.31    2.36    0.018
sexm           -3.03       0.72   -4.18   <0.001
```

Only 2/150 males had tumors as compared to 40/150 females, leading to an estimated hazard ratio of exp(-3.03)= .05 for males as compared to females, while the carcinogenic treatment is estimated to have increased the tumor rate by exp(0.73) = 2.1 fold.

At this point many readers will be protesting (correctly) that the 0/1 status variable is not a Poisson realization, i.e., what would be obtained if t were fixed and y counted radioactive decays, for instance, over some interval. This will be most obvious in a mortality study that followed everyone to death; the status in that case will be identically 1, with no variance at all. In this latter case, it is the time t which is the random quantity. As was pointed out by Berry [13], the log-likelihood from a GLM Poisson model is correct in this latter case up to a fixed constant. This was a forerunner of the modern and more general argument for this fact, which is based on counting processes and martingales [3]. Since a fixed constant factors out of the log-likelihood, correct coefficients, hypothesis tests and confidence intervals can be obtained using GLM software.

Poisson models are closely related to the rate estimates that were proposed in **??** as simple summaries. If there is a single binary covariate and time is in years, then the predicted value $\exp(X\hat{\beta})$ from a Poisson model, for each group, will exactly equal the simple estimate r= (total events)/(total observation time). A formal test for rate = 0 is rarely interesting (immortality is unlikely a priori), but the model does give a simple way to compare rates. If t_i is replaced by the expected number of events for each subject, calculated from known rates for some population, then the predicted value will be the standardized mortality ratio comparing the set of subjects to said reference population. This has a long history in epidemiology [56].

Stratification by time

The most limiting aspect of Poisson models is the assumption of a constant underlying rate. For a study with long follow-up, this will almost certainly not be true; Figure 2.2 shows that the baseline population mortality increases 100 fold from age 50 to 100 (while the male/female hazard ratio stays nearly constant). An early period of higher or lower mortality is common even in shorter studies. Figure 2.8 shows the estimated cumulative hazard for the lung cancer study, along with a linear approximation for the first 6 months. The death rate clearly rises after the initial period. This can be accommodated in the Poisson model by breaking each subject's time into disjoint epochs,

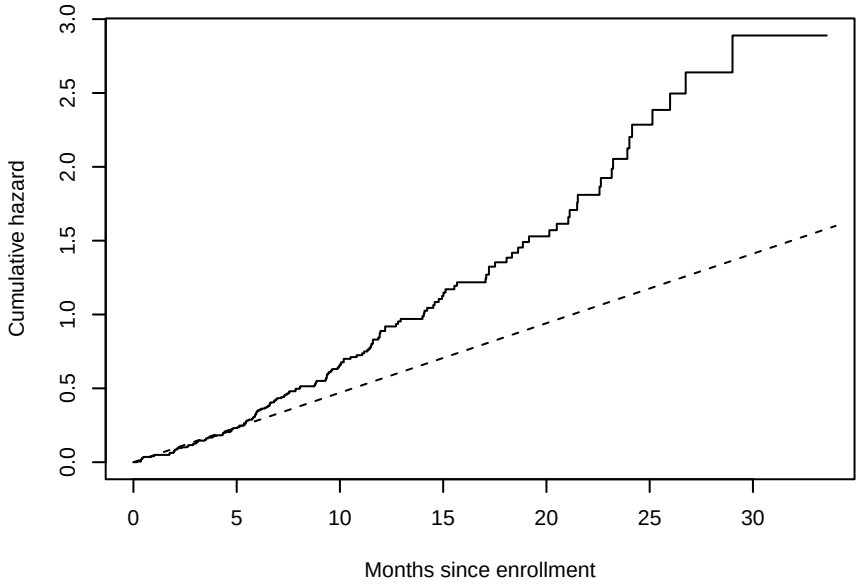

Figure 2.8 *Cumulative hazard curve for the lung cancer study. The dashed line is an approximation to the first 6 month's experience.*

with the hazard rate (slope of the cumulative hazard) assumed to be constant within epoch. In the model, this corresponds to a separate intercept term for each epoch. For instance, divide the lung subjects into 3 epochs of 0–6, 6–18 and 18+ months. Table 2.3 shows the result of Poisson and Cox model fits. The coefficient effects were attenuated when each subject's rate is assumed to be constant over time, but the results for a 3 epoch Poisson fit are nearly identical to the Cox model results, both in terms of coefficients and standard errors. We will return to this topic in a later example; it is a consequence of the fact that three connected line segments provide a close fit to the cumulative hazard in Figure 2.8.

As a second example, look at the follow-up for the MGUS dataset, as shown in Figure 2.9. In this case the hazard is larger over the first several months after enrollment. Such early periods of either higher or lower hazard are the rule rather than the exception in clinical studies. For MGUS the study used opportunistic enrollment: all those for whom a particular laboratory test had been ordered. This selects subjects at a time when they are currently visiting the health system, which is not surprisingly an interval of poorer health status. In the cancer study, as in many clinical trials, patients at risk for immediate mortality are often explicitly excluded, or if not, may be too ill to desire participation.

	Intercept			Age	Female	Performance
	1	2	3	(Decades)	Sex	Score
Poisson 1	0.54			0.06 (0.06)	-0.32 (0.11)	0.29 (0.07)
Poisson 2	-0.80	-0.29	0.16	0.10 (0.09)	-0.54 (0.17)	0.46 (0.11)
Cox				0.11 (0.09)	-0.55 (0.17)	0.46 (0.11)

Table 2.3 *Coefficients (Standard error) from two Poisson models and one Cox model fit using the lung data including the terms age (divided by 10), sex, and ECOG performance score. Poisson 1 includes log(time) as an offset whereas Poisson 2 splits follow-up time into 3 time periods (epochs) and includes the epochs in the model. The coefficients for the Poisson 2 model, which accounts for the different underlying rates in the 3 epochs, are similar to the coefficients in the Cox model. (Note that baseline rates are exp(intercept)).*

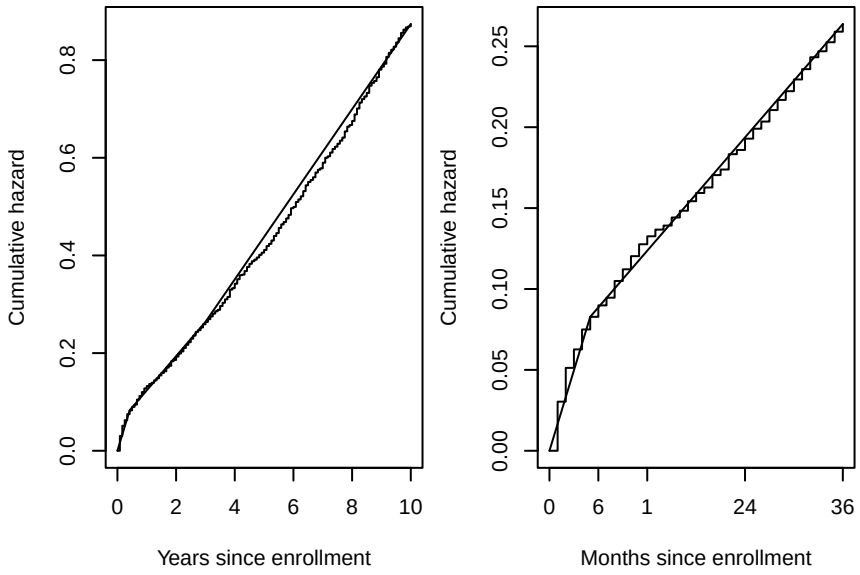

Figure 2.9 *Cumulative hazard curve for the MGUS data over 10 years post enrollment, and an expansion of the first 3 years.*

Connection with the Cox model

As seen above in Table 2.3, the coefficients of the Poisson model are very close to those for a Cox model. In fact, if the time scale is divided sufficiently fine, coefficients will be identical to a Cox model fit. Looking at the two model definitions of Poisson (2.12) and Cox (2.13) just below, a Poisson model fit using partitioned time is an approximation of $\beta_0(t)$ with a step function.

$$\lambda_i(t) = \exp(\beta_0 + X_i\beta) \tag{2.12}$$

$$\lambda_i(t) = \exp(\beta_0(t) + X_i\beta) \tag{2.13}$$

	Age group					
	< 50	50–59	60–69	70-79	80–89	90+
<5 m	82	155	337	479	293	38
5-24 m	81	152	333	473	285	44
> 24 m	61	154	370	616	554	160

Table 2.4 *Time since enrollment and age strata for the expanded Poisson data set.*

The formal identity requires a separate intercept (interval) for every unique event time in the dataset, in which case the Poisson model has completely reconstructed the time-dependent intercept $\beta_0(t)$ of the Cox. This approach will be computationally inefficient, however, due to the large number of intercept coefficients. If a smaller number of intervals is chosen, such that the straight line approximation is close enough, then the Poisson approach with a smaller number of intervals will be sufficient.

This approach first appeared as an approximation when Cox models were not yet a part of the standard statistical packages, see for instance Whitehead [115] or Laird and Olivier [60]. The approach has become useful again with the advent of machine learning software, which may support GLM models but not (directly) the proportional hazard model, or in Bayesian computations, for the same reason. Appendix A.7 shows more fully how to get near perfect approximations to the proportional hazard model using either Poisson or binomial models. This needs to be copied from the vignette.

Multiple time scales

The MGUS dataset has very long follow-up time, and it is important to model both the effects of age and the initial hazard after enrollment. Unlike the Cox model, where one time scale or the other must be chosen as primary, Poisson models allow for both by using a matrix of intercepts. This is done by creating a matrix of time strata. As an example we will use post enrollment periods of 0–5, 5–24, and 24+ months, and break up age by decade. Death rates are a function of *current* age, so it is not sufficient to partition subjects by their age at enrollment. Table 2.4 tabulates the 4667 pseudo-observations that are created for the 1384 MGUS subjects.

For analysis, the primary work is to partition each subject in the dataset into multiple parts. The first subject, for instance, is enrolled at age 88 years and has 30 months of follow-up. They will partition into 3 intervals: the first 5 months are in the 0-5 months x 80-89 years bin, then next 19 months in the 5-24 months x 80-89 years bin, and the last 6 months in the 24+ months x 90+ years bin. Assume that the resulting dataset has a variable `time` which contains the number of months within each interval, the modeling covariates of interest, and categorical variables giving the age and time group for each interval.

Though "build the appropriate data" is a simple statement, in actual prac-

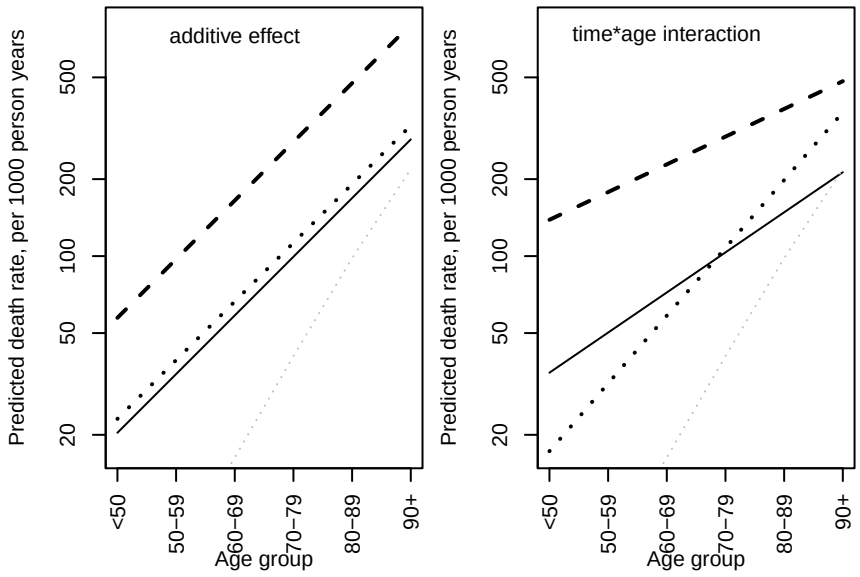

Figure 2.10 *Absolute risk estimates from two Poisson models that include age group and time since enrollment for a male with a median hgb value of 13.5. The two panels show predictions from different models with and without an interaction, and the lines show different time periods (dashed= < 5 months, solid= 5-24 months, dotted=24+ months). The thin dotted line in each panel shows Minnesota 2000 death rates.*

tice this step can often consume over half of the analysis time. Creating a *correct* dataset is the single most important step for valid results. Once this is done, the model fitting is straightforward.

Assume that the final dataset is `mdata2`, `tgroup` is coded as 1–3 and `agegrp` as 1–6. R code and results for the three different models is shown below.

```
> pfit1 <- glm(death ~ offset(log(time)) + factor(agegrp) + factor(tgroup) +
                  sex + hgb, family=poisson, data= mdata2)
> pfit2 <- glm(death ~ offset(log(time)) + agegrp + factor(tgroup) +
                  sex + hgb, family=poisson, data= mdata2)
> pfit3 <- glm(death ~ offset(log(time)) + agegrp * factor(tgroup) +
                  sex + hgb, family=poisson, data= mdata2)
```

```
       male se(male)    hgb se(hgb)
pfit1 0.481    0.067 -0.148   0.017
pfit2 0.479    0.067 -0.150   0.017
pfit3 0.486    0.067 -0.151   0.017
```

The first fit treats the age group as a categorical variable, the second fit treats the age group as a continuous variable, and the third fit adds an

interaction between continuous age group and the time since diagnosis. In all 3 models, time since diagnosis is treated as categorical. The table just below the code shows the estimated coefficients and standard errors for the covariates of interest, while Figure 2.10 shows the absolute risk estimates from the model for a male with the median hgb value of 13.5. The example and results illustrate 3 points:

- Once the data is set up, it is easy to fit a selection of models. The code shown is for R, but it is equally simple in other packages.
- Examination of the baseline rates can reveal interesting patterns. The figure shows that death rates for this cohort remain higher than the underlying Minnesota population across time, with the biggest effect in the time period from 0-5 months after diagnosis. The interaction model shows that 24+ months after diagnosis the excess is fairly constant across age groups, while the excess early mortality is greatest for younger subjects. The additive model treats these as parallel lines; a likelihood ratio test comparing the two models shows that it is substantially inferior (χ^2=19.0 on 2 degrees of freedom).
- Coefficients for the other covariates, however, are almost completely unaffected when the 'wrong' age/time baseline rates are used. Estimated absolute hazards may change, but relative hazards are remarkably robust.

A disadvantage of the Poisson approach is that the time scales must each be divided into discrete segments over which the baseline hazard will be approximately constant, i.e., the hazard is constant and the cumulative hazard is linear over each time segment. The advantage is that it is easy to look at interactions (e.g., is the first period particularly deadly for the oldest or youngest subjects) or at covariate by time interactions.

2.3 Accelerated Failure Time models

An accelerated failure time (AFT) model can be used to directly model the time-to-event. Parametric AFT models such as survreg in R or lifereg in SAS maximize a standard log-likelihood, with terms of

$f(t)$	an observed event time
$\int_t^\infty f(t)dt$	right censored at t
$\int_{-\infty}^t f(t)dt$	left censored at t
$\int_s^t f(t)dt$	interval censored between s and t

Advantages and disadvantages of AFT models relative to proportional hazard models are:

- Advantages

  - They naturally accommodate left, right, and interval censored data
  - A range of distributions f is available
  - Predictions can be extrapolated beyond the time range of the data
- Disadvantages
  - The models do not accommodate time-dependent covariates
  - A correct distribution f needs to be chosen

The models are normally parameterized as location-scale alternatives

$$\frac{g(t_i) - \eta_i}{\sigma} =\sim F$$
$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots$$

where $g(t)$ is either the identity or $\log(t)$. A good reference for the parameterization is chapter 2 of Kalbfleisch and Prentice [54], which also introduces the common distributions for f of Gaussian, log-normal, logistic, log-logistic, Weibull, and exponential.

For survival data the transformation $g(t) = \log(t)$ is nearly always chosen; this transforms survival times, which must be > 0, to the entire real line from negative to positive infinity, and removes constraints on the range of the linear predictor $X\beta$. Table 2.5 compares the coefficients of a Cox model, Weibull, and log-logistic fit, using the lung cancer data. We use age in decades to make the coefficients more similar in size.

	Intercept	age(decades)	female	PS	(scale)
Cox		0.111	-0.553	0.464	
Weibull	6.273	-0.075	0.401	-0.340	0.731
log-logistic	5.937	-0.081	0.487	-0.405	0.536

Table 2.5 *Comparison of models of the lung cancer data predicting survival using the covariates age, sex, and ECOG performance score (ps). The models are: Cox proportional hazards, accelerated failure time (AFT) using a Weibull distribution and AFT using the log-logistic distribution.*

The first thing to notice is that the coefficients have all changed signs: in an AFT model a positive coefficient translates to a longer lifetime, while in a hazard model a positive coefficient predicts a higher death rate, what is 'good' in one model is 'bad' in another. Coefficients from the Weibull and log-logistic models are similar. For the AFT models, the intercept and scale parameters capture the underlying shape of the predicted survival curve, whereas for the Cox model this is contained in the baseline hazard $\lambda_0(t)$. Figure 2.11 shows the predicted survival curves for a 'low risk' (age 50, female, performance score 0) and a 'high risk' (age 70, male, performance score 2) subject under all 3 models. When the underlying survival pattern fits well to a Weibull shape, as is the case here, then it will also be true that $-\beta/\sigma$ from the Weibull model will closely approximate the Cox model coefficients.

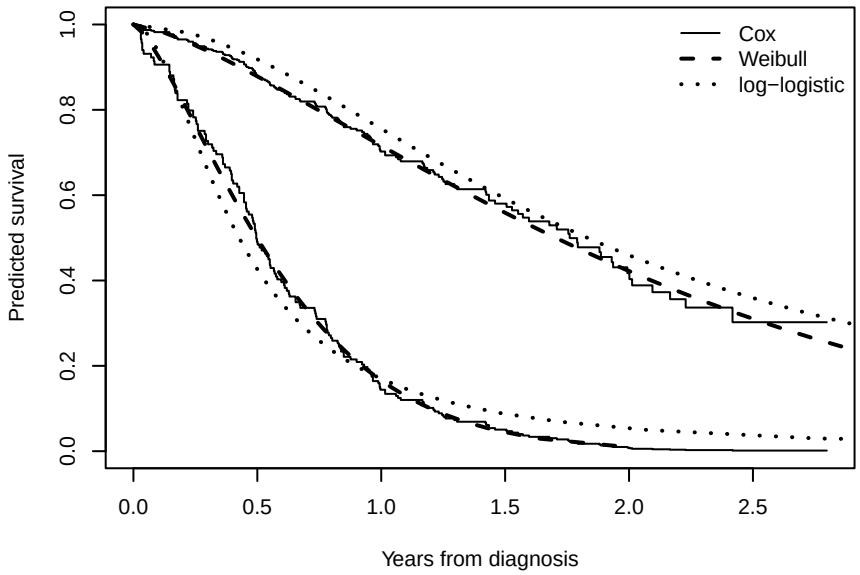

Figure 2.11 *Predicted survival curves from 3 different fits to the lung cancer data for 2 subjects: a 'low risk' (age 50 years, female, performance score 0) and a 'high risk' (age 70 years, male, performance score 2).*

The figure also illustrates the primary weakness of parametric AFT models, which is that one needs to choose an appropriate distribution. Coefficients and their standard errors are normally robust to a bad choice, but as seen here, predicted values can be biased. An alternate approach is to fit the semi-parametric AFT model, which allows for a general survival shape. The solution of this model has turned out to be a computationally challenging problem, resulting in several potential estimators. The model is still not available in the majority of statistical packages and as such has not gained significant traction in practical analyses. An example using the R `aftgee` package is included in the companion examples document.

Non-survival uses

Censored data models have several uses outside of the survival context as well. One well known case is *tobit* regression, which is an ordinary Gaussian linear model using data that has been left censored at 0. The model is quite popular in econometrics, however its relationship to survival data and/or the use of standard survival programs to fit the model is often unrecognized. Here is simple code using the original data from Tobin's 1958 paper [108].

```
> tfit <- survreg(Surv(durable, durable>0, type='left') ~age + quant,
                  data=tobin, dist='gaussian')
```

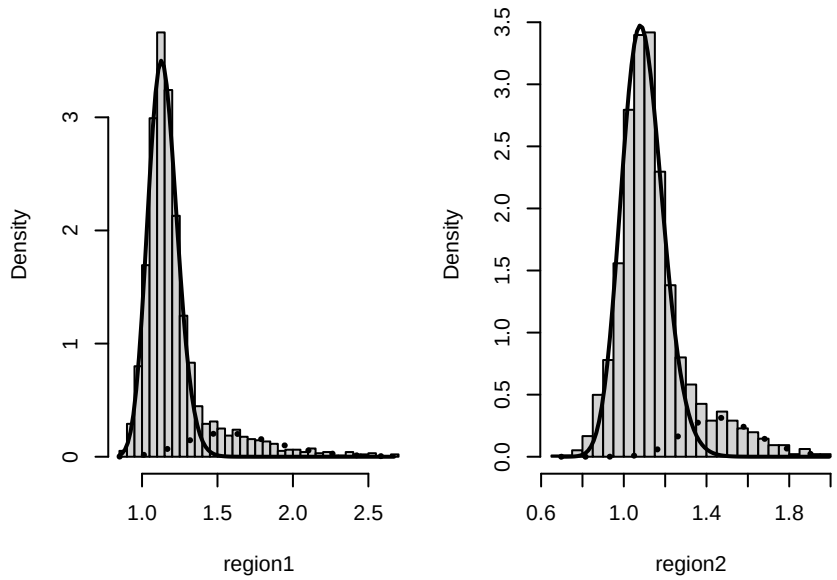

Figure 2.12 *Two component log-normal mixture model, fit to data from two different brain regions.*

A second example of censored linear regression might be the analysis of a laboratory test whose assay has a restricted range. Thus, we might only know that $y < 100$, say, or $y > 450$.

Move this next example to the companion?

A final example is the use of artificial censoring in fitting a mixture model. This example emphasizes that although we usually think of censoring as a shortcoming in the data that has to be worked around, censoring can also be an analysis tool. The raw data in this case was a set of measurements of the amount of tau protein in 86 brain regions defined by an atlas, from 1925 PET scans of participants in the Mayo Clinic Study of Aging (MCSA). The MCSA is a population-based cohort that contains an age- and sex- stratified random sample of subjects from Olmsted County, Minnesota. The particular analysis in question involved fitting a two-component log-normal mixture model to each of the regions, separately. Figure 2.12 shows the data and fitted distributions for two of the regions.

The left panel shows a fit that looks as expected: the larger leftmost peak corresponds to participants without tau deposition in that region, with a smaller and broader peak to the right containing those who have begun to accumulate the abnormal protein. Using a log-normal mixture helps to accommodate the (expected) right skew in the second peak. In order to deal with differences in per-subject bioavailability of the tracer, the standard preprocessing pipeline has scaled each participant's values by the reading from a

brain region known to have only non-specific binding, thus the lower threshold near 1 is also as expected. Nearly all the regions fit this pattern, with the fraction of subjects in the 'abnormal' peak ranging from 0 to 15%; it is well known that abnormal accumulation of tau starts earlier in some brain regions than others.

The right panel in Figure 2.12 shows one of the regions where the fit was not satisfactory. The resulting posterior probability of being in the 'abnormal' peak has the counterintuitive prediction that those with abnormally high *or* low measurements have tauopathy; the total fraction estimated to be in the second peak is also too large to be credible. The problem appears to be the subset of subjects with very low values, less than .85 or thereabouts. The stretched left hand tail does not fit neatly into the overall distribution of normals, and it is left to the other component to pick up the slack by becoming overly wide.

There are several possible approaches to address the issue. The crudest is to simply remove all values less than some cutoff, .85 say, but the result from this process will then systematically underestimate the fraction of normals. Another is to add a third mixture to the estimate, though there is no guarantee that it will align where we want. A third is to recognize that although it is important to accurately estimate the *fraction* of subjects with measured values < 1, for the purpose of the study we do not really care about the shape of the lower distribution. This leads to the using a censored mixture distribution: values < 1 are treated as left-censored at 1. The underlying structure of the EM algorithm remains unchanged, with an AFT model estimate of a censored log-normal replacing the usual log-normal parameter estimates. Full code is found in the companion data, it is quite simple and short. Figure 2.13 shows the original and censored results. For the first region the results are essentially unchanged, while for the second region the censored regression result is better aligned with our biological expectations. The censoring has relaxed the likelihood's need to assign positive probability at the leftmost (outlying) values. To reprise the earlier comment, censoring can sometime be a useful tool in our arsenal. Artificial left truncation of follow-up time is another example.

2.4 Binomial models

A simple binomial model is perhaps the least satisfactory way to proceed with time-to-event data. A primary reason is that it is a very inefficient use of the information at hand. To quote an early mentor "a single yes/no is the least possible information that can be measured on a patient", and in clinical research patients are normally our scarcest resource. Due to the familiarity of logistic regression, this approach is often taken, however, but even beyond the loss of information there are pitfalls.

The simplest (and least satisfactory) approach is to simply ignore censoring. As an extreme example consider the MGUS data, which contains follow-up on all 241 subjects who had been diagnosed with a particular blood abnormal-

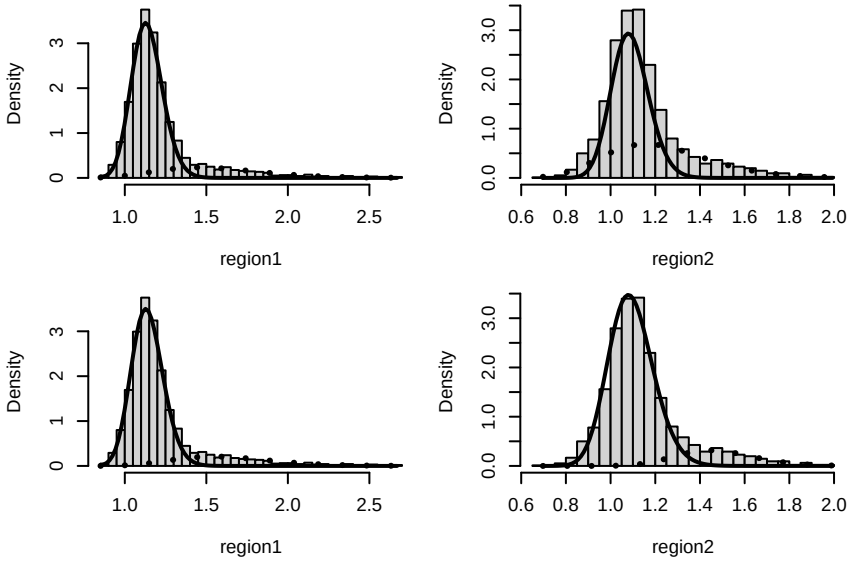

Figure 2.13 *Two component log-normal mixture models, fit to data from two different brain regions. The top panels use a standard log-normal mixture, the bottom panels artificially left censor the data at 1.*

ity over a 10 year period, who are then followed forward in time. This dataset has extremely good follow-up: only 16 of the 241 subjects are censored, all between 31 and 40 years after the initial lab test (all 16 were still under active follow-up when the dataset was created). A hazard or AFT model will address the question of what factors most influence survival time, whereas a binomial model tries to predict which 16 subjects will be censored. This last will be related to survival probability, of course, but is also influenced by whether a subject's test was early or late in the 10 year window. The binomial fit also has much lower power. Power for a hazard model is related to the total number of events, while for a binomial model it is related to the smaller of the event and non-event count.

A somewhat less naive method, but still incorrect, is to choose a targeted time point, e.g., one can use "1 year survival" for the lung cancer data. Of the 228 subjects 121 have died by one year and 65 lived longer than one year. But what is to be done with the 42 whose last status is alive, but at less than one year? Omitting these subjects is easily seen to lead to bias: the naive survival estimate of $65/(165 + 65) = .35$ is substantially below the Kaplan-Meier's 1 year estimate of .41. This overall bias in the mean has long been recognized, e.g., see the 1952 paper by Berkson and Gage [12].

It turns out that all of the coefficients, not just the intercept, will be unreliable using this approach. A common form of this error is found in studies

of a new biomarker or test. One group, named "poor prognosis" is composed of patients who fail (death, recurrence, etc.) within a short period of follow-up, 1 year say. Another group labeled "good prognosis" is composed of those with long-term success, no failure within 3 years, say. Simon [29] points out that the majority of microarray studies take this approach and discusses the shortcomings.

For the lung example above, one solution to the issue of censoring before 1 year is to use redistribute-to-the-right (RTTR) weights. The 42 subjects censored before 1 year are given a weight of 0, as with the naive method, but in this case each censored subject's weight is redistributed to all those at risk at the time of their censoring. The death at day 165 gains case weight from the two subjects censored before 165, the death at day 201 from the 12 censorings before 201, etc. The RTTR algorithm was introduced in Section **??** as a way to explain how the Kaplan-Meier is constructed.

A remaining issue is which timepoint to use: 6 month survival, one year, two years? The joint use of multiple cutpoints will be closely related to ordinal logistic regression. In Appendix A.7 we show that when taken to its limit, binomial models based jointly on multiple timepoints become an approximation to the Cox model. The association between Cox models and continuation ratio binomial models is even stronger, see for instance Chapter 13 of [44]. In Chapter 7 we will explore a more useful approach based on pseudovalues.

Summary

In summary, many different approaches can be used to model time-to-event data. The Cox model is by far the most commonly used, but is clearly not the only possible choice. In addition, the flexibility of the Cox model allows additional modeling choices, such as alternate time scales. Model diagnostics are important and will be discussed in the next chapter.

Chapter 3

Model assessment

> "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful." George E.P. Box

As stated in Chapter 2, one of the goals when building a model, perhaps the most important one, is simplification.

3.1 Model checks

An important first step in examining a model is to critically examine the assumptions. For a standard Cox model

$$\lambda(t) = \exp(\beta_0(t) + \beta_1 x_1 + \beta_2 x_2 + \ldots)$$

we have assumptions of

1. Functional form: each covariate x is included using a simple linear form.
2. Additivity: each covariate stands alone in the linear predictors, not relying on other covariates.
3. Proportional hazards: the coefficients β are constant over time: if a covariate x_1 doubles the hazard of an event on day 10, it also doubles it on day 100.
4. Stability: no single observation, or small set, has undue influence on the results.

Nearly the same list will hold for an accelerated failure time, Poisson, additive hazards or binomial model as well, though we will mostly use Cox models for illustration.

Model checks always bring to mind residuals. There are four major types of residuals from a Cox model, i.e., martingale, deviance, score and Schoenfeld, and two others that are derived from these: dfbeta and scaled Schoenfeld residuals. Martingale residuals have been used to examine functional form, Schoenfeld and scaled Schoenfeld to examine proportional hazards, and dfbeta for influence. For accelerated failure time models the `residuals` function in R can return 9 different types of residuals, motivated by their use in a well known textbook on reliability methods [72]. However, with modern computing we can also employ more direct assessments, and we find these easier to interpret.

3.1.1 Functional form

For a continuous covariate, the Cox model assumes the ratio of hazards (hazard ratio or HR) is constant for similar increments in the continuous variable. For instance, for age as a covariate the HR between a 30- and 40-year-old is assumed to be the same as the HR between a 70- and 80-year old. However, in biological data both upper and lower threshold effects are common. For example, the model for end-stage liver disease (MELD), which is used in the US to score subjects waiting for a liver transplant, each rise in serum creatinine over a value of 1.0 increases the risk score, but values below 1.0 do not decrease it, as there is no medical advantage to a value that is less than is usual for a normal (undiseased) individual. The risk of a particular outcome might not begin to rise until age 50 years, or even decline after age 80 years. Thus investigation of the proper functional form is important, rather than blindly assuming the loglinear relationship inherent in the default specification of the Cox or AFT model is correct.

There are several methods to check for appropriate functional form, i.e., non-linearity of the effect for one or more predictors. Some early methods involve using martingale residuals. For example, Therneau et al. [104] suggested plotting martingale residuals from a null model against each covariate separately along with a scatterplot smoother line. This plot is similar to the y vs. x scatterplots used in linear models. Of course, the choice of smoother function and its tuning parameters to control 'wiggle' can have an impact on the visual assessment of non-linearity.

The martingale residual is O - E, where O is the observed number of events (usually 0 or 1 for each subject) and E is the expected number of events. For datasets with few events and heavy censoring, the martingale residuals can be almost binary with a large proportion of values near zero. This can make it difficult for smoothers to detect non-linear effects.

Correlated variables also present difficulties when assessing non-linear effects using residuals. One common idea that does not work is to plot the residuals from a model with all variables except the variable of interest against the variable of interest. This approach does not improve upon the simple plot using residuals from a null model, and may introduce additional visual artifacts. In the linear models setting, an adjusted variable plot can be used in this instance: let y, x, and z be the response, variable of interest, and adjusting covariate(s), respectively. Then plot the residual of y on z versus the residual of x on z. However, extending this approach to survival models is not straightforward. Thus, experience has revealed that martingale residual plots can be misleading, in particular when there is a high level of censoring and/or strong correlations between predictors.

The direct fit of a smooth curve within the Cox model fit is much more resistant to these effects, and is easy to accomplish with current software. That is, our linear predictor becomes $g_1(x_1)+g_2(x_2)+\ldots$, where each g is a smooth function. (Such generalized additive models have supplanted residual plots in

simple linear models as well.) There are of course many possible smooth curves that could be fit, which include:

- Regression splines. These are fit by creating special predictor covariates. These methods are particularly easy since it is "just an X matrix", and so they work with any model: linear, logistic, accelerated failure, or hazard models.
  A spline is a thin flexible strip of metal or wood, traditionally used to create the smooth curve for laying out the hull of a boat, an airplane wing, or other smooth object. A set of pins are fastened to a board and the spline threaded through them. Being a physical object, the resulting curve has derivatives of all orders. Numeric splines used in fitting almost always have derivatives of order 3 ("cubic splines") since that is sufficient for the eye to see them as smoothly continuous.
  The "knots" of a regression spline are the x-axis locations of the set of pins, while the y-axis location of each pin is chosen by the fitting routine so as to maximize the fit. A primary choice when fitting a spline is the number of knots, trading off flexibility and stability. Generally 3 or 4 knots are sufficient. A disadvantage of regression splines is the arbitrary choice of the knot locations. Changing the location of the knots can change the resulting smooth curve. For sudden features, like a change-point, knot locations can enhance or mask the feature. However, smoother transitions are often moderately insensitive to the knot locations. A common choice for the knots are quantiles of the data.
- "Restricted" cubic splines or "natural" splines. A physical spline will be linear past the last knot. Natural splines are cubic splines with this extra restriction added.
- Smoothing splines. These have certain statistical advantages over regression splines, though the gains are often small. One advantage is that the user only needs to specify the number of degrees of freedom. A disadvantage is that they must be built into the particular fitting routine. For instance, in the R survival library the Cox model function supports smoothing splines while the function for accelerated failure time models does not.
- Polynomial. It is easy to add polynomial terms of bili, bili^2, bili^3, etc. However, it has been repeatedly shown that polynomial terms can be unstable.
- Fractional polynomials. One of the critiques of splines is that the resulting function is difficult to transport from one software package to another, as the basis functions used by different routines are not consistent. Fractional polynomials pick powers from a limited set (-2, -1, -.5, 0, .5, 1, 2, 3), where x^0 is understood to be log(x). Two terms often suffice and can be found with a simple automated search.

In general, the smoothing spline may be preferred because it is the easiest for the user to specify. Capabilities built-in to the Cox model function in the R survival library make this the easiest option to implement. The output also provides tests for linear and non-linear effects, which simplify interpretation.

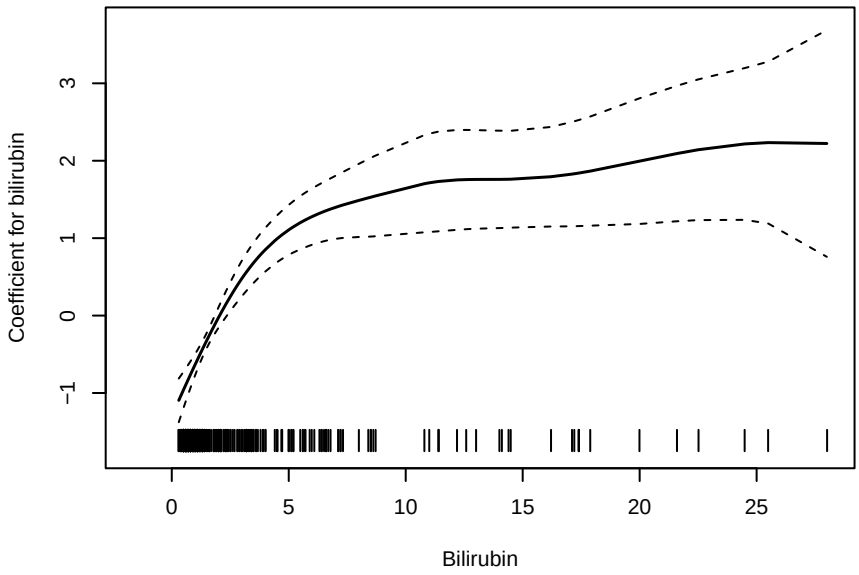

Figure 3.1 *A spline fit to investigate the functional form of bilirubin.*

Below is an example investigating the effect of bilirubin on survival in patients with primary biliary cirrhosis (PBC). Implementing a penalized smoothing spline is as simple as adding `pspline` as a function call surrounding a single variable in the model. The basic print of the model shows tests for the linear effect and for the non-linear effect. Note that the degrees of freedom for the non-linear effect defaults to approximate 3, which is often adequate, but this can be modified. If a `summary` of the model is requested, then the lower portion of the output shows 12 coefficients for the spline in the Cox model. These coefficients correspond to each of the basis functions in the spline and can be difficult to interpret. A plot is the easiest way to visualize the non-linear effect.

```
> coxph(Surv(time, status==2) ~ age + ascites + edema +
        pspline(bili), data=pbc)

                       coef     se2 Chisq   DF       p
age                   0.031 0.00863 12.73 1.00 3.60e-04
ascites               0.823 0.28270  8.33 1.00 3.90e-03
edema                 1.282 0.29401 18.90 1.00 1.38e-05
pspline(bili), linear 0.125 0.01751 50.29 1.00 1.33e-12
pspline(bili), nonlin               41.14 3.04 6.49e-09
```

Figure 3.1 shows the result of using the smoothing spline to model the functional form of bilirubin adjusting for age, ascites and edema. One major

advantage of the direct fit is that confidence intervals are easily obtained for the curve, which in this case helps to show that the curvature is significant. The default plotting method for non-linear terms, used in Figure 3.1, subtracts out the mean prediction. Thus the $\hat{y}$ values of the plot sum to zero. (If a predictor x appears multiple times, the point is only plotted once, so the average of the plotted points may be non-zero.) Since the Cox model is a relative hazard model, we can choose whatever center we like. Figure 3.2 shows the same plot using bilirubin=1.2 (the upper limit of normal) as the reference, i.e., the hazard ratio is 1 for someone with a bilirubin value of 1.2. This plot is more interpretable. The figure uses a log scale for the y-axis, which is appropriate for the question of whether a non-linear bilirubin effect was even necessary for this dataset. In this example, the x-axis (bilirubin) is also presented on the log scale. The plot is nearly linear, showing that log(bili) is a good choice, with some attenuation of the slope at the lowest values: having a "better than normal" bilirubin does not reduce risk as much as a high bilirubin increases risk.

Of note, the `pspline` function in the survival package has several arguments to control the line "wiggle", including the desired degrees of freedom (`df=4`), the number of basis functions (default, `nterm=2.5*df`), and the degree of splines (`degree=3`). There are times when it is desirable to have a certain amount of wiggle in the figure, but retain a monotonic curve. Optional parameters for the `pspline` function allow the user to retain that monotonic relationship. See the `splines.pdf` vignette in the survival package for more details.

3.1.2 Interactions

A second question, often thornier to answer than functional form, is whether the effect of each variable can be treated as a simple sum $\beta_1 x_1 + \beta_2 x_2 + \cdots$. In other words, does the effect of a 1 unit change in x_1 depend on the concurrent level of x_2? This is often referred to as an interaction, and assessing interactions in a Cox model is similar to assessing them in other types of models.

1. If `x1` and `x2` are categorical predictors, then it suffices to replace `x1 + x2` by `x1 * x2` in the R model formula which automatically is expanded to `x1 + x2 + x1:x2`. Other software packages will have different syntax, but the same idea. For instance, if `x1` has 3 levels and `x2` has 4 levels, the additive model will have (3-1) + (4-1) = 5 coefficients for the two variables (plus one for the intercept) and the interaction model 3*4 - 1 = 11 coefficients, one for each possible combination of `x1` and `x2`. This large increase in the number of coefficients, and correspondingly the number of degrees of freedom can be problematic for datasets with small sample size. In addition, interpretation of this many coefficients can be complicated, and interesting sub-structures may be missed.
2. If `x1` is categorical with 3 levels and x2 is continuous then the model `x1`

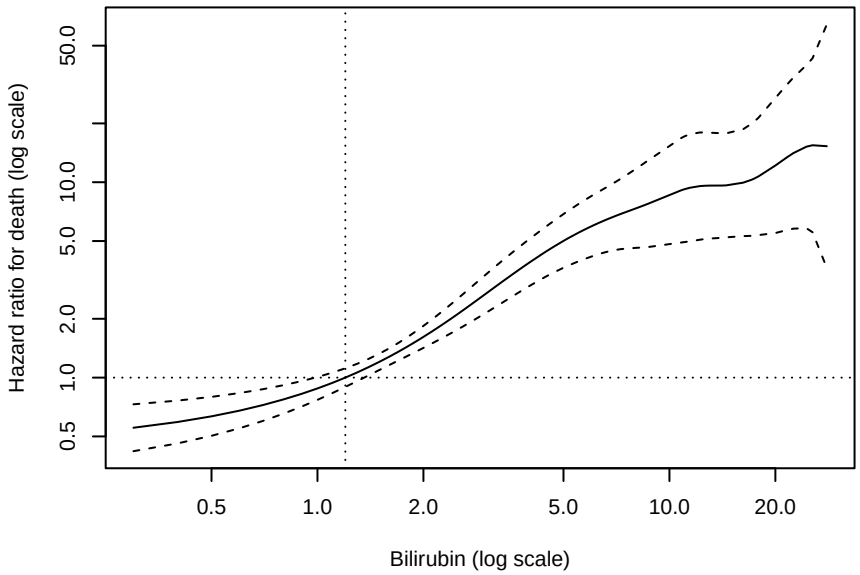

Figure 3.2 *A spline fit to investigate the functional form of bilirubin after centering the results at bilirubin=1.2.*

`+ x2` will have $2 + 1 = 3$ coefficients, while `x1*x2` will have $2 + 3 = 5$ coefficients, including a separate `x2` coefficient for each level of `x1`. This may be the simplest interaction to interpret. However, this approach defaults to linear effects of the continuous variable, and non-linear effects should also be considered.

3. If both are continuous then the `x1 + x2` model has 2 coefficients and the `x1*x2` model has 3 coefficients, adding only the product of the two variables to the model. This simple multiplication represents only one of the ways that continuous variables might have an interactive effect.

In the categorical-continuous situation there are two ways to parameterize the covariates. The first and perhaps more common way is to have a coefficient for the slope of `x2`, for the changes from the first level of `x1`, along with *change* in slope between the first level of `x1` and each of the other levels of `x1`. This way lends itself to a formal test of whether A and B are different, but is not always easy to interpret. The second approach is to estimate separate `x2` coefficients for each level of `x1`, which is often easier to understand. In R and most other packages the default action of * is to code in the first form. A handy alternative in R is `x1/x2` which leads to the second form. Coding "dummy" variables "by hand" is the most general approach, which can be used to obtain custom parameterizations and comparisons. These three approaches are illustrated in the code below. Note the use of the `I()` function which indicates to R that the (*) should indicate multiplication and not an interaction. The second

and third models produce identical results, but the third approach allows for more customization. Similar coding may also be useful when there are two categorical variables with multiple levels.

```
> coxph(Surv(time, status==2) ~ sex * log(bili), data=pbc)
> coxph(Surv(time, status==2) ~ sex/log(bili), data=pbc)
> coxph(Surv(time, status==2) ~ sex + I(log(bili)* (sex=='m')) +
                                      I(log(bili)* (sex=='f')), data=pbc)
```

The fact that the underlying code does three different things for these different types of interactions, yet the user types the same symbol (*) is the source of much confusion. Cases 1 and 2 can be truthfully labeled as "interaction", i.e., the effect of `x2` depends on the level of `x1`. However, case 3 is simple multiplication and does not fully explore all the ways that continuous variables might have an interactive effect.

Often interactions are assessed with a goal of "checking the box" to state that they have been considered and were not found, in order to strengthen the message of the simple story. This is not ideal, as the intent should be to understand the data rather than to keep the story simple at all costs. It is important to remember that statistical tests and p-values are not sufficient. Graphs can be very useful to help assess whether potential interactions between covariates exist or are meaningful.

3.1.3 Proportional hazards

The Cox model assumes that each coefficient is constant over time, which is known as the proportional hazards (PH) assumption. That is, if a particular factor doubles your risk on day 1, it also doubles it on day 100 or day 1000. This corresponds to survival curves that cannot cross and which spread apart in a very controlled way. First, be clear that on a long enough time scale the PH assumption is *never* true. The question we need to assess is whether, for a given dataset and time frame, representation of a covariate's effect with a single coefficient value is a defensible simplification.

Tests for proportional hazards can be based on either Schoenfeld or martingale residuals, both are readily available in packages. If there is evidence for important non-PH, a more careful assessment of the shape of non-proportionality can be achieved by explicitly fitting a model with time-dependent coefficient(s) $\beta(t)$ for the covariate(s) in question. This topic is explored more in chapter 4.

At each event time, the Cox model compares the observation which failed to all others in the risk set for failure. The Schoenfeld residual is defined as

$$x_d(t) - \overline{x}(t)$$

the difference between the covariate vector of the failure and a weighted mean of those at risk, with weights of $\exp(\hat{\beta}X)$. These were suggested as a check of proportional hazards by Schoenfeld [92]. Grambsch and Therneau [42] showed

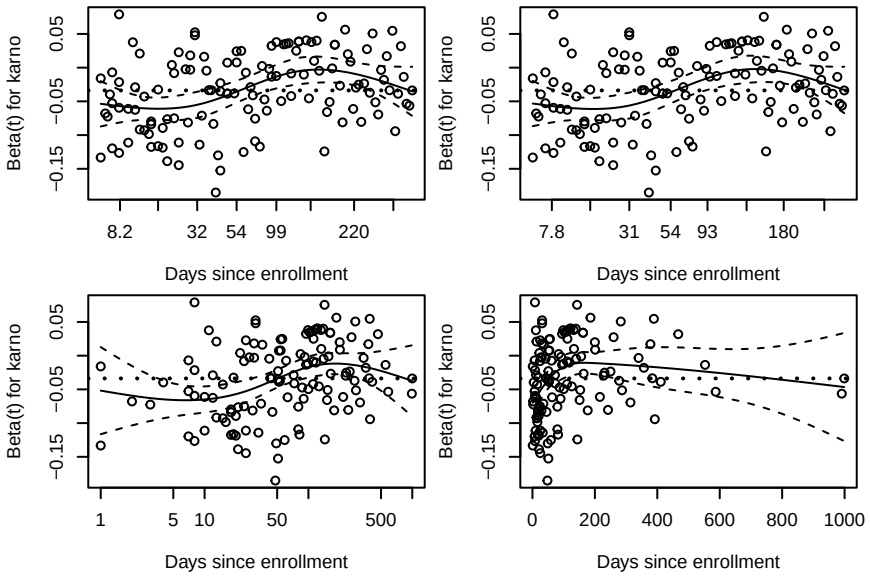

Figure 3.3 *This figure depicts 4 plots to examine the proportional hazards assumption using different time transformations. Clockwise from the upper left are the KM, rank, identity, and log transforms. The dotted horizontal line is placed at the coefficient estimate from the Cox model. The strength of the* `karno` *variable is higher at the start of the follow-up and then decreases.*

that if the covariates are correlated, it is important to scale the residuals by the inverse of the variance-covariance matrix of X, leading to scaled Schoenfeld residuals. If $\beta(t) = a + bg(t)$ for some pre-specified function g, then a plot of the scaled residuals versus $g(t)$ will be approximately linear with slope b, and a test for $b = 0$ can be used as a test for PH. It is important to accompany the test with a plot, as outliers in either the covariate or the time values can have a large impact. Figure 3.3 shows an example using the `veteran` dataset, which demonstrates potential issues with the Karnofsky score. Karnofsky score is a very important predictor, but its effect is not constant over time. Early on it has a large negative effect; the risk of someone at the first quartile of `karno` is approximately exp((75 - 40)*.03377) = 3.2 times that of someone at the third quartile, but by 100 days this effect has waned and is not much different from zero. In this example, the p-value for a non-proportional effect of `karno` is $< .001$ for the first three time transforms and .03 for the "identity" transformation.

An advantage of this approach is its speed and simplicity, a score test for $b = 0$ can be computed very efficiently, e.g., the `cox.zph` function in R. The disadvantage is that a particular time-transform g must be pre-specified: should it be $g(t) = t$, $\log(t)$, $\log(1 + t)$, or something else? A correct or near correct choice will have high power, a bad choice may miss important non-

proportionality. Common choices for the time transform are $g(t) = t$, $\log(t)$, rank(t), and KM(t). The last is the transform of time such that the Kaplan-Meier will be a straight line, and is in essence a variant of rank(t) that is less affected by censoring. In the R survival package the KM transform is the default, purely for safety reasons: the resulting x-axis of the plot never contains outliers. However, the resulting plot itself is harder to interpret than one using t or $\log(t)$.

The cumulative sum of martingale residual approach of Lin, Wei, and Zing [65] is based the fact that if the model is correct, then

$$U(t) = \sum_i \int_0^t [x_i(s) - \overline{x}(s)] dM_i(s)$$

will be a Brownian bridge. To test this, replace the observed martingale process $dM_i(s)$ with a random mean zero martingale process, multiple times, and compare the observed trace $B(t)$ to the family of random traces, both visually and using a formal test, such as the Kolmogorov-Smirnov or Anderson-Darling. Since the formal test is based on the maximum deviation, which is invariant to time transforms, the user does not need to make a choice of $g(t)$, making this a more robust choice. Figure 3.4 shows this approach for Karnofsky score in the veteran dataset.

Disadvantages of the test are

1. It will in general be less powerful than the score test. (This may be an advantage if one's goal is to keep the analysis simple by not detecting non-PH.)
2. The test operates on each coefficient separately, making the assessment of a categorical variable more complex.
3. When there is non-PH, the plot is harder to assess with respect to the actual shape of the non-PH effect.

3.1.4 How to deal with non-proportionality

Here are several possible solutions for dealing with non-proportional hazards:

- Check functional form - often there is a relationship between these concepts. This relationship was demonstrated by Abrahamowicz and Mackenzie, who proposed a product model that simultaneously relaxes the assumptions of PH and linearity for a continuous predictor [1].
- Stratify the analysis (e.g., the non-proportionality may be caused by different center/group effects).
- Partition the time axis into early/late effects (e.g., 1st year is different).
- Model non-proportionality by time-varying covariates (e.g., age + age*time). More details on this approach will be discussed in Section 4.6.
- Check whether there are outliers influencing the non-proportionality.
- Use a different model (e.g., accelerated failure time model)

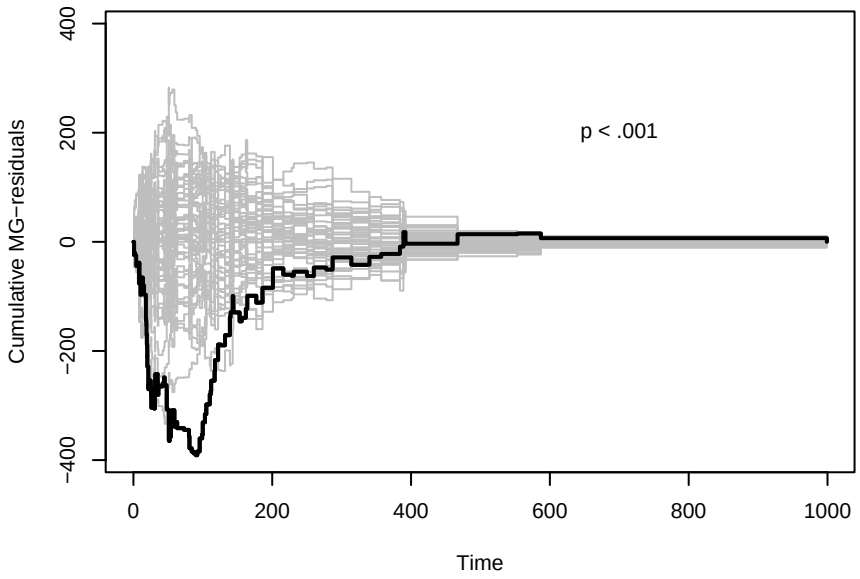

Figure 3.4 *Cumulative sums of martingale residuals plot for the Karnofsky effect, in the veteran data. Solid line = observed trace for $B(t)$, lighter lines = 50 random realizations.*

3.1.5 Influential points

Outliers and high leverage points are often overlooked in time-to-event analyses. A common type of outlier is the rare patient who miraculously lives much longer than all the others in the cohort of patients with a certain disease. In the real world, this type of outlier occurs more often than one might expect, but always warrants checking for accuracy. Another type of outlier involves extreme covariate values (e.g., age $>$ 100 years). The influence of each data point is related to whether it is an extreme value or not; a point has to be both far from the mean `x` value and have a large residual to impart a significant influence. Dfbeta residuals offer one way to check for influential points. These residuals estimate how much β would change if one point was deleted. Dfbetas are approximated using an infinitesimal jackknife approach. A limitation of this approximation is that it ignores the change in the variance associated with the observation that is removed. Because a large outlier will often increase the variance of the estimated coefficient, the dfbeta residual will be an underestimate of the influence for a point with large influence. This underestimation may be more pronounced for small datasets, since in larger datasets the influence of any single observation will likely be small. Figure 3.5 shows the dfbeta residuals for a 10 year increase in age plotted against the age values using the `lung` dataset. The figure shows 2 influential points, one on the upper left corner of the plot and the other on the lower right corner of the plot. Note

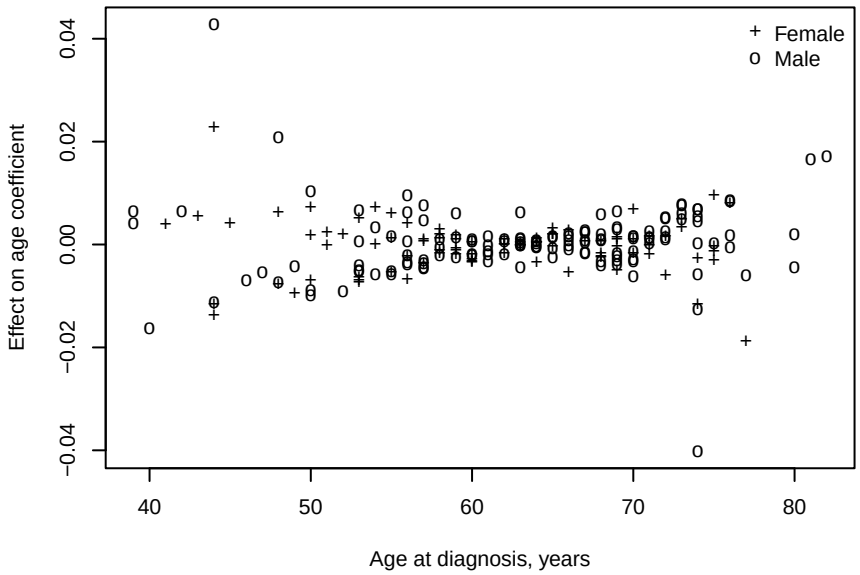

Figure 3.5 *This figure depicts the dfbeta residuals versus age for the lung dataset. These dfbeta values correspond to a 10 year increase in age, and result from a model that also includes sex and ECOG score. It shows 2 influential points, one on the upper left corner of the plot and the other on the lower right corner of the plot.*

that a positive dfbeta residual indicates that inclusion of that observation is associated with an increase in the coefficient. Concordantly, removal of an observation with a positive dfbeta residual would be associated with a decrease in the coefficient. Highly influential data points warrant double-checking, as they are often associated with data errors.

3.2 Model performance

When reporting linear regression models, it is common to report a general goodness-of-fit measure such as R^2, which is a measure of how well the model explains the observed variability of the endpoint. With the Cox model there are several different goodness-of-fit measures, each of which is affected in its own way by censoring and the measure's ability to accommodate time-dependent covariates.

With time-to-event data there are several natural goodness-of-fit targets that can be estimated.

1. The event rate (intensity or hazard).
2. Whether an endpoint happened or not (yes/no) by some prespecified time t. More generally we have the predicted survival curve $S_i(t)$ for any combination of predictors.

3. The time to an event.

Even though the hazard ratio is a primary focus of the Cox model, it turns out to not be a good focus for model assessment. (The same is true for the time acceleration $\exp(\beta)$ parameter in an accelerated failure time model.) A primary reason is that although the model provides a predicted rate $\hat{\lambda}_i(t)$ for each subject, there is no corresponding observed datum to compare it to. Time and status are observed values, but a per-subject rate is not observable. At a more theoretical level, it has been shown that hazards are not a *causal* parameter [48], which also makes them unsuitable at a more fundamentally level.

3.2.1 Brier score

The Brier score is a measure of accuracy similar to a mean squared error, which was designed for use with predicted probabilities. It has been adapted from the binomial probability setting to the time-to-event setting by using probabilities from survival functions in place of binomial probabilities. So in the time-to-event setting, the Brier score is a function of time. It can be computed at a pre-specified timepoint of interest or it can be computed at multiple timepoints and depicted using a figure.

The squared difference between the observed per-subject survival function (i.e., a step function indicating whether a subject has or has not experienced the event of interest at each point in time) and the predicted survival function is a measure of accuracy at any given timepoint. The average of this over all subjects is known as the Brier score for the model, defined as

$$B(t) = (1/n)\sum_{i=1}^{n}(\delta_i(t) - \hat{S}_i(t))^2$$

where $\delta_i(t)$ is the 0/1 indicator of an event by time t and $\hat{S}_i$ is the predicted survival function for each subject.

is delta an indicator of event as in the formula or an indicator of non-event as in the figure? Since we are comparing it to the survival function, I think it should be a non-event.

Figure 3.6 shows the individual and predicted curves for 4 subjects. Suppose we want to compute the Brier score at 6 years; how should the subject in the lower right panel of the plot, censored 4.8 years, appear in the sum? The usual approach is to employ the redistribute-to-the-right (RTTR) algorithm: beyond the point of censoring, the censored subject will be represented by the average Brier score of all those who were still under observation on the day the patient was censored. This is done by using a time-dependent case weight for each subject, and redistributing a censored observation's weight, at the point of censoring, onto all those who survive past that point. These other subjects represent the censored subject in the future. The RTTR algorithm was discussed in more detail in Section **??**.

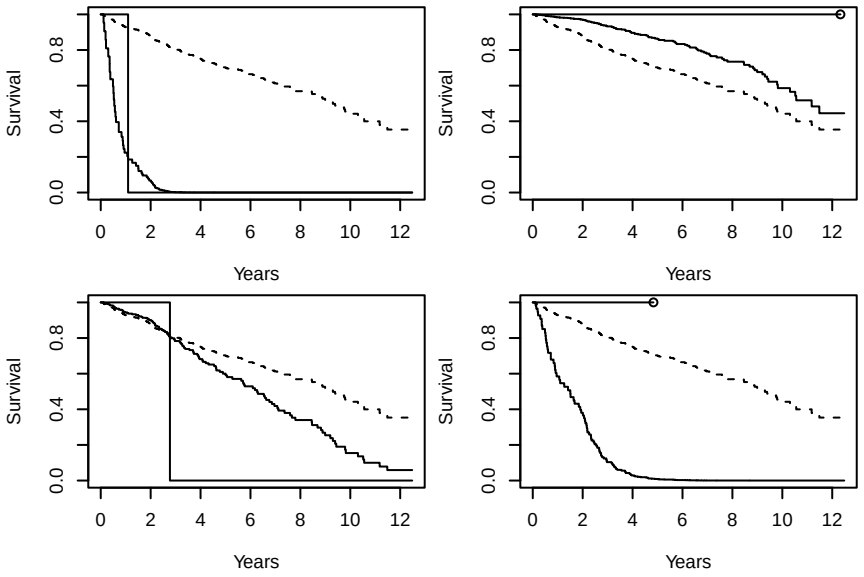

Figure 3.6 *Predicted and actual (step-function) curves for four individual patients using the PBC dataset. The predicted curves are based on a model including age, ascites, edema, and bilirubin with predictions specific to each patient's covariate values. The dashed line is the overall Kaplan-Meier for the dataset, which is the same for all 4 patients and can also be thought of as the NULL model.*

The Brier score under a null model is

$$B_0(t) = 1/n \sum_i (\delta_i(t) - S(t))^2$$

Where $S(t)$ is the overall survival for the dataset, ignoring covariates. If the ordinary Kaplan-Meier is used for the NULL model $S(t)$, then because it also can be computed via the redistribute-to-the-right algorithm, the Brier score at time t will be $B(t) = nS(1-S)^2 + n(1-S)S^2 = nS(1-S)$, which we recognizes as the variance of a binomial random variable. Proof: the sum of weights for those still alive will be nS, since KM(t) itself is the sum of (time-dependent) weights for those still alive at t, divided by n.

Figure 3.7 shows Brier scores for the PBC model and for the NULL model. The absolute value of the Brier score will be small when the survival curves are close to 0 or 1, and largest when they are near 1/2. Hence, it is more informative to look at a relative Brier score, i.e., the fraction of error that is 'explained' by the fitted model. If we think of the 0/1 survival status at a fixed time t as the 'response' y, it is easy to see that the Brier score is simply the residual sums of squares for $\hat{y}$, the NULL model score is the same quantity for a model with only an intercept, and the usual R^2 formula for this 0/1 outcome

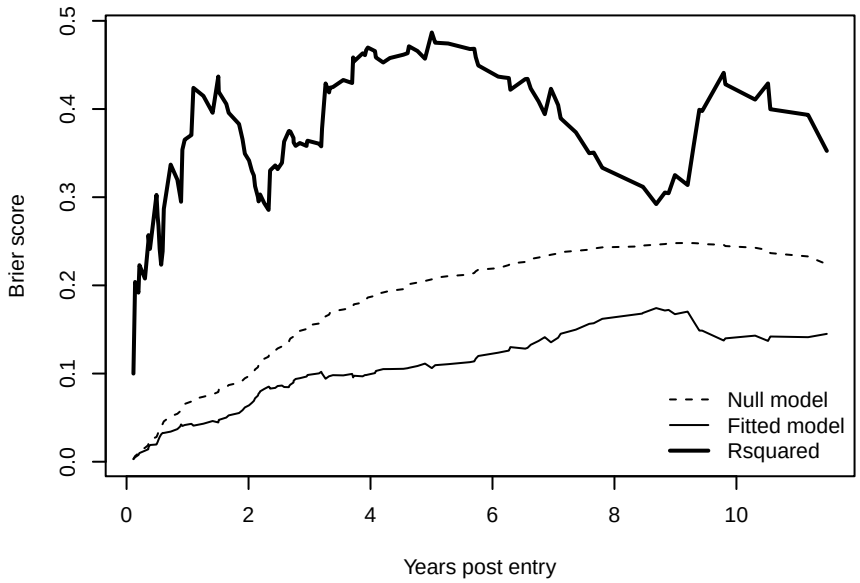

Figure 3.7 *Brier scores for the PBC data using the null model (dashed line) and a model with age, ascites, edema, and a spline term for bilirubin (solid line). The thick solid line shows a relative Brier score (or R-sq) obtained from the difference between Brier scores for the null and fitted models divided by Brier score for the null model.*

becomes

$$R^2_B(s) = \frac{B_0(s) - B(s)}{B_0(s)} \tag{3.1}$$

Kattan and Gerds [55] call this the index of prediction accuracy, while Van-Houwelingen and Putter [111] favor the simple difference, which is the numerator of (3.1). The latter also suggest that the integral of the difference, over time, could be used as a single number summary.

Either the Brier scores, the change between fitted and NULL, or the relative change can be used as a summary. This can be reported for a particular timepoint or points of interest, the integral over a region of interest, or the entire curve can be shown as has been done here.

3.2.2 R^2

Continuous time R^2 Definition of an R^2 statistic for the time variable is a difficult problem for censored data. The first issue is that the Cox model itself does not predict a mean time $\hat{t}_i$ for each subject, and a second issue is what to do with censored observations of the response. These could perhaps be approached by using the restricted mean survival time (RMST) as the target, along with the redistribute-to-the-right algorithm for censoring. A third issue, discussed by Korn and Simon [57], is the choice of scale. It is not clear that

an (observed, predicted) pair $(t_i, \hat{t}_i)$ of (4 months, 8 months) and one of (124 months, 128 months) should be treated as the same size of prediction error, and in fact for most medical questions the first would be considered a large mistake in prediction and the second a trivial one. For all three reasons, a direct analog of the linear model R^2 has received little attention.

For the proportional hazards model, an intuitive choice would be to use the martingale residual. However, the semiparametric nature of the model along with the absence of a scale parameter means that the average size of a residual does not change between a null model and one with strong covariates. Figure 3.8 shows the martingale residuals for a fit to the advanced lung cancer data for a NULL model and for fit using age, sex, and Karnofsky score (p<.001). The mean residual is 0 in both cases, of course, but the standard deviation actually increases slightly from .84 to .9 (the expected value is 1), and the overall shape of the residual distribution hardly changes. This is expected: under both H_0 and H_a the theoretical distribution of the martingale residuals is a centered exponential distribution, and the expected sum of squares is identical under both.

Almost all of the work in survival has focused on synthetic R^2 measures, which are based on various identities that hold for linear models. One of the simplest is the measure of Nagelkerke [75] which is based on the fact that the chi-square statistic T for the regression model can be written as $T = g(R^2, n)$. Inverting this gives $R^2 = f(T, n)$, and application to the proportional hazards model leads to

$$R^2_N = 1 - e^{2[PL(0) - PL(\hat{\beta})]/n}$$

This proposal has been found to have poor performance in the presence of censoring, however.

A large number of such statistics have been proposed. Overviews can be found in [84, 22, 23, 89]. Two measures that fare reasonably well in the reviews are R^2_{PM} and R^2_D [88]. Let $\eta = X\beta$ be the vector of linear predictors from the model. Kent and O'Quigley [50] use the linear model identity of

$$\begin{aligned} R^2 &= \text{var}(\hat{y})/\text{var}(y) \\ &= \frac{\text{var}(\hat{y})}{\hat{\sigma}^2 + \text{var}(\hat{y})} \end{aligned}$$

to create a Cox model value of

$$R^2_{PM} = \frac{\text{var}(\eta)}{\pi^2/6 + \text{var}(\eta)}$$

Royston and Sauerbrei [88] use a Gaussian smooth of the linear predictors:

- Define s as the vector of normal scores corresponding to rank(η); s will have mean 0 and variance 1.
- Refit the model using s as the single predictor. By definition the resulting coefficient β is the standard deviation of the revised linear predictor.

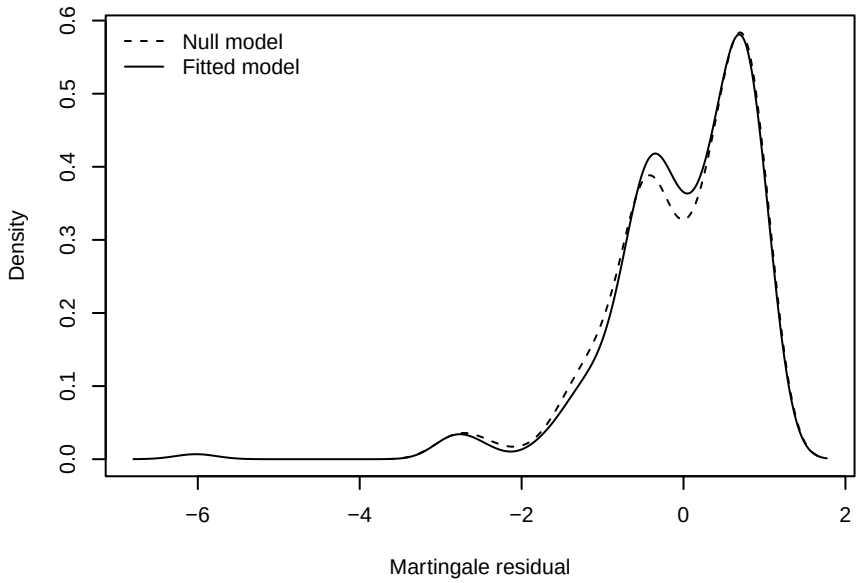

Figure 3.8 *Martingale residuals for a fit to the advanced lung cancer data for a NULL model and for a fit using age, sex, and Karnofsky score (p<.001).*

$$D \equiv \sqrt{8/\pi}\beta R_D^2 = \frac{\beta^2}{\pi^2/6 + \beta^2}$$

Using the lung data we have $R^2_{PM} = .065$ and $R^2_D = .003$. not calculated correctly

3.2.3 Concordance

Concordance is a measure of discrimination (i.e., can the model distinguish between low and high risk?). This is equivalent to assessing whether the patients are ranked correctly from low to high risk. For continuous data the concordance between two variables x and y is defined as $P(x_i > x_j | y_i > y_j)$, i.e., the probability that the two variables share the same ordering for some randomly chosen pair of subjects i and j. The method plays a particularly strong role in survival analysis because a model's intercept term, or equivalently the baseline hazard in a Cox model, is not required: the predicted survival probability will be larger for observation i than for observation j if and only if the linear predictors are ordered: $\eta_i(t) > \eta_j(t)$ implies $S_i(t) < S_j(t)$.

For continuous data, four common measurements of the concordance are the Kendall's τ_a and τ_b, the Goodman-Kruskal γ statistic, and Somers' d; they differ only in how ties are handled. The four measures each lie between -1 and 1, similar to R^2, while the concordance is defined on the range from 0

to 1, but transformation between the two scales is simple: $C = (d+1)/2$. The concordance is closely related to several other statistics:

- For binomial data the most popular measure of performance is the area under the receiver operating curve (AUC); AUC $= C = (d+1)/2$.
- When x is a 0/1 treatment indicator, the concordance is equivalent to the Wilcoxon rank-sum test. If y is a survival time, it is equivalent to the Gehan-Wilcoxon test.
- For survival data, let $r(t)$ be the rank of each observation, relative to all observations which are at risk at time t. Then the concordance is equivalent to the score statistic for a Cox model using $r(t)$ as a time-dependent covariate.[105]
- Common variants of the AUC are equivalent to fitting a Cox model with time-dependent risk weights, which are in turn related to the Peto-Wilcoxon and Fleming-Harrington tests.[96]

add refs for Peto-Wilcoxon (Should this be Gehan-Wilcoxon?) and Fleming-Harrington above

Details of all these can be found in the formula appendix. Suffice it to say that this leads to a rich, and often confusing, set of interdependencies, along with multiple ways of viewing and understanding the concordance. (In all these cases ties are handled in the same way as Somers' d; ties in y are considered incomparable; pairs that are tied in x, but not y, score as 1/2.)

The concordance has a natural interpretation as an experiment: present pairs of subjects one at a time to the physician, statistical model, or some other oracle, and count the number of correct predictions that are returned. Pairs that have the same outcome (y) are not put forward for scoring (they are uninformative with respect to the oracle's accuracy), and if the oracle cannot decide (ties in $\hat{y}$) it makes a random choice. This leads to $C + T_x/2$ correct selections out of $C + D + T_x$ choices, which is easily seen to be equal to $(d+1)/2$.

This hypothetical experiment gives a baseline insight into the concordance. A value of 1/2 corresponds to using a random guess for each subject. Values of $< .55$ are not very impressive: the death ordering for some subject pairs will often be obvious, and a rater with no medical knowledge at all might do nearly as well by marking the obvious pairs and using a coin flip for all the rest. Values of $< .5$ are possible when a model is applied to new data. Values $> .8$ have proven to be elusive in clinical prediction models, except for very short-term predictions or models that are over-fit.

To extend the concordance statistic to survival data, any pairs which cannot be ranked with certainty are also considered incomparable, not just those with tied event times t. For instance, assume that t_i were censored at time 10 and t_j is an event at time 20. Subject i may or may not survive longer than subject j, and so it is not possible to score whether $\hat{y}$ has ranked them correctly or not. Note that if t_i is censored at time 10 and t_j is an event at time 10 then we do know that $t_i > t_j$, though the precise event time for subject

i is still uncertain. For hazard models we also need to flip the definition of concordant/discordant, since in this case a large value of $X\beta$ predicts a larger hazard rate, which equates to an earlier death.

Concordance is automatically computed in `coxph` and displays in the summary. It can also be obtained using the `concordance` function.

```
> pfit2 <- coxph(Surv(time,status==2)~age+ascites+pspline(bili), data=pbc)
> concordance(pfit2)
Call:
concordance.coxph(object = pfit2)

n= 312
Concordance= 0.8295 se= 0.01897
concordant discordant     tied.x     tied.y    tied.xy
     20735       4262          0          3          0
```

An advantage of concordance is that it is defined for linear models, generalized linear models, and survival models, allowing a common metric for all three. More details regard options to re-weight concordance and how to measure concordance in stratified models will be discussed in the chapter on developing and evaluating prognostic risk models.

3.2.4 Comparing models

The likelihood ratio test is the most common method to compare two models. This test can easily be obtained using the `anova` function. However, this test can only be used to compare nested models.

```
> pfit1 <- coxph(Surv(time,status==2)~age+ascites+pspline(bili), data=pbc)
> pfit1
Call:
coxph(formula = Surv(time, status == 2) ~ age + ascites + pspline(bili),
    data = pbc)

                          coef se(coef)      se2    Chisq   DF
age                    0.02821  0.00862  0.00855 10.70103 1.00
ascites                1.09964  0.27127  0.26946 16.43218 1.00
pspline(bili), linear  0.14143  0.01681  0.01672 70.79288 1.00
pspline(bili), nonlin                            39.76244 3.04
                            p
age                    0.0011
ascites                5.0e-05
pspline(bili), linear < 2e-16
pspline(bili), nonlin 1.3e-08

Iterations: 4 outer, 12 Newton-Raphson
```

```
     Theta= 0.727
Degrees of freedom for terms= 1 1 4
Likelihood ratio test=168  on 6.01 df, p=<2e-16
n= 312, number of events= 125
   (106 observations deleted due to missingness)
> pfit2 <- coxph(Surv(time,status==2)~age+ascites+pspline(bili)+edema, data=pbc)
> pfit2
Call:
coxph(formula = Surv(time, status == 2) ~ age + ascites + pspline(bili) +
    edema, data = pbc)

                          coef se(coef)      se2    Chisq   DF
age                    0.03102  0.00870  0.00863 12.72782 1.00
ascites                0.82253  0.28502  0.28270  8.32832 1.00
pspline(bili), linear  0.12475  0.01759  0.01751 50.28779 1.00
pspline(bili), nonlin                            41.13794 3.04
edema                  1.28238  0.29496  0.29401 18.90178 1.00
                             p
age                    0.00036
ascites                0.00390
pspline(bili), linear 1.3e-12
pspline(bili), nonlin 6.5e-09
edema                  1.4e-05

Iterations: 4 outer, 12 Newton-Raphson
     Theta= 0.734
Degrees of freedom for terms= 1 1 4 1
Likelihood ratio test=184  on 7 df, p=<2e-16
n= 312, number of events= 125
   (106 observations deleted due to missingness)
> anova(pfit1, pfit2, test='F')
Analysis of Deviance Table
 Cox model: response is  Surv(time, status == 2)
 Model 1: ~ age + ascites + pspline(bili)
 Model 2: ~ age + ascites + pspline(bili) + edema
   loglik Chisq     Df Pr(>|Chi|)
1 -556.01
2 -548.01    16 0.9972  6.299e-05
```

Concordance can also be used to compare two models. The variance of the concordance statistic can be computed using an infinitesimal jackknife (IJ)

argument. Define

$$d_i = \frac{\partial C}{\partial w_i}$$
$$\text{var}(C) = \sum_i w_i^2 d_i^2$$

If there are k C statistics on the same set of subjects, using different prediction models, the variance-covariance matrix is $D\text{diag}(w_i^2)D'$, where D is the n by k matrix of IJ values. This allows for testing equality of concordance values. One advantage of the IJ variance is that the set of models presented to the function need not be nested.

The gain in concordance from adding a new predictor will often be small, even if the predictor is highly "significant" in the underlying model. In the PBC data, for instance, the Cox models with and without the edema variable have a concordance of 0.829 and 0.834, respectively; a tiny gain, while the underlying model has $p < .001$ for edema. This helps to temper our enthusiasm about the actual size of an improvement. A disadvantage is that, as a rank statistic, the concordance is less powerful than the likelihood ratio test used in the underlying model. Thus, the p-value comparing concordance for 2 different models will often be less impressive than the likelihood ratio test comparing the models.

```
> concordance(pfit1, pfit2)
Call:
concordance.coxph(object = pfit1, pfit2)

n= 312
      concordance     se
pfit1      0.8295 0.0190
pfit2      0.8335 0.0197

      concordant discordant tied.x tied.y tied.xy
pfit1      20735       4262      0      3       0
pfit2      20836       4161      0      3       0
>
> # compare concordance values of pfit1 and pfit2
> ctest<-concordance(pfit1, pfit2)
> contr <- c(-1, 1)
> dtest <- contr %*% coef(ctest)
> dvar  <- contr %*% vcov(ctest) %*% contr
>
> c(contrast=dtest, sd=sqrt(dvar), z=dtest/sqrt(dvar))
   contrast          sd           z
0.004040485 0.005009428 0.806576041
```

Summary

In summary, as with all models, checking assumptions is important. In addition to the common checks of influence, functional form and additivity, it is also important to check the proportional hazards assumption when using Cox models. Concordance is the most popular measure of model performance in the survival setting. Another important model performance measure is the accuracy of the predictions (i.e., calibration). Calibration was not included in this chapter because models are always calibrated to the data used to fit them. Information on calibration, as well as a more in-depth discussion of discrimination can be found in the chapter on Developing and evaluating prognostic risk models.

Chapter 4

Time-dependent covariates

"Life can only be understood backwards; but it must be lived forwards."
Søren Kierkegaard

One of the strengths of the Cox model is its ability to encompass covariates that change over time. These are referred to as time-dependent or time-varying covariates. The practical reason for this is that the partial likelihood factors into a term for each event, and that term only involves the subjects who are at risk at that event time plus their covariate values *at that time*. We can think of the Cox model as a lottery model, one drawing at each event, which assesses the probability that the subject who experienced the event (i.e., the 'winner') should have done so.

4.1 Counting process data

At the computational level, a primary task is to communicate to the Cox model program who is at risk, at each timepoint, and what their covariates are. One of the more popular ways to do so uses the *counting process* form of a dataset. In this form each subject is represented by one or more observations. Each observation represents an interval of time `(time1, time2]`, the covariates that apply over that interval, and whether the interval terminates with an event.

Consider a subject with follow-up from time 0 to death at 185 days, and assume that we have a time-dependent covariate (creatinine) that was measured at day 0, 90 and 120 with values of 0.9, 1.5, and 1.2 mg/dl. The data might look like the following

```
  subject time1 time2 death creatinine
1       5     0    90     0        0.9
2       5    90   120     0        1.5
3       5   120   185     1        1.2
```

We read this as stating that over the interval from 0 to 90 the creatinine for subject 5 was 0.9 mg/dl, and that this interval did not end in a death. The underlying code treats intervals as open on the left and closed on the right, e.g., $(0, 90], (90, 120], (120, 185]$. This convention is critical for the validity of the proportional hazards model, which requires that events are predicted using

only data that was known *before* the event. To use a gambling analogy, all bets must be placed before the dice are rolled. While the creatinine level may change every day in actuality, the model can only make use of the last known value for a subject. A more thorough discussion of this is found in the section on immortal time bias.

A well-known example with a time-dependent covariate is the Stanford Heart Transplant study. Subjects who were enrolled in the study were treated with standard therapy until such time as a donor heart became available. Because of the need for a close tissue match between donor and recipient, the transplanted subject can be considered as a random selection from all those on the waiting list. A subject is managed using medical therapy until the time of transplant, and is on the transplant treatment arm thereafter. Here are the first few rows of the analysis dataset:

```
id   age  surgery transplant  start  stop  event
 1   30    0        0            0     49    1
 2   51    0        0            0      5    1
 3   54    0        1            0     15    1
 4   40    0        0            0     35    0
 4   40    0        1           35     38    1
 ...
```

Subjects 1 and 2 died before receiving a transplant and are on the medical treatment arm (transplant = 0) for their entire follow-up. Subject 3 received a transplant on the day of enrollment, so is on the transplant arm from day 0 forward. Subject 4 received a transplant on day 35 after enrollment, so they have transplant=0 for the first 35 days and transplant=1 for days 35–38. None of the first four subjects had a prior heart surgery (i.e., surgery = 0).

Once the dataset is constructed, fitting a Cox model is easy:

```
> coxph(Surv(start, stop, event) ~ age + surgery + transplant,
         data=heart)
Call:
coxph(formula = Surv(start, stop, event) ~ age + surgery + transplant,
    data = heart)

               coef exp(coef) se(coef)      z      p
age          0.03054   1.03101  0.01389  2.198 0.0279
surgery     -0.77333   0.46147  0.35967 -2.150 0.0315
transplant1  0.01610   1.01623  0.30859  0.052 0.9584

Likelihood ratio test=10.72  on 3 df, p=0.01335
n= 172, number of events= 75
```

Building the dataset requires some thought.

1. One subject died on the day of entry. However (0,0) is an illegal time interval for the `coxph` routine. It suffices to have them die on day 0.5. An alternative

is to add 1 day to everyone's follow-up, e.g., subject 2 who enrolled on Jan 2 1968 and died on Jan 7 would be credited with 6 days. (This is what Kalbfleisch and Prentice did in their textbook.) The resulting model fit is identical for either strategy.

2. A subject transplanted on day 10 is considered to have been on medical treatment for days 1–10 and on "transplant" treatment starting on day 11. That is, except for patient 38 who died on the same day as their procedure. They should be treated as a transplant death; the problem is resolved by moving this transplant back .5 day.

In our experience most datasets contain one or two subjects which require careful thought; the Stanford data is not unique. Since time is in days, the fractional time of .5 could be any value between 0 and 1; our choice will not affect the results of a Cox model fit. Thinking again of this as a lottery model, what matters is the decision about *who* is in the risk set at each death time and *what* their covariates were just prior to that timepoint. The timepoints themselves can be shifted: Cox models using time, time + 10, and log(time+3) as the scale all give identical results, and this is true for any transformation that leaves the time values in the same order.

When fitting a model with time-dependent covariates, we often find that 80% of the effort is spent building a proper dataset. Once that is done, the analysis itself is simple. One of the attractions of using counting process data is that the process of building the dataset can use any data manipulation tools that are at hand: data manipulation has a wide scope and each user/package will have their own favorite tools. All are valid; survival analysis only enters in after the data is completed. Unfortunately, it is all too easy to create an invalid dataset. The final result of the build needs to satisfy three basic tenets.

1. Each subject's timeline is well defined.
   - If the subject is at risk for an event at some time t, the dataset contains an interval for them that includes t.
   - Only 1 copy of the subject exists at any given time, i.e., there are no overlapping intervals.
2. All intervals are of positive length.
3. Outcomes are distinct; a subject cannot have two events at the same time.

Condition 3 is a consequence of condition 2.

In practice this means that a given subject will normally have a contiguous string of time intervals (a,b], (b,c], (c,d], ... In rare cases there can be a "hole" in the follow-up when they were not at risk, but these are rare enough that datasets which exhibit such behavior should be viewed with suspicion. An example where we used a gap concerned a follow-up study where one subject was lost to contact after 2 years, only to then reappear for an appointment at year 4 and maintain contact until the end of the study at year 7. After discussion it was decided to code their intervals as (0,2), (4,7). The rationale was that they were not at risk for an *observed* event during the interval from

2–4; if an event had happened, there was no guarantee that the study would have found out about it.

In R the `survcheck` routine is an aid for such validation, but we emphasize that it is an aid and not a guarantee. It is important to print out the data for a subset of subjects, actually read that printout, and *think*. In particular also look at ties. Both the death on day 0 for the heart transplant study and the one that occurred during transplant surgery required study-specific knowledge for a proper resolution.

One common question with the start-stop data setup is whether we need to worry about correlation, since a given subject now has multiple observations. The answer is no, we do not. The reason is that the (time1, time2] representation is essentially a programming trick; each subject only occurs once in a given risk set. When a subject has time-dependent values, the program needs to know which value to use at which time, and this is a way of communicating that information. The likelihood equations at any event time will use only one copy of any subject; the program picks out the correct row of data at each time. There are two exceptions to this rule:

- When a subject has multiple events. In this situation the rows for the events are correlated within subject and a cluster variance is needed. This will be discussed later in Chapters 12 and 11 .
- When a subject appears in overlapping intervals. However, as stated above this is almost certainly a data error.

4.2 Immortal time bias

As mentioned in Section 1.4.1, a pervasive problem in survival analysis is immortal time bias. The key rule for time-dependent covariates in a Cox model is simple and essentially the same as that for gambling: *you cannot look into the future*. A covariate may change in any way based on past data or outcomes, but it may not look forward in time. Below we review several common mistakes.

Responders vs. non-responders

One of the more well-known examples of immortal time bias is analysis by treatment response. There are multiple examples in the literature of survival curves comparing those who had a response to treatment (shrinkage of tumor, lowering of cholesterol, etc.) to those who did not, and they invariably show that responders have a better curve. A Cox model fit to this type of data will demonstrate a strong "significant" covariate effect. The problem arises because any early deaths, those that occur before response can be assessed, will all be assigned to the non-responder group, even deaths that have nothing to do with the condition under study.

Figure 4.1 shows simulated survival curves based on the advanced lung cancer dataset. Using the covariates and follow-up time from the data, we added a simulated parking variable. That is, assume that the subjects come

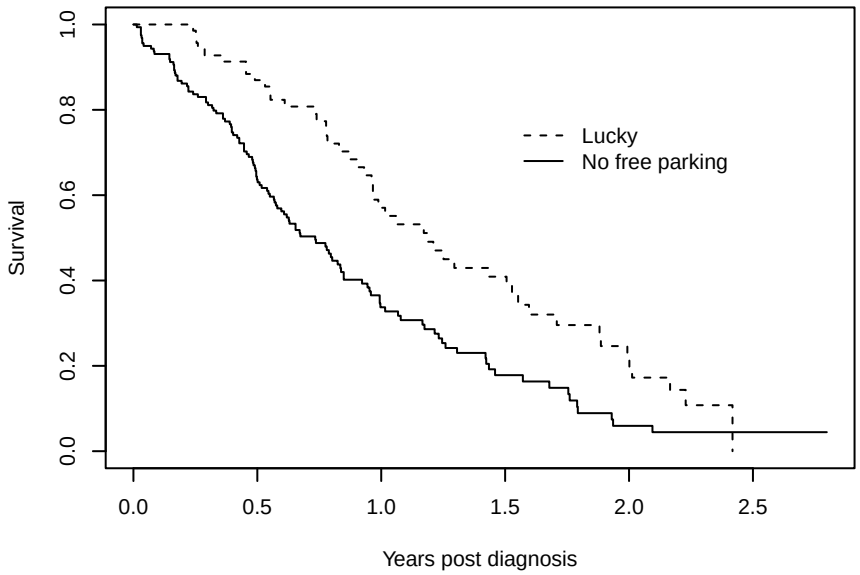

Figure 4.1 *Example illustrating immortal time bias using the lung dataset.*

in monthly for 12 cycles of treatment; at each visit 5% of the subjects were randomly chosen as the winner of a free parking pass for the day; 69/228 of the simulated subjects are marked as recipients of at least one of these. Now create a response variable `winner` that is 1 for anyone who ever received a pass and 0 for the others. A Cox model using this covariate reports a hazard ratio (death rate) of 0.53 fold for winners, with a p-value $< .001$. What is most surprising about this error is the *size* of the false effect that is produced. The issue is that there is a reverse causality: those with longer survival are more likely to win a parking pass, not that a parking pass confers longer survival. If the response variable is instead coded as a time-dependent covariate (depending only on the past), then the problem disappears.

This is exactly the issue with "ever responded" as a covariate. Even for a placebo treatment such an analysis will show an apparent benefit; the approach can never answer the question of whether response is a marker of success. An initial analysis of the Stanford heart transplant data made exactly this error, using "ever transplanted" as a fixed covariate, and yielded a 5-fold advantage for transplant. Gail [36] showed the correct approach and found a hazard ratio of 0.99.

The correct analysis for an evolving covariate, such as response, is to use a time-dependent covariate $X(t) = 1$ if the subject responded *before* time t. Such a covariate provides valid assessment of whether response predicts *future* events. The arrow of time must always point in the right direction. In

our simulated example, the time-dependent covariates "any free parking prior to today" shows no association with survival.

The alarm about this incorrect approach has been sounded often [6, 18, 36, 101] but the analysis is routinely re-discovered. An equivalent form of the issue is a table comparing responders and non-responders. Any covariate that is related to survival (e.g., age) will appear to also be related to response, even if response were in actuality a random covariate, e.g. young people are more likely to respond (or get a free parking pass) only because they live longer. A Cox model with response as the outcome might or might not show an association with age.

Compliance

A common and valid question is whether those who comply with a treatment regimen will do better than those who do not comply. Done incorrectly, this is a repeat of the responders versus non-responders analysis, but can actually have a bias that goes the other way. The longer someone survives on study, the more chances they have to be labeled as non-compliant.

Oakes [78] gives an example where even the correct analysis can be surprising. Using data from a large randomized trial, he showed that those who dropped out of the trial had a significantly higher death rate in the year after dropping out than those who remained on study. However, the size of this effect was identical for patients on the placebo and on the active treatment arms.

Total dose received

Another form of immortal time bias is discussed in Redmond et al. [86], written in response to a study that had divided breast cancer chemotherapy patients into three groups based on whether the patient eventually received $> 85\%$, 65–85% or $< 65\%$ of the full protocol dose [15]. (The planned dose is tailored to each patients' size.) The motivation for the analysis was the important question of whether dose reductions, usually in response to toxicity, are detrimental to eventual cure. The chemotherapy regimen spanned 12 weeks of treatment and the early deaths, not surprisingly, did not get all their dose. Hence, the "low dose" group is guaranteed to look worse and, as in the simulation example above, the estimated coefficient was large. Redmond reprized the analysis using a time-dependent dose variable "fraction received to date" and found no effect.

A subtler version of the error occurred when using the covariate "fraction of expected dose". Each subject's total dose received was divided by the expected amount over their observed follow-up time rather than by the protocol target dose at 12 weeks. This was then used as a single time-fixed covariate. The bias in this case went the other direction. Consider two subjects, one of whom died after 4 of the 12 cycles of therapy and a second who completed the regimen. Because toxicity is cumulative, a subject's chance of a dose reduction increases

with the number of cycles. Indeed, most subjects were at 100% of the expected dose at cycle 4 but over half had their dose reduced by the end. The model fit declared that a smaller fraction predicts longer life. The problem again is with the covariate definition of fraction of dose that the subject will *eventually* receive. Its value early in the study depends on future information, which is not yet available.

Markers of imminent death

A covariate that changes just before an event should always be treated with caution. As a very simple case, variables like "family has been called to the bedside" are not actual predictors of death, but are rather early notification of a cascade that is already in progress. Transfer to hospice care or the scheduling of a hospice care consult are of a similar flavor. Such predictors reveal no new medical insights. (Because they are such strong prophets of imminent demise, this type of variable will quickly come to the fore when using automated methods such as stepwise regression or machine learning, however.)

As a particular, and somewhat more subtle example, consider a study of the possible effect of digitalis treatment on cardiac mortality [79], for which one of the authors was a part of the analysis team. The cardiac patients in the cohort often went on or off medications, so significant effort was devoted to abstraction of the patient record. Digitalis was then coded as a time-dependent covariate, much like treatment in the Stanford Heart study. Despite strong prior suspicion that digitalis would result in increased mortality, the first models showed a huge benefit for the drug — over 4 fold. On further investigation it was discovered that during a patient's final week or days, many of their optional medications (e.g., digitalis) were stopped, something that the study nurses had diligently recorded. As a consequence, nearly no one died while actually on digitalis, and the Cox model declared it to be a wonder drug! A covariate of "digitalis within the last 14 days" erased the effect.

A variant of this can occur due to nearly tied dates. The medical visit of a seriously diseased subject may, for instance, span multiple days with laboratory tests, imaging, and consultation with the physician and other caretakers. Imagine a cancer visit where tests performed on Tuesday are the basis for a physician declaration of progression on Wednesday. With an automated data system, it is quite possible to have a time-dependent laboratory result which "predicts" progression with near certainty.

Another example from the authors' experience involved a secondary endpoint analysis in a placebo controlled drug trial. A small number of the treated subjects ($\sim$5 out of 100) were expected to have an adverse reaction to the agent and the study form contained an extra variable "weeks on drug", which was used to capture this information for subjects on the active arm. Entry and last follow-up were recorded as dates, as is common. The following pseudo code was used to create a time-dependent covariate of 1=active, 0=placebo. Can you spot the error?

1. For subjects without any adverse reaction, create a single data row with `time1=0`, `time2 = (fu_date - entry_dt)`, the event indicator and the treatment arm.
2. For subjects with a value for "weeks on drug"
   - If `drug_weeks < futime/7` this is a subject with an adverse reaction before the secondary endpoint. Create two lines
     - First line: `time1=0`, `time2 = drug_weeks*7`, event = 0 (no event) and treatment arm
     - Second line: `time1= drug_weeks*7`, `time2= (fu_date - entry_dt)`, event indicator, and treatment arm.
   - Otherwise the adverse event was after the secondary endpoint, treat them the same as no adverse reaction.

Imagine a subject with exactly one year of follow-up: 365 days. Since $52 * 7 = 364$, the above code will cross this subject over to placebo on day 364. Indeed, due to this rounding error, approximately half the deaths on the treatment arm are crossed over to placebo 1-3 days before death, leading to a hazard ratio of around 2 for the placebo arm. The correct hazard ratio was just less than 1.0, i.e., no treatment effect at all.

Even if the variables do not arise out of error, as in our examples, one has to ask whether such a predictor is even interesting. If a variable can only predict imminent events, then of what clinical use is it? One validation strategy for any time-dependent variable is to delay the covariate change by a small amount. Based on experience, 7 days seems to be a reasonable choice, since that is normally long enough to avoid date/timing issues within a single clinical visit. If including this delay in the analysis measurably changes the results of the fit (as it did in all three datasets above), then a deeper investigation is in order.

Conditional inclusion

In a 1952 paper, Berkson and Gage [12] pointed out the error with a common method of computing 5-year survival after surgery. This was to count only subjects whose status at 5 years was known, i.e., those with a death at $\leq$ 5 years plus those alive for 5 or more years. A subject whose surgery was 4 years ago, say, but is still alive and under active follow-up was ignored in the calculation. They showed that this approach consistently under-estimates the true survival of subjects.

The error is more subtle than an improper covariate but of the same type: a subject is included only if certain future data is available. The proper approach to analysis is the familiar Kaplan-Meier curve, which does its computation sequentially: each time that a subject dies, they are compared to all those still alive *at that time*. Subjects are retained in the calculation based on information that is currently known, without reference to the future.

A variant of this is the comparison of "cases" to "controls", subjects who

experience some particular endpoint versus those who do not. If the outcome is uncommon, and in particular if further data will need to be gathered on the subjects that are included in a study (reading histories, abstraction from pathology reports, further tests on stored blood samples, etc.), we may not want to include all of the non-cases in the analysis. A particular example was a study of the occurrence of rejection symptoms sometime in the first 4 years after receiving a liver transplant. A temptation is to match each subject with rejection (case) to 1–3 patients who had no rejection in 4 years (matched on age, sex, or other patient features) and extract data on only that subset of cases and controls. This approach is biased, as it turns out. The pool of potential controls for a subject who had rejection at 21 months after transplant must include *all* subjects who are under observation and rejection free at 21 months. Selection of controls based on any outcome or measurement beyond this index date leads to incorrect analysis, e.g., a later rejection, loss to follow-up before 3 years, a later diabetes diagnosis, etc.

Conditional exclusion

Perhaps a more subtle example of immortal time bias occurs when subjects are excluded based on co-existence of diseases. For example, in a study comparing two disparate diseases that can co-occur, such as a comparison of cardiovascular disease occurrence in patients with rheumatoid arthritis and diabetes mellitus, any patients with both rheumatoid arthritis and diabetes mellitus were excluded. The appropriate analysis would include these patients in the group for whichever disease occurred first and then move them to a combined group at the time when the second disease was diagnosed.

Confirmatory diagnosis

In this era of large medical record datasets, covariates and endpoints are often identified electronically. One issue with these datasets is that diagnostic codes are sometimes used to justify screening for a condition prior to actual diagnosis. One common solution is to require two diagnosis codes within some defined period, e.g., two high blood sugar values to declare diabetes. The appropriate analysis sets the date of diabetes at the second date, however the authors have seen examples where the date was set at the first occurrence, thus using future information.

Summary

There are many variations on the error of looking into the future; here are few more in addition to the longer discussions above:

- Interpolation of the values of a laboratory test linearly between observation times.
- Removal of subjects at baseline who do not finish a treatment plan.
- Imputing the date of an adverse event as midway between visit dates.

Each of these creates covariates at time t whose numeric value depends on information from the future. In each of these examples, the authors have seen cases where the error produced a negative, positive, or essentially no bias in the resulting coefficient. We find it nearly impossible to predict a priori how much bias will occur in any given dataset.

4.3 Predictable time-dependent covariates

A special type of time-dependent covariate is a predictable covariate, of which the two most common are age and time since enrollment. Unlike weight or blood pressure, the subject need not come into the office to know the value of the covariate: given the age at entry, one can calculate the current age for each subject at any time in the future.

It may often be the case that "current age" is a more natural predictor of outcome than the age at baseline. In a long-term study with death as the outcome, the underlying death rate is a function of current age, not someone's age when they enrolled. However, age is entered as a linear term, then the effect of changing age can be ignored in a Cox model. This is due to the structure of the partial likelihood. Assume that subject i has an event at time t, with j denoting other subjects in the risk set at that time, and let with a_j denote the age at study entry for each subject. At time t the ages will be $a_i + t$ and $a_j + t$, and the partial likelihood term for this death is

$$\frac{e^{\beta*(a_i+t)}}{\sum_{j\in R_i} e^{\beta*(a_j+t)}} = \frac{e^{\beta*a_i}}{\sum_{j\in R_i} e^{\beta*a_j}}$$

Using either the time-dependent age (the left hand version) or age at baseline (right hand), the partial likelihood term is identical since $\exp(\beta t)$ cancels out of the fraction. However, if the effect of age on risk is *non-linear*, this cancellation does not occur.

Figure 2.2 shows overall US death rates for the year 2010 for ages 40–90 years. Over this age range, which encompasses a large fraction of clinical research, the rates are nearly log-linear and additive with respect to age and sex; a relationship that also holds for cause-specific death rates for many of the major sub-causes of death. Thus, much of the time we actually can treat age as a linear term in the Cox model. This is not always the case, however, and should always be checked before blithely assuming simplicity.

Suppose, however, that we wish to fit current age as a non-linear effect. Since age changes continuously, we would in theory need a very large counting process dataset; one that had one interval per day for each subject. This is overkill, however, for two reasons. The first is medical: although we all grow older, risk does not increase so rapidly that we need to know age to the nearest day. In nearly all cases using a subject's age in years (integer) will suffice. A second reason is that even if we wanted perfect age, the underlying Cox model only needs the covariate values at the event times. A subject with 12 years

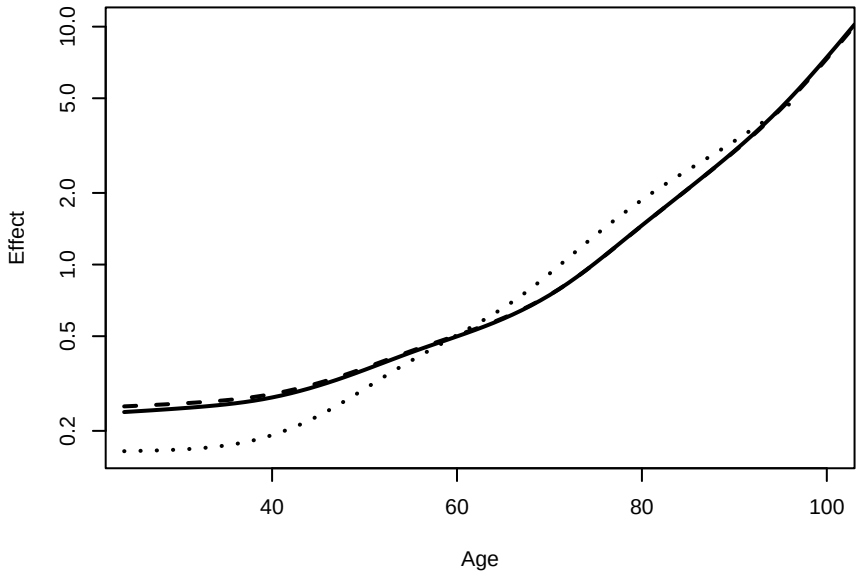

Figure 4.2 *Three models were fit using the MGUS data. Estimated effects are shown for baseline age (dotted), and current age using either yearly updates (solid) or updates at each death time (dashed).*

of follow-up, which spans 50 study deaths, will need only 51 (time1, time2] intervals, not 4383 of them (one per day).

Using the monoclonal gammopathy (MGUS) dataset we created fits using baseline age, age updated yearly, and age updated only at the death times. Figure 4.2 illustrates the functional form of age from these three models. We see that the fits using a yearly update and a per-event update hardly differ, though the dataset for the latter is over 10 times larger. There is some evidence for curvature in the age effect, and as a consequence the fits using baseline age and current age are not the same, though the difference in concordance is small (.666 versus .668). Further examination of this dataset (not shown) suggests that some of the curvature is due to a selection effect. Subjects entered the MGUS data set by virtue of having a certain laboratory test ordered, which in turn marks them as a high risk patient.

Since the Cox model computes relative hazard, the y-axis in the plot has an arbitrary intercept. By default, we have centered it so that the mean risk score $X\beta$ is 0. A horizontal line at y=1.0 intersects the curve near the mean age of 70. One could, if desired, decide set the risk at age 65 to the reference value of 1.0, which will shift the y axis labels downward.

4.4 Building time-dependent datasets

Building a counting process dataset for a time-dependent analysis is deceptively simple, but easy to do incorrectly. The desired final dataset looks like the following:

```
     id time1 time2 status sbp
1 Smith     0   120      0 110
2 Smith   120   320      0 125
3 Smith   320   360      1 115
4 Jones     0   100      0  99
5 Adams    50   140      0 142
6 Adams   140   160      0 150
7 Adams   160   300      0 145
8 Adams   300   410      1 148
```

- Each subject is represented by a sequence of time intervals (time1, time2], open on the left and closed on the right.
  - Intervals cannot overlap (only one copy of a subject at any given time). Exception: if a Cox model has multiple strata, then each is a separate time scale: a subject can 'start over' in a new stratum.
  - Intervals should be sequential, without gaps. There are valid exceptions to this, but they are rare.
  - When there is delayed entry, the first interval will not start at 0.
- A 0/1 event variable `status` indicates whether a given interval ended in an event.
- Covariates x1, x2, etc. describe the covariate value(s) that apply over the interval.
- An identifier variable shows which rows belong to each subject. This is not strictly necessary for all fits, but always useful.

Importantly, a covariate measured at time t applies to intervals *after* time t, e.g., in the above example the systolic blood pressure (sbp) of 110 for subject Smith measured at baseline applies to the (0, 120] interval, the value of 125 measured on day 120 applies to the (120, 320] interval, and the value of 115 measured on day 320 applies to the final time interval. For each interval, events happen at the end of an interval and covariates apply over the entire interval.

When creating a new time interval (i.e., the place where an interval is split into 2 intervals), available information may fall somewhere before, during, or after the existing intervals. Figure 4.3 depicts some types of scenarios that may occur.

- early: information occurs prior to the start of the subject's timeframe.
- late: information occurs after the end of a subjects timeframe.
- gap: information occurs in a gap between intervals at risk. If someone is

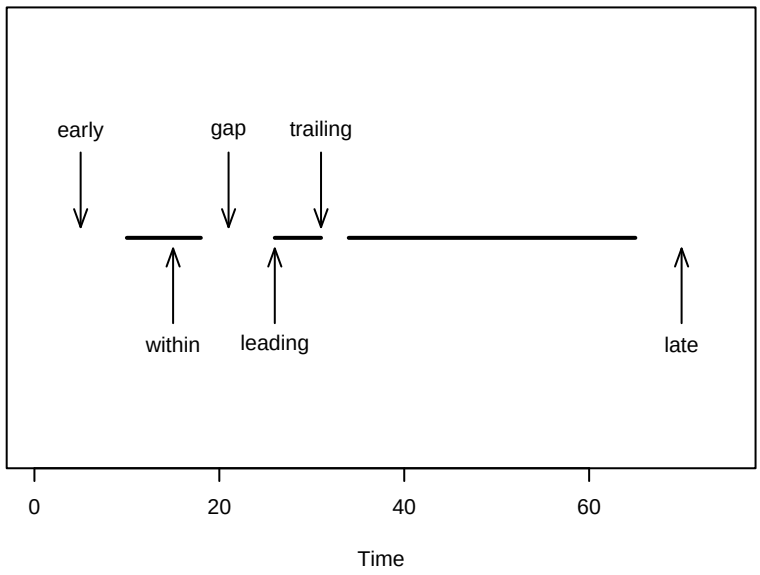

Figure 4.3 *Examples of where new information may be available when updating an existing dataset.*

at risk of the event from day 0 to 30, and then from day 50 to day 70. A count would occur under 'gap', if the covariate changed on day 40.

- within: information known in an existing time interval.
- boundary: information known at the same time as the beginning/end of an interval pair such as (0,10), (10, 25).
- leading: information occurs at the start of an interval.
- trailing: information occurs at the end of an interval.
- tied: two values of the same covariate, at the same time, for the same subject (e.g., multiple blood pressure values on the same day).
- missid: subjects are in the new information and are not in the existing dataset.

Things to remember when adding new information into an existing dataset:

- Time-dependent covariates that occur before the defined follow-up interval (i.e., early) or during a gap in time do not generate new splits, but do set the covariate for future times. Rationale: if diabetes was diagnosed when the patient was not under observation, the patient will still have diabetes when s/he returns to observation.
- Events that are early, in a gap, or late are not used in the analysis. The rationale for this is that we don't know who the appropriate comparison group is for these events, so they must be ignored.

	Baseline covariates		Time-dependent covariates	
	β	se(β)	β	se(β)
log(bilirubin)	0.94	0.10	1.18	0.11
albumin	-0.76	0.23	-1.60	0.19
edema	0.88	0.28	0.89	0.23
enrollment age (decades)	0.40	0.08	0.44	0.09
loglik	193.6		452.1	

Table 4.1 *Coefficients, standard errors, and partial log likelihood for PBC models with baseline and time-dependent covariates.*

- Date errors are one of the more common causes for early/late events; so they are always worth checking.
- "Tied" values, i.e., values measured at the same time for the same subject, are also noted. For instance, assume that the input dataset has two triglyceride observations, for the same subject Smith, on day 135. Since we can't create a data row with times (135, 135) a decision needs to be made about which triglyceride value to use.

The R survival package has several tools, such as `tmerge`, which can aid in building counting process data sets; several examples are given in the companion document. The most useful tool, however, is simply to look at the data. Print out the rows for a handful of subjects, and read it, to check that all the conditions in the list above have been satisfied.

4.5 Survival curves

How do we create the predicted survival curve from a Cox model with time-dependent covariates? It turns out to be quite difficult to create a correct curve, and easy to create a misleadingly useless one.

As an example of the simple, but flawed approach, consider modeling sequential laboratory tests in the PBC study, with time to death as the endpoint. PBC is a chronic autoimmune disease, resulting in a slow progressive destruction of the small bile ducts of the liver. Markers of liver function such as bilirubin and albumin levels give insight into the current stage of disease. Not surprisingly, time-dependent markers work better than baseline. Table 4.1 compares the two models. The fit with time-dependent laboratory measurements has a log-likelihood that is over 2x that for the fit using only baseline.

Figure 4.4 compares predicted survival curves from the two models, each is for a hypothetical subject with early disease: age 40 years with a bilirubin of 1.5, albumin of 3.0, no edema. The prediction using baseline covariates is clinically reasonable, while that using time-dependent covariates is completely different. The code for the two is essentially identical: define a single row

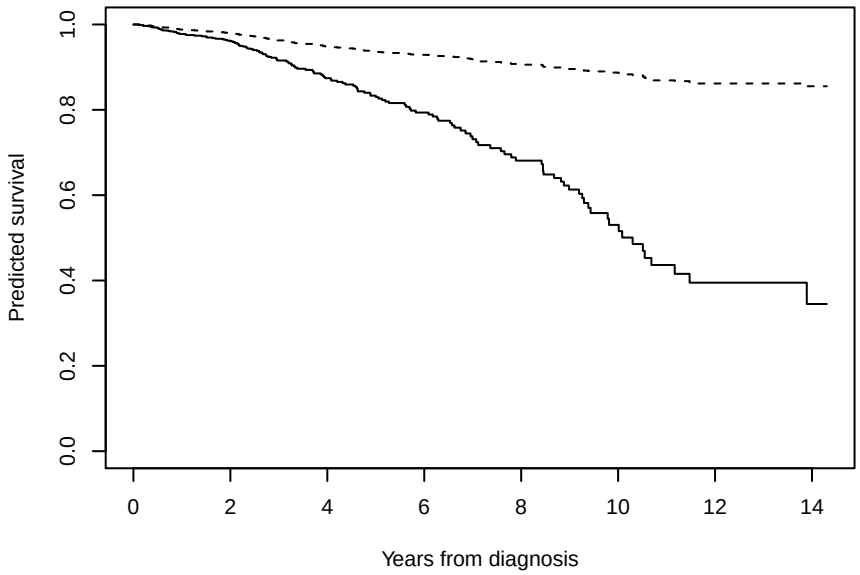

Figure 4.4 *Predicted survival curves from two PBC models. Solid lines are predictions from a model based on baseline covariates and dashed line are predictions from the model based on time-dependent covariates.*

dataset with the target values for bilirubin, albumin, edema, and age, then ask for model prediction for that hypothetical subject. What is going wrong?

The time-dependent curve shows the prediction for a subject who starts with mild disease *and their laboratory values never change*. Given the nature of this disease, such a subject does not exist. Remember, model 1 makes use only of baseline value: it is a prediction of future survival given those baseline values. However, model 2 makes use of *current* biomarker values. To create a true time-dependent curve, we would need to define a future covariate path; one can then create a predicted survival for that covariate trajectory. One can create such a path "by hand", but even then the interpretation is unsure. Say that the proposed path has a bilirubin level of 3 at 10 years. A strict view is that because only living subjects can have a bilirubin level, a "survival fraction" for this postulated path of values is nonsense: survival at 10 years must be 1, by definition.

There are 4 ways to proceed: recognize that this is a difficult problem and forego absolute risk curves with time-dependent covariates, use landmarking as a partial solution, multistate models as discussed in section **??**, or joint models. Options 2 and 4 are discussed below.

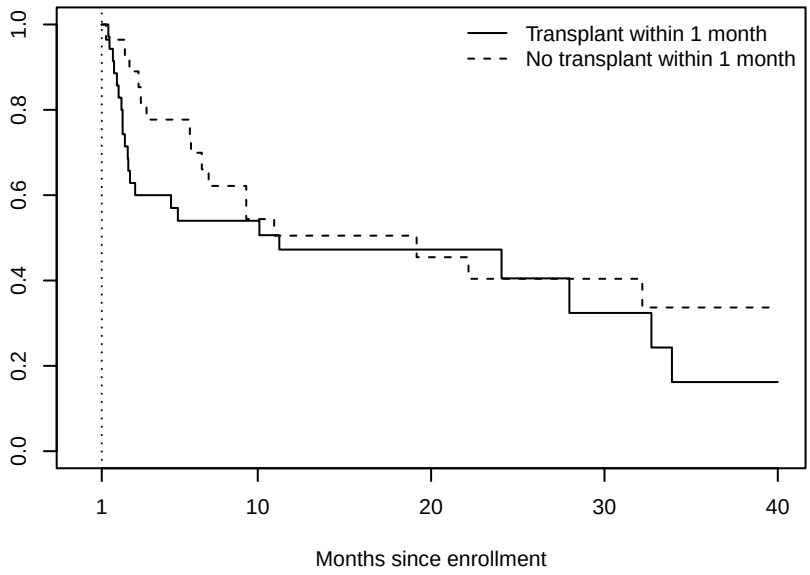

Figure 4.5 *Landmark curves of transplant in the first month after waitlisting for subjects in the Stanford heart transplant study.*

4.5.1 Landmarking

A valid approach to drawing predicted survival curves for a dataset with time-dependent covariates is landmarking. Landmarking involves choosing a specific timepoint of interest and basing the survival curves on the values of the time-dependent covariate(s) at that timepoint. For example, in the Stanford heart transplant study, choosing the 30 day timepoint, survival curves could be drawn for the subjects alive at 30 days, divided by whose who have and have not received a transplant before that time. The starting point for these Kaplan-Meier curves would be the chosen timepoint of 1 month after enrollment (Figure 4.5). Subjects who died prior to 1 month are excluded from this analysis. Subjects without a transplant at 1 month would remain in the "no transplant at one month" group for the rest of their follow-up, even if they later received a transplant.

Advantages are:

- This approach avoids looking into the future.
- The subjects plotted in each survival curve can be clearly described.

Disadvantages are:

- The choice of a timepoint of interest can be arbitrary. Different timepoints will yield different plots, which may lead to different conclusions. A very early or very late landmark timepoint may yield one group with a small

sample size or a small number of events, which can lead to unstable or imprecise Kaplan-Meier curves.

- Future changes in the time-dependent covariate are not accounted for, which leads to attenuation of the effect. For example, patients in the "no transplant" group may later receive a transplant, which may affect the event rate. This makes the "no tranplant" group more similar to the "transplant" group, which leads to attenuation.
- The exclusion of subjects who experience the event prior to the chosen landmark can results in a loss of power.

4.5.2 Joint models

One valid way to create survival curves in the presence of time dependent covariates is to model the entire process using joint models and dynamic predictions. This essentially combines information from a mixed effects model of the longitudinal measurements and a Cox model for the survival endpoint using an MCMC approach. More details can be found in the book by Dimitris Rizopoulos **??** and in the 'JMbayes2' package vignettes.

The plot below illustrates the wide variety of paths that a subject may take. It uses those with an initial bilirubin in a $\pm.8$ window around 1.5, to line up with prior examples. (This might go in the companion.)

include dynamic prediction figure - only include one of the above panels?

4.6 Time-dependent coefficients

Time-dependent covariates and time-dependent coefficients are two different extensions of Cox models, as shown in the two equations below.

$$\lambda(t) = \lambda_0(t)e^{\beta X(t)} \tag{4.1}$$

$$\lambda(t) = \lambda_0(t)e^{\beta(t)X} \tag{4.2}$$

Equation (4.1) is a time-dependent covariate, which is a common usage. Equation (4.2) has a time-dependent coefficient, which is much less common. Time-dependent coefficients are one way to deal with non-proportional hazards: the proportional hazard assumption is precisely that the coefficient does not change over time, and was discussed earlier in Section 3.1.3.

Figure 3.3 used the veterans cancer data to illustrate a violation of the proportional hazards assumption. Karnofsky score has a large negative effect early on, but by 100 days this has waned and is not much different from zero. One explanation is that in this very acute illness, any measure that is over 6 months old is no longer relevant. Another is that the original population was a mixture of subjects who were sensitive or not to Karnofsky status; the biology has not changed but rather that the susceptible fraction has been lost.

Proportional hazards is never perfectly true, but over a short to medium time scale it is often reasonable. In this case, however, the departure is quite

large and a time-dependent coefficient is a more useful summary of the actual state. The `cox.zph` plot is excellent for diagnosis of the problem, but does not produce a formal fit of $\beta(t)$. What if we want to fit the model?

Piecewise time-dependent coefficients

One of the simplest extensions is a step function for $\beta(t)$, i.e., a piecewise fit with different coefficients over different time intervals. One way to accomplished this is by breaking the dataset into disjoint time ranges, and then fit a separate Cox model to each portion. An easier approach is to use the `survSplit` function to break the dataset into time-dependent parts. We will arbitrarily divide the veteran's data into 3 epochs of the first 3 months, 3-6 months, and greater than 6 months. The `survSplit` function call below creates a new covariate, `tgp`, with 3 levels for the 3 time periods.

```
> vet2 <- survSplit(Surv(time, status) ~ ., data= veteran, cut=c(90, 180),
                    episode= "tgp", id="id")
> vet2[1:7, c("id", "tstart", "time", "status", "tgp", "age", "karno")]
  id tstart time status tgp age karno
1  1      0   72      1   1  69    60
2  2      0   90      0   1  64    70
3  2     90  180      0   2  64    70
4  2    180  411      1   3  64    70
5  3      0   90      0   1  38    60
6  3     90  180      0   2  38    60
7  3    180  228      1   3  38    60
```

The first subject died at 72 days; their data in the derived dataset `vet2` is the same as in the original data. The second and third subjects contribute time to each of the three intervals. At this point we could fit 3 separate models to the intervals. Below we fit all three at once; an advantage of this latter approach is that we can use a single coefficient for treatment and prior chemotherapy while allowing the coefficient for Karnofsky score to vary by interval.

```
coxph(formula = Surv(tstart, time, status) ~ trt + prior + karno:strata(tgp),
    data = vet2)
                           coef exp(coef) se(coef) Pr(>|z|)
trt                     -0.0110      0.99   0.1891  9.5e-01
prior                   -0.0061      0.99   0.0204  7.6e-01
karno:strata(tgp)tgp=1 -0.0488      0.95   0.0062  4.6e-15
karno:strata(tgp)tgp=2  0.0081      1.01   0.0128  5.3e-01
karno:strata(tgp)tgp=3 -0.0083      0.99   0.0146  5.7e-01
```

A fit to the revised data shows that the effect of baseline Karnofsky score is essentially limited to the first three months. A formal test of proportional hazards on this revised fit (not shown) shows no further time-dependent effect of Karnofsky score. This last is of course no surprise, since we used the

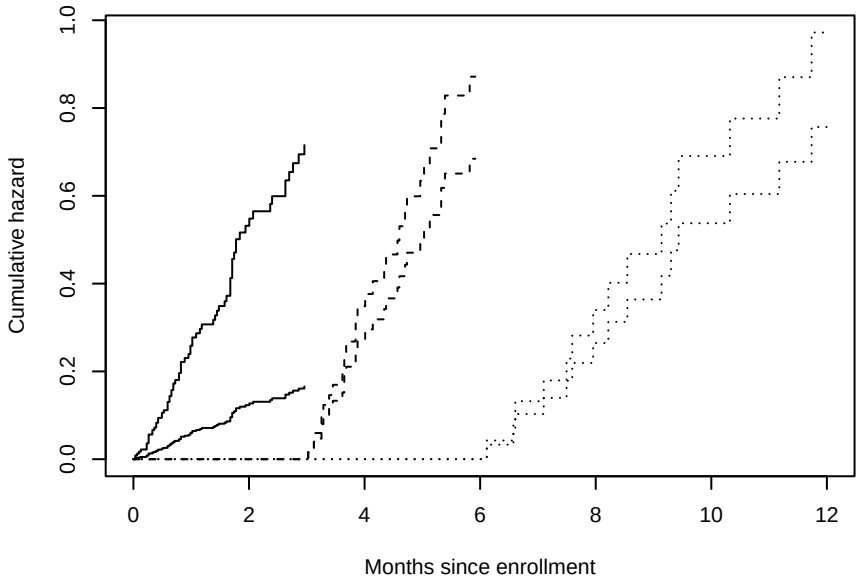

Figure 4.6 *Cumulative hazard curves from a Cox model that includes treatment, prior chemotherapy status, and an interaction of Karnofsky score with time strata (3 months, 3-6 months, and greater than 6 months). The curves are for 2 subjects with Karnofsky score of 60 or 90, active treatment, and no prior chemotherapy. The solid, dashed and dotted lines show the cumulative hazard for the first, second and third strata, respectively.*

original graph to pick the cut points. A "test" that the coefficients for the three intervals are different is biased by this sequential process and should be viewed with skepticism.

In contrast to time-dependent covariates, survival curves for a model with time-dependent coefficients are well-defined. Creation of the curves requires a bit more care — this is one of the cases where it is legal to create a future covariate path. Reminder: a survival curve generated from a Cox model is specific to a particular set of covariates, as is true for other predictions such as a risk score. The first step is to create a dataframe with the covariates of interest. Figure 4.6 shows the cumulative hazards for two subjects, one with low and another with high Karnofsky score.

This generates curves for each stratum. To create a single curve combining strata, we need to create a covariate path, i.e., a dummy subject whose value of `tgp` will change at pre-specified times. Details for producing this figure are laid out in the examples document. The figure shows the Kaplan-Meier curves, splitting the data at the median Karnofsky score of 60. The overlaid predictions are for two hypothetical subjects at the 25th and 75th percentile of Karnofsky score. We do not expect the curves to exactly overlay: each Kaplan-Meier curve is for a cohort of subjects containing variable levels of Karnofsky

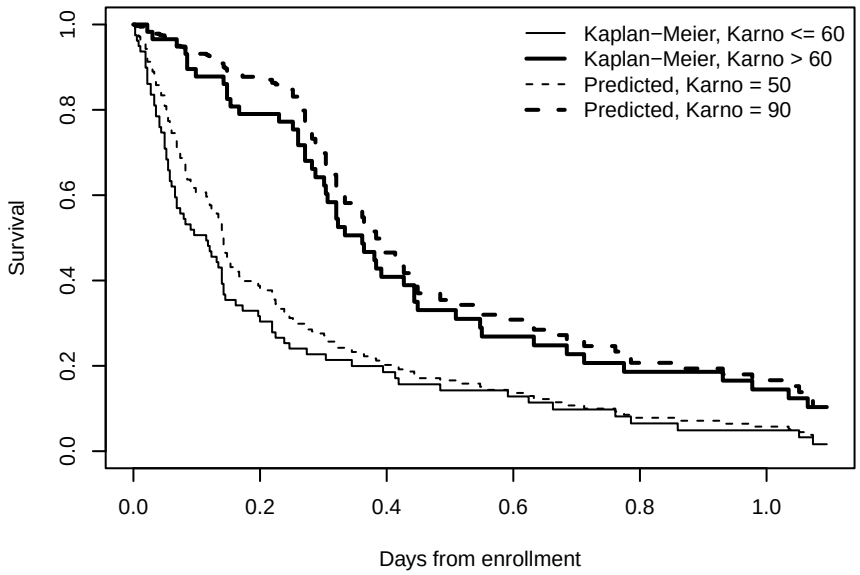

Figure 4.7 *Predicted survival curves from a Cox model that includes treatment, prior chemotherapy status, and an interaction of Karnofsky score with time strata (3 months, 3-6 months, and greater than 6 months), along with Kaplan Meier curves for groups of subjects with Karnofsky score above and below 60.*

score, treatment, and chemotherapy prior to the study, while the predicted curves from the Cox model are each for a single individual. More comparable curves from the Cox model can be obtained using adjustment, which will be discussed in Chapter 5.

Continuous time-dependent coefficients

Instead of fitting piecewise time-dependent coefficients, it may be desirable to fit continuous time-dependent coefficients. This avoids the issue of choosing cutpoints, but has the issue of choosing the correct functional form. If $\beta(t)$ is assumed to have a simple functional form, we can fool an ordinary Cox model program into doing the fit. For instance, the particular form $\beta(t) = a+b\log(t)$ is often recommended. Then $\beta(t)x = ax + b\log(t)x = ax + bz$ for the special time dependent covariate $z = \log(t)x$.

A problem with this approach is that it tests for a particular form of the fit (log) without any verification that the form is correct. As shown in Grambsch and Therneau [42], the resulting test is equivalent to fitting a line to the `cox.zph` plot with log(time) as the horizontal axis along with a test for zero slope. A test of $\beta(t) = a + bt$ is equivalent to the plot with `time` as the horizontal axis, etc. Plots of these time scales are shown in Figure 3.3. The p-values for the proportional hazards tests are $< .05$ for both plots. However,

the linear time scale fit is particularly problematic due to the small number of points on the far right, which highly influence the fit.

Despite the problem of choosing the form without verifying it is correct, let's assume we want to forge ahead with a simple function and fit a time-dependent coefficient on the log-transformed time scale. An obvious but incorrect approach is

```
> coxph(Surv(time, status) ~ trt + prior + karno +
                            karno:log(time), data=veteran)
```

This mistake has been made often enough that the `coxph` routine has been updated to print a warning message for such attempts. The issue is that the above code does not actually create a time-dependent covariate. We are asking the routine to distinguish between the variable `time` in the dataset, which is a fixed value for each subject denoting their final observation time, and the underlying continuous time scale used by the model. Being unable to read our mind, the computer treats both instances of the word `time` as the first type, placing a constant value on the right hand side. This variable most definitely breaks the rule about not looking into the future, and one would quickly find the circularity: large values of `time` predict long survival, because long survival creates large values for `time`.

The solution is to create an expanded dataset, split at each of the observed event times; now the two meanings coincide.

```
> uniq.times <- unique(veteran$time[veteran$status==1])
> vet3 <- survSplit(Surv(time, status) ~ ., data= veteran, cut=uniq.times)
> coxph(Surv(tstart, time, status) ~ trt + prior + karno + karno:log(time),
                data=vet3)
Call:
coxph(formula = Surv(tstart, time, status) ~ trt + prior + karno +
    karno:log(time), data = vet3)

                     coef exp(coef)  se(coef)      z        p
trt              0.031858  1.032371  0.189614  0.168  0.86657
prior           -0.008606  0.991431  0.020266 -0.425  0.67109
karno           -0.083008  0.920344  0.017113 -4.850 1.23e-06
karno:log(time)  0.013269  1.013357  0.004322  3.070  0.00214

Likelihood ratio test=53.19  on 4 df, p=7.766e-11
n= 5959, number of events= 128
```

The time-dependent coefficient is estimated to be $\beta(t) = -0.083 + 0.013 * \log(t)$. We can add said line to the `cox.zph` plot (Figure 4.8). Not surprisingly, the result is rather too low for time > 100 and underestimates the initial slope. Still the fit is better than a horizontal line, as confirmed by the p-value for the slope term. (The p-value from cox.zph is nearly identical, as it must be, since the tests in cox.zph are for a linear effect on the chosen time scale.)

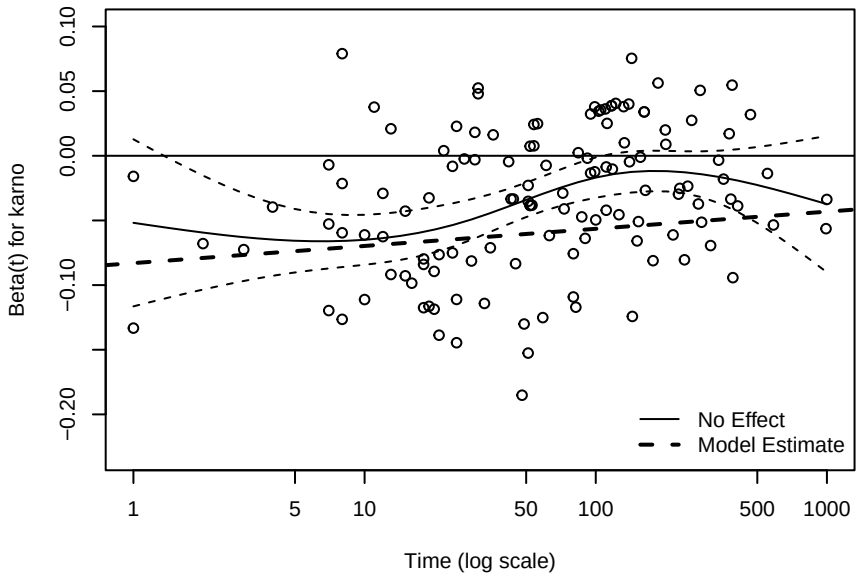

Figure 4.8 *This figure depicts the proportional hazards plot for Karnofsky score in the veterans dataset using a log-transformed time scale. This shows non-proportional hazards with a negative effect for the first 100 days that wanes thereafter. The thick dashed line shows the continuous time-dependent coefficient fit with a log scale.*

Summary

In summary, time-dependent covariates are often a useful extension of Cox models. They can be easily fit with existing software, but care must be taken to properly build the dataset, particularly if there is more than one time-dependent covariate of interest. Care must also be taken in the interpretation of time-dependent effects, as they represent the instantaneous effect of the covariate of interest. Time-dependent covariates have no use in prediction of future outcomes, because they use the current value of each measurement, which changes throughout follow-up. They are most useful for investigating the biologic effects of a covariate on an outcome, or for ensuring thorough adjustment for the covariate of interest. Another concern is the frequency of measurement of measurements, such as lab values. Andersen and Liestol demonstrated the infrequent measurement of time-dependent covariates can lead to attenuation of effect size [4].

Graphical representations of time-dependent covariates are also problematic, but landmarking can help. Finally, the chapter would not be complete without another admonition about the insidious and pervasive nature of immortal time bias. Even seasoned analysts often fall victim to this bias. Detailed study of the concept is encouraged to aid in early recognition and avoidance

of this bias. The good news is that this is a bias that can be avoided and addressed in nearly all instances.

Chapter 5

Absolute risk

> "Comparative experiments are mandatory in order to not view coincidences as cause-effect relationships. ... The comparative experiment requires, to be of some value, to be run in the same time and on as similar as possible patients, else the physician walks at random and becomes the sport of illusions." C.Bernard, Introduction à L'Etude de la Médicine Expérimantale, 1866

In survival analysis, attention has often focused on hazard ratios (HR), which have a long reign as the primary and often only Cox model summary that is reported. However, there are several shortcomings to HR, particularly HR in the absence of anything else, and there is recently a renewed emphasis on also reporting measures of absolute risk [48]. One well recognized example of the difference between relative and absolute measures is the impact of smoking on health. Smoking increases the hazard ratio for lung cancer by approximately 24 fold for men, and public perception of smoking risk is almost entirely focused on cancer. However, lung cancer in the general (non-smoking) population is a very rare disease, and even with the increased risk, only 1 in 10 lifetime smokers are projected to get lung cancer [21]. But smoking also increases coronary heart disease and stroke hazard ratios by 2–4 fold and the *absolute* impact of these is much greater, as 40% will die of these two causes. On the relative (HR) scale the effect of smoking on cancer risk is 10 times its effect on cardiovascular disease, but on absolute scale effect on CVD risk is 4 times that of the effect on cancer. In terms of which of the two is better to report we would argue the proper answer is 'both'.

In this chapter, we will use the outcomes from a population-based study of free light chain (FLC) levels as a running example. For illustration we have divided the continuous FLC levels into three groups, as was done in the original clinical paper [28], which are those below the 70th percentile of FLC (low), 70–90th percentile (medium) and above the 90th percentile (high). Division into three groups is convenient to illustrate the methods, but we do not make any claim that such a categorization is optimal or even sensible statistical practice. Figure 5.1 shows the overall Kaplan-Meier (KM) curves for subjects in each group, showing an obvious association of FLC with survival. The effect is apparent in a fitted model as well: a fitted Weibull

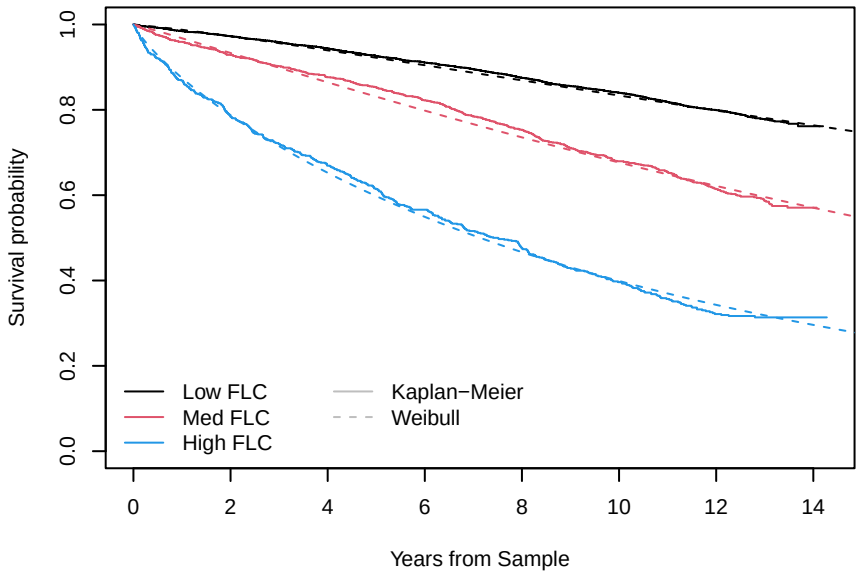

Figure 5.1 *Kaplan-Meier curves from the free light chain (FLC) study, separated by subjects' level of FLC. Overlaid are predictions from a fitted Weibull model.*

	50–59	60–69	70–79	80+
Low FLC	2587 (47)	1690 (30)	969 (17)	314 (6)
Med FLC	442 (29)	446 (29)	424 (28)	215 (14)
High FLC	121 (16)	188 (25)	225 (30)	224 (30)

Table 5.1 *Comparison of the age distributions (row percents) for each of the three groups.*

regression shows accelerations of approximately 2 and 5 fold for the death times in the medium and high FLC groups, respectively, as compared to the FLC low group. Predicted survival curves from the model are overlaid on the Kaplan-Meier and fit very well.

5.1 Covariate adjustment

A key problem with the simple curves of Figure 5.1 is confounding; both the KM and the Weibull curves are misleading due a strong age interaction with age and sex. Average FLC amounts rise with age, at least in part because it is eliminated through the kidneys and renal function declines with age. The female/male balance of the population also changes with age, so sex is also a potentially important confounder. Table 5.1 shows the distribution of ages for each of the FLC groups; the low group has very few 80+ year olds while for the high FLC group, 80+ year olds comprises nearly a third. "Young people

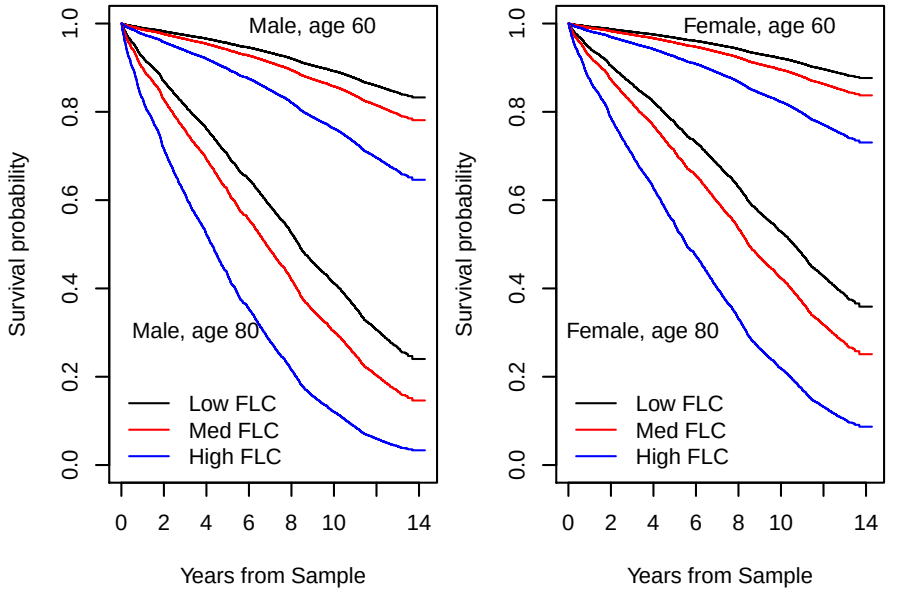

Figure 5.2 *Predicted survival curves for the FLC data, from a Cox model containing FLC group, age, and sex.*

live longer" is an unsurprising finding; it is much more interesting to look at the effect of FLC after adjustment for age and sex. Adjusting for age and sex in the model fit is easy: simply add those variables to the model formula. For the Weibull model of Figure 5.1 the coefficient estimates for the FLC medium and FLC high groups are reduced by more than half. Much, but not all, of the apparent FLC effect is explained by age and sex. Adjustment of the absolute risk estimates is not quite as simple.

Subgroup curves

Since it is straightforward to obtain curves from a fitted model, a common approach to covariate adjustment is to create a "flock" of curves such as those shown in Figure 5.2. To create the covariate-specific curves, one first creates a dataset which has one row for each desired curve, specifying values for each of the covariates found in the model. Comparing these curves to those in Figure 5.1, an attenuation of the FLC effect is immediately apparent; this agrees with the estimated HR for FLC group from a Cox model without adjustment of 1 (for the reference group, Low FLC), 2.1 and 5.2 as compared to 1, 1.5, and 2.4, respectively, from a model with adjustment for age and sex.

Further flexibility in the shape of the curves can be obtained by stratifying on the variable of interest (e.g., FLC) while adjusting for the others. That is, replace `group` with `strata(group)` in the model where `group` is the marker for

the three FLC groups. This allows the curves for the three FLC groups to have a different *shape*, the cost being that no coefficients or p-values are available for the stratification variable. In this particular example, stratification leads to almost no change.

The problem with this approach is how to decide *which* covariate values to choose for the other variables, and how many levels of them. When there are multiple confounders to adjust for, the resulting number of curves will become unmanageable.

"mean subject"

A recurring, erroneous approach is to deal with the multiplicity by using the mean for each of the confounding covariates, leading to obtain a single curve for each FLC group. Unfortunately, many packages produce $\hat{S}(t, \overline{x})$ as the default "predicted survival" if no covariate set is specified by the user. Most simply the issue is that

$$E[S(t; x)] \neq S(t; E[x])\,;$$

expected values cannot be moved inside a nonlinear function, and the "mean covariate" curve corresponds neither to any subject nor to a population. With respect to the first, in the FLC dataset the 0/1 dummy variable for sex has a mean value of 0.45. Who exactly does this represent?

Treating the $\overline{x}$ curve as a population curve is as large an error, though more subtle. Figure 5.3 reprises an example from Therneau and Grambsch [103]. Consider a set of grandfathers and grandsons at a baseball game, with mean ages of 60 and 10, respectively. The predicted survival $E[S(t; x)]$ for this cohort dips twice as first the grandfathers and then later the grandsons reach old age. It looks nothing like the predicted survival $S(t; E[x])$ of a 35 year old.

This error has been recurrently highlighted in the literature, but is still widespread [40]. One of its more common variants is to fit and plot a stratified Cox model:

```
> cfitc1 <- coxph(Surv(futime, death) ~ strata(group) + age + sex,
                  data=flchain2)
> sfitc1 <- survfit(cfitc1)
```

This is a very fast and easy way to produce a set of three curves, one for each stratum. But, as just stated, these are curves for some hypothetical subject of the mean study age (64.3) and indeterminate sex. A Cox model that treats the FLC group as a covariate imposes an additional constraint of proportional hazards across the 3 FLC groups and is even less satisfactory.

Population average estimates

To avoid the messiness of multiple covariate-specific curves and to provide an illustration of the difference between groups after adjustment for confounders, it would be useful to create a single overall curve for each FLC group. These

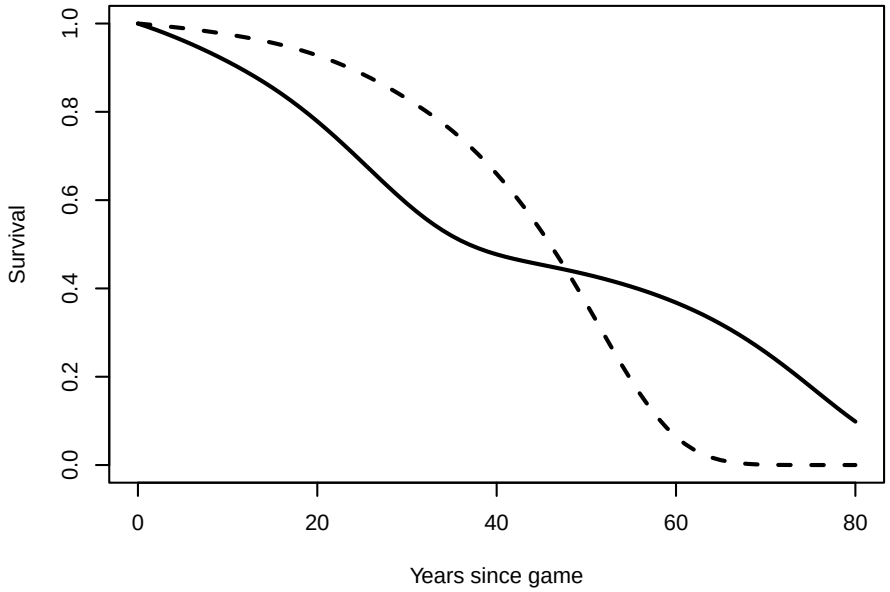

Figure 5.3 *Predicted survival for a fictional cohort of 10 and 60 year olds (solid line); the dashed line shows the predicted survival for a cohort of 35 year olds.*

curves need to be both adjusted for the other covariates and properly calibrated (i.e., the overall average value is correct). The key idea is to impose balance. To illustrate the idea we start with a simple example.

Consider the hypothetical data shown in Figure 5.4 comparing two treatment arms, A and B, with age as a confounder. What is a succinct but useful summary of the treatment effect for arms A and B and of the difference between them? One approach is to select a fixed *population* for the age distribution, and then compute the mean effect over that population.

More formally, assume we have a fitted model. We want to compute the conditional expectations

$$m_A = E_F\left(y|\text{trt} = A\right) \tag{5.1}$$

$$m_B = E_F\left(y|\text{trt} = B\right) \tag{5.2}$$

where F is some chosen population for the covariates other than treatment.

The most important question is what choice to give for the population F, and this depends critically on what question we want to answer. For instance, in the simple example of Figure 5.4, if we were considering deployment of these two treatments in nursing home patients, then it would make sense to use an average that gives larger weights to older ages, e.g., a known age distribution for nursing homes. Three common choices for F are:

1. Empirical: the dataset itself or a specific subset.

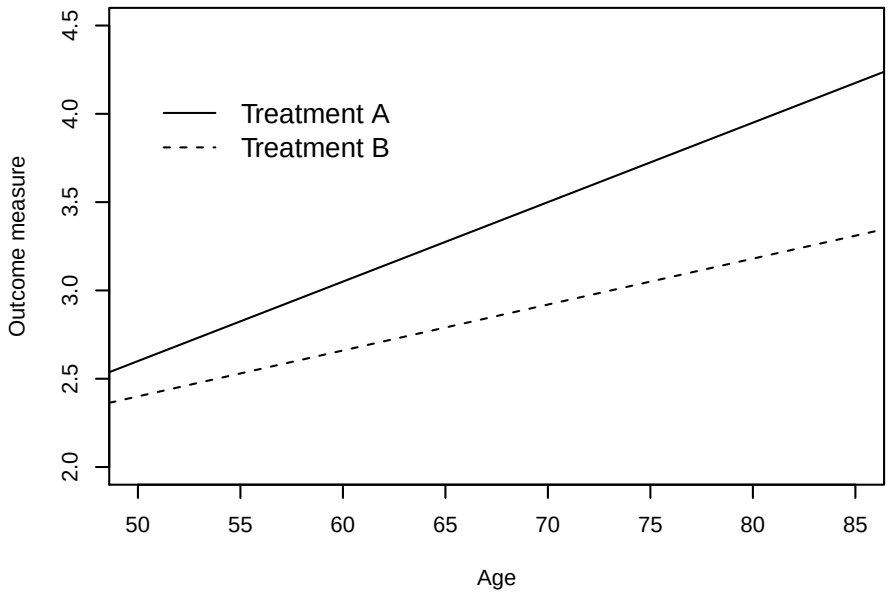

Figure 5.4 *Hypothetical clinical trial comparing 2 treatment arms (A and B)*

- For the simple example above, this would be the distribution of all n ages in the dataset, irrespective of treatment.
- For a case-control study, it is common to use the distribution of the cases.

2. External: An external reference population, such as:
   - A fixed external reference, e.g., the age/sex distribution of the 2000 US Census. This approach is common in epidemiologic studies.
   - Data from a prior study. This can be useful for comparison of one study to another.
3. Theoretical:
   - A fixed statistical distribution, e.g., using the Gaussian for a random effect term.
   - Factorial or Yates: the dataset for a balanced factorial experiment. This is only applicable if the adjusting variables are all categorical; the population consists of all unique combinations of the adjusters.

This basic idea has been explored and re-discovered many times in statistics, which should be no surprise: one of a statisticians' first instincts is to wrap "E()" around some expression, and a second is to balance experiments with respect to ancillary covariates (or confounders).

There are two main approaches to adjust for imbalance, as illustrated in Figure 5.5.

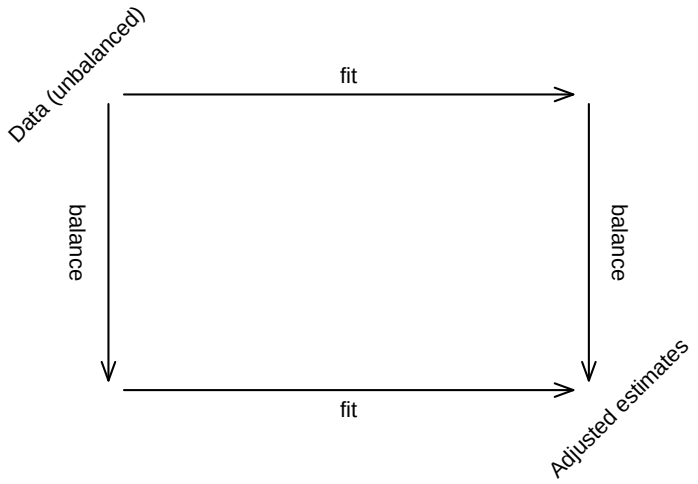

Figure 5.5 *Diagram showing computational approaches to obtain adjusted estimates from unbalanced data.*

1. Balance, then fit. First weight the data such that each subgroup's weighted distribution matches that of the target population. Then proceed with a simple analysis of survival using the weighted data, ignoring the confounders.
2. Fit, then balance. First develop a comprehensive overall model including all of the confounders. Then obtain predicted values for each combination of confounders, and use these to create population averages of the predicted values.

An analysis might use a combinations of these, balancing on some factors and modeling others. The first approach has a long history in survey methods, with a more recent revival under the rubric of propensity scores, the second can be traced back to (at least) Yates' 1934 paper on analysis of unbalanced data [116].

5.2 Adjusted survival curves

A natural target for population averaging is the survival curve $S(t)$ due to its direct interpretation as an absolute risk. The next sections explore first two variants of the "balance then fit" approach to creating adjusted survival curves (i.e., inverse probability weighting and matching), followed by two "fit then balance" approaches to adjustment (i.e., stratification and model-based

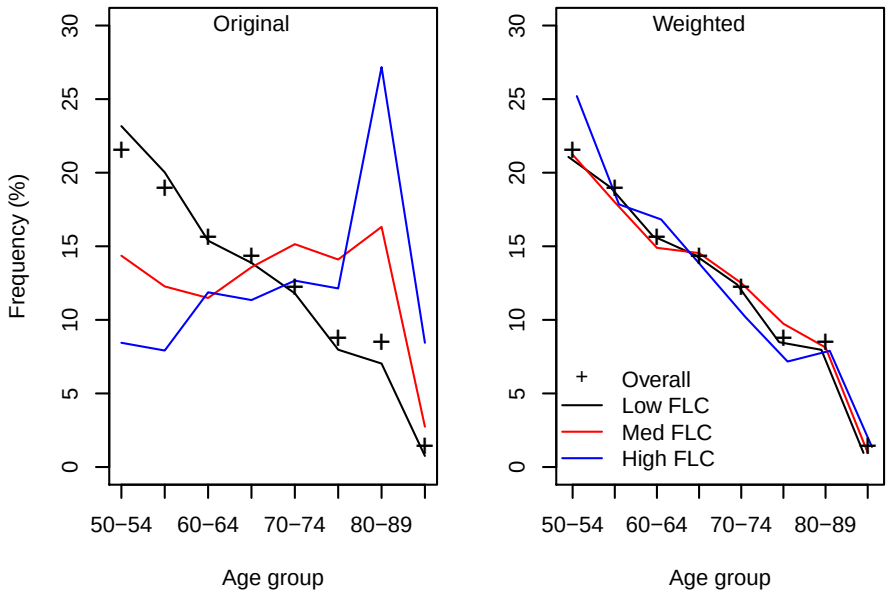

Figure 5.6 *The left panel shows the original unadjusted distribution, and the right panel shows the weighted age distribution using logistic regression models with continuous age, for females, in the low (1), medium (2) and high (3) FLC groups. The overall age distribution is depicted with "+" in both panels.*

adjustment). Any estimators that are a direct consequence of the survival curve such as the RMST or the median time to death will also be available.

5.2.1 Inverse probability weighting

The most direct way to balance the data is to use simple case weights. Using Table 5.1 consider first the age 50-59 stratum. The low, medium and high FLC groups have 46.5 (2587/5560), 28.9 (442/1527) and 16 (121/758) percent of their subjects in the age 50–59 category. The overall percentage of subjects in the 50–59 category is 40.2. Apply the case weights of 40.2/46.5, 40.2/28.9 and 40.2/16 to subjects in the low, medium and high FLC subsets of that column, and then repeat the same calculation and weighting for each of the other three columns will yield a weighted age distribution for each of the three FLC strata that will match the single overall distribution F for the age strata. The resulting age distribution is precisely the *empirical* population described in the previous section of this chapter.

Since the goal of weighting is to balance the ages of the groups, a reasonable check is to compute and plot the weighted age distribution for each FLC group.

Figure 5.6 shows the result. The weighted age distribution is not perfectly

balanced, i.e., the low, medium and high FLC groups do not exactly overlay one another, but in this case these simple weights have done an excellent job.

A more flexible method than this 'by hand' approach to computing the weights is to use three logistic regression models, one for each of the FLC groups. This method will also easily accommodate more than one adjuster. For instance, the code below adjusts for both age and sex. Three logistic regression models will be used because we have 3 FLC groups. However, in the simpler case where there are only two groups, a single logistic fit will suffice, using `phat` and `1-phat` as the predicted values for the two groups.

```
> ipw <- double(nrow(flchain2))  # vector of new weights
> gfrac <- table(flchain2$group)/nrow(flchain2) # fraction in each group
> for (i in 1:3) {
     y <- (as.numeric(flchain2$group)==i)
     tfit <- glm(y ~ agegp*sex, data=flchain2, family=binomial)
     phat <- predict(tfit, type="response")
     ipw[y] <- gfrac[i]/phat[y]
     }
> sfitw1 <- survfit(Surv(futime, death) ~ group,
                   data=flchain2, weight= ipw, robust=TRUE)
>
```

Figure 5.7 shows both the original and IPW population adjusted KM curves. Adjustment for age and sex has removed over half the apparent difference between the curves for the different FLC groups.

As previously introduced, the restricted mean survival time (RMST) provides a convenient one number summary of a KM curve. In the curves above the unadjusted RMST values with $s = 14$ years were 12.4, 10.9 and 7.7. The adjusted RMST values with standard errors are 12.1(0.1), 11.5(0.1) and 9.6(0.2), respectively, which are interpreted as an expected survival of 12.1 out of 14 years for someone in the low FLC group.

Propensity scores

The weighting method just shown is the same approach that is used for calculation of propensity scores. In other words, the creation of weights via the use of propensity scores is equivalent to choosing an empirical population as the reference distribution F. Perhaps the most important part of this realization is that the rich literature on propensity scores provides guidance for more complicated problems [8][9][71]. For instance, if the number of covariates were larger or age more finely divided, then the simple approach could easily fail due to an empty cell and the consequent division by zero. Similarly, outlier covariate values can sometimes result in large weights, which is extensively discussed in the IPW literature [9].

Advantages of the regression framework are that it can easily accommodate more variables, and it can contain continuous as well as categorical predictors. The disadvantage is that such models are often used without the necessary

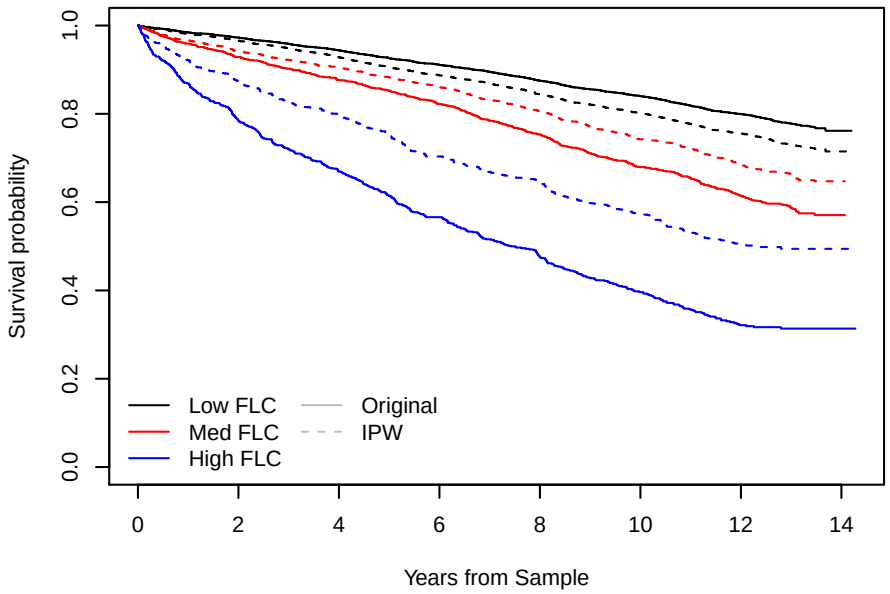

Figure 5.7 *Survival curves for the three FLC groups both before (solid lines) and after (dashed) population adjustment for different age/sex distributions. Adjustment, based on inverse probability weighting (IPW), is to the empirical age/sex distribution of the dataset as a whole.*

work to check their validity. For instance, models with `age + sex` alone make the assumption that the odds of being a member of each group is linear in age and with the same slope for males and females. How well does this work? We emphasize that it is obligatory to check the weighted distributions to ensure balance is achieved, as it is quite common that a simple model is not sufficient and the resulting weight adjustment is inadequate.

Stabilized weights

We have chosen to use the stabilized weights $\pi_i/\hat{p}_i$, where π_i is the overall fraction of subjects in each $i = 1, 2, 3$ FLC group and $\hat{p}_i$ is the fraction of subjects in each age and sex category for each FLC group. The stabilized values have a mean weight of 1 within each FLC group (i.e., they will sum to the original sample size of each FLC group), while the simple IPW weights of $1/\hat{p}_i$ will have a common sum of weights within each FLC group (i.e., the sum of the weights in each FLC group will equal the total sample size). Stabilization has no impact on the individual survival curves or their standard errors, since within each group all the weights have been multiplied by a constant. When comparing curves across groups, however, the stabilized weights can reduce the standard error of the test statistic. This results in increased power for the

robust score test, although in this particular dataset the improvement is not very large.

Standard errors and hypothesis tests

A proper pointwise variance for the IPW curves is obtained by using the robust option, which computes an infinitesimal jackknife (IJ) estimate, and a test of significance can be obtained from a Cox model and the robust score (logrank) test.

```
> coxph(Surv(futime, death) ~ group,
       data=flchain2, weight=ipw, robust=TRUE)
```

Survey sampling

Estimation based on weighted data is a common theme in survey sampling. Correct standard errors for the curves are readily computed using methods from that literature. In R, the `svykm` routine in the `survey` package handles both this simple case and more complex sampling schemes. For simple case weights, the robust variance produced by `coxph`, based on an IJ, is identical to the standard Horvitz-Thompsen variance estimate used in survey sampling [14], which is the basis for the robust log-rank test. The `survfit` routine uses an IJ estimate of variance for either a grouped or robust variance, which will be valid for weighted data. Interestingly, when there is a single observation per subject and case weights are 1, the IJ estimate is identical to the usual Greenwood estimate of variance for a KM estimate A.5. More formally, one can compute the estimate and its variance using the `survey` package in R, which also allows for a much wider range of sampling strategies than simple weighting.

```
> library(survey)
> sdes <- svydesign(id = ~0, weights= ~ipw, data=flchain2)
> dfit <- svykm(Surv(futime, death) ~ group, design=sdes, se=TRUE)
```

Some concluding advice here on IPW, model size, etc? Reprise the literature.

Direct standardization

One classic reference population commonly used in epidemiology studies is the national age/sex distribution at a specific time point (e.g., 2000). The method of *direct standardization* corresponds to the use of π_i based on this national age/sex distribution as the numerator of our IPW [20]. The weighted age/sex distribution for each of the groups will then equal that for the target distribution π. An obvious advantage of this approach is that the resulting curves represent a tangible and well-defined group. For example, we could adjust our curves to match the age/sex distribution of the 2000 US population

using the `uspop2` dataset found in the survival package in R. This exercise is demonstrated in the examples, and barely differs from the adjusted curves found in Figure 5.7. This is not surprising since the overall age/sex distribution for Olmsted County is nearly identical to that for the US as a whole.

5.2.2 Matching

Another approach for balancing is to select a subset of the data such that the age/sex distribution is the same in each FLC group. A common example of this is case-control studies, where one of the groups (i.e., the 'cases') is rare and precious. An ideal sample would retain all of the subjects in this group. A subset that matches their age/sex distribution is then selected from the more common group. Thus, the overall distribution F for the target population is based on the case group. As an example we take a case-control like approach to the FLC data, with the high FLC group as the "cases" since it is the smallest group.

For each age/sex category, the balanced subset will include all patients from the high FLC group, and equal numbers of randomly sampled subjects from the low and medium FLC groups. We cannot *quite* compute a true case-control estimate for this data since there are not enough "controls" in the female 90+ category to be able to select one unique control for each case, and likewise in the male 80-89 and 90+ age categories. To get around this, we will sample with replacement in these strata.

The survival curves for the matched data are shown in Figure 5.8. The curve for the high FLC group is unchanged, since by definition all of those subjects were retained. We see that matching on age and sex has reduced the apparent survival difference between the groups by about half, but a clinically important effect for high FLC values remains. The curve for group 1 has moved more than that for group 2 since the age/sex adjustment is more severe for that group.

In actual practice, case-control study designs arise when matching and selection can occur *before* data collection, leading to a substantial decrease in the amount of data that needs to be gathered and a consequent cost or time savings. When a dataset is already in hand, a disadvantage is that the approach wastes data. Throwing away information in order to achieve balance is wasteful.

One (small) advantage of matched subsets is that standard variance calculations for the curves are correct. The values provided by the usual KM need no further processing. We can also use the usual statistical tests to check for differences between the curves.

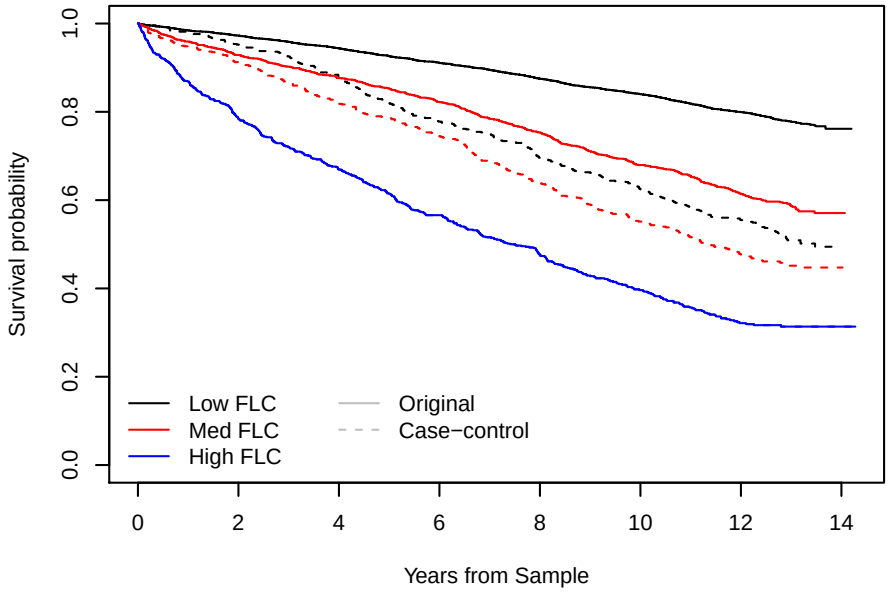

Figure 5.8 *Survival curves from a matched case-control sample are shown as dashed lines, curves for the original unweighted dataset are solid lines.*

5.2.3 Stratification

The simplest "fit then balance" approach is to subdivide the data into homogeneous age/sex strata, compute survival curves within each strata, and then combine results.

Computing KM curves for all the combinations is easy. The resultant survival object has 48 curves: 8 age categories * 2 sexes * 3 FLC groups. To get a single curve for the low FLC group we need to take a weighted average over the 16 age/sex combinations that apply to that group, and similarly for the medium and high FLC groups.

```
> sfits1 <- survfit(Surv(futime, death) ~ group + agegp + sex,
                    data=flchain2)
> temp <- summary(sfits1)$table
> temp[1:6, c(1,4)] #abbrev printout to fit page
                                      records events
group=Low FLC, agegp=50-54, sex=F         738     28
group=Low FLC, agegp=50-54, sex=M         658     45
group=Low FLC, agegp=55-59, sex=F         638     52
group=Low FLC, agegp=55-59, sex=M         553     58
group=Low FLC, agegp=60-64, sex=F         490     52
group=Low FLC, agegp=60-64, sex=M         427     63
```

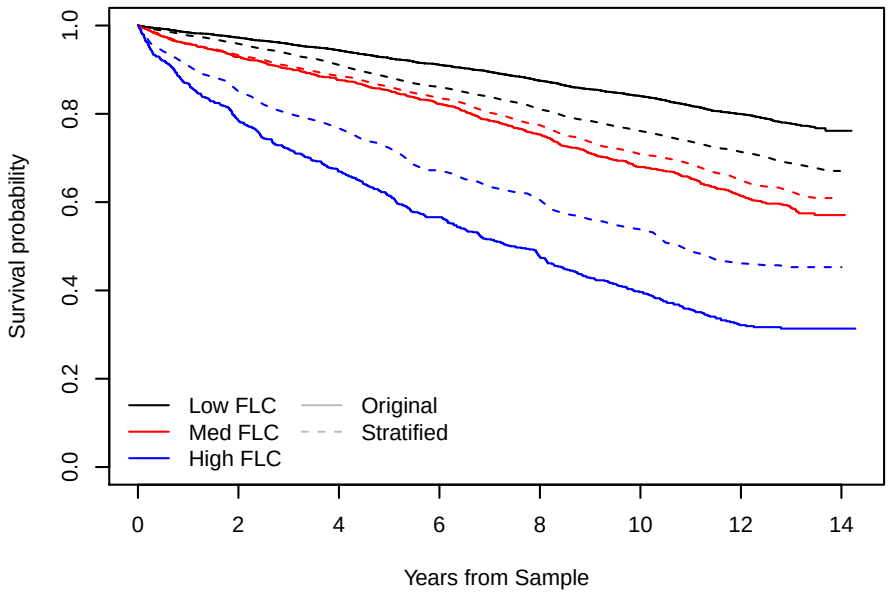

Figure 5.9 *Estimated curves from a stratified approach (dashed lines), along with those from the original data (solid lines).*

Combining the curves is a bit of a nuisance computationally, because each of them is reported on a different set of time points. A solution is to use the `summary` function for survfit objects along with the `times` argument of that function. This feature was originally designed to allow printout of curves at selected time points (6 months, 1 year, ...), but can also be used to select a common set of time points for averaging.

The overall curve is a weighted average chosen to match the original age/sex distribution of the population, shown in Figure 5.9. Very careful comparison of these curves with those obtained using IPW shows that they have almost identical spread, with just a tiny amount of downward shift.

There are two major disadvantages to the stratified curves.

1. When the original dataset is small or the number of confounders is large, it is not always feasible to stratify into a large enough set of groups that each will be homogeneous. The approach also does not accommodate continuous covariates.
2. The overall standard error becomes undefined if any of the component curves fall to zero. Since the curves are formed from disjoint sets of observations, they are independent and the variance of the weighted average is then a weighted sum of variances. However, when a KM curve drops to zero, the usual standard error estimate at that point involves 0/0 and becomes undefined, leading to the NaN (not a number) value in R. In the

above example, this happens at about the halfway point of the graph. Some software packages carry forward the standard error from the last non-zero point on the curve, and others define the variance as 0 when $\hat{S}(t) = 0$, but the statistical validity of either of these is uncertain.

To test for overall differences between the curves, we can use a stratified test statistic, which is a sum of the test statistics computed within each subgroup. The most common choice is the stratified log-rank statistic which is shown below. The score test from a stratified Cox model would give the same result.

Add in example showing too many strata and how this approach can break down when there are some strata with small numbers

5.2.4 Model-based adjustment

Our last approach is to create predicted curves based on a fitted model, also known as "covariate adjustment." The first step is to fit a model including the confounders (e.g., age and sex) and the groups of interest (e.g., FLC groups). The next step is to obtain covariate-specific predictions for each combination of covariates, which will then be averaged. The underlying fitted model must be rich enough to provide such predictions. It is almost required that it include group by covariate interactions, or even to fit a separate model for each group. Variables that have an important effect are retained, where "importance" is based on the size of a covariate's effect rather than its p-value. (Similar advice applies to propensity score models.)

A natural choice for the fits is to use a Cox model. Using the data distribution as our reference, this involves obtaining predicted survival curves for all n age/sex pairs in the dataset. So we will obtain n predictions from the model for every subject in the reference population, assuming they were in the low FLC group, and then take the average of these curves. Then do the same process again with all n subjects in the medium FLC group, and yet again with everyone placed in the high FLC group. Essentially one has 3 copies of the reference population, identical in age and sex distribution, differing only in the variable for which we desire adjusted curves. Computation of these averages is both completely straightforward and deathly tedious. The R survival library includes a `yates` which automates this task, details are given in Section xxx of the examples. The default printout for the function focuses on the RMST estimate rather than the entire curve, for brevity. A plot of the 3 averaged curves is shown in Figure 5.10.

```
> cfit4 <- coxph(Surv(futime/365.25, death) ~ age + sex + group,
                 data=flchain2)
> yate4 <- yates(cfit4, ~group, predict="survival")
> print(yate4)
    group    pmm      std       test chisq df      Pr
  Low FLC 11.826 0.039401      group 167.1  2 < 1e-08
```

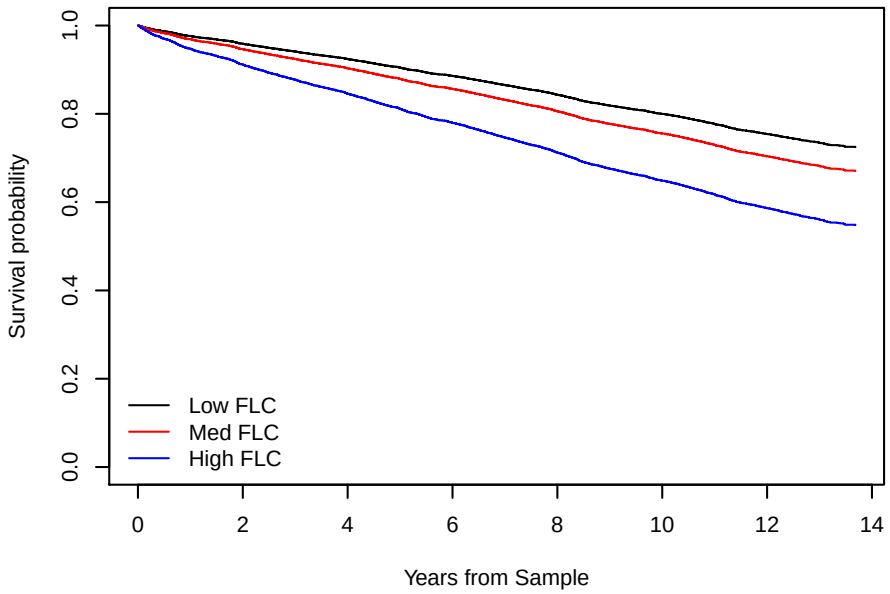

Figure 5.10 *Predicted population curves using a simple Cox model.*

```
  Med FLC 11.365 0.089188
 High FLC 10.271 0.125377

> plot(yate4$summary, col=c(1,2,4), xlab="Years from Sample",
      ylab="Survival")
```

Comparing this adjustment (Figure 5.10) to the IPW and stratified adjustments shown in Figures 5.7 and 5.9 we see that the results are closer together. Also, the estimated RMST values of 11.8, 11.4 and 10.3 for the low, medium and high FLC groups, respectively, are much closer together than the values of 12.0695224, 11.5173698 and 9.6133291 found from the IPW approach.

So where exactly does the model go wrong? Since this is such a large dataset, we have the luxury of looking at subsets. This would be a very large number of curves to plot, so an overlay of the observed and expected curves by group would be too confusing. Instead, we will summarize each group according to their observed and expected number of events, using the same 48 subgroups of FLC by age by sex as the stratified fit (Figure 5.11).

The excess risks, defined as the ratio of observed to expected number of deaths, are mostly modest ranging from 0.8 to 1.5. If the model fits the data well, there should not be obvious patterns in the excess risk. However, for both males and females, the model has substantially underestimated the expected number of events (i.e., overestimated the survival) for high FLC subjects that are less than 70 years old. Since these younger age categories have the largest

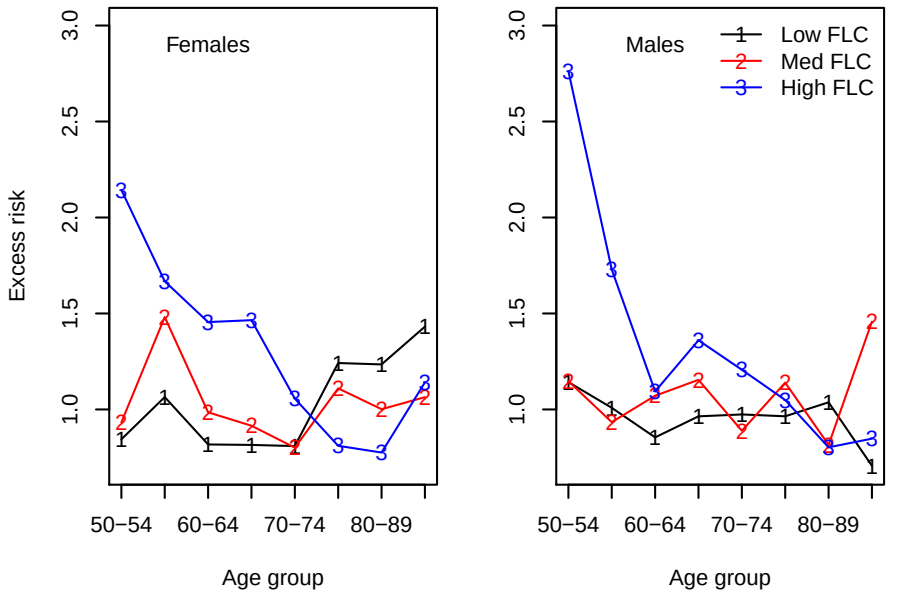

Figure 5.11 *Excess risk, defined as the ratio of observed to expected number of deaths by category, where the expected count is based on predictions from a simple Cox model.*

number of observations in the overall age/sex distribution, this overestimation will have a large impact on the adjusted curve for the high FLC group. There is also some evidence of excess risk in the low FLC group for older females. Altogether this suggests that the model might need to have a different age coefficient for each of the three FLC groups.

The deeper issue is that the simple additive model above was the wrong approach. The raw material for an average predicted value needs to be $\hat{y}$ values that are good estimates of each *per group* survival curve, e.g., the actual survival of FLC high subjects of each age and sex. To do so, it will be almost imperative to include group by covariate interactions, i.e., a much richer model with respect to the grouping variable. These individual estimates will have higher variance, which is then dampened by taking a population average across age and sex. This overall message that a small bias is more important than small variance (and certainly more important than a small p-value), is an echo of similar statements found in the propensity score literature. In the case of a survival model, allowing for non-proportional group effects is also important.

[The IPW in some sense automatically included group * covariate interactions, since each group was a separate fit. In some ways I guess that makes it easier. It's not quite the same to have one model with group 1 vs 2 vs 3, and

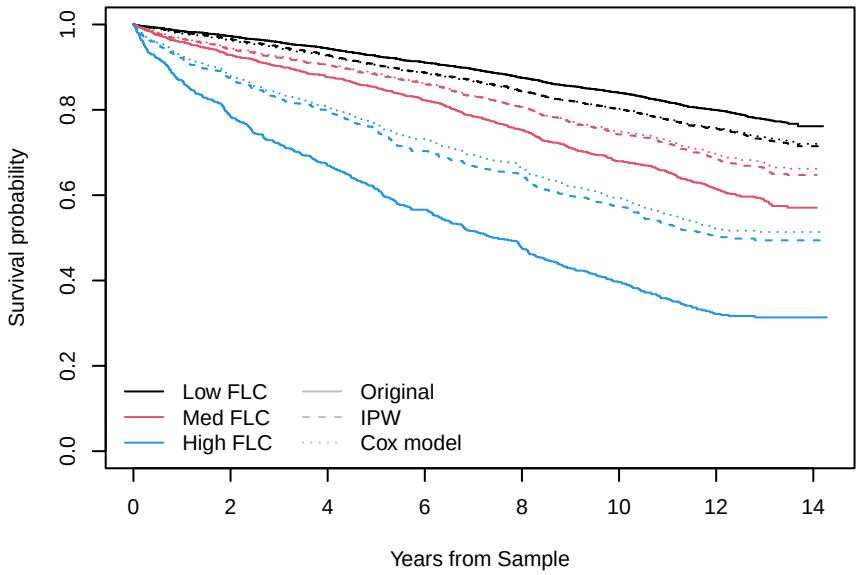

Figure 5.12 *Survival curves by FLC group from two adjustment approaches along with the unadjusted curves. Solid = unadjusted, dashed = adjusted for age and sex using IPW weights, dotted = adjusted for age and sex using the final Cox model including sex and an interaction between age and the group strata.*

two models with 1 vs (2-3), 2 vs (1 and 3), ... and the logic breaks down in the 2 group case. I'm not sure how much discussion to toss in, if any.]

With this guidance we next choose a more focused model of strata of group interacting with age and sex, which can be denoted as `strata(group)/(age + sex)`. The resulting RMST values are 12.1, 11.6 and 9.9. Figure 5.12 shows the resulting adjusted curves, along with IPW curves and the original, unadjusted estimates. The IPW and covariate adjusted estimates now agree fairly closely. A recheck of the observed/expected ratios now shows a much more random pattern, though some excess remains in the upper right corner (not shown). As a footnote we also point out that a fully stratified Cox model, i.e., the 48 separate strata of section 5.2.3 and no other covariates, leads to an identical prediction to what was shown there.

One problem with the model-based estimate is that standard errors for the curves are difficult to estimate. Standard errors of the individual curves for each age/sex/FLC combination are a standard output of the survfit function, but the collection of curves is correlated, since they all depend on a common estimate of the model's coefficient vector β. Curves with ages that are on opposite sides of the mean age will be anti-correlated (i.e., an increase in the age coefficient of the model would raise one and lower the other), whereas those for close ages are positively correlated. A proper variance for the unweighted average has been derived by Gail and Byar [37], but it is complex and has

	Unadjusted		IPW weights	
	Events	Rate	Events	Rate
FLC low	1110	18.6	1355.7	23.3
FLC med	562	39.2	452.4	30.5
FLC high	562	39.2	452.4	30.5

Table 5.2 *Simple counts and rates, before and after adjustment for IPW IPW weights. Rate is expressed as events per 1000 subjects per year.*

not been implemented in any of the standard packages, nor extended to the weighted case. A bootstrap estimate would appear to be the most feasible.

5.3 Event rates

Another candidate for adjustment is the simple event rate. We have previously argued for using this statistic for simple tabulations and overviews of a dataset.

5.3.1 IPW and selection estimates

Table 5.2 shows counts and rates for the original data, and after reweighting each observation using the IPW weights calculated in Section 5.2.1. In the unadjusted data the rates differ by 4.5 fold between the low and high FLC groups; this is reduced to approximately 2.5 fold after adjustment. In the low FLC group subjects at the oldest ages receive an increased weight, reflecting the relative under-representation of low FLC in this age group, while the younger ages get a smaller weight. The effect is to increase the total "effective" number of counts in low FLC group, while a converse weighting in the high FLC group reduces the effective number of events.

A formal test of equivalence of the rates can be based on Poisson regression. For the unadjusted rates the response variable in the model is the 0/1 death indicator, the predictor the FLC group, and `log(futime)` is an offset. For the weighted data a generalized estimating equation (GEE) variant of the poisson model will compute a correct variance.

The data can also be balanced by selection of a matched subset of subjects. As in Section 5.2.2, for this data the overall result is nearly the same as the IPW fit, though less statistically efficient. Because there are no case weights, a formal test to compare rates can be based on simple poisson regression.

5.3.2 Stratification

For a stratified estimate, we first create multi-way tabulations of the data, by FLC group, age, and sex strata, first of the event counts and second of the total years of follow-up. (R code can use the `pyears` routine to both of these at once.) The "model" in this simple case is simply the ratio of events/time, within each cell. A weighted average of the individual rates for

each group is then computed, using the overall age/sex distribution of the study as weights. The final values from this computation are 28.4, 37 and 69.5 deaths per thousand per year for the low, medium, and high FLC groups, respectively.

These values are higher than the IPW results of 23.3, 30.5 and 56.2. A primary reason for this that we are taking the average of a convex function, and Jensen's inequality guarantees that

$$E\left(\frac{X}{Y}\right) > \frac{E(X)}{E(Y)}$$

for any two random variables X and Y, in this case the event counts and the observed follow-up time. The high FLC group has smaller denominators, on average, which makes the effect more acute in that group. Is this statement correct?

5.3.3 Modeling

Per-subject event rates can be predicted using a poisson regression model that has age, sex, and FLC group as covariates. In fact, if these three predictors are categorical and the model includes all interactions, the result will be identical to the stratified model. The poisson fit in that case is a saturated model and predicts each cell by its individual rate. In our example we will use a model with a smooth age term, represented by a spline with 3 degrees of freedom, sex, and FLC group. Based on the lessons learned when using a Cox model to predict individual survival curves, interactions between the variable of interest (FLC) and the other terms are also added. The resulting estimates of 30.2, 37.5 and 68.7 are very similar to the stratified results. The model fit suffers from the same non-linearity inflation of an average ratio versus a ratio of averages.

Because it allows for a varying baseline hazard, a Cox model is much less natural as a model of absolute rates. The strategy is to pick a particular follow-up time, compute individual survival $S_i(t)$ at that time, and then estimate the per-subject hazard as $h_i = -\log(S_i(t))$. Since -log is a convex function we see that averages of h do not behave the same as averages of S.

Why not use models for the absolute number of deaths? You have to force the person-years in each cell to be the same, or it doesn't make a lot of sense.

5.3.4 Population rates

The above has been focused on comparing a sample of subjects within itself, while adjusting for any unbalanced covariates. There is a long history of comparing the event rate within a group to external rates. Carrying our example forward, let μ_{ij} be the 2000 United States death rate for a subject of age i and sex j and m_{ij} the corresponding death rate in the FLC study. The two obvious choices for the population are the empirical age/sex distribution of

the data p_{ij} and the year 2000 United States population distribution π_{ij}. Since the FLC study enrollment was limited to subjects 50 years of age or older π will be used over the same range.

Using the balance $\rightarrow$ model paradigm we first compute a weighed average of the rates and then take the ratio, leading to

$$\begin{aligned} SMR &= \frac{\sum_{ij} p_{ij} m_{ij}}{\sum_{ij} p_{ij} \mu_{ij}} \\ &= \frac{\sum_{ij} d_{ij}}{\sum_{ij} p_{ij} \mu_{ij}} \end{aligned} \tag{5.3}$$

$$CMF = \frac{\sum_{ij} \pi_{ij} m_{ij}}{\sum_{ij} \pi_{ij} \mu_{ij}} \tag{5.4}$$

The standardized mortality ratio (SMR) a long history of use in epidemiology. When the total amount of observation time in each cell is used as the weight p_{ij}, which is a more accurate measure of the amount of information in each cell than the number of subjects, then the numerator of the SMR collapses to the total observed number of deaths, and the denominator is the number of deaths we would expect to observe in a matched population followed for the same duration. The comparative mortality ratio CMF is the same estimate, but adjusted to the age/sex distribution of the reference population.

reference.

Using the model $\rightarrow$ balance paradigm we first compute the rate ratio within each cell and then take the average, leading to

$$\begin{aligned} abc &= \sum p_{ij} \frac{m_{ij}}{\mu_{ij}} \\ def &= \sum \pi_{ij} \frac{m_{ij}}{\mu_{ij}} \end{aligned} \tag{5.5}$$

The abc also has a long history, but precise interpretation of a mean of ratios is acknowledged to be difficult. I think that abc above is related to the indirect estimate on Keiding page 530, but I don't quite have it. Thus the vague label. It certainly is related to the 'ratio of averages or average of ratios' discussion on page 537. Keiding calls the balance then model direction a marginal approach, and model then balance a conditional approach. Perhaps its best

5.4 Other estimands

The population average methods given above can certainly be applied to other quantities from the model, e.g., computing a population averaged hazard ratio or log(hazard ratio) for a Cox model. However, the predicted survival curve and quantities derived from it have natural advantages of decomposition and interpretability. Consider a mixture distribution $F_0 = \sum w_k F_k$ where F_k are the underlying populations. Then the survival curve for F is also decomposable as $S_o(t) = \sum w_k S_k(t)$. Because of this the balance$\rightarrow$fit and fit$\rightarrow$balance

estimators of S estimate essentially the same quantity: the first the survival of the mixture distribution and the second a mixture of survivals. This equivalence makes the result easier to interpret for both cases. The same holds true for summaries that are derived from S, such as the median survival or the RMST.

Hazards and hazard Ratios

A deeper look at the hazard reveals some of the issues. The hazard $\lambda_0(t)$ of a mixture population has a time-dependent value

$$\lambda_0(t) = \frac{\sum_k w_k S_k(t)\lambda_k(t)}{\sum_k w_k S_k(t)} \tag{5.6}$$

$$\leq \sum w_k \lambda_k(t) \tag{5.7}$$

A population mean of the hazard (5.7) is easy to calculate but difficult to interpret, since the computed average is equal to the true population hazard (5.6) only for a moment at time $t = 0$. From that point forward the high risk subpopulations die faster, i.e., the mixing weights change. It has the further non-intuitive property that changing a treatment covariate from (0=control, 1=treatment) to (1=control, 0=treatment) reciprocates the HR for the Cox model from r to $1/r$, but does not change the ratio of average population hazards in the same way. The PMM of the log-hazard $\hat{\beta}$ does not share this second flaw, but remains difficult to describe. What exactly does the quantity represent? The hazard ratio also turn out to be problematic for causal modeling [25], while absolute estimates are appropriate for that task.

Odds ratios

The same argument can be made for logistic regression, ofttimes a companion to survival models. For population average values of the absolute risk $\hat{p}$, the IPW and PMM approaches target the same quantity. The IPW of the odds ratio depends on underlying prevalences, and population averaged estimates of both the odds ratio and log odds are very difficult to interpret.

Type III tests

In an interesting 1934 paper Yates [116] proposed a solution to a linear models problem with unbalanced data and interactions, consisting of a weighted sum of predicted values. The issue he addressed is essentially that of figure 5.4, and his solution exactly the fit→balance algorithm, using the linear predictor $\eta = X\beta$ as the statistic and a full factorial design as the population. This approach persists in the form of *least squares means* (population averages of η) and *type III tests* (tests of equality between those averages) [**?**]. Though related to this approach they have three significant problems

1. The target statistic η is difficult to interpret outside of a linear model. For a

Cox model for instance this estimates log(hazard ratio). What does a mean of log hazard ratios represent?

2. A full factorial is used as the reference population, which is often uninteresting: an answer to the question that nobody asked.
3. Mystery. Given the computational constraints of the time, Yates could not do a brute force computation of all nk = number of subjects x number of groups predicted values, and needed to collapse the test into a more manageable shortcut. Several variants of this survive, but most users have no intuition about what they mean. Some of the currently used formulas are wrong.

5.5 Connections to other work

Keiding and Clayton [56] provide a comprehensive review of standardization, starting with the SMR and tracing the ideas forward. The use the label 'marginal summaries' for the balance $\rightarrow$ model approach and 'conditional' for the model $\rightarrow$ approach, revealing that each has been rediscovered multiple times and how they relate to other approaches. Debate about the meaning and interpretation of conditional estimates such as (5.5), an *average of ratios*, has a particularly long history.

In the modern age the issues have re-emerged in area of causal modeling. In the potential outcomes framework, each subject has a potential outcome under each treatment, with $Y_i^{(A)}$ the outcome for subject i under treatment A and $Y_i^{(B)}$ the outcome under treatment B. Since each subject only received one of the treatments, only one of these outcomes is observed, yielding the data element y_i. The outcome for the other treatment is unobservable, and is referred to as the counterfactual. Then $d_i = Y_i^{(A)} - Y_i^{(B)}$ is the causal effect of treatment A, treating B as the null or '0' treatment, and $E(d)$ is the overall causal effect. The ideal way to assess the average causal effect $E(d)$ is via a controlled clinical trial. Of course this is not always possible, and assessments using observational data are needed. In this literature the same two practical estimates turn up, weighting in the form of IPW and averaged predictions in the form of g-computations [56].

This connection helps further clarify the modeling for a PMM estimate. An assumption for the causal approach is that d_i can be approximated by $\hat{d}_i = \hat{y}_i^{(A)} - \hat{y}_i^{(B)}$, two predictions for y_i that differ only in the treatment assignment. For this to work, the model needs to have a differential treatment effect which depends on the values of the other variables, i.e., at least some treatment by covariate interaction terms.

5.6 Conclusions

Population averaged estimates are a useful addition to our statistical toolbox. When employing them the two important questions to answer are what statistic to average and what population to use as reference.

With respect to which statistic to use, absolute risk values are the most interpretable and useful. Population averages of both HR and log hazards are difficult to grasp and others have raised deeper theoretical concerns. Perhaps as importantly, the HR, though extremely useful, has been overused to exclusion. Hernan's clever title "The hazards of hazard ratios" [48] points out the flaws of relying on them exclusively. We need to add an absolute risk viewpoint to our summaries.

In terms of which population to use, the primary answer is to be clear about what question you want to answer. Using the observed dataset as the population is the most common, although one valid critique of this is the shifting sands aspect: if an analysis is run multiple times during the course of a trial's accrual, each will have a different reference. Studies from two institutions are likely to vary even more. The use of a fixed factorial population addresses this, but that population will often be completely unsatisfactory. In the FLC example, this reference population would have equal numbers of subjects in the 50–54, 55–59, ..., 90+ age groups. In reality such a population does not now and never will exist; how could averages over such a population be of any interest? There are a few cases where the factorial population makes complete sense, such as a balanced laboratory trial which has become imbalanced through loss of a few cases, but these are a small minority. External references, such as the US population, are excellent in the few cases where they apply, but will often not include all of the covariates that we wish. In the end, the observed data may be the "least worst" choice. Senn has pointed out the connection between this population and random effects models [95].

When two populations need to be adjusted and one is much larger than the other, the balanced subset method has been popular. It is most often seen in the context of a case-control study, with cases as the rarer group and a set of matched controls selected from the larger one. This method has the advantage that the usual standard error estimates from a standard package are appropriate, so no further work is required. This approach may be of particular interest when the distribution of the adjusters is very different from the general population (e.g., a study of patients with lupus, which predominately affects women), where adjusting to the overall age/sex distribution of the US population may not be of interest. However, in the general situation it can lead to a correct answer but for the wrong problem, i.e., for a population in which we are not interested.

The population weighted estimate is flexible, has a readily available variance in many statistical packages (but not all), and the result is directly interpretable. It is the method we recommend in general. The approach can be extended to a large number of balancing factors by using a regression model

to derive the weights. Exploration and checking of said model for adequacy is an important step in this case. The biggest downside to the method arises when there is a subset which is rare in the data sample but frequent in the adjusting population. In this case subjects in that subset will be assigned large weights, and the resulting curves will have high variance.

Risk set modeling is a very flexible method, but is also the one where it is easiest to go wrong by using an inadequate model; also variance estimation is more difficult. To the extent that the fitted model is relevant, it allows for interpolation and extrapolation to a reference population with a different distribution of covariates than the one in the training data. It may be applicable in cases such as rare subsets where population weighting is problematic, with the understanding that one is depending heavily on extrapolation, which is always dangerous.

Inverse probability of censoring (IPC) weights

There is a close relationship between the redistribute-to-the-right (RTTR) algorithm and inverse probability of censoring weights (IPC) 1/G(t). The latter are often created by invoking the Kaplan-Meier function with a reversed status of 0=event and 1=censored. That is, compute $G(t)$ as the KM for *censoring* rather than death, and use weights of $1/G$ for the events and 0 for censored observations. One immediate advantage of this approach is that it gives an avenue to address informative censoring. For instance, say that for a particular study censoring was related to treatment arm. One can fit per-treatment estimates of G, and reweight each event by the appropriate G for that event. An extension of this approach using a multivariable model for $G(t)$ plays an important role in marginal-structural models.

There is a technical detail, often overlooked, which is necessary to make this approach equivalent to the RTTR. If there are any tied censoring and event times, e.g., times 2 and 5 in the small example above, all survival methods assume that the censoring occurs after the event. This is a reflection of how data is gathered; if subject Smith is observed to still be alive on day 108, and Jones perishes on day 108, we nevertheless know that Smith dies after Jones, even though the recorded time values are the same. When the status variable is switched from 0/1 to 1/0 in order to compute G, it is necessary to preserve this ordering. A second hazard occurs when a subject's observations have been split into multiple rows, e.g., for a time-dependent covariate; the simple algorithm can mistakenly treat all of the intermediate rows as a censoring event.

It is also important to note that $G(t)$, like the at-risk indicator $Y_i(T)$, should be left continuous, whereas the ordinary survival curve is right continuous, weights should be based on $1/G(t+)$. If all these details are carried out correctly, then the sum of weights before and after IPC reweighting will be the identical. Much statistical software is not cognizant of these issues, however, and uses the naive "recode censoring" algorithm. Fortunately, for many

datasets the total number of such ties is small and the subsequent error often ignorable. (The censor-after-death argument breaks down when time has been coarsened into intervals. For example, some datasets in the R survival package have time rounded to months, in order to preserve subject anonymity. We do not know the relative ordering of subjects within the same month.)

If the dataset included delayed entry for some subjects, then the exact equivalence between the RTTR, Kaplan-Meier estimate, and IPC weights breaks down. As well, the best approach to IPC weighting is no longer completely clear. One approach is to give a weight of $G(t_{i0}+)/G(t_i+)$, where t_{i0} is the entry time for subject i. (If a subject is broken into multiple intervals, this is not the starting time of the most recent interval, but of the first interval). If a give subject has a hole in their risk period, e.g., intervals of (0, 10) and (20, 50), then we have no suggestions for a proper IPC weight.

Chapter 6

Developing and evaluating prognostic risk models

Risk prediction models are becoming commonplace in the medical literature, reflecting the fact that a multitude of clinical care decisions are made every day based on the probability of disease or the probability of future events. The methodology for deriving and assessing risk scores has received much attention in the statistical literature [98, 49, 19, 64, 27]. In fact, guidelines for the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) have been published [35].

6.1 Developing risk scores

There are many issues to consider when developing a risk prediction model and entire books have been written on this subject [99, 111]. This brief section will introduce the issues to consider from a time-to-event perspective.

Purpose of risk prediction

It is important to clearly understand the purpose of the risk prediction model and how it will be used before developing it. For example, if the purpose of a cardiovascular disease risk score is to aid in medical decision-making regarding whether to prescribe statins to lower cholesterol, then it is important to consider what to do with data from patients who are already taking statins or other medications to lower their cholesterol. For this reason, historic cohorts with data collection prior to the availability of statins are often used to develop the risk scores. These cohorts have the advantages of long follow-up and do not have any biases related to statin use. However, the cardiovascular disease event rates in these cohorts may not reflect current event rates, as there have been many advances in recent years in addition to statins that have improved cardiovascular disease outcomes [117]. Another option would be to use more contemporary data and exclude (i.e., censor) subjects at the time when statins were initiated, but this informative censoring could equate to systematic exclusion of higher risk patients, which would result in unrealistic cardiovascular disease event rates.

Population of interest

It is also important that the data used to develop a risk score is representative of the patients who will be assessed by the risk calculator. For this purpose, population-based data may be preferred to ensure that the entire spectrum of risk is represented in the data. Referral bias can have an unpredictable impact on the estimated risk rates. Patients who are referred to a specialist may have more severe disease that could not be adequately managed by their general practitioner, or they could be patients who are well enough to travel to see a specialist. Thus, referral bias may lead to either over- or under-estimation of the true risk rate in the general population. Electronic health records are increasingly being used to develop risk scores for patients seen at a particular institution. These may be useful for that particular patient population, but might not be generalizable to other institutions who see patients with different demographics or different disease severity.

Outcome definition

The choice of outcome for a risk score influences its applicability and interpretation. The outcome should be associated with serious consequences, so that it is something clinicians and patients care about and wish to avoid. In addition, the outcome should be defined in a standardized way that is objective and easily applicable to external populations [26]. For these reasons, "soft" outcomes (e.g., angina) are less desirable. "Hard" outcomes (e.g., myocardial infarction, stroke or fatal events) may be preferred. The Systematic COronary Risk Evaluation (SCORE) developers chose to include only fatal events to maximize portability to other populations, as fatal events were thought to be the easiest to define [24]. However, this resulted in low calculated risk estimates (i.e., $< 5\%$ chance in the next 10 years), which may not be high enough to convince patients to undertake lifestyle modifications like weight loss or smoking cessation. Furthermore, rare outcomes are often difficult to predict. For example, a model to predict an outcome occurring in 2% of patients would have 98% accuracy if it predicted 100% of patients would not experience the event. Depending on the strength of the associations between candidate risk factors and the outcome, it may not be possible to build a model that improves on just predicting "no one gets the event".

Time horizon

The time horizon of the risk score is also an important consideration. The majority of cardiovascular disease risk scores compute a 10-year risk estimate because atherosclerotic changes accumulate over several years before an event occurs. Thus, preventive measures need to start long before the event will occur in order to be effective. However, 10-year risk estimates are low for young patients (age < 50 years), so 30-year or lifetime risk estimates may be preferable for risk assessment in the young [68, 67]. Disadvantages of lifetime risk estimates include the higher likelihood of inaccuracy due to future changes

in risk, and the need for long-term data for derivation, which may not reflect contemporary event rates by its very nature.

The available length of follow-up data is another consideration for the time horizon. If the available follow-up is not sufficient to develop a risk score with a clinically relevant time horizon, then a risk score should not be developed. For example, a 1-year cardiovascular disease risk calculator would not be clinically useful, because the 1-year risk estimates would be too low in most patients to justify pharmacologic treatments or lifestyle modifications. Also, those with a high 1-year risk may be easily identifiable without a risk score.

Once the time point for the risk score has been selected, it is often advantageous to truncate the follow-up in the dataset to a time point soon after the time point at which the risk estimate will be calculated. For example, the Framingham risk score, which calculates risk at 10 years uses a dataset for model-building that includes follow-up through 12 years. This approach provides a stable estimate at 10 years by not cutting off the follow-up too close to the time point of interest. It also minimizes issues with risk factors losing their predictive strength over time. While it is always tempting to make use of all available data, risk factors measured at baseline are often stronger predictors of short-term outcomes and have less ability to predict outcomes occurring long after they were measured. So including 20 years of follow-up in a model intended to predict 5-year outcomes can result in underestimated coefficients for risk factors of interest.

Risk factor selection

Risk scores are predicated on the notion that multiple risk factors may interact to increase the risk of developing a disease. Patients with modest increases in several risk factors may be at an equivalent or higher risk of disease than patients with one highly elevated risk factor. Therefore, clinical management decisions should account for the constellation of risk factors in order to optimize preventive strategies. Considerations for potential risk factors should include whether the risk factor can be objectively measured and whether its measurement is reliable and reproducible. In addition, the cost of recording the risk factor in terms of both money and clinician time/effort should be considered. Furthermore, causality is important, as risk factors included in a risk score are assumed to result in lower risk when they are improved. However, causality is difficult to prove except in randomized clinical trials.

Sample size considerations

Risk score development requires a large dataset with a large number of events. The usual rule of thumb of 10 events per variable is inadequate for developing a risk prediction model [81]. A recent article recommended a minimum of 200 patients with events and 200 patients without events for development and validation of a risk score [19]. When risk scores are developed in smaller datasets, it is often difficult to properly assess non-linear effects and interac-

tions to ensure optimal fit of the risk factors, and it is also more likely that the model will be over-fit and will suffer from over-optimism, which results in reduced predictive ability when the risk score is applied to external data [94].

For rare diseases, single centers may have too few patients to develop a risk score. However, combining data from multiple centers can be challenging. Retrospectively collected data is often used to avoid waiting for follow-up to accrue in a prospective study, but it is difficult to ensure that study-specific factors (e.g., variable definitions, surveillance methods, identification of subjects, participation rates, etc.) are similar across centers. Assessment of heterogeneity across centers is necessary. Registries often include larger numbers of patients than single center studies and are thought to be less heterogeneous due to standardized data collection procedures, but selection bias for enrolled subjects is an important issue that is often ignored. Heterogeneity can make it difficult or even impossible to develop a risk score depending on whether the sources of the heterogeneity can be adequately explained and counteracted or not.

Model specification

Cox proportional hazards models are the most common choice for development of a prognostic risk score because they do not make any assumptions about the functional form of the underlying baseline hazard function. However, the proportional hazards assumption must be checked when using Cox models.

Missing values in potential predictor variables should be examined and imputation methods should be considered, if appropriate based on the mechanism of the missingness. For continuous risk factors, categorization and functional form are also important considerations. Continuous variables contain more information than categorized versions, so categorization of continuous variables is not recommended. However, many risk factors, such as blood pressure and lipids, have commonly used cut points for categorization that have been recommended in national and international guidelines for treatment. Functional form must be assessed for continuous risk factors, as variables may have non-linear effects, such as J- or U-shaped relationships with the outcome of interest. Smoothing splines or restricted cubic splines can be used to model variables with non-linear effects.

Stepwise selection methods are a widely used and long-standing method to select a subset of variables for inclusion in a risk score. However, these methods have many disadvantages including over-estimation of the magnitude of the regression coefficients [97]. Modern techniques, such as shrinkage, have been developed to try to limit over-fitting of a model. One such method is least absolute shrinkage and selection operator (LASSO) [107]. Machine learning methods, such as random forests and gradient boosting models, have also been adapted for use with time-to-event data. These methods may provide better predictions, but often have the disadvantage of not providing a formula for computation and to aid in understanding of the influence of the risk factors

and the interplay between risk factors in the model. Interactions between potential predictors should also be considered in the risk model development.

Absolute risk complexities

Risk scores typically estimate absolute risks (e.g., 5% risk of cardiovascular disease in the next 10 years) instead of relative risks (e.g., a 50-fold increased risk of cardiovascular disease among patients aged <40 years with systemic lupus erythematosus compared to similar persons without SLE) [70]. The relative risks associated with particular risk factors may be very similar across different populations (e.g., the 1.5-2 fold increased risk of cardiovascular disease among patients with diabetes), but the absolute risk levels may vary between countries or over time. Cardiovascular disease risk scores developed for the general population in a particular country often do not perform well in other countries [17]. Efforts to re-calibrate risk scores for use in other populations have sometimes been successful [66, 47]. Risk scores designed for use in the general population may not work well in certain sub-populations, such as minorities or rare high-risk groups (i.e., patients with HIV). Finally, the presence of competing risks should also be considered. Ignoring competing risks can yield risk estimates that overestimate the true risk of an outcome. We will revisit this issue in a later chapter.

Dynamic prediction

Another issue is that risk prediction models are often constructed using risk factor data available at a single 'baseline' timepoint, which could correspond with a date of diagnosis or of study entry. However, in practice, risk predictions are desired at multiple timepoints, perhaps at every visit or during each annual check-up. Obtaining these "dynamic predictions" requires incorporation of risk factor data that varies over time (i.e., time-dependent covariates). As discussed in Section 4.5.1, landmarking can be useful for obtaining absolute risk predictions in the presence of time-dependent covariates for a cohort of subjects beginning at a specific timepoint. The concept of landmarking can be exploited by using truncation and administrative censoring along with a sliding time window to create multiple landmarking intervals for each subject. For example, if building a 10-year prediction model, a subject who has 20 years of follow-up could contribute multiple assessments to the model, starting with baseline and ending at 10 years, then starting with 1 year and ending at 11 years, and so on. All of these intervals can be stacked to create a 'super prediction dataset' [111]. This dataset can be analyzed using standard survival software with the `cluster(id)` option to obtain a robust variance accounting for the dependency of the multiple observations per subject. More details of this approach, as well as other ways to obtain dynamic predictions, can be found in van Houwelingen and Putter's book [111].

6.2 Validating risk scores

Validation tries to answer the question of how well a particular model will work in practice. While validation is a step that comes after risk score development, it is useful to consider how the model will be validated ahead of time in order to adequately prepare for this step, as it may require partitioning the dataset before beginning model development or procuring an external dataset for validation.

We will illustrate model validation concepts using the datasets from two breast cancer studies that are discussed in Royston and Altman [87]. The dataset used to create a risk score is comprised of 2982 primary breast cancer patients whose records were included in the Rotterdam tumor bank, 1546 of whom had node positive disease. The outcome, recurrence free survival time, is defined as the time from initial diagnosis until the earlier of disease recurrence or death from any cause. The final Cox model using the Rotterdam data includes age, menopausal status, tumor size, a linear effect for the number of positive lymph nodes up to 10, and tumor grade (Table 6.1). The main time points of interest are 3, 5 and 10 years following diagnosis with underlying baseline survival estimates of 0.687, 0.564, and 0.394. The validation dataset contains results of a trial in primary breast cancer conducted by the German Breast Cancer Study Group (GBSG); 720 patients with node positive breast cancer were recruited into a study with a 2x2 design to investigate the effectiveness of 3 vs. 6 cycles of chemotherapy and of additional hormonal treatment with estrogen. The available dataset contains 686 subjects who have full information on a set of prognostic factors.

term	coef	se(coef)	p.value
size20-50	0.282	0.055	<0.001
size>50	0.455	0.084	<0.001
grade	0.313	0.060	<0.001
pmin(nodes, 10)	0.131	0.007	<0.001
pmin(age - 55, 0)	-0.024	0.005	<0.001
pmax(age - 65, 0)	0.031	0.006	<0.001
meno	0.278	0.084	0.001

Table 6.1 *Coefficients for the prognostic model based on the Rotterdam data*

When there is no validation data, the risk score approach is susceptible to overfitting. This can be illustrated by way of a model fit using multiple random variables. Figure 6.1 shows fits to the Rotterdam dataset; the left panel uses predictions from the Rotterdam model, and the right panel uses predictions from a model with 120 random binomial variables. (This latter is within the common rule of 10–15 events per covariate.) Each figure shows the predicted survival from the Cox model along with the Kaplan-Meier for that tertile. Why does the random covariate fit look good even when it should not? First of all, the Cox model's flexible baseline hazard will ensure that the

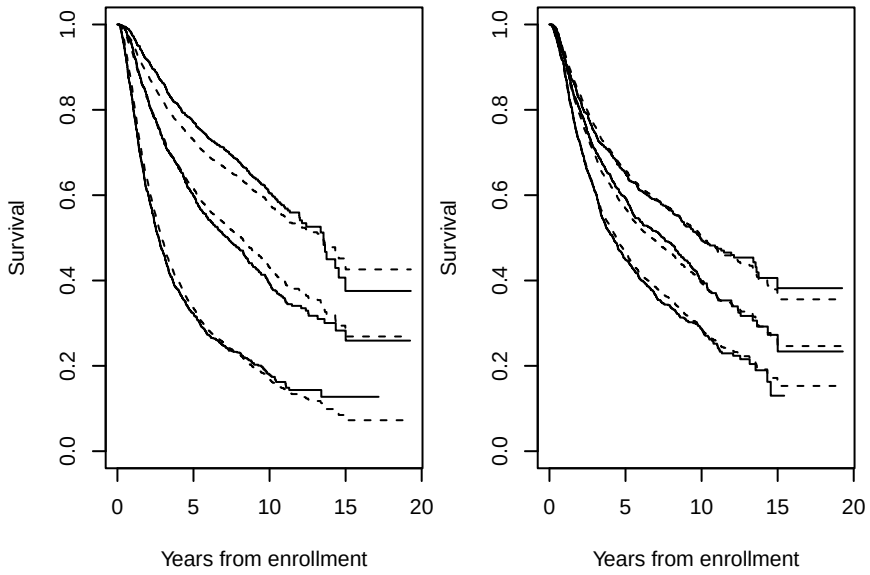

Figure 6.1 *Observed survival curves for the Rotterdam dataset (solid lines), along with predicted curves from a Cox model (dashed lines). Left panel: a fit using the 6 variable model explored earlier. Right panel: a fit with 120 random binomial predictors.*

mean of the predicted curves and the mean of the 3 Kaplan-Meier curves will agree at every time point: the average is always right. This invariably leads to an eyeball test of "looks about right". The only possible difference is one of shape: the three predicted curves must have proportional hazards but the KM curves might not. This approach of comparing curves from a post-fit partition of the data is not worthwhile.

6.2.1 Internal validation

Internal validation refers to validation using the existing cohort. It can be performed by dividing the original cohort into derivation and test sets. This approach is common, but is sub-optimal unless the original cohort is extremely large because it reduces the available sample size for model development [100]. Cross-validation techniques and bootstrap sampling are better choices because the majority of the data are used for model development. For example, in 10-fold cross validation, 10 models are developed with one decile of data removed from each model fit. Thus 90% of the data is used for each model and predictions are obtained on the remaining 10% to assess model performance. Bootstrap sampling involves sampling with replacement to achieve multiple datasets with the same sample size as the original dataset. Therefore, each model is developed using data with the same sample size as the original data

set. However, subdividing the data into derivation and test sets is becoming the preferred method because it seems more intuitive to readers and because it is easier to perform correctly. Cross-validation methods for model validation are more difficult to perform correctly, as they need to be kept separate from cross-validation used for variable selection. This often requires an "inner" cross-validation for variable selection and an "outer" cross-validation for model validation [85].

While internal validation seeks to simulate external validation, the fact that the original dataset is subdivided for validation helps to ensure comparability and reduces the need for thorough data checks.

6.2.2 External validation

External validation involves assessing the performance of a risk calculator in a different cohort than was used for its derivation. External validation is considered to be a stronger assessment of performance of a risk score than internal validation, because it addresses transportability to another cohort of patients, and not just reproducibility in the same cohort used for its development. Different types of external validation include temporal validation (i.e., studying patients who were treated more recently), geographic validation (i.e., studying patients from other geographic areas) and full external validation (i.e., studying patients treated in fully different settings).

There are multiple considerations to ensure comparability and meaningful model assessments when using external datasets for validation.

Covariate checks

Any validation should start with simple checks of the covariates. Normally, summaries of the variables in the prognostic model are provided in an initial table, including counts for categorical predictors and the mean and standard deviation for continuous ones. If, for instance, a lab measure were reported in mg/dl in the original dataset and mmol/L in the validation dataset, then that would cause prediction errors. Comparisons of the Rotterdam and GBSG variables are shown in Table 6.2. We see that the distributions of age, size and menopausal status are very similar. However, in the Rotterdam data nearly half the subjects are node negative while none are node negative in the GBSG cohort, and the Rotterdam subjects have systematically higher grade. This last bears further scrutiny as there are multiple grading scales available. We might argue that grade should be omitted, but remember that the development model will often have appeared in an earlier publication and modifications of the model will not be possible. In this case we proceed, but with caution.

	Rotterdam (N=2982)	GBSG (N=686)
Postmenopausal	1670 (56%)	396 (58%)
Age		
... <40	412 (14%)	73 (11%)
... 40-50	796 (27%)	216 (31%)
... 50-65	1047 (35%)	331 (48%)
... 65-80	679 (23%)	66 (10%)
... >80	48 (2%)	0 (0%)
Size		
... <20	1387 (47%)	180 (26%)
... 20-50	1291 (43%)	453 (66%)
... >50	304 (10%)	53 (8%)
Grade		
... 0	0 (0%)	81 (12%)
... 1	794 (27%)	444 (65%)
... 2	2188 (73%)	161 (23%)
Nodes		
... 0	1436 (48%)	0 (0%)
... 1-5	1008 (34%)	474 (69%)
... 6-15	477 (16%)	175 (26%)
... >15	61 (2%)	37 (5%)

Table 6.2 *Comparison of Rotterdam and GBSG covariates*

Event rate checks

A second important check is to look at the overall Kaplan-Meier curve for the validation data and compare it to the development data's curve (see Figure 6.2). The most serious issue would be a change in scale, e.g., one study has time in years and the other in months, which is easily correctable. A major change in the event rate may indicate a calibration problem unless there is also a shift in the risk predictions due to a different case-mix.

It is also important to make sure that the validation data has enough follow-up to correspond to the prognostic model's main timepoints. The GBSG dataset has adequate follow-up for a 3 or 5-year prediction score, but has no follow-up at 10 years. In fact, by 7 years there are only three subjects under observation. If the development paper showed grouped curves, by tumor size say, it will be easiest to mimic that breakdown in the validation data. But we emphasize that we do not view this comparison as a formal validation, but simply an important data check before validation.

Limiting the time scale

Another important issue to consider is limiting the time scale. When assessing a particular survival model, we should first ask over what time period the model will actually be applied. If, for instance, the model will be used to

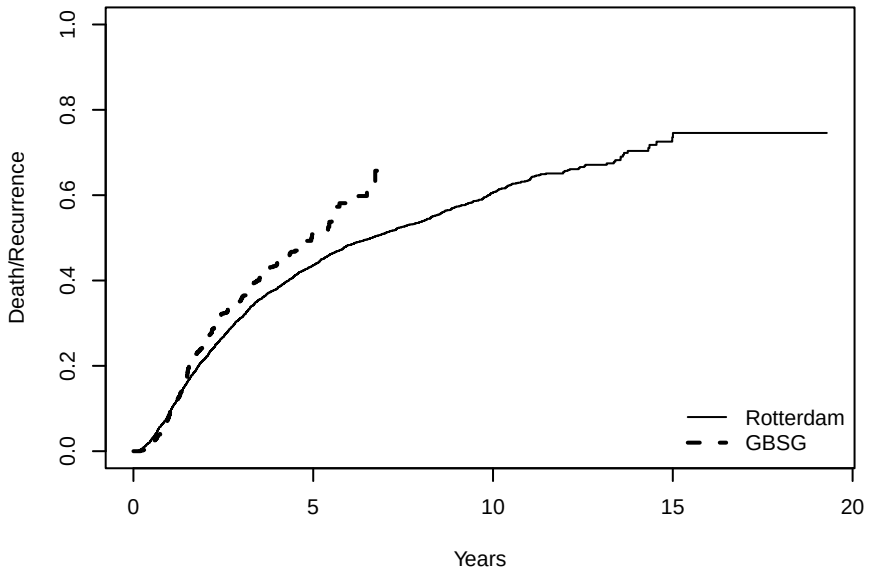

Figure 6.2 *Comparison of time to death or recurrence using the Rotterdam and GBSG datasets*

assign higher risk subjects to a more aggressive treatment, where by high risk we mean more than a 50% chance of failure within 2 years, then whether the risk score can correctly order two subjects whose failure times are at 4 vs. 5 years is essentially immaterial. A second example would be a prediction that is based on laboratory values from each patient, values that will be reassessed at each yearly visit. The risk score based on today's values might perform very poorly 4 years from now simply because those measurements will be badly out of date by that time. More importantly, no sensible physician would use 4 year old values in preference to current results: an old risk score becomes medically irrelevant after 1–2 years. The underlying principle is to ensure that our measure of model performance is *fit for purpose*: it answers a question we actually care to ask. For survival data, this will often mean restricting the follow-up to times less than some upper limit τ; this can be done by treating values $> \tau$ as censored.

Baseline hazard

A key issue in validating a published Cox model is that the baseline hazard is normally not reported, and without the baseline hazard an evaluation of calibration is not possible: calibration requires estimates of absolute risk. The same problem would occur in a logistic or linear model if the intercept β_0 were not reported; in a Cox model, the baseline hazard $\lambda_0(t)$ plays the role of an intercept. If the report includes one or more predicted survival curves,

however, an approximate baseline hazard can be recovered; the baseline hazard was after all required for creation of the presented curve. Limited progress may be possible if only an unadjusted Kaplan-Meier curve is included.

Linear predictor

Another check that is sometimes done is to compute the linear predictor (LP) $\eta_i = X_i\beta$ using the covariates X_i for each subject in the risk set along with the coefficients β from the prediction model, and display these as a histogram. The value η is also known as the prognostic index. However, since such a figure is rarely if ever available for the development data, its utility is questionable.

6.2.3 Performance measures

Validation tries to answer the question of how well a particular model will work in practice. There is an important asymmetry between the model and its development data on the one hand, and the methods and data used for validation on the other. For this task, the assumptions and possible uncertainties in the model are of *no concern whatever*. The goal is only to understand how well said predictions perform, irrespective of whether they came from a sophisticated statistical model or the daily horoscope. With respect to the actual validation metrics, on the other hand, we care deeply about any assumptions. We would like to make as few of them as possible, and understand the possible consequences of any assumptions that we do make.

One of the key tenets of validation is a desire to assess if the fitted model is both useful and reliable in the new dataset. The two primary measures used to assess the performance of a risk prediction tool are discrimination and calibration. Discrimination is the extent to which the risk estimates from the model separate observations into groups with different prognoses. Observations with a higher predicted risk should exhibit a higher event rate and those with lower predicted risk a lower rate. Calibration measures prediction accuracy. A model might exhibit high discrimination and miscalibration, e.g., two groups with predicted rates of 10 and 20% have actual event rates of 15 and 30, or it could have calibration without discrimination, e.g., all groups are predicted to be at 25%, which happens to match the overall observed rate. A model that has discrimination without calibration may still be useful (e.g., to divide subjects into low vs. high risk for stratified treatment assignment in a trial).

6.2.3.1 Discrimination

The primary method used to assess discrimination is the concordance statistic which was introduced in Section 3.2.3. The method plays a particularly strong role in survival analysis because a model's intercept term, or equivalently the baseline hazard in a Cox model, is not required: the predicted survival probability will be larger for observation i and for observation j if and only

if the linear predictors are ordered: $\eta_i(t) > \eta_j(t)$ implies $S_i(t) < S_j(t)$. Since the baseline hazard is often omitted from published reports, the concordance may be the only readily available validation method for a published study.

The most straightforward way to accommodate censoring into the concordance computation is to consider certain pairs to not be evaluable. That is, if subject i is censored at 2 years and subject j has an event at 4 years we cannot assess whether $t_i > t_j$. Note that if t_i is censored at time 10 and t_j is an event at time 10 then we do know that $t_i > t_j$, though the precise event time for subject i is still uncertain.

- This approach is known as Harrell's C-statistic; it counts the fraction of evaluable pairs that are correctly ordered. For a 0/1 predictor, Harrell's C is equivalent to the Gehan-Wilcoxon test.
- Mantel showed that the Gehan-Wilcoxon can be computed most efficiently by first ordering the data by time, and at each death time computing a single term $0 \le g(t_i) \le 1$ which is the the rank of the subject who experience the event at t_i among all those still at risk at time t_i. (Formally, the rank of their predicted risk). The concordance is a sum over the deaths of $n(t_i)g(t_i)$ where $n(t)$ is the number at risk at time t.
- Therneau and Watson [105] showed the same computational approach applies for the concordance in general, and also that the resulting calculation is equivalent to the score statistic from a Cox model using time-dependent ranks. This provides a natural extension of C to a Cox model with time-dependent covariates, and also a link to the wider literature of Cox model theory.
- Multiple other weights have been proposed to replace $n(t)$ in tests, including the Peto-Wilcoxon[82][83], the Fleming-Harrington $\gamma\rho$ family, the Schemper [90] and the Tarone-Ware family. Any of these can be applied to the concordance computation, with the same arguments with respect to which one is 'best' under what circumstances. Uno's [109] proposed weighting is essentially identical to Schemper. For all of the datasets used in this book, the change in concordance is of an ignorable size. Also, it is not clear how to carry forward the weighting arguments when data is subject to left truncation, such as will occur with delayed entry. Multistate models will face the same issue, compounded. correct refs? 14. Tarone RE, Ware J. On distribution-free tests for equality of survival distributions. Biometrics. 1977;64:156–160. Fleming TR, Harrington DP, O'Brien PC. Design for group sequential tests. Control Clin Trials. 1984;5:348–361. or Harrington DP, Fleming TR. A class of rank test procedures for censored survival data, Biometrika, 1982;69:553-565

Table 6.3 shows concordance values for the Rotterdam data using the full time scale of 19 years or upper thresholds of $\tau = 3$, $\tau = 5$, or $\tau = 10$ years, and for the GBSG validation data using $\tau = 3$ and $\tau = 5$. It is not uncommon to see the overall concordance decrease slightly with longer follow-up.

Remember it may be necessary to truncate the follow-up of the validation

Time	Rotterdam	GBSG
3	0.703 (0.009)	0.686 (0.017)
5	0.689 (0.007)	0.672 (0.016)
10	0.681 (0.007)	
19	0.681 (0.007)	

Table 6.3 *Concordance measures (standard errors) for the Rotterdam (development) and GBSG (validation) datasets with different specified upper limits.*

cohort beyond some upper limit of interest in order to ensure discrimination is assessed over a relevant time period. Most concordance functions include this upper limit as an optional argument. When there are time-dependent covariates, the time window for relevance may increase, but the interpretation of the concordance becomes more local, i.e., how well do current risk scores predict current events.

Stratified models. In describing x as the predicted value, we have been cavalier about whether this is the linear predictor η, a predicted survival at 5 years, the predicted median survival, cumulative hazard, etc. The reason is that we don't have to be detailed: if $\hat{y}_i > \hat{y}_j$ on any one of these scales, it will also be true on the others, and the concordance depends only on the ordering. The same holds true for the predicted response in a logistic regression model. An exception to this is stratified Cox models. Since two different strata will have separate estimated baseline survival curves, the predicted curves for two subjects, one from each stratum, could cross. Use of η as a shortcut for comparison is no longer possible, since there is no longer a monotone relationship. There are two possible solutions. The first, which is the one adopted by the survival package, is to compute the totals of concordant and discordant pairs separately for each stratum, and then add up across strata at the end. This approach parallels what is done for the log likelihood and coefficients of the Cox model. Normally, interest centers on how well covariates other than the strata predict outcome, and not on between stratum differences, so this seems to be a sensible approach. An alternate approach is to choose a particular outcome: the predicted survival or RMST at some specified time τ, say, compute that prediction for each subject, and use this value to compute a concordance.

The same issue and approach applies to parametric survival models that contain a strata term in the formula. This leads to models that have a different estimated scale factor in each stratum, giving predicted survival curves that have a different shape and thus can cross. The same default action seems sensible.

6.2.3.2 Calibration

Assessing calibration focuses on how accurately the model predicts absolute risk. Calibration assessment is sometimes separated into mean, weak, moderate, and strong calibration [19]. The first simply states that the overall average

is correct, and the second that a linear regression of observed vs. predicted has slope 1. Moderate calibration assesses a possible non-linear relation between observed and predicted, and strong calibration that predicted and observed match for all subsets of predictors. These four levels require increasing sample size in the validation dataset in order to be effectively assessed. While methods for testing and visualizing calibration have been well thought out in the binary regression setting, there is a wide misunderstanding in the medical literature of how to assess calibration in the Cox model setting.

To assess calibration for time-to-event data, it is useful to view the data as a counting process [27]. This is closely related to the *martingale residual* $M_i(t)$ [104] defined as the observed - expected number of events for each subject up to time t. As the name would suggest, if the development model is correct the $M_i(t)$ processes in the validation dataset are each a mean zero martingale, the O portion being simply the 0/1 status indicator for each observation and the E the predicted event count for that observation and their covariates, *at the observed follow-up time t_i for the observation.* (Martingale residuals also underlie the mathematics of several targeted measures of model adequacy, e.g., the checks for proportional hazards discussed in Section 3.1.3.)

The predicted number of events E_i incorporates both the predicted relative risk for a subject and their follow-up time, thus dealing naturally with the common case of a subject who drops out from the validation study almost immediately after enrollment; such subjects add little or no information. (Nearly every study has 1 or 2 of these). Such an observation will have $O_i = 0$ and E_i of nearly zero, and ends up having essentially no influence on the assessment of calibration, which is what we want.

Calibration-in-the-large. An overall measure of fit is the comparison of $\sum O_i$ to $\sum E_i$, the total expected number of events. For the GBSB validation dataset we have 285 observed vs. 248.4 predicted events at five years; the recurrence free survival (RFS) rate is slightly higher (1.15) than predicted. There is a close connection to the standardized incidence ratio (SIR), a measure with a long history in population science and epidemiology.

The overall O/E value is also called calibration-in-the-large, since it gives a single overall value for the model. Because it involves an absolute prediction (expected number of events) the statistic is a measure of calibration; this also means that the baseline hazard is required for its computation. For a Cox model, calibration-in-the-large is always exact for the development model since the martingale residuals sum to zero, by definition, and is thus uninteresting for goodness-of-fit, as mentioned in Section 3.2.

Modeling approach. An immediate statistical question is how to derive a confidence interval for this O/E value. A simple solution is provided by Berry [13]. When the validation model is correct, it turns out that the likelihood function has the same form as a Poisson, up to a constant. The consequence is that an ordinary GLM Poisson model with the 0/1 observed status as the response and $\log(E_i)$ as an offset returns correct estimates, standard errors,

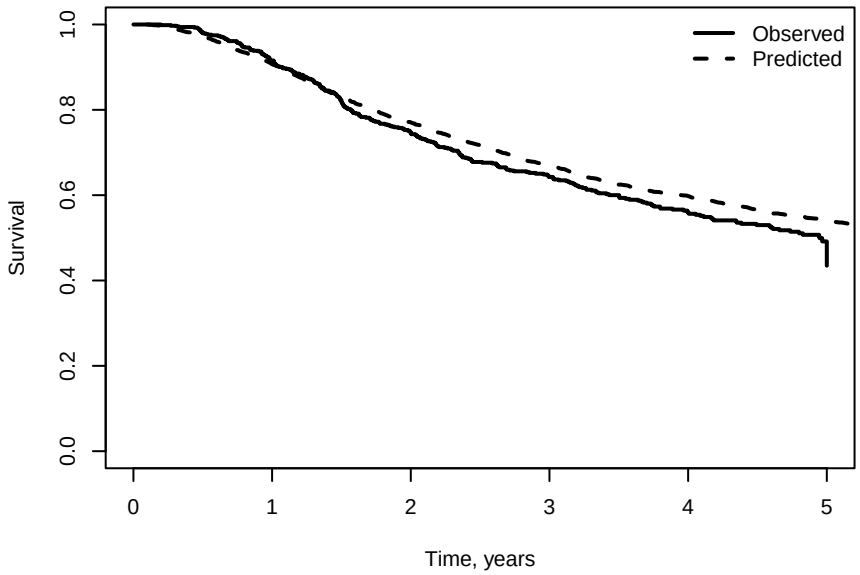

Figure 6.3 *Observed and predicted survival using the GBSG cohort and the prognostic model created using the Rotterdam data.*

and log-likelihood, the last with a dataset dependent constant added. This last also means that likelihood ratio tests are correct.

This fit does not assume that the data is Poisson, a common mistake users make when describing the approach. For example, if the validation dataset had complete follow-up to failure for all subjects and no censoring, then the y value in the `glm` fit will be a constant value of 1 for all observations, which is certainly not Poisson. The `glm` code is no more than a handy trick that allows us to use existing programs. It does depend on an assumption that the development model is correct, failure of which is precisely what we are trying to test.

The Poisson model approach can also be useful in assessing other levels of calibration, including the calibration slope, calibration according to risk levels, and calibration according to covariate levels. Results for four such Poisson models are shown in Table 6.4.

1. The first model contains an intercept only, and assesses the calibration of the predictive model. For the GBSG validation data, an intercept-only model has coefficient and standard error of 0.14 and 0.06 giving an estimate and confidence interval for $\exp(\beta_0)$ of 1.15, (1.02, 1.29). The p-value from this comparison is an appropriate test for the visual difference between observed and predicted curves that was noted in Figure 6.3. This is referred to as overall calibration, mean calibration, or calibration-in-the-large.
2. The second fit assesses calibration slope. Note that the p-value for slope

is for a test of *slope* = 0, but we want to test *slope* = 1. This test is $z = (0.969 - 1)/0.102 = -0.2941$, which yields $p = .82$. Note that these first 2 models will show good calibration by definition when assessing a model using the data that was used to develop it. So at a minimum, cross-validation will be needed to get a fair assessment of calibration when an external dataset is not available. This is a test of weak calibration.

3. The third fit uses categories of the risk score to partition patients; in this case there is not a relationship between the predicted death rate and the decreased risk seen in the second cohort. An analogous procedure for logistic regression models is to place the 0/1 outcome and the prediction (again as an offset) within an extended logistic regression model; this process provides a replacement and extension for the well known Hosmer-Lemeshow test [19] and is a test for moderate calibration. Figure 6.4 shows a calibration plot based on deciles of predicted risk and is a common plot used for assessing calibration [98].
4. An immediate follow-up question is whether this decreased risk is focused on one particular subgroup; the fourth fit shows that, rather surprisingly, almost all of the excess risk appears to fall on the middle tumor size. This type of assessment can also be done for continuous variables by including a smooth term in the Poisson models, as shown in Figure 6.5. Note that the coefficients from the Poisson models are often exponentiated, resulting in standardized incidence ratios (SIR), which are the ratios of the observed to predicted events. An SIR > 1 indicates the observed event rate is higher than predicted, and an SIR < 1 indicates the predicted risk is higher then the observed event rate. Exploring a wide range of variables is an assessment of strong calibration.

6.2.3.3 *Alternative approaches*

The performance measures described above for assessing discrimination and calibration are sufficient for model validation. This section describes other methods proposed in the literature for assessing the validity of a model. Advantages and disadvantages of these approaches are discussed.

R^2 measures. One way to proceed is to use the prognostic index as input to one of the R^2 measures that are based on variation in the risk scores. This includes the Kent and O'Quigley and Royston and Sauerbrei R^2 measures and the Goen and Heller variant of concordance. An attraction of this approach is that they produce values on a familiar scale, while sidestepping the censoring issues. In this case we have 0.17 for the Kent and O'Quigley measure, 0.18 for the Royston and Sauerbrei proposal, and 0.65 as the indirect measure of concordance.

In our opinion these transformed values, though more familiar, are not very helpful. The Royston and Sauerbrei measure is particularly problematic since it involves a normal scores transformation of the prognostic index $f(\hat{\eta})$.

	Estimate	Std. Error	P
Intercept only			
Intercept	0.186	0.058	0.001
Intercept and slope			
Intercept	0.204	0.083	0.014
Slope*	0.969	0.102	< 0.001
Risk groups			
Low	0.169	0.119	0.153
Medium	0.329	0.104	0.002
High	0.105	0.086	0.221
Tumor size			
< 20	0.267	0.124	0.031
$20 - 50$	0.176	0.070	0.012
> 50	0.085	0.183	0.641

Table 6.4 *Coefficients from four Poisson models to assess calibration, using predicted values from a model fit to the Rotterdam subjects, applied to the GBSG data. Note: ignore p-value for slope, as this tests slope* $= 0$*, but the appropriate test for calibration is slope* $= 1$*.*

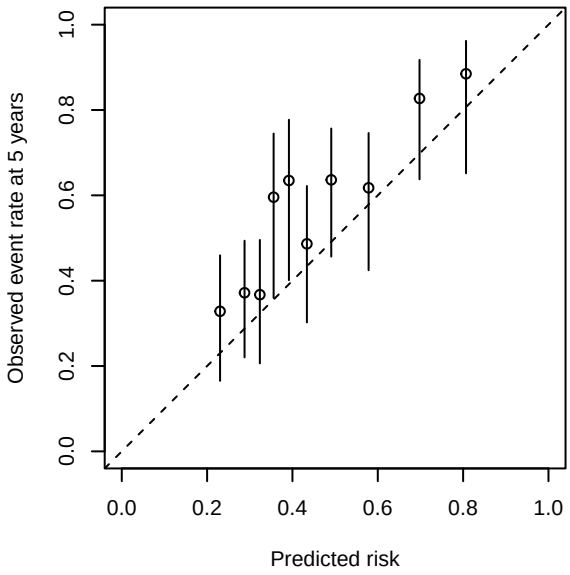

Figure 6.4 *Calibration plot for 5 year risk from a Cox model fit to the* `rotterdam` *data, using predictions from the GBSG cohort. Risk scores based on the first model were used to divide the subjects into deciles of predicted risk. Observed event rates are KM values at 5 years for the GBSG with 95% confidence intervals. Dashed line is the identity line. Observed event rates are approximately the same as predicted for most deciles.*

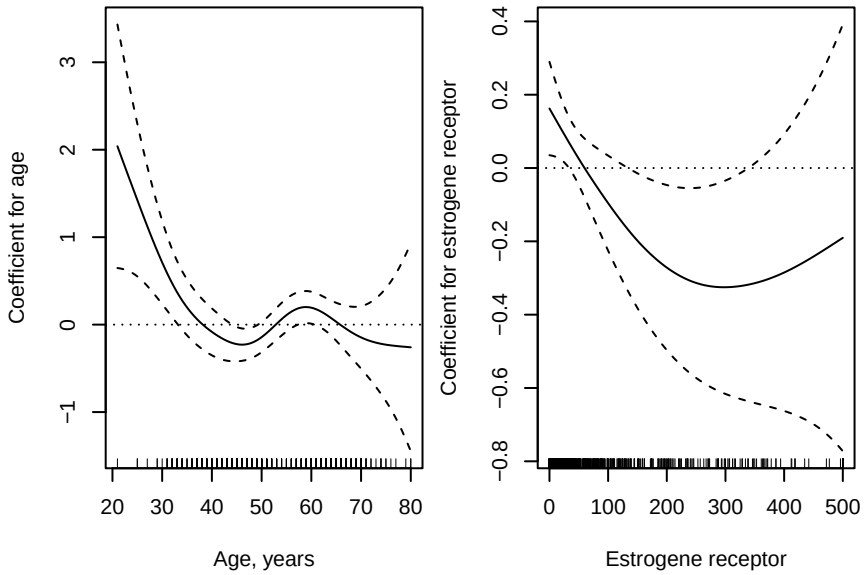

Figure 6.5 *Calibration plot for 5 year risk from a Cox model fit to the* `rotterdam` *data, using predictions from the GBSG cohort. The left panel shows the coefficient from a smooth fit for age. Calibration is poor at lower levels of age. The right panel shows the coefficient from a smooth fit for estrogen receptor (not in the original model). Calibration is poor for most values.*

This transformation is dataset dependent, but more importantly it no longer answers the central question of how well η_i predicts in future subjects. It is $\hat{\eta}$, after all, that will be used for future patients as they come one by one to the clinic, not a sample-dependent transform $f(\eta)$.

Other modeling approaches. Several modeling alternatives to our Poisson model approach have been proposed. These include a more flexible Cox model using a spline of the prognostic index [10] or a more flexible non-parametric fit as utilized in the `calibrate` function of the R `rms` package. Primary trade offs between the various modeling options are flexibility vs noise (bias/variance) and few vs. many underlying assumptions.

Redistribute to the right. When the focus of validation is a single time point τ, such as the 10-year time point that is commonly used for cardiovascular disease risk prediction, the key problem is what to do with validation data points which are censored prior to τ. The predicted value $\hat{y}_i(\tau)$ is readily available from the model for all subjects but the observed datum $y_i(\tau)$ is lacking.

One approach is to use the redistribute-to-the-right (RTTR) algorithm to reassign the weights of these observations. Application of the RTTR algorithm, or equivalently inverse probability of censoring weights (see Section **??**) results

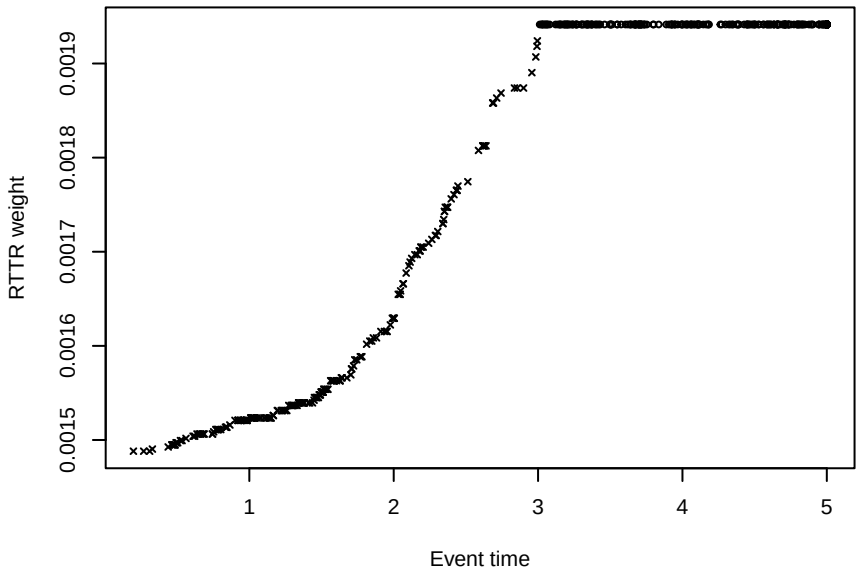

Figure 6.6 *Redistributed case weights for the events (Xs) and non-events (circles) in the GBSG data using a 3 year target.*

in a modified dataset where all observations censored before τ have a weight of zero. It is then easy to proceed forward with standard methods. There is a price for this simplification, however, in terms of both statistical efficiency and robustness to assumptions, which we discuss later.

For the GBSG dataset 131 of the 686 subjects are censored before our 3 year target. Figure 6.6 shows the distribution of weights for the remaining 555. The "cases" (y=1) are events at ≤ 3 years, each receives a portion of the weight for each censor that was earlier; all the "controls" end up with the same weight, and are those whose event time is > 3 years or censoring time ≥ 3 years.

Uno et al [110] define the standard binomial measures of sensitivity, specificity, etc. using RTTR weights. Computations are straight-forward from that point forward. Note that the concordance at a time cutpoint may differ from the continuous concordance. Like a choice between mean and median, deciding which measure of concordance is 'best' is less important than remembering that they are not the same, and cannot be directly compared.

Brier score. One measure of agreement is a simple correlation coefficient between the binary response of $y_i(\tau) = 0/1$ and the predicted probability $\hat{p}_i(\tau)$ of an event at or before τ. This can be written as

$$1 - R^2 = \frac{(1/n)\sum w_i(y_i - \hat{p}_i(\tau))^2}{(1/n)\sum w_i(y_i - p_0(t))^2} \tag{6.1}$$

Where $p_0(t)$ is the estimated probability under a null model, e.g., from the Kaplan-Meier, and w_i are the RTTR weights. (The sum of the RTTR weights, if computed properly, is n for any cutpoint τ). The numerator of (6.1) is called the Brier score, and the fraction itself has been labeled as the index of predictive accuracy [55]. van Howelingen and Putter [111] suggest using an average of the Brier score over a range of τ values.

Inverse probability of censoring weights. The RTTR weights make the assumption that censoring is completely uninformative. Informative censoring may occur in any study. In the GBSG dataset, for instance, patients who receive hormone therapy have both a lower censoring rate and a lower failure rate. This can be particularly acute in population studies where there is a change over calender time. Verhuel [112] describes 20 years' experience with aortic valve replacement in an Amsterdam medical center, where in the first 10 years only 3% of the recipients were over age 70, and over the second 10 years 27% were. Oakes [78] shows that even in the controlled environment of a double blind clinical trial a patient's decision to drop out may be predictive of outcome, their example shows important (but equal effects) for the treated and placebo arms of a particular trial.

These informative censoring patterns can be addressed by explicitly model the censoring process using available covariates to obtain inverse probability of censoring weights (IPCW). (As shown in Section **??**, IPCW without covariates reproduce the RTTR algorithm). This results in the more reasonable assumption of uninformative censoring conditional on the covariates of interest.

Ignore censored observations. An option which should be avoided is to simply ignore those validation observations which are censored before time τ. The resulting estimates of survival at time τ will be biased low, as pointed out 50 years ago by Berkson. The impact on estimates of sensitivity, specificity, and AUC is more complex, depending on the association between censoring and survival, and between censoring and the risk score.

Pseudovalues. A semi-binomial approach is to compute the pseudovalues at time τ for each observation. These will be based on the Kaplan-Meier of the validation dataset. An advantage over the RTTR approach is that each observation retains its own value of the prognostic index $\hat{\eta}$.

Goodness-of-fit tests for calibration. Many alternatives to the Hosmer-Lemeshow test have been proposed for use in time-to-event datasets. Many of these tests did not appropriately account for censoring, so the tests did not perform well. The modified Nam-d'Agostino test using the Greenwood standard error seems to perform well. However, it still has the same limitations as the Hosmer-Lemeshow test. So it may not perform well in large samples where statistical significance is easy to obtain. In addition, the test result is only a p-value, which does not provide any measure of the effect size. For these reasons, we prefer the Poisson model approach described earlier in the

chapter, which provides a measure of effect size and is a very flexible approach that can accommodate multiple levels of calibration assessment.

Net reclassification index and the integrated discrimination index. Other measures of performance include the net reclassification index (NRI) and the integrated discrimination index (IDI), which are used to compare two risk scores (including comparisons of an original and updated risk score). The NRI is most useful for assessing the number of patients reclassified above or below a pre-specified treatment threshold [64]. Reclassification alone is not enough to show improvement of one risk calculator over another, as it is important to increase the risk prediction for patients who will have events and to decrease the risk prediction for patients who will not have events. The NRI is a weighted sum of the probabilities that patients who have higher risk predictions by the new calculator actually have higher observed event rates and the probability that patients who have lower risk predictions by the new calculator actually have lower observed event rates. The IDI is the difference in predicted probabilities between those who do and do not develop the outcome. When computing either the NRI or the IDI, it is important to first assess calibration, as these measures can be artificially inflated in the presence of poor calibration [49, 64].

Updating risk scores

Rather than always developing new models from scratch, it is often of interest to improve an existing model by re-calibrating or adding additional predictors when it is believed that the underlying event rate has changed or new information (e.g., biomarkers) are now standardly available. Remember that this updated risk score should be validated as well. Possible modifications include:

- Adjust the underlying baseline survival
- Re-estimate the model coefficients using the same predictors
- Use the prediction rule as an offset and add in additional predictors
- Estimate the coefficients for the original and new predictors

More information on re-calibration and other important issues when updating risk scores can be found elsewhere. add refs

Summary

In summary, risk score development requires careful consideration and planning before starting the modeling process. Model validation also requires careful consideration and planning. When using external validation sets, additional checks are needed to ensure comparability. Time-to-event data creates some challenges for assessing model performance, as it is important to properly account for censoring and to make sure the assessments involve the relevant time period of interest.

Chapter 7

Pseudovalues

"Any darn fool can make something complex; it takes a genius to make something simple." — Pete Seeger

The reason special methods are needed to analyze time-to-event outcomes is that the data are incomplete. Since not all subjects have reached the outcome, two variables are needed to specify an outcome. One is the observation time and the other is an indicator of whether the event has been reached or the observation is censored. Pseudovalues provide an attractive alternative that captures specific time-to-event outcomes in a single outcome variable that allows covariate effects to be examined using standard regression methods for continuous outcomes. The basic idea is to replace the incomplete observations with an appropriate estimator. Pseudovalues can be created for the survival function, the restricted mean survival time (RMST), and the cumulative incidence function for competing risks data [5].

For example, assume that the statistic of interest is 5 year survival, $\theta = S(5)$ where S is the survival curve, and we want to understand how that estimate depends on covariates. Because y is censored, the standard approaches of scatterplots, means, and regression are not applicable, and we have to use other statistical methods (the entire focus of this book). Because the Kaplan-Meier provides a valid estimate $\hat{\theta}$ of $S(5)$, however, it can be used as a basis for pseudovalues $y_{(i)}$ for the response y_i. The resultant pseudovalues are *not* censored, and can therefore be plugged into ordinary statistical methods. More importantly, this allows for simple approaches to other statistics such as the RMST, which can be estimated from the KM, but for which specialized censored data methods are not widespread.

Andersen and Pohar Perme [5] give an overview of methods and results for survival-based pseudovalues. We largely follow their approach, with two exceptions. The first is notational, in that they use $\hat{\theta}_i$ for the ith pseudovalue and we use $\hat{\theta}_{(i)}$. The (i) notation for leaving one out is borrowed from Efron [32]. The second is that we make use of infinitesimal jackknife (IJ) based pseudovalues, as provided by the `pseudo` function in the survival package.

7.1 Definition

Let θ be a statistic of interest, $\hat{\theta}$ an estimator of the quantity based on the data, $\hat{\theta}_{-i}$ the estimate computed using all but the ith observation, and $\hat{\theta}(.) = (1/n)\sum \hat{\theta}_{-i}$ be the mean of the removed observations. Then the jackknife based pseudovalues $\hat{\theta}_{(i)}$ and the jackknife estimate of variance V_J are defined as [32]

$$\begin{aligned}
\hat{\theta}_{(i)} &= \hat{\theta} + (n-1)(\hat{\theta} - \hat{\theta}_{-i}) \\
&= n\hat{\theta} - (n-1)\hat{\theta}_{-i} \\
V_J &= \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{-i} - \hat{\theta}_{(.)})^2 \\
&= \frac{1}{n(n-1)}\sum(\theta_{(i)} - \hat{\theta}_{(.)})^2 \\
\hat{\theta}_{(.)} &= (1/n)\sum \theta_{(i)}
\end{aligned}$$

In the jackknife variance formula, pseudovalues play the same role as the original observations y in an ordinary variance. The pseudovalues "stand in" for y. For the simple mean $\hat{\theta} = \sum y_i/n$, it is easy to verify that $\hat{\theta}_{(i)} = y_i$ and that V_j equals the usual estimate of variance. One of the great advantages of the jackknife is that it provides a variance estimate in cases where a "usual" variance estimate is not available.

The IJ replaces the difference $\hat{\theta} - \hat{\theta}_{-i}$ with the first order Taylor approximation from the weighted version of θ

$$U_i = \left.\frac{\partial\hat{\theta}}{\partial w_i}\right|_{w=1} \tag{7.1}$$

$$V_{IJ} = \sum U_i^2 \tag{7.2}$$

$$\tilde{\theta}_{(i)} = \hat{\theta} + nU_i \tag{7.3}$$

When θ is the simple mean, then $\tilde{\theta}_{(i)} = y_i$, which motivates the use of n vs. $n-1$ in the definition. Efron [32] shows that in several cases the IJ estimate of variance can have a large negative bias in small samples, while the jackknife is more reliable. In the case of the survival curve, however, this concern is ameliorated by the fact that for the Kaplan-Meier, the IJ estimate of variance is exactly equal to the Greenwood estimate of variance; see section A.5 for details. Multiple simulation studies and many years' experience have solidified the reliability of the Greenwood estimator.

The package's choice of the IJ method hinged on two issues:

1. Consistency: The survival package already makes extensive use of IJ values to compute robust variance estimates for the Cox model; IJ based variance estimates for the Aalen-Johansen estimate are then a natural follow on,

as are IJ estimates for post-Cox survival curves. (The ability to leverage existing test suites for the computer code was an added bonus).

2. Speed: IJ values can be assembled more quickly. The jackknife computations are at best $O(nd)$ where n is the number of observations and d the number of events, while the IJ can be computed in $O(n+d)$. (This becomes important for very large datasets; but was is not an issue for any of the data examples the book.)

One feature of the IJ values is that $\sum_i U_i = 0$. This means that the IJ pseudovalues have a slight negative correlation, and also guarantees that the pseudovalues sum to the underlying estimate $\hat{\theta}$.

Informative censoring

An important assumption for validity of pseudovalue estimates is that the underlying Kaplan-Meier estimate is an (approximately) unbiased estimate of the true survival distribution. This condition, however, may not hold if there is informative censoring. Andersen and Pohar Perme [5] show that if censoring depends on a categorical covariate such that the stratified KM curves are unbiased, then pseudovalues based on the per-group curves will successfully remove the bias.

7.2 Restricted mean survival time

One of the more compelling uses for pseudovalues is the RMST, providing simple modeling tools for the mean survival time. As noted in Chapter 2, for any positive probability distribution, a well known identity is that the mean is equal to the area under the survival curve.

Since the Kaplan-Meier gives an unbiased estimate of $S(t)$ we can use the area under the KM to estimate the mean time to death. However, since the entire $S(t)$ curve is usually not available, i.e., the KM terminates before reaching $S(t) = 0$, we instead estimate the RMST, the area under the KM up to some specified point τ. This is interpreted as the expected number of life years out of the first τ. If $\tau = 10$ and the area under the curve were 8.1, the phrase would be an expected lifetime of "8.1 out of 10 years". Common choices for τ are either a fixed time of interest such as 2, 5 or 10 years, the last observed event time, or the last timepoint in the data.

Here is a simple example using the lung cancer dataset. For this cohort of subjects with advanced disease, the strongest predictor is the ECOG performance score, a widely used functional status scale with values ranging from 0 = asymptomatic to 4 = bedbound (Figure 7.1).

For a single categorical predictor such as this, it is easy to obtain the RMST and its standard error directly from the same routine used to create the KM: example R code is shown below.

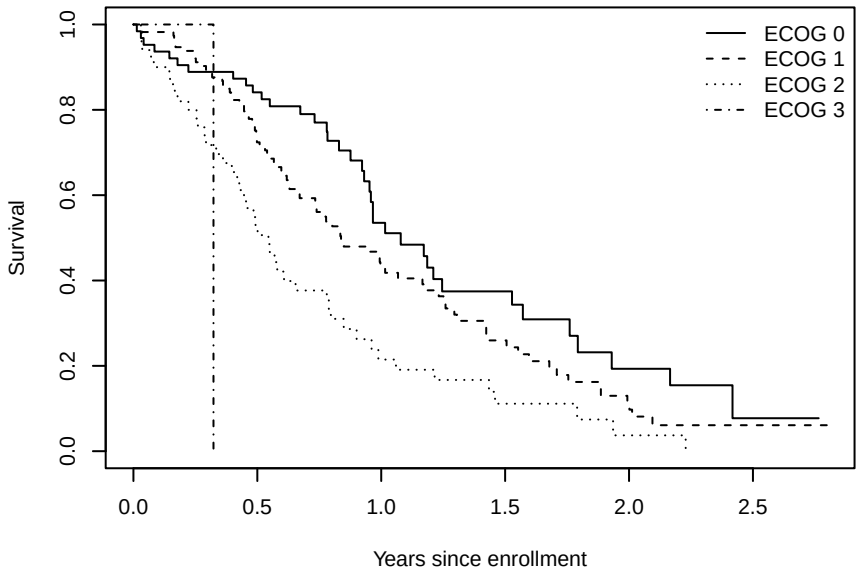

Figure 7.1 *Kaplan-Meier curve using the lung cancer data stratified by the ECOG performance score.*

```
> lfit1 <- survfit(Surv(time/365.25, status) ~ ph.ecog, data = lung)
> print(lfit1, rmean = 2.5)
Call: survfit(formula = Surv(time/365.25, status) ~ ph.ecog, data = lung)

   1 observation deleted due to missingness
             n events rmean* se(rmean) median 0.95LCL 0.95UCL
ph.ecog=0   63     37  1.251    0.1085  1.079   0.953   1.572
ph.ecog=1  113     82  1.035    0.0692  0.838   0.734   1.175
ph.ecog=2   50     44  0.708    0.0855  0.545   0.427   0.789
ph.ecog=3    1      1  0.323    0.0000  0.323      NA      NA
    * restricted mean with upper limit =  2.5
```

Using $\tau = 2.5$ years, subjects with ECOG score 1 and 2 have about .2 and .5 fewer years of survival, on average, than subjects with ECOG score 0.

Pseudovalues for modeling are based on the overall Kaplan-Meier, not broken down by ECOG score, and can be estimated using the `pseudo` function available in the R survival package or the STATA stpmean command, for instance. The resulting values are then used as the response variable in an ordinary regression, along with a robust variance estimate (known variously as Eicker-Huber-White, GEE, or Horvitz-Thompsen). Results of a fit using ECOG score, age and sex are shown in Table 7.1. After adjusting for age and sex, the RMST is estimated to decrease by 0.26 years for each 1 point increase

	Estimate	LS SE	Robust SE	P
Intercept	1.837	0.335	0.336	0.000
Performance Score	1.837	0.335	0.336	$< .001$
Male	-0.322	0.099	0.100	0.001
Age (decades)	-0.061	0.054	0.006	0.272

Table 7.1 *Coefficients from a linear model fit to RMST pseudovalues, for the lung cancer data.*

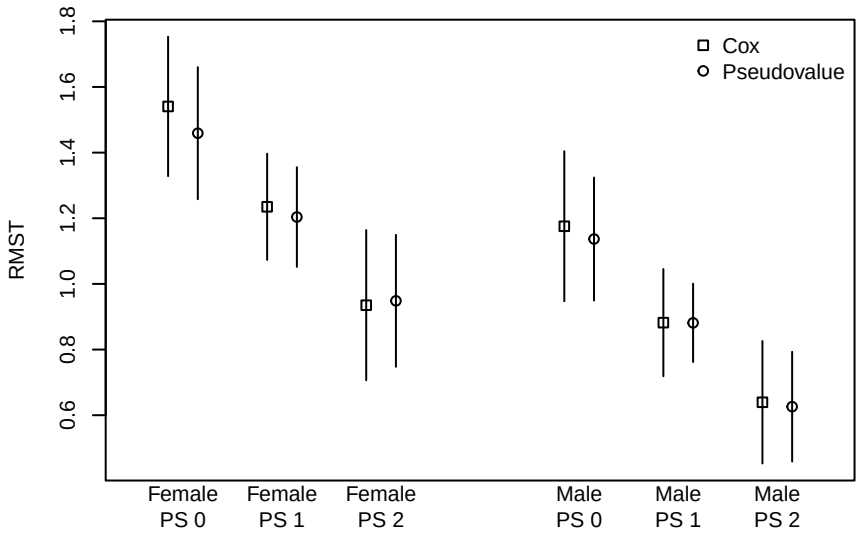

Figure 7.2 *RMST estimates and 95% confidence intervals from a Cox proportional hazards model and from a linear model fit of the pseudovalues. All predictions are for a 62 year old subject.*

in ECOG score, and to increase by 0.32 years for females compared to males. The effect of age is negligible.

We can also get RMST values from a Cox model fit: for any combination of the predictor variables, generate the corresponding survival curve, and integrate its area. Figure 7.2 compares estimates from the Cox model and psuedovalue predictions for 6 subjects with performance score of 0, 1, or 2, male or female, age 62. The estimates and standard errors are nearly identical.

Why do the Cox and linear model RMST estimates agree so closely? For a Cox model, the survival curve follows the prescription that $S(t) = \exp(\Lambda_0(t)\exp(X\beta))$, which is not a linear function of the covariates. But as shown in Figure 7.3, the function $f(x) = \exp(-a\exp(x))$ is remarkably *close* to linear over a wide range. The linear predictor in a Cox model most com-

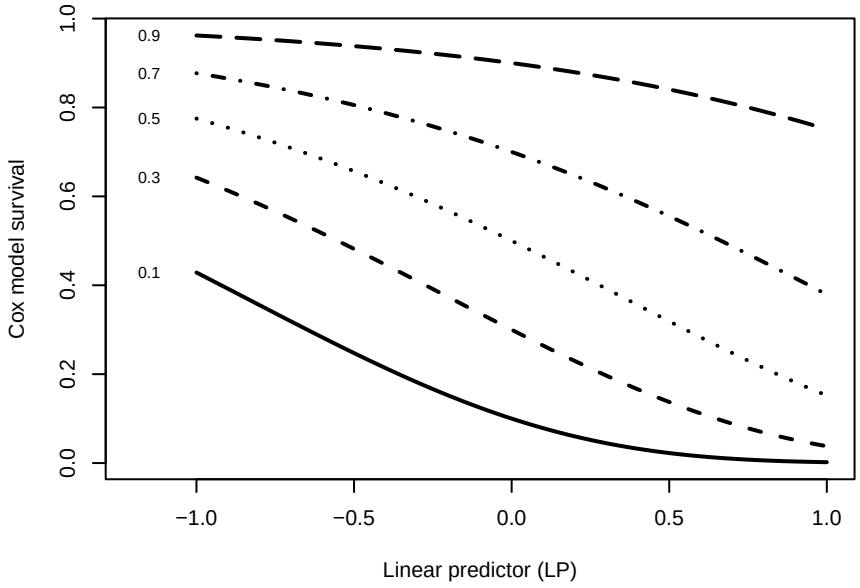

Figure 7.3 *The linear predictor for a Cox model versus predicted survival, for select values of the baseline survival.*

monly falls in the range from -.5 to .5; which corresponds to a 3-fold difference in risk between the best and worst covariates. A model with particularly strong covariates may have a linear predictor that ranges from -1 to 1, which correspond to hazard ratios of 1/3 up to 3 times the "average" risk. The overall survival curve often does not drop below 0.3, other than for studies with very long follow-up. Over these ranges, the figure shows that estimated survival is approximately linear in the linear predictor. If all the quantiles of S are linear, then the RMST will be a weighted sum of linear components and must also be linear.

Thus, if the Cox model holds and the covariate effects and follow-up time are bounded, a linear model for RMST pseudovalues will capture nearly the same information as RMST predictions from the Cox model. Two further advantages of the pseudovalue model are that the coefficients are directly interpretable as years of life, and that the model may still be valid when the proportional hazards assumption does not hold.

Robust variance

The IJ residuals for observations that are censored early in the study are of necessity small, as they have less opportunity to affect the results, which in turn means that the pseudovalues are not equivariant. (An observation censored before the first event has residual 0, so its pseudovalue has no variance at all.) Early, mid, and late events can have large positive, near 0, and small

	ECOG score	Female sex	Age (decades)
Linear model	-0.2554 (0.0683)	-0.3224 (0.0989)	-0.0609 (0.0543)
GEE model	-0.2554 (0.0670)	0.3224 (0.0994)	-0.0609 (0.0552)
Survey model	-0.2554 (0.0672)	0.3224 (0.0996)	-0.0609 (0.0553)

Table 7.2 *Pseudovalue estimates (standard errors) for restricted mean survival time, using an ordinary linear model, a GEE linear model, and a survey sampling linear model.*

negative overall influence, due to the interplay between their effect on the denominator (makes S larger) and their effect on the numerator (makes S smaller). In such a case, theory argues for using an alternate variance for the linear model, e.g., White's estimate, which was first proposed as a correction for possible heteroscedasticity [113]. For a linear model, this is identical to the "working independence" estimate of a GEE model, and/or the variance estimate from a survey sampling approach. Table 7.2 shows the estimates from all three models side by side. By design, all three approaches will give the same coefficient estimates, differing only in the estimated standard errors. (The survey sampling and GEE based errors differ by a factor of $\sqrt{n/(n-1)}$ which has roots in how each counts degrees of freedom.) In this example standard errors from the "naive" approach hardly differ from those using corrected methods; we will see this repeated in further examples whenever results from a single timepoint are used. This suggests that although the formally correct method should be used for a final publication, working models for preliminary analysis can often use the simple linear model.

Multiple timepoints

It is possible to extract the RMST at multiple timepoints of the curve, and then jointly fit all of the time values at once. The question is whether this is worth the trouble. Two features argue against it at the start: RMST values at multiple timepoints will be highly correlated, so much so that the effective information in the multitime model may hardly be any larger than using the latest of the time points, and secondly that covariate effects may not be constant over time.

As an example of the second, look at the RMST values for females vs. males in the lung cancer dataset, at 0.5, 1, 1.5, 2, and 2.5 years as shown in Table 7.3, based on the simple Kaplan-Meier. The spread increases over time, as one would expect, since the curves continue to spread apart. The difference at 2.5 years is 8 times that at 6 months, and twice the difference at 1 year. Percentage increases are somewhat better behaved, but also not stable, ranging from 1.1 to 1.4 fold. Nevertheless, as a trial, fit a model using both the 1.5 and 2.5 year RMST values. The resulting dataset will have 2 observations per subject, so using a robust variance is crucial. The results are found in Table 7.4. The fits used a log-link, so represent multiplicative effects. As expected the coefficients for the joint fit are midway between those from

a 1.5 and 2.5 year estimates. Standard errors from the fit that uses "more" data are no better than the single year fits, however. Since coefficients are not identical, this is easiest to see in the t-statistics, which are the coefficient to standard error ratios.

	0.5	1	1.5	2	2.5
Female	0.47	0.81	1.04	1.19	1.24
Male	0.42	0.66	0.79	0.85	0.88
difference	0.05	0.15	0.25	0.34	0.35
ratio	1.11	1.23	1.31	1.40	1.40

Table 7.3 *RMST values for females versus males in the lung cancer dataset at 0.5, 1, 1.5, 2 and 2.5 years.*

7.3 Survival probability

AML data

As a first illustration for a Pr(death) as the pseudovalue target, use a small dataset which records the time to relapse for 23 pediatric patients with acute myelogenous leukemia. Relapse times range from 5 to 48 months, 5 of the 23 patient times are censored, and pseudovalues for the survival probability at 12, 24 and 36 months are listed in Table 7.5.

The first censoring is at 13 months, and so at the first reporting time of 12 months the pseudovalues behave exactly like pseudovalues for a simple mean, and thus have recaptured the (uncensored) 0/1 response exactly. At 24 months, after censoring enters, the pseudovalues are no longer constrained to lie in $(0, 1)$. Since $S(t)$ assesses the probability of *no event* by time t, the pseudovalue approximates an indicator variable of 1=no relapse, 0=relapse. It is easy to see that the pseudovalue for $(1 - S)$ is 1 – the pseudovalue for S. Since we find the coefficients from models for Pr(relapse) more natural to interpret, our regressions will most often use $1 - \hat{\theta}_{(i)}$ as the response.

What happens if we use these values in an ordinary regression? Using the tabled values as the response and month 12 vs. 24 vs. 36 as a categorical predictor, a linear model fit exactly reproduces the Kaplan-Meier estimates at 12, 24 and 36 months. The coefficients of the regression have exactly reproduced

	ECOG score			Female sex			Age in decades		
	β	se(β)	t	β	se(β)	t	β	se(β)	t
Time 1.5 only	-0.21	0.05	-3.87	0.25	0.07	3.46	-0.04	0.04	-0.88
Time 2.5 only	-0.26	0.07	-3.76	0.31	0.10	3.20	-0.06	0.05	-1.08
Both	-0.23	0.06	-3.89	0.28	0.08	3.38	-0.05	0.05	-1.02

Table 7.4 *Coefficients, standard error, and t-statistic from fits using the RMST at year 1.5, that using year 2.5, and both. All models using a log link.*

	time	pseudo12	pseudo24	pseudo36
12	5	0.00	0.00	0.00
13	5	0.00	0.00	0.00
14	8	0.00	0.00	0.00
15	8	0.00	0.00	0.00
1	9	0.00	0.00	0.00
16	12	0.00	0.00	0.00
3	13+	1.00	0.79	0.40
2	13	1.00	0.00	0.00
17	16+	1.00	0.79	0.40
4	18	1.00	-0.11	-0.06
5	23	1.00	-0.11	-0.06
18	23	1.00	-0.11	-0.06
19	27	1.00	1.03	-0.06
6	28+	1.00	1.03	0.58
20	30	1.00	1.03	-0.13
7	31	1.00	1.03	-0.13
21	33	1.00	1.03	-0.13
8	34	1.00	1.03	-0.13
22	43	1.00	1.03	1.14
9	45+	1.00	1.03	1.14
23	45	1.00	1.03	1.14
10	48	1.00	1.03	1.14
11	161+	1.00	1.03	1.14

Table 7.5 *Pseudovalue estimates at 12, 24, and 36 months using the AML dataset.*

the Kaplan-Meier; the standard errors differ by assuming a constant variance for all 3 time points, as shown in Table 7.6.

Model	time	estimate	se
Pseudo	12	0.26	0.10
Pseudo	24	0.45	0.10
Pseudo	36	0.72	0.10
KM	12	0.26	0.09
KM	24	0.45	0.11
KM	36	0.72	0.10

Table 7.6 *Comparison of the relapse rate based on linear model coefficients from survival function pseudovalues and Kaplan-Meier estimates at 12, 24 and 36 months using the AML dataset.*

A natural next step is to add treatment as a covariate, which has levels of Maintained and non-Maintained; results are shown in Table 7.7. The Cox model estimates that the subjects on the nonmaintained arm have a hazard rate of about 2.5 fold higher than the maintained arm, $p = .07$. The linear

model	term	coef	se	stat	p.value
Cox	x-Nonmaintained	0.92	0.51	1.79	0.074
Pseudo	Intercept	0.15	0.12	1.29	0.203
Pseudo	x-Nonmaintained	0.21	0.12	1.79	0.078
Pseudo	factor(time)24	0.19	0.14	1.35	0.183
Pseudo	factor(time)36	0.46	0.14	3.24	0.002

Table 7.7 *Comparison of the relapse rate based on a Cox model and linear model coefficients from survival function pseudovalues at 12, 24 and 36 months using the AML dataset.*

model estimates the probability of relapse at 12 or 24 months to be about 21% higher for the nonmaintained group, $p = .08$. The overall probabilities of relapse are shown in Table 7.8.

	Cox12	Pseudo12	Cox24	Pseudo24	Cox36	Pseudo36
Maintained	0.16	0.15	0.29	0.34	0.53	0.61
Nonmaintained	0.35	0.36	0.58	0.55	0.85	0.82
Difference	0.19	0.21	0.29	0.21	0.32	0.21

Table 7.8 *Probability of relapse using the AML dataset for subjects in the two treatment arms, as estimated using a Cox model and using pseudovalues.*

The linear model contains the strong assumption that the difference in survival (i.e., relapse in this example) is the same at 12, 24, 36 months; the Cox model the equally strong one that hazards are proportional across all timepoints, which in turn implies that the curves separate further over time. Figure 7.4 shows the Kaplan-Meier estimates for the two treatment arms, and reveals differences that are large/small/medium at 12/24/36 months, respectively, but mostly it reveals that the sample size is simply too small to address the question of which assumptions are better.

Colon cancer

To better understand pseudovalues for $S(t)$, we turn to a larger dataset involving time to failure (recurrence or death) in the colon cancer study.

Figure 7.5 shows that the Levamisole + 5-FU treatment is markedly superior to either observation or 5-FU alone (the standard treatment at the time), with the majority of events occurring in the first 3 years. Cox model results in Table 7.9 show that the number of positive lymph nodes and greater local spread of the disease (extent) are also potent predictors. The number of positive lymph nodes ranges from 0 to 33 with quartiles of 1, 2, and 5. The form used in the model, linear up to a threshold of 10 (as indicated by `pmin(nodes, 10)`), was chosen for simplicity and suggested by a spline model.

To assess survival, which values should we use: year 4 alone, years 1–6, or some subset? One important consideration for all of this is the choice of a

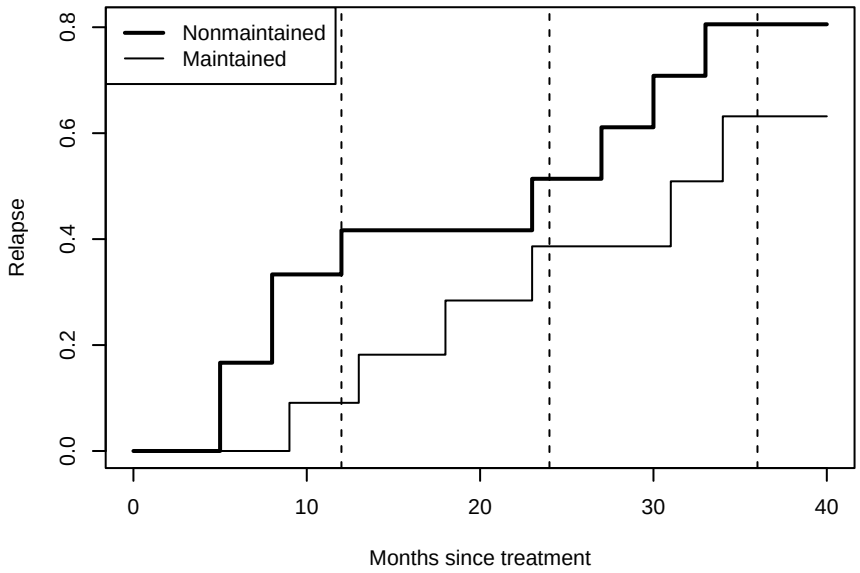

Figure 7.4 *Kaplan-Meier curves of the probability of relapse by treatment arm for the AML dataset.*

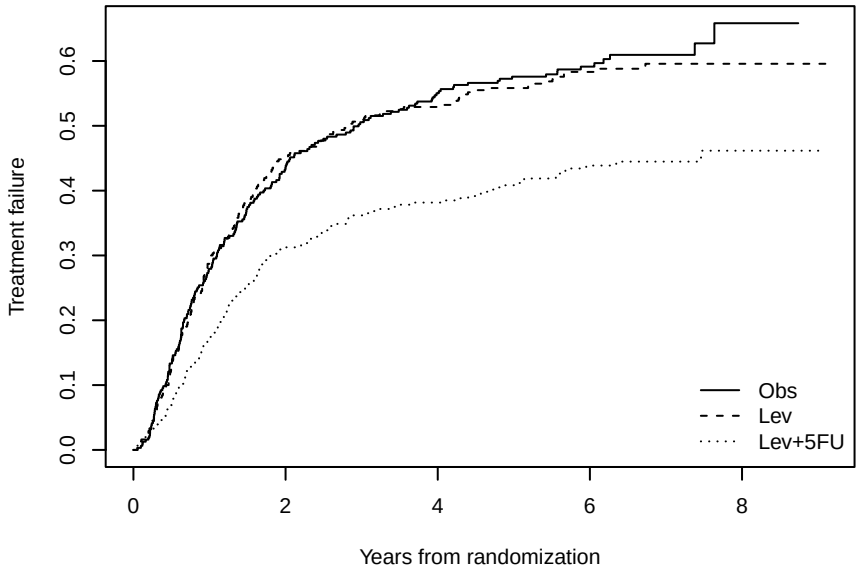

Figure 7.5 *Overall time to failure curves for the Levamisole colon cancer trial by treatment arm.*

	coef	se(coef)	z	p-value
rxLev	-0.073	0.105	-0.70	0.484
rxLev+5FU	-0.488	0.115	-4.25	<0.001
sex	-0.089	0.091	-0.98	0.325
age	0.002	0.004	0.57	0.570
extent	0.497	0.111	4.46	<0.001
pmin(nodes, 10)	0.136	0.015	9.17	<0.001
obstruct	0.205	0.112	1.83	0.067
perfor	0.075	0.255	0.29	0.769
adhere	0.198	0.122	1.62	0.106

Table 7.9 *Cox model for time to failure using the Levamisole colon cancer trial data.*

transformation function f, where we assume that $E(y) \approx f(\eta) + \epsilon$. (Generalized linear model literature normally focuses on the *link* function $g = f^{-1}$.) The ideal function f will

- Transform treatment effects to a common scale over time. That is, a separate intercept for each timepoint will suffice. No interactions between covariates and time are needed.
- Cause multivariate effects to be additive. For example, the effect of treatment is the same for those with all 4 levels for disease extent. No covariate*covariate interactions are needed.
- Normalize the variance so that ϵ is constant across time and across covariates. Alternatively, choose a distribution that properly maps between the predicted mean and variance.
- Bound predicted values to the range of 0–1.

Satisfying all four of these at once is likely to be impossible. Logistic regression, most users' immediate response to estimation of a yes/no question, fails directly. First, it does not accommodate response values outside the range of 0–1, and secondly the variance for a predicted value near 0 or 1 is assumed to drop to zero. The minimum 6 year pseudovalue of -.4 challenges both of those.

As a first pass, look at the difference between the two 5FU arms over time, based on the simple KM, using absolute, logit, and log(-log(x)) scales (Table 7.10. For this particular dataset and these timepoints, the absolute difference between the curves is, surprisingly, more stable across time than logit, log, or log-log differences. Each of the latter expand the differences in the earlier years, when failure probabilities are small, but do so by too much. (Survival curves that split early and then remain largely parallel, as these do, are admittedly an uncommon case.)

Based on this, do a first model with linear effects. The next question is which timepoints to use, and how many. First, look at 6 models with a single timepoint, each containing treatment, extent and nodes, but tabulate only the levamisole coefficient (Table 7.11.

	Y1	Y2	Y3	Y4	Y5	Y6
absolute	11.30	13.90	14.50	14.70	15.00	14.40
logit	49.90	36.80	33.60	32.70	31.30	28.50
loglog	56.90	47.20	45.20	44.90	44.30	41.60
log	14.70	22.60	25.70	27.20	29.20	29.80

Table 7.10 *Comparison of the difference between the two 5FU arms over time (Years 1 - 6) using the simple KM, using absolute, logit, and log(-log(x)) scales.*

	coefficient	se.glm	se.survey	coef/se
Year 1	-0.10	0.03	0.03	-3.09
Year 2	-0.12	0.04	0.04	-3.10
Year 3	-0.14	0.04	0.04	-3.58
Year 4	-0.16	0.04	0.04	-4.22
Year 5	-0.16	0.04	0.04	-4.20
Year 6	-0.14	0.04	0.04	-3.76

Table 7.11 *Coefficients for treatment from separate models fit using pseudovalue estimates at a point in time (Year 1, Year 2, ...). Each model includes treatment, extent and nodes.*

When using a single timepoint, we see that the robust variance is not strictly necessary, and also that the best timepoint in terms of z statistic or power, by an admittedly small margin, is at 4–5 years when the results have largely matured. This happens to be the largest estimated gain for levamisole, reducing the absolute failure rate by 16%. Now consider combinations of 3 time timepoints (2, 4, 6) or all 6 timepoints.

The use of 1, 3, or 6 timepoints hardly changes the t-statistic or the estimate for the 3 important predictors.

Link function and variance considerations

Although a linear link was indicated for the colon data, the logit link and complementary log-log links are more commonly used. Predicted values for the linear model using only year 4 are between -0.01 and 1.016, which is outside the (0,1) range for valid predictions. Use of one of the standard binomial link functions will correct this. (Failures are few: only 3/929 predictions are ≤ 0 and only 1 is ≥ 1 when using a single timepoint. This increases to 47 and 3 for the model based on pseudovalues from all 6 years, however.)

At the same time, we want to think about the variance structure. The normal approach to 0/1 data is to use a binomial variance $p(1-p)$, which recognizes that the variance varies with the mean μ. At the same time, the fact that this variance function is undefined for predictions outside of (0,1) absolutely forces the use of a link function. Figure 7.6 shows the estimated mean/variance relationship for the colon cancer fit, along with the theoretical binomial variance and the estimated variance from the linear fit. Because

Table 7.12 *Comparison of coefficients using 3 different sets of pseudovalues.*

Term	Estimate	Std.Err	t-value	p.value
Year 4 only				
rxLev	-0.030	0.038	-0.8	0.427
rxLev+5FU	-0.162	0.038	-4.3	¡0.001
extent	0.152	0.031	5.0	¡0.001
node10	0.045	0.006	8.2	¡0.001
Years 2, 4, 6				
rxLev	-0.012	0.035	-0.3	0.728
rxLev+5FU	-0.140	0.035	-4.0	¡0.001
extent	0.151	0.028	5.4	¡0.001
node10	0.046	0.005	9.0	¡0.001
Years 1-6				
rxLev	-0.011	0.034	-0.3	0.736
rxLev+5FU	-0.136	0.033	-4.1	¡0.001
extent	0.139	0.027	5.2	¡0.001
node10	0.045	0.005	9.2	¡0.001

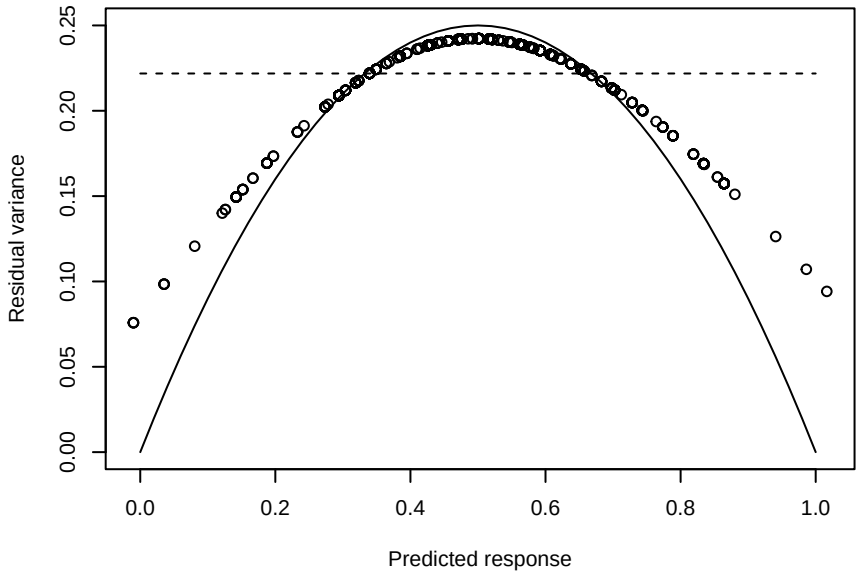

Figure 7.6 *Spline smooth of the squared residuals versus the predicted value, from the linear model using 4 year pseudovalues from the colon cancer dataset. The solid line shows the binomial variance function $p(1-p)$ and the dashed line the estimated variance from the linear model.*

censored data pseudovalues are not constrained to (0,1), the variance does not drop to zero at the endpoints. Use of a binomial variance for this data risks creating artificial outliers through variance inflation, e.g., at $\hat{y} = .5$ the scaled residuals have a variance of approximately .24/.25 = .96 while scaled residuals near $\hat{y} = .05$ would have a variance of .1/(.05*.95) = 2.1.

When using a working-independence variance via GEE or regression using survey sampling weights, inferences are still valid when the wrong variance structure is chosen, but at a possible cost in statistical efficiency. A common compromise for $S(t)$ pseudovalues has been to use a link function to constrain the predictions to (0,1), but not to assume the binomial variance. The simplest form of this is a GLM fit using a Gaussian distribution and either logit or log-log link. All this to the companion.

However, in R at least, directly using this will fail, i.e., the following code will lead to an out-of-range error.

glm(pseudo rx + extent + node4 + factor(year), design=cpdesign, family= gaussian(link = "logit"))

The same out-of-range error occurs with `svyglm`. The issue is that the glm function uses the logit link of $f(y) = \log(y/(1-y))$ to create starting estimates for the iteration, and values outside of $(0, 1)$ lead to a missing value. This is actually the only place in the code that the link is used; all other computations use the inverse link. Two choices are to give explicit initial values or to define our own link. Both of these are shown in the examples document.

Using the log-log link, we get very similar coefficients to the `coxph` fit, while standard errors are just a little bit larger. (Theoretically, coefficients could be closer yet if all timepoints were used.) Comparison of the single timepoint to the multiple timepoint fits raises another interesting point. For both fits, the log-log fit makes an additivity assumption, that on this particular scale the covariate effects are additive — no covariate*covariate interactions are needed. The fit using multiple timepoints has an additional assumption of proportional hazards, as does the Cox model, namely that covariate effects are the same at all timepoints. At the cost of slightly higher standard error, the single timepoint pseudovalue fit frees us from the proportional hazards assumption. On the one hand this makes it an attractive alternative when proportional hazards does not hold; on the other hand it requires explicitly acknowledging that the reported result will depend on which timepoint we choose.

Should we move the coefficients and standard errors for the fits into a table? First 4 columns contain the two treatment coefficients, extent, and node10, followed by the intercept columns for years 1-6 (blanks for models that use a subset of years). For space reasons give beta (se) for the first 4 and only beta for the others. Then rows for linear with 1, 3, 6 times, logistic with 1,3,6, and log-log with 1,3, 6. Then the code all moves out of the main text.

Another note: One wants to use the pseudo value for P(dead) along with logit, probit, or complementary log-log links. That results in coefficients that have the expected sign, large values give higher prob of death.

Summary

In summary, pseudovalues are a useful concept that captures time-to-event outcomes at specific timepoints in a single outcome variable. This allows the use of standard regression methods to model the covariate effects on the time-to-event outcome. We did not touch on model checks, but it is worth mentioning that standard model checks are still important. Pseudo-residuals have been discussed in the literature [5]. These are just the ordinary residuals from the model fit to pseudo y. Scatterplots of residuals are not very useful for the same reason that they are not useful for binary data: you get two stripes at $0 - \hat{y}$ and $1 - \hat{y}$; smoothers are needed. this is not surprising, as residual plots are not very useful in Cox models either. is this the best way to address model checks? Or should we add something to the prior examples?

Modeling RMST is popular in the clinical trial literature, as it avoids the proportional hazards assumption of the Cox model, and provides more meaningful summaries of the differences between treatment arms in terms of years of difference instead of the relative risk measure provided by the hazard ratio. Pseudovalues are also an attractive option for use in machine learning, as they will allow time-to-event outcomes to be analyzed using existing machine learning methods for continuous outcomes. However, some additional work will be needed to ensure the proper variance is accounted for when using pseudovalues with machine learning methods.

We focused on simple examples to introduce this concept, but more complex ways to use pseudovalues have also bee developed. For instance, stratified pseudovalues can also be computed, i.e., where each is based on the observation's leverage within a subset. For example, the `lung` dataset comes from a multi-center study, and contains an identifier for the institution from which the subject was recruited. We might want to adjust for the possibility that different institutions recruit from different patient populations. One way to do so is to add `factor(inst)` to the model, but another is to base the result off per-institution curves.

The simplicity and appeal of this revolutionary concept may make the reader wonder if Cox models can become obsolete. However, pseudovalues do have some limitations. There are many uses for Cox models that cannot be represented using pseudovalues. Time-dependent covariates are an obvious issue that cannot be addressed using pseudovalues. In addition, the flexibility of the Cox model provided by the counting process construct ensures it will continue to be used to model complex data, such as multistate models, which will be discussed in the second part of this book. The use of pseudovalues in the settings of competing risks and multistate models will be discussed in later chapters.

Part II

Multiple endpoints

Chapter 8

Introduction to multistate processes

8.1 Overview

> "We have the duty of formulating, of summarizing, and of communicating our conclusions, in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions." – Ronald Fisher

Disease processes are often more complex than just a single event of interest. A few examples are studies focused on achieving remission with a competing risk of death, studies where patients may experience the outcome (e.g., hospitalization) multiple times, or studies where patients have several states

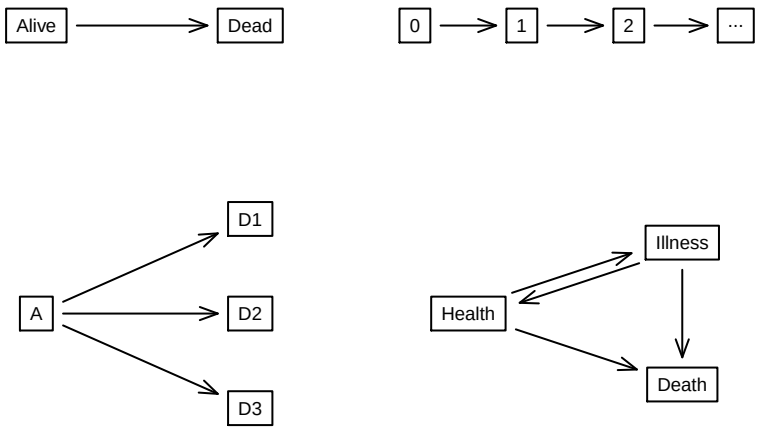

Figure 8.1 *Four multistate models. The upper left panel depicts simple survival, the upper right panel depicts sequential events, the lower left panel illustrates competing risks, and the lower right panel shows a multistate illness-death model.*

(e.g., remission, recurrence, death). In Figure 8.1 the "classic" single outcome model is shown in the upper left, the situation to which the first half of the book was addressed. In this second portion we deal with the more general case. In the author's own work, studies with a single dominating outcome have actually an exception; we think of the data as a more comprehensive process.

The remaining panels of Figure 8.1 show three other illustrative cases; in clockwise order they are

- Repeated or sequential events. An example of this would be recurrent infections, discussed in Chapter 12.
- The illness-death process: subjects can transition from a state of health to a state of illness or to death, and those who are ill to death. The reverse arrow from the illness state back to the health state may also be possible. A variation without recovery will be discussed in Chapter **??**.
- In the classic competing risk process, all subjects start in the same state and each subject can make a single transition to one of several mutually-exclusive, terminal states. This is examined in Chapter 9.

The initial step in any multistate analysis is create a box and arrow plot which will describe the target of the analysis, and help guide it. In this figure each box is a state and each arrow a possible transition between two states. Think this through, as it is key. (We often suggest that the programmer print this out and tack beside their terminal). Not uncommonly, we may want to perform analysis using more than one structure for the states, each of which may shed a different light on the study, revealing different nuances. Chapter **??** gives an extended example of this type.

8.2 Aalen-Johansen estimate

An important component of any analysis will be the Aalen-Johansen estimate, which is the natural extension of the Kaplan-Meier to the multistate case. We use the notation $p(t)$ = probability in state. At each time point t, $p_k(t)$ estimates the probability of being in state k at time t. The Aalen-Johansen method estimates *all* the states at once. Like the Kaplan-Meier, the estimate changes only when there is an observed transition and stays constant between those times; in practice the estimate will be a matrix with m columns and one row for each transition or censoring time.

Figure 8.2 illustrates a small example data set with 5 subjects from an illness-death model, and the table below shows the result of the Aalen-Johansen estimate. At time 0 all of the subjects are in the health state, $p(0) = (1, 0, 0)$. At time 2 subject 2 transitions to the illness state and $p(2) = (.8, .2, 0)$. At time 3 subject 1 transitions, and etc.; at time 11 the AJ estimates that all will be in the death state, $p(11) = c(0, 0, 1)$.

1. The AJ provides estimates for all the states at once.
2. At any given time point $\sum p(t) = 1$ by definition, since p is an estimated

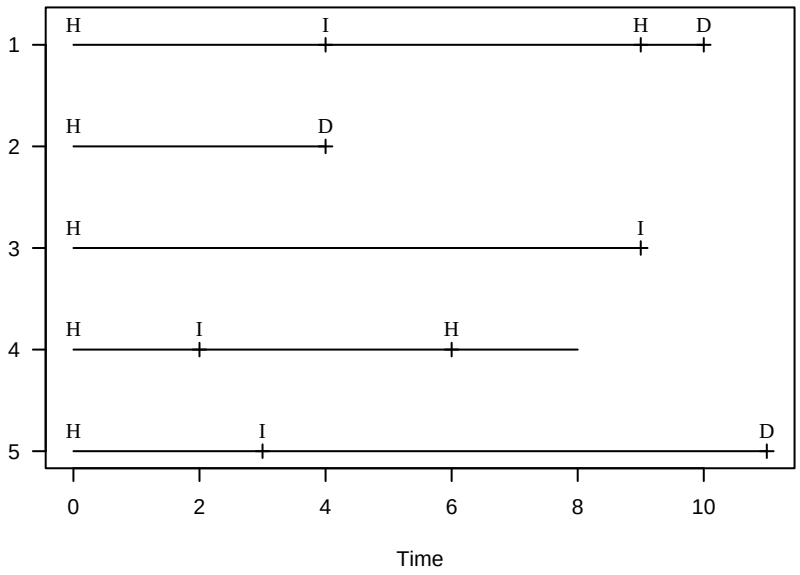

Figure 8.2 *Illustration of a simple illness-death data set. All subjects start in the healthy state (H), and may transition to illness (I), death (D), or from illness back to health (subject 2).*

distribution across the states. Each row of the matrix sums to 1. Put another way, everyone has to be somewhere.

3. The estimate changes only at transition times; when plotted each individual per-state curve will be step function. When there is a censoring, e.g., subject 2 at time 8, it is common practice for programs to add a line to the matrix at that time even though the estimate does not change; this is seen below. (The same is true for the Kaplan-Meier).

```
Time Health Illness Death
  0    1.0     0.0   0.0
  2    0.8     0.2   0.0
  3    0.6     0.4   0.0
  4    0.2     0.6   0.2
  6    0.4     0.4   0.2
  8    0.4     0.4   0.2
  9    0.2     0.6   0.2
  10   0.0     0.6   0.4
  11   0.0     0.0   1.0
```

Computation of the AJ

The AJ estimate is a sequential product $p(t_1) = p(0)T(t_1)$, $p(t_2) = p(t_1)T(t_2)$, $p(t_3) = p(t_2)T(t_3), \ldots$; where $p(0)$ is the initial distribution and each $T(s)$ is the m by m transition matrix at time s. For the simple example above $p(0) = (1, 0, 0)$, everyone in the data set starts in the Health state. The first transition matrix T is at time 2, when subject 2 transitions to the Illness state. The first 4 transitions matrices are

$$T(2) = \begin{pmatrix} 4/5 & 1/5 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad T(3) = \begin{pmatrix} 3/4 & 1/4 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$T(4) = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad T(6) = \begin{pmatrix} 1 & 0 & 0 \\ 1/3 & 2/3 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Row k of $T(s)$ describes the transitions at time s for all those in state k just before that time, ignoring observations that are in other states (each row of T is created independently). Elements are non-negative and each row sums to 1 (everyone in state k has to go somewhere). For tied times, the computational rule is that transitions happen before censoring; further detail is found in **??**.

There are two special cases where this simple matrix multiplication collapses to a more compact form; the simple 2 state model where $p_1(t)$ reduces to the Kaplan-Meier estimate, and competing risks where $p_k(t)$ reprises the cumulative incidence estimator for cause k. (These is left as an exercise for the reader; do the first few multiplications and it becomes clear.) In general, however, there is no simple sub-formula for a particular one of the states, ignoring all others; the natural approach is to compute all states at once. Attempts to use the ordinary Kaplan-Meier for this task via the addition of artificial censoring are doomed to failure. An unfortunately common example is a KM estimate for cause of failure k, censoring all other failure types, in the case of competing risks. There are explanations of exactly why this fails, but the bottom line is that, much like pushing the wrong button on a calculator, you have simple used the wrong formula.

8.3 Multistate hazard model

A second step on multistate data is to model the transition rates between states, the arrows on the multistate diagram. In contrast to the probability in state estimate $p(t)$, each rate can be modeled separately; our primary tool in the remainder of the book will be a multistate hazard (MH) model in which each of the rates is modeled using proportional hazards. For each individual Cox model the risk set at time t is all those who are currently in the box (state) at the foot of the arrow at time t, these are the observations which could experience the transition. The event will be the occurrence of the transition.

The coefficients from a fit will be a matrix with one row for each predictor and one column for each transition. Interpretation of the resulting hazard ratios is no different than for an ordinary Cox model. Probability in state estimates based on these estimated hazards can be created after the PH fits; that computation is closely related to the Aalen-Johansen.

We will also make use of variations of this overall pattern, e.g., one might decide to constrain two coefficients to be equal, force a coefficient to be 0 for some transitions, or constrain the baseline hazards. These will be discussed as they arise.

8.4 Software and data

It will come as no surprise to practitioners that properly creating data set(s) for analysis often takes more time and more work than the analysis itself. For the MH model in particular, there could be a separate "Cox model" data set for every transition. Our examples primarily make use of the R survival library, which takes a particular approach to the process. That is, first create a multistate data set with the following properties.

- Each subject is represented as a set of rows, each of which covers a range of time `(time1, time2]`. Covariates in that row are the values that apply over that time range. This setup is widely used for time-dependent covariates.
- Additional variables describe the current state and next state; the latter contains the state to which this observation will transition at `time2`, or a censoring code if there is no transition.
- For each subject, the data describes a valid path through time
  - A subject cannot be two places at once (no overlapping intervals)
  - Each subject describes a continuous time course (no holes in the followup)
  - Time in any entered state must be > 0. (No intervals of length 0. If a subject enters state "X" at time s, the next time interval starting at s must have X as the current state.)

Essentially, each subject has to describe a path through through the states that does not violate physics. If the above holds, then the routines for the Aalen-Johansen and MH fits can make necessary inferences correctly, without further data manipulation. Nevertheless, creation and validation of appropriate multistate data sets consumes a significant portion of the companion document. One particular warning is that multistate analyses create more opportunities for immortal time bias. The biases which result are as impactful as in the single state case, but sometimes more difficult to spot. Be both vigilant and cautious in this regard.

	Events per subject	
	Single	Multiple
One event type	Standard survival	Repeated events
> 1 event type	Competing risks	Full multistate

Table 8.1 *Four multistate model types*

8.5 Models

For expositional convenience, the remainder of the book will address three varieties of the multistate model in separate chapters.

Chapter 9

Competing Risks

> The statistician cannot evade the responsibility for understanding the process he applies or recommends. -– Sir Ronald A. Fisher

One of the simpler situations where more than one event is encountered is competing risks (CR). Competing risks involve multiple events of different types, but each subject can only experience one of the events. The general case and a simpler one are shown in Figure 9.1. The label comes from the case where the endpoints in the left panel of the figure are distinct causes of death, e.g., cancer, heart disease, etc., only one of which can occur. Often there is an event of interest and a competing risk of death, as shown in the right panel. CR methods are useful in the analysis of data in these situations and also in a much wider context.

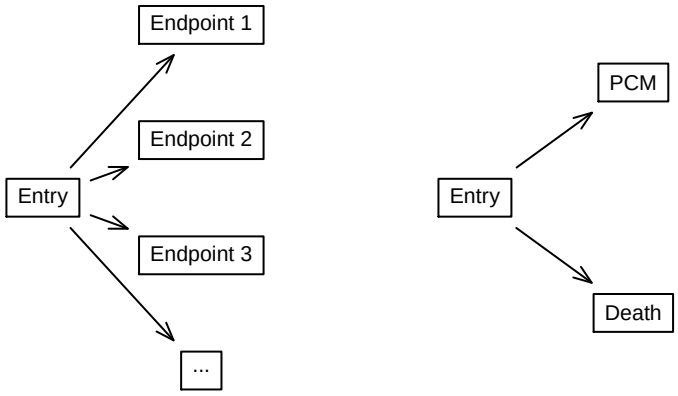

Figure 9.1 *A general competing risks framework is shown on the left, and the simple competing risk framework for the MGUS study is shown on the right.*

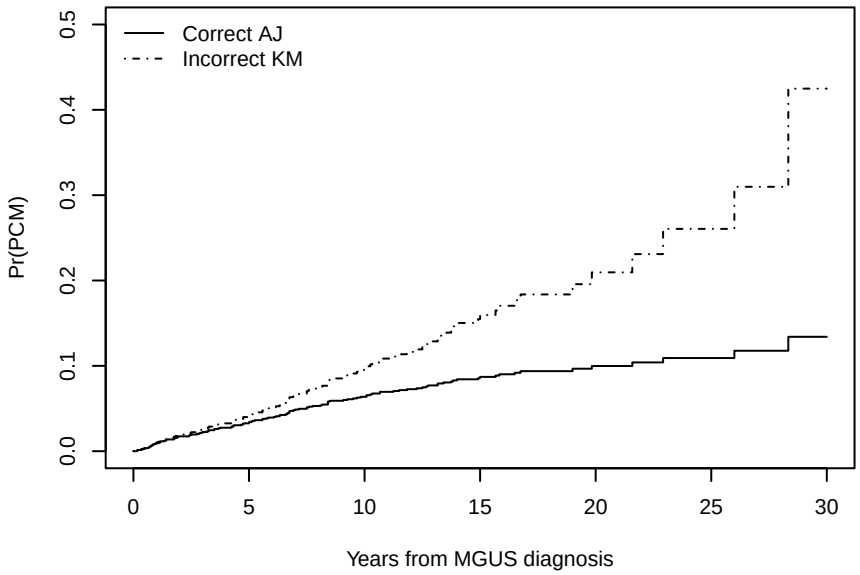

Figure 9.2 *Aalen-Johansen curve using the MGUS data for PCM (solid) and and an invalid estimate (dotdash): the Kaplan-Meier for PCM, treating death as censored.*

Correct analysis tools are needed for working with competing risk data. For instance, consider the MGUS data and the diagram illustrated in the right panel of Figure 9.1. These are subjects who were identified with a serum protein electrophoresis test which revealed a monoclonal spike, but no overt signs of plasma cell disease. The primary question motivating the study was whether these subjects had an increased risk of multiple myeloma or other plasma cell malignancy (PCM) going forward. In this case PCM is the outcome of primary interest, with death as the competing risk. The study observed 115 PCM events, of which the majority were multiple myeloma, and 860 deaths without PCM. This is a case where the competing risk has many more events than the endpoint of primary interest.

Perhaps the most frequent computing error with competing risks data is the use of an ordinary KM estimate for each of the endpoints separately, treating the other outcome as censored. This may be particularly enticing when only one of the event types is of primary interest and others may not be included in the final figure. The result of this approach will be biased upward, however, sometimes substantially so. Figure 9.2 shows the substantial difference between the simple KM estimates for PCM and the correct (Aalen-Johansen) approach. The KM fit attempts to estimate the expected occurrence of PCM if death from other causes did not occur. In that hypothetical world, it is indeed true that many more subjects would progress to PCM (the incorrect curve is higher), but it is also not a world that any of us will ever inhabit.

The competing risk curve correctly estimates the fraction of MGUS subjects who *will experience* PCM, a quantity sometimes known as the lifetime risk, and one which is actually observable.

9.1 Survival probabilities or probability in state

9.1.1 Aalen-Johansen method

Review A.4.2 Aalen-Johansen and see what content here is still required - still includes lots of formulas for this audience.

As with a single survival endpoint, standard summaries of competing risk data are estimates of the probability of being in a given state over follow-up (i.e., probability in state m where there are now multiple states). Within the competing risk literature this if often called *cumulative incidence* (CI) and this can be estimated using the Aalen-Johansen (AJ) method. Like the KM, the CI is simply a special case of the AJ, and we will use AJ as our label for all the non-KM instances of its use.

The AJ method provides estimates of $p(t)$, which is a vector containing the probability of being in each of the states at time t. If there were no censoring, then p could be computed as a simple tabulation at time t of the current state for each subject, in the same way that the KM is equivalent to the empirical cumulative distribution function (CDF) when there is no censoring.

Mathematically the AJ estimate is simple. For each unique time s that an event (transition) occurs, form a transition matrix $T(s)$ whose jk element is the fraction of those in state j at time $s-$ who transitioned to state k at exactly time s. Each row of T will sum to 1, i.e., everyone has to go somewhere.

The AJ estimator is then

$$p(t) = p(0) \prod_{s \le t} T(s) \tag{9.1}$$

where $p(0)$ is the initial distribution of subjects, and $p(t)$ is a vector with one element per state containing the estimated distribution of states at time t.

Let's work this out for the simple two-state alive:dead model with state 1= alive and state 2 = dead. Use the same set of 20 hypothetical subjects as was used to illustrate the KM in Section **??** consisting of 1, 2, 2+, 3, 4+, 4+, 5, 5+, 8, 8, 9, 10+, 11+, 12, 14+, 15+, 16, 16+, 18, and 20+, where the numbers are the follow-up times for each subject and the '+' indicates censoring. The rows of the transition matrix represent the current state; the first row for those currently in state 1 (alive), the second row for those currently in state 2 (dead). All subjects start in the alive state and thus $p(0) = (1, 0)$. The first

few transition matrices are

$$T(1) = \begin{pmatrix} 19/20 & 1/20 \\ 0 & 1 \end{pmatrix}$$
$$T(2) = \begin{pmatrix} 18/19 & 1/19 \\ 0 & 1 \end{pmatrix}$$
$$T(3) = \begin{pmatrix} 16/17 & 1/17 \\ 0 & 1 \end{pmatrix}$$
$$T(5) = \begin{pmatrix} 13/14 & 1/14 \\ 0 & 1 \end{pmatrix}$$

At time 1, one out of the 20 alive subjects died while 19/20 stayed in state 1 (alive). The second row of the matrix is always (0, 1) since there are no transitions from dead to alive, death is referred to as an *absorbing* state.

In general, let $n(s)$ be the number of subjects still alive (in state 1) at time $s-$ and $d(s)$ the number of deaths at time s. Writing out the matrices for the first few transitions and multiplying them leads to

$$p_1(t) = \prod_{s \le t} \left[n(s) - d(s) \right] / n(s) \tag{9.2}$$

which we recognize as the KM estimate of survival. For the two state alive-dead model, the AJ estimate has reprised the KM, or equivalently, the KM is a special case of the AJ.

For competing risks, the probability in state has traditionally been estimated using the cumulative incidence (CI) function

$$CI_k(t) = \int_0^t \hat{\lambda}_k(u) S(u-) du \tag{9.3}$$

where $\hat{\lambda}_k(s) =$ (number of observations experiencing endpoint k at time s)/ (number still at risk at $s-$) is the observed incidence function for outcome k, and S is the overall survival curve for "time to any endpoint".

The AJ transition matrix for the competing risks case looks very similar to the KM. For the 3-state version shown on the right side of Figure 9.1:

$$\hat{\lambda}_{1k}(s) = \left(\sum_i dN_{i1k}(s) \right) / \left(\sum_i Y_{i1}(s) \right) \tag{9.4}$$

$$T(s) = \begin{pmatrix} 1 - \left(\hat{\lambda}_{12}(s) + \hat{\lambda}_{13}(s) \right) & \hat{\lambda}_{12}(s) & \hat{\lambda}_{13}(s) \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$p(t) = p(0) \prod_{s \le t} T(s) \tag{9.5}$$

The individual hazard estimates (9.4) are equivalent to the ordinary Nelson-Aalen formula. Multiplying out the matrices, we see that the first element of $p(t)$ is the KM estimate for "any transition", while the remaining elements of $p(t)$ are exactly the CI estimates for each of the terminal states. That is, the CI estimate is also a special case of the AJ estimate. The label "cumulative incidence" is one of the more unfortunate ones in the survival lexicon, in our opinion, since we normally use 'incidence' and 'hazard' as interchangeable synonyms, but the CI is *not* an integral of the hazard. We will purposely avoid this label whenever possible and refer to this as an AJ estimate. Some authors define the CI using the exponential form $\exp(-\Lambda(t))$ instead of the KM form for S. The result becomes a hybrid of the AJ and the exponential form of the AJ, as shown in Section 11.3.

The standard error of the AJ estimate can be computed using an infinitesimal jackknife. Let $D(t)$ be a matrix with one row per subject and one column per state. Each row contains the estimated *change* in $p(t)$ corresponding to subject i, i.e., the derivative of p with respect to the ith subject's case weight, i.e.,

$$D_{ij}(t) = \frac{\partial p_j(t)}{\partial w_i}$$

Then $V(t) = D'WD$ is the estimated variance-covariance matrix of $p(t)$, the diagonal matrix W of sampling weights will normally be the identity. Mathematical details are in the formula appendix. Interestingly, for an ordinary KM with case weights of 1, V is equal to the usual Greenwood variance.

The $p(t)$ vector obeys the obvious constraint that $\sum_j p_j(t) = 1$, i.e., at any given time point t, each observation has to be somewhere. One possible label for p is the *current prevalence* estimate, since it estimates what fraction of the subjects are in any given state across time. However the word "prevalence" is certain to generate confusion whenever death is one of the states, due to its historic use as the fraction of *living* subjects who have a particular condition. We will use the phrase *probability in state* or simply $p(t)$ from this point forward.

In the simple two state model, $p_1(t)$= Pr(alive) is the usual KM survival estimate, and we have $p_1(t) = 1-p_2(t)$, or equivalently Pr(alive) = 1 - Pr(dead). Plots for the 2 state case sometimes choose to show Pr(alive) and sometimes Pr(dead). For simple survival we have gotten used to the idea of using Pr(dead) and Pr(alive) interchangeably, but that habit needs to be left behind for multistate models. In multistate models, curves of $1 - p_k(t)$ = probability(any other state than k) are not useful.

Example

The companion document shows how to obtain the curves using the R survival package; the essential step in this case is to create an augmented status variable that is a factor and has levels: censor, PCM and death. In R, it suffices to use this three level code as the "status" variable in a `survfit` call,

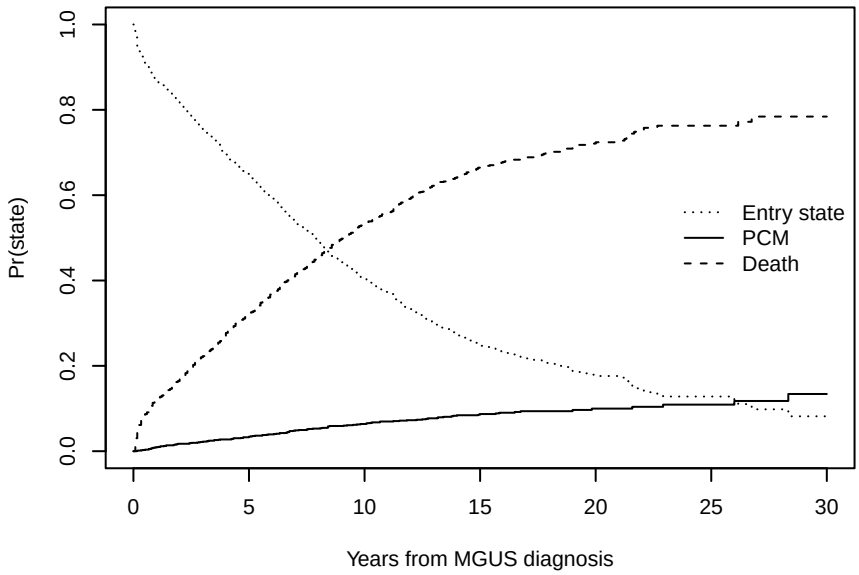

Figure 9.3 *Competing risk curves for the MGUS data with the events PCM (solid line) and death (dashed line), plus the entry entry state (dotted)*

replacing the usual 0/1 indicator used for an ordinary KM. (In other packages a separate routine may be needed to compute the AJ estimate.) Figure 9.3 shows AJ curves for the data. At 30 years post enrollment, about 13% of the MGUS subjects have experienced plasma cell malignancy (PCM), 78% have died without PCM, and 8% are still alive without PCM. Since probabilities for the three states of entry, PCM, and death must add to 1, it is common to only display two of the three when plotting the outcomes; Pr(still in the entry state) is the least interesting so it is almost uniformly omitted, however it is shown in the summary of the fit as seen in the output below.

```
> summary(msurv, scale=12, times=c(0,10,20,30)*12)
Call: survfit(formula = Surv(etime, event) ~ 1, data = mdata)

 time n.risk n.event Pr((s0)) Pr(PCM) Pr(death)
    0   1384       0   1.0000  0.0000     0.000
   10    424     781   0.4045  0.0637     0.532
   20     57     177   0.1762  0.0998     0.724
   30      3      15   0.0818  0.1340     0.784
```

Note it is important to have the same amount of follow-up for all of the competing events. Often there may be additional data on death outcomes available beyond the last time point where information about the outcome of interest was collected or available. This additional information on a single

outcome is not helpful; the subject will have to be censored at the last time point when information about *all* possible outcomes was available. Otherwise, it is possible that the outcome of interest occurred prior to death and was not recorded.

The redistribute-to-the-right computation illuminates why this is the case. Observations "hand off" their case weight to the remaining uncensored observations at the point they are censored; those remaining observations act as representatives to reflect the future probability of an event for the censored subject. But a subject who experiences death does not have a future probability of a PCM event, and their weight should not be redistributed.

9.1.2 Plotting multiple competing risks

A more classic competing risk example is found with the `flchain` dataset, which contains long term follow-up of 7874 subjects who were assayed for free light chain immunoglobulin. It is an interesting dataset in that the original population included nearly all subjects over the age of 50 years in Olmsted County, Minnesota, so it provides a snapshot of overall population mortality in the region. There are 2169 observed deaths that were annotated using the ICD chapter of the primary cause of death. For illustration purposes we will divide the deaths into 4 groups: Blood and Circulatory, Neoplasms, Respiratory, and Other. We will follow the analysis of Dispenzieri et. al [28] and divide subjects into two groups who are above/below the 90th percentile of overall FLC level, and also divide by sex.

Figure 9.4 shows results of the AJ fit. At the right hand edge the four causes will add to 1.0 since everyone has died. The interesting aspects are the relative proportions and timings of the four death endpoints, e.g., at age 80 years the circulatory deaths are 6% for females and 13% for males in the left column, but the final circulatory tallies are quite similar. A primary message, however, is that such displays are complex, and sound graphical design can play as large a role as statistical estimation. It is not feasible to place all 16 curves in a single panel.

Another arrangement that may be useful is the cumulative or stacked format shown in Figure 9.5. The proportion with death due to neoplasm are presented by the difference between the first and second curves, the respiratory deaths by the difference between the second and third, and other deaths by the difference between the third and the fourth. The upper curve marks the occurrence of any death. The cumulative curve is preferred by some, but the visual impression is strongly affected by the order in which the endpoints are presented.

Figure 9.6 shows the probability in state and estimated hazards, by sex, for the MGUS data. The death rate for males is higher than for females as would be expected for subjects in this age range where the median detection age is 72 years. The hazard rates for PCM appear to be essentially identical for males and females at approximately 1% per year. Though the hazards are

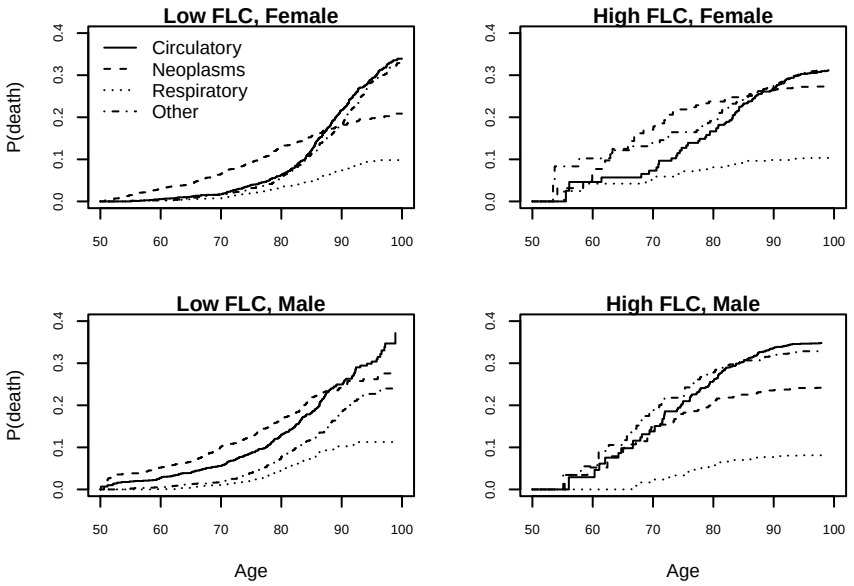

Figure 9.4 *Competing risk curves for causes of death (circulatory, neoplasms, respiratory, or other) using the FLC study. The right column shows the subset of subjects in the upper 10 percent of FLC values and the left column shows the bottom 90 percent. The top row contains females and the bottom row contains males.*

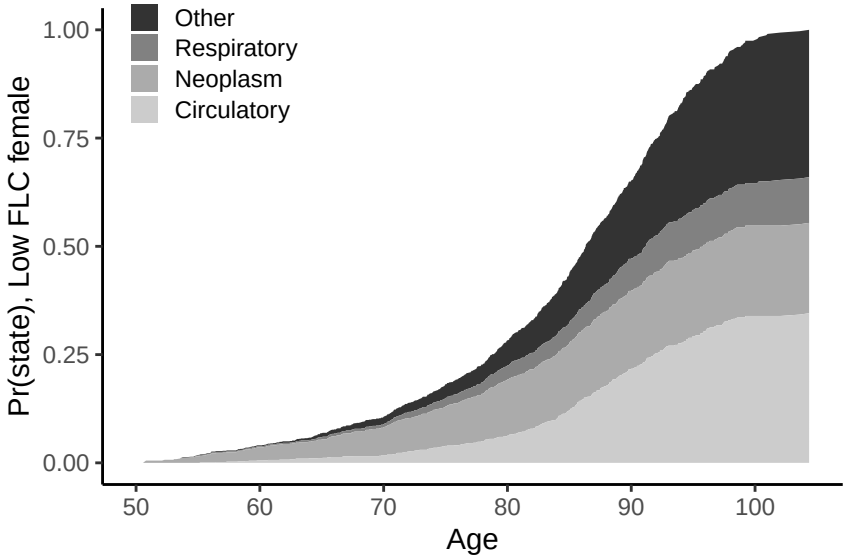

Figure 9.5 *Stacked plot of death probabilities by cause for females with low FLC.*

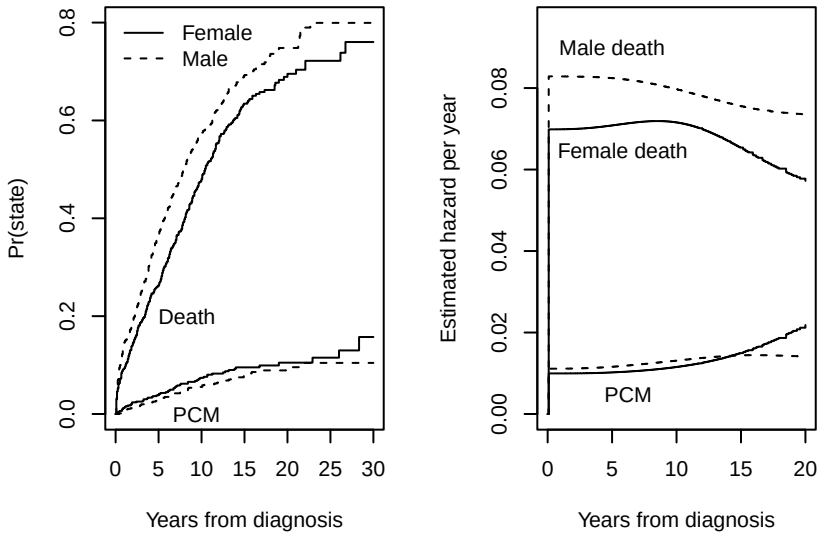

Figure 9.6 *Left panel: absolute risk estimates stratified by sex for death and PCM using the MGUS data. Right panel: estimated hazards.*

equal, a female's absolute risk of PCM at 30 years post diagnosis is about 5% higher than a male's. Females live longer, and thus accumulate more exposure to the 1%/year PCM risk.

9.2 Models

As noted in Chapter 2, an ideal model satisfies the three conditions of simplicity, validity, and reliability. While competing risks add additional complexity to the modeling process, several approaches to modeling competing risk outcomes can potentially satisfy these conditions.

1. Multistate hazard model: An extension of the usual Cox model and our preferred approach. This approach is also referred to as the cause-specific hazard model.
2. Fine-Gray model: this approach was specifically developed to model competing risks using sub-distribution hazards. However, there are deficiencies with the biological plausibility of this model, as well as issues with the proportional hazards assumption.
3. Pseudovalues: As mentioned previously in Chapter 7, pseudovalues can also be used in the competing risk setting.

Determining the best choice of model for a particular dataset involves multiple considerations, which will be discussed in the rest of this chapter.

9.2.1 Multistate hazard model

Perhaps surprisingly, hazards can be estimated singly, each computed without reference to any other transitions [43]. Quite simply, a separate Cox model can be applied to each of the transitions in a multistate model, in this case the transition from entry to PCM and the transition from entry to death without PCM. This can be done using two disjoint Cox models, each with all n subjects, using 2 separate event variables. This means that any statistical package can be used to fit a multistate hazard (MH) model, simply by creating the required stacked dataset with a row of data for each transition for each patient. Packages such as mstate in R also provide tools for creating such a dataset.

The `coxph` function in the survival package takes an alternate approach for MH models: the user is expected to create a valid counting process dataset, containing an identifier variable that labels subjects and a multistate outcome status. In the competing risk setting, this dataset will typically have one row of data per patient, and the multi-level status variable needs to be a factor. The routine then does all of the necessary bookkeeping internally. The same dataset and approach is used for Cox models and for the AJ curves. This avoids the need to create the stacked dataset, which can be error-prone.

Our analysis uses three covariates from the MGUS data: age, sex, and the magnitude of the monoclonal spike (from the SPEP test). The `mdata` dataset is a copy of the `mgus2` data found in the survival package, with additional variables `etime` which is the time in months to the first of PCM or death and `event`, a factor variable with the levels of censor, PCM, and death. The variable `age10` is age in decades; this makes the age coefficient similar in size to the others. The code below then fits the default model, which has a separate baseline hazard and set of coefficients for each transition.

```
> mfit1 <- coxph(Surv(etime, event) ~ age10 + sex + mspike, mdata, id=id)
> print(mfit1, digits=2)
Call:
coxph(formula = Surv(etime, event) ~ age10 + sex + mspike, data = mdata,
    id = id)


1:2        coef exp(coef) se(coef) robust se   z     p
  age10   0.164     1.178    0.084     0.069 2.4  0.02
  sexM   -0.005     0.995    0.188     0.188 0.0  0.98
  mspike  0.884     2.421    0.165     0.168 5.3 2e-07


1:3        coef exp(coef) se(coef) robust se    z      p
  age10   0.652     1.919    0.036     0.037 17.4 <2e-16
  sexM    0.389     1.475    0.070     0.067  5.8  5e-09
  mspike -0.059     0.942    0.064     0.062 -1.0    0.3
```

```
 States:  1= (s0), 2= PCM, 3= death

Likelihood ratio test=419  on 6 df, p=<2e-16
n= 1373, number of events= 969
   (11 observations deleted due to missingness)
```

The effect of age and sex on non-PCM mortality is profound ($p < .001$), which is not a surprise given the median starting age of 72 years. Risk rises 1.9 fold per decade of age and the death rate for males is 1.5 times as great as that for females. [1] The size of the serum monoclonal spike has almost no impact on non-PCM mortality; a 1 unit increase changes mortality by only 6%.

The mspike size has a major impact on progression, however; each 1 gram change increases risk of PCM by 2.4 fold. The interquartile range of `mspike` is 0.9 grams, so this risk increase is clinically important. The effect of age on the progression rate is much less pronounced, with a coefficient only 1/4 that for mortality, while the effect of sex on progression is completely negligible.

Model checks

Before proceeding further it is important to check the key assumptions of proportional hazards and linearity. The table below shows the result of a score test for proportional hazards individually for each variable.

```
               chisq df       p
PCM: age10     2.766  1   0.096
PCM: sex       1.093  1   0.296
PCM: mspike    0.284  1   0.594
Death: age10  33.110  1 8.7e-09
Death: sex     0.746  1   0.388
Death: mspike  0.392  1   0.531
```

There is clearly a failure of PH for age with respect to the death outcome. Further examination shows that death rates for the first two years after MGUS diagnosis are actually quite a bit higher than the known population rates. This is visible in Figure 9.3 which shows a steep upward step in the first year for death. This is due to a selection effect: SPEP is a test that will be ordered when the physician either suspects or needs to rule out one of several serious conditions, i.e., it tends to be ordered when there is a serious health issue. An internal study (not shown) showed that this early excess applied to everyone who had the test ordered, regardless of the outcome of the test. In the first 1–3 years after the test is ordered everyone has a high death rate, independent of their age, i.e., being younger doesn't help. A simple approach is to start

[1] Yearly death rates for Minnesota can be found in the `survexp.mn` dataset, and show values of (1.8, 4.6, 1.1, 2.7) percent for sex/age of male/65, female/65, male/75 and female/75, respectively in 1995; the MGUS hazard ratios for death closely mimic these values.

the analysis at 3 years post detection so as to avoid this selection. This gives the fit below. Further tests for PH or for non-linearity are not significant.

```
Call:
coxph(formula = Surv(etime, event) ~ age10 + sex + mspike, data = mdata,
    subset = (etime > 36), id = id)


1:2         coef exp(coef) se(coef) robust se     z        p
  age10  0.08770   1.09166  0.09611   0.08128 1.079 0.280575
  sexM   0.02695   1.02732  0.22088   0.21897 0.123 0.902041
  mspike 0.75521   2.12805  0.20013   0.20312 3.718 0.000201


1:3          coef exp(coef) se(coef) robust se      z        p
  age10   0.81138   2.25101  0.04805   0.04741 17.115  < 2e-16
  sexM    0.41788   1.51874  0.08713   0.08448  4.946 7.56e-07
  mspike -0.01253   0.98755  0.08199   0.08283 -0.151     0.88

 States:  1= (s0), 2= PCM, 3= death

Likelihood ratio test=365.4  on 6 df, p=< 2.2e-16
n= 1035, number of events= 634
   (6 observations deleted due to missingness)
```

Absolute risk

Although the basic fit can be done using any Cox model program, obtaining the predicted *absolute* risk from these fits requires further code. As outlined in Section 11.3, the necessary algorithm is an extension of the AJ estimate, i.e., a product of matrices that have a row and column for each state. Elements of the matrices are the predicted hazards from the appropriate Cox models. Using the (standard) predicted survival from any single one of those Cox models makes exactly the same error as using the KM instead of the AJ estimate, and the result will likewise be an overestimate of the probability of that particular outcome.

Figure 9.7 shows proper predicted risks for four hypothetical subjects, a male and female with monoclonal spikes of 0.5 and 1.5 grams , each of the four aged 72 years at MGUS detection (the median age). The initial horizontal segment extends to the landmark point at year 3. Even though sex has no effect on the hazard of PCM ($p > .9$) the predicted lifetime risk of PCM is higher for females; .02 for mspike = .05 (.074 vs .053) and .04 for mspike =1.5 (.148 vs .108). Likewise, spike size has an effect on the probability of death, even though its hazard impact is small.

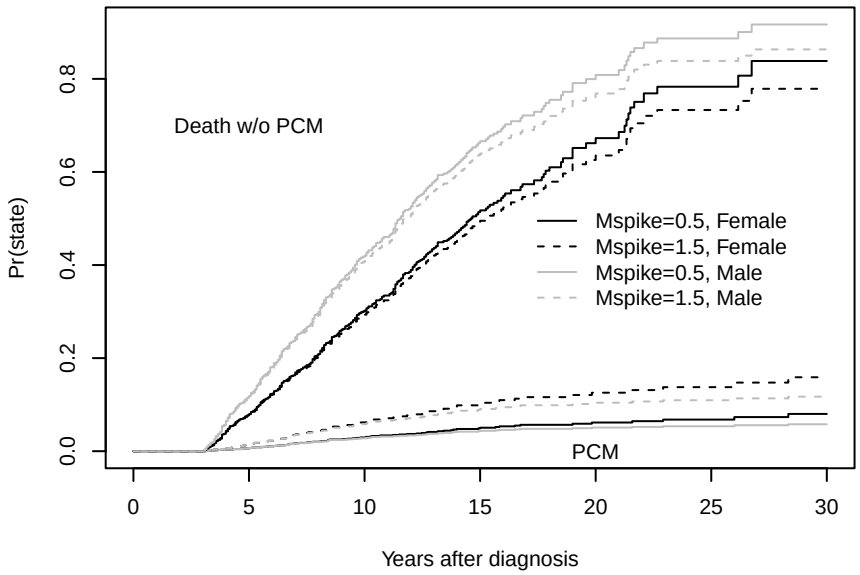

Figure 9.7 *Predicted absolute risk of a 72 year old using the multistate model fit with the covariates age, sex, and mspike. The data used in this model begins 3 years after detection of MGUS.*

9.2.2 Fine-Gray model

The Fine-Gray (FG) model provides an alternate way of looking at competing risks data. One of the motivations for the FG approach was that the hazard approach does not directly answer the question of absolute risk, e.g., a single numeric coefficient giving the effect of sex on the lifetime risk of PCM (and a p-value). The FG approach models competing risks based on the assumption shown in equation 9.6, that is, that the complimentary log-log transformation leads to a simple additive model.

$$\log(-\log(1 - p_j(t))) = \beta_0(t) + \beta_1 x_1 + \beta_2 x_2 + \dots \tag{9.6}$$

In practice, the FG approach transforms the multistate problem into a collection of single transitions. In the left panel of Figure 9.1, draw a circle around all of the states except the chosen endpoint and collapse them into a single meta-state, leading to one model for "transition to PCM" vs. (death or entry) and another for "transition to death" vs. (PCM or entry).

The computational complexities of the FG model relate to censoring: the natural denominator at time t includes those who have not yet experienced either an event or censoring, and those who have experienced a type 2 transition but *would not yet* have been censored. Estimating this last quantity requires a strong assumption with respect to independence of censoring and outcomes. In an ordinary Cox model, we assume that censoring is independent, condi-

tional on the linear predictor $X\beta$. In the FG model, the assumption is that censoring for event type 2 is independent conditional on the linear predictor for event type 1. In rare cases this assumption can be explored directly, i.e., the event of interest is non-fatal, and ongoing follow-up continued after the event.

Geskus [39] has shown that the fit can be done as a two stage process: the first stage creates a special dataset with case weights, then the second fits a weighted `coxph` or `survfit` model to the data. In R the dataset can be created using the `finegray` command. In SAS it is done behind the scenes as part of the `phreg` procedure. In both cases, an advantage is the ability to use standard Cox model code for the numerical work. This approach also allows the use of standard methods to test for PH, which translate to a test of equation (9.6), and to create predicted curves for $p_1(t)$.

One minor downside is that the intermediate dataset will be much larger than the base. For instance, the MGUS baseline dataset has 1,384 observations, whereas the FG data has 41,775 observations for PCM and 12,910 observations for death. These odd risk sets are perhaps the most frequently listed issue with the FG model, but this is actually a minor complaint. The state probabilities $p(t)$ in a multistate model are implicitly fractions of the total population we started with: someone who dies in month 1 is still a part of the denominator for the fraction of subjects with PCM at 20 years. In the FG formulas this subject explicitly appears in risk set denominators at a later time, which looks odd but is largely an artifact. Interestingly, if we compute a weighted KM using the expanded datasets, this will reconstruct the AJ estimate.

Another concern with the FG model is that a physical manifestation of it is hard to visualize. A compound containing two radioisotopes, for instance, is a simple example that matches the multistate hazard model. Concrete examples such as this can help us think through what a given estimate or procedure actually means. We have yet to find a physical or disease process which would, in theory, obey proportional subdistribution hazards.

When there is only a single categorical 0/1 covariate, the FG model reduces to Gray's test of the subdistribution function, in the same way that a `coxph` fit with a single categorical predictor is equivalent to the log-rank test.

Delayed entry raises another wrinkle in the process. Need to fill in, once I understand what Geskus is thinking when he sets this up. I think it's another riff on the censoring issue but not sure. Those who die before entry also need a portion in the denominator?

Proportional hazards

One of the major issues with the FG model is the PH assumption. The table below shows the tests for PH from the two FG fits - one for PCM and one for death. As with the multistate hazards (MH) model, we see significant non-PH for age and death, for the same patient selection issue identified then. There is however also strong non-proportionality for age and the PCM endpoint.

	Multistate		Fine-Gray	
	PCM	Death	PCM	Death
Age (decades)	0.09 (0.10)	0.81 (0.05)	-0.30 (0.08)	0.70 (0.05)
Male sex	0.03 (0.22)	0.42 (0.09)	-0.23 (0.22)	0.37 (0.09)
Spike size	0.76 (0.20)	-0.01 (0.08)	0.78 (0.19)	-0.17 (0.08)

Table 9.1 *Coefficients from the multistate hazard model and the Fine-Gray model, for the MGUS example.*

```
               chisq df       p
PCM: age10    21.336  1 3.9e-06
PCM: sex       0.142  1  0.7067
PCM:mspike     0.037  1  0.8474
Death: age10   8.670  1  0.0032
Death: sex     1.411  1  0.2349
Death: mspike  0.568  1  0.4510
```

Table 9.1 shows coefficients from the MH and FG fits for both endpoints. We clearly see the difference in focus: male sex has no effect on the hazard of PCM, but a negative effect on the subdistribution PCM hazard, and spike size has no effect on the death hazard, but a negative effect on the subdistribution death hazard. Standard errors of the estimates are comparable. The difference in the age effect on PCM is interesting, slightly positive for the MH model and strongly negative for the FG model. Higher age increases death rates (undoubtedly) therefore leading to less chance of PCM.

Unlike the MH model, using a landmark of 3 years does not ameliorate the PH issue: strong non-PH is still evident for age with the PCM endpoint. There is clearly something other than patient selection at work. Figure 9.8 shows the estimated coefficient of age over time; PH corresponds to a horizontal line. At the start, age has little effect on accumulation of PCM events; by 30 years, each decade is estimated to reduce it by 3 fold. An underlying reason for this effect will become more clear when we look at predictions of absolute risk. Nevertheless, for consistency with the multistate results, we continue with the landmark data for both models.

Absolute risk

Unlike the MH model, the curve produced from a single FG (pseudo Cox) fit is the correct one; the underlying model (if correct) has already dealt with the multistate issues. However, in Figure 9.9, the predictions at specific ages are quite different for the FG model: new PCM cases are predicted 20+ years after diagnosis in both the old and young ages, while new PCM cases are predicted to cease in the MH fit. The average of all four curves is nearly the same at each age for the MH and FG models — they both estimate the mean correctly — but the global PH assumption of the FG model forces the curves to remain parallel, even though this is unrealistic. For a cohort of 80 year olds,

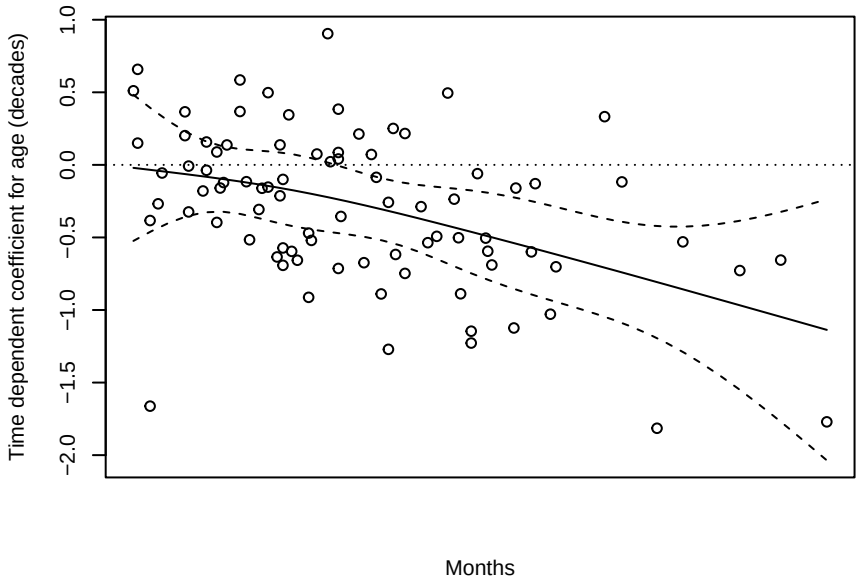

Figure 9.8 *Estimated time-dependent coefficient for age, Fine-Gray model for PCM. The data used in this model begins 3 years after detection of MGUS.*

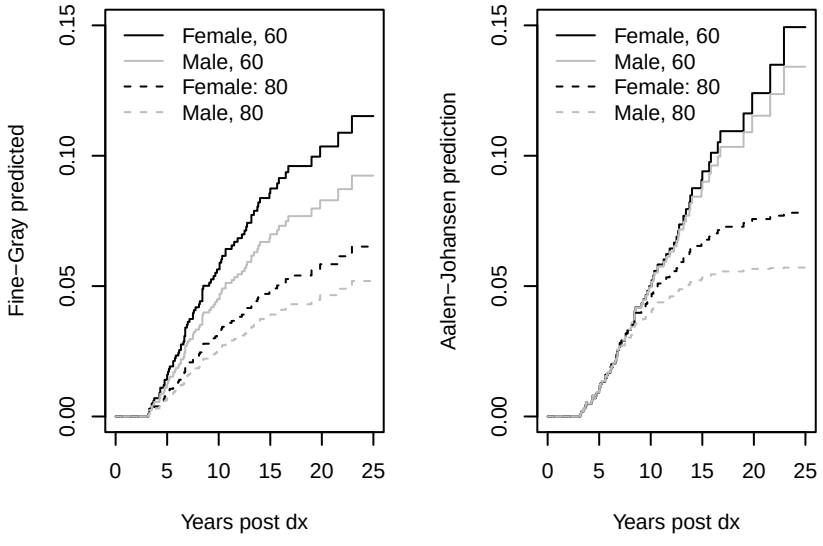

Figure 9.9 *Comparison of Fine-Gray (left panel) and Aalen-Johansen (right panel) estimates of absolute risk of PCM in the MGUS dataset, by sex, for ages at diagnosis of 60 and 80 years and a serum mspike of 1.2 grams.*

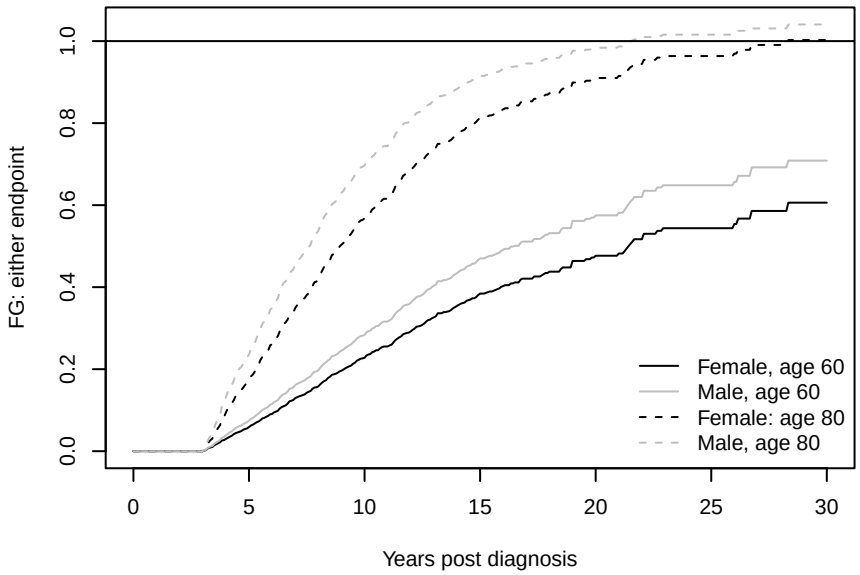

Figure 9.10 *Absolute risk estimates for both outcomes combined using the Fine-Gray model. Note that the combined probability of an event can exceed 1.*

there can be no further PCM cases after 20 years since all the subjects will be dead: the 80 year olds' curve *must* become horizontal, entirely due to the depredations of the other outcome. This is why a time-dependent age effect becomes so strongly negative.

A further weakness of the FG approach is that since the two endpoints are modeled separately, the results do not have to be consistent [11]. Figure 9.10 shows the predicted fraction who have experienced either endpoint. For subjects diagnosed at age 80 years, the FG model predicts that more than 100% will either progress or die by 30 years. Predictions based on the MH approach do not have this issue.

The primary strength of the FG model with respect to the MH approach is that if lifetime risk is a primary question, then the model has given us a simple and digestible answer to that question: "females have a 1.26 fold higher lifetime risk of PCM, after adjustment for age and serum m-spike". This simplicity has a high price, however; lacking PH, such simplicity is only an illusion. The authors are not proponents of the FG approach.

Mathematics of Fine-Gray

The mathematics behind the FG estimate starts with the functions $F_k(t) = p_k(t)$, where p is the same probability-in-state function estimated by the AJ estimate. This can be thought of as the distribution function for the improper random variable $T^* = I(\text{event type} = k)T + I(\text{event type} \neq k)\infty$. Fine and

Gray refer to F_k as a subdistribution function. In analogy to the survival probability in the two state model define

$$\gamma_k(t) = -d\log[1 - F_k(t)]/dt \tag{9.7}$$

and assume that $\gamma_k(t; x) = \gamma_{k0}(t)\exp(X\beta)$. In a 2 state alive $\longrightarrow$ death model, γ becomes the usual hazard function λ. In the same way that our multivariate Cox model `mfit1` made the simplifying assumption that the impact of male sex is to increase the hazard for death by a factor of 1.48, independent of the subject's age or serum mspike value, the Fine-Gray model assumes that each covariate's effect on $\log(1 - F)$ is a constant, independent of other variables. Both models' assumptions are wonderfully simplifying with respect to understanding a covariate, since we can think about each covariate separately from all the others. This is, of course, under the assumption that the model is correct, i.e., that additivity across covariates, linearity, and PH all hold.

The model can also include interactions and transformations of the covariates so more complexity than (9.6) is possible, but the key feature remains, which is that $\beta_0(t)$ factors out as a linear term. This is directly parallel to the PH assumption in the Cox model. It is also true that model (9.6) cannot hold simultaneously for more than one endpoint, but most analyses focus on only one chosen outcome.

In a simple Cox model with states of 0=entry and 1=event, $\Lambda(t) = -\log(1 - p_1(t))$ is the cumulative hazard function, and equation (9.6) is precisely the proportional hazards assumption. Fine and Gray define $-\log(1 - p_1(t))$ to be the cumulative subdistribution hazard; and the model assumes proportional subdistribution hazards. If there were no censoring, estimation of the subdistribution is fairly clear. It is based on ratios of (number of transitions of type 1)/(number who have not yet had a type 1 transition); the denominator includes those who have experienced other transitions. This quantity appears odd at first, but it corresponds to the fact that $p_1(t)$, in the uncensored case, is the number of events of type 1 divided by the total sample size; $(1 - p_1)$ includes both those who have not yet had any transition and those who transitioned to state 2.

9.2.3 Pseudovalues

An attractive way to estimate the effects of a given covariate on the absolute risk is to fit a model directly to the absolute risk. To allow for censoring, we make use of pseudovalues as discussed in Chapter 7. As before, the starting point is to ask what results we are interested in, e.g., lifetime risk or an average covariate effect over time.

Pseudovalues for the competing risk case have the same form as was explored in Chapter 7. The difference is that they will, in this case, be based on the AJ estimate; each observation will have an influence/pseudovalue on the absolute risk of PCM and a second one for death. Computation and use of the

	30 year linear	6 point linear	6 point cloglog	Fine-Gray
Age (decades)	-0.053 (0.021)	-0.019 (0.007)	-0.175 (0.092)	-0.169 (0.056)
Male sex	-0.054 (0.044)	-0.019 (0.016)	-0.253 (0.256)	-0.213 (0.180)
Serum M spike	0.119 (0.040)	0.077 (0.017)	0.889 (0.214)	0.888 (0.149)

Table 9.2 *Results for 3 fits using the pseudovalues for PCM. The first column uses only the 30 year follow-up, the second and third the 5, 10, 15, 20, 25 and 30 year values. The first two fits use a linear link and the last a complementary log-log link in the GLM fit. The last column shows coefficient from the Fine-Gray model.*

resulting values is no different than the single state model. Important questions, as before, are what time point(s) to select for the summary, whether an average over multiple time points is useful (and if so which time points), and which transformation of the data is best: identity, logit and complimentary log-log are three common choices. Pseudovalues can also focus on the sojourn time in each state.

Table 9.2 shows the results of 3 pseudovalue fits to the probability in state values. The first uses only the value at 30 years, and not surprisingly reproduces the difference of 5% at 30 years that is seen in the AJ curves of Figure 9.6, given that age and mspike values are reasonably balanced with respect to sex. The second model uses all 6 time points of 5, 10, ..., 30 years jointly, and again an average male/female difference of .02 is what we would expect given the AJ plot. It is interesting that for both of these models, the predicted effect of a 10 year increase in age is about equal to the male/female effect.

The third fit mimics the assumptions of a FG model found in equation (9.6), using all 6 time points and a complimentary log-log transform as the link function; notice how closely the coefficients track the FG results, though with a larger standard error. Figure 9.11 shows the predicted values from this fit, plotted atop predicted PCM curves from the FG model. These also agree quite well, reprising the assumed (and inaccurate) parallelism between ages 60 and 80 years. Modeling the effects of covariates on a single outcome does not work well when those covariates also have a strong effect on the competing outcome.

I did this with the full data. I am fairly sure that using the landmark AJ (add start.time option) the pseudovalues will then track the landmarked FG. Both in the companion? Which one here?

We could also use pseudovalues to explore the restricted mean time in state. For the MGUS data this estimate does not make medical sense, however, at least when the data is treated as a competing risks problem. Say we set the upper limit for the RMTS at 30 years = 360 months. Now consider subject 124, who progressed at 80 months and died at 82 months: the RMTS estimate will credit this subject with 360-80 = 260 months in the PCM state. The competing risk curves of Figure 9.3 are sensible, the upper plateau 16% PCM

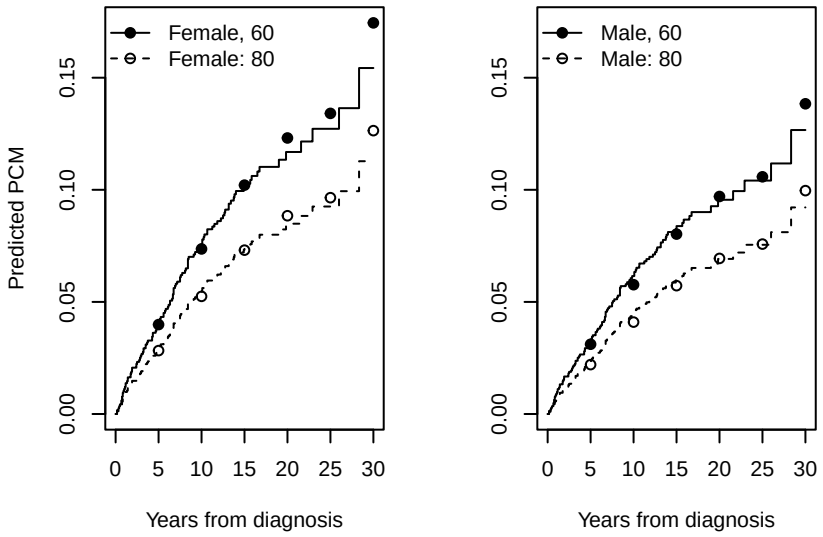

Figure 9.11 *Predicted curves from the Fine-Gray MGUS model and points from the pseudo-values analysis at 6 time points for 4 hypothetical subjects.*

at 35 years represents the cumulative fraction of subjects who had a PCM before death. The area under the curve, the AJ estimate of the RMTS, is however not interpretable as an average sojourn time in that state. Since subjects in the MGUS study continued to be followed for death after PCM, for this dataset we can add the transition from PCM to death to close the loop and instead estimate the multistate model; such models are the topic of Chapter 11, where this dataset will be re-visited.

9.3 Adjusted curves

It is possible to estimate AJ curves that are adjusted for both competing risks and other covariates of interest to ensure comparability. This involves utilizing the methodology described in this chapter along with the methodology shown in Chapter 5. The simplest approach is to use inverse probability weights, which can be calculated and used to re-weight the AJ curves. Model based adjustments are also possible, though they tend to require additional work. See the companion example document for sample code demonstrating how to do this using the FLC dataset. more here?

9.4 Conclusions

Competing risks are perhaps the most common type of data involving multiple events of different types, and it is important to use the right tools to analyze

the data. For estimating absolute risk, the KM will not provide the correct values and the AJ is required. However, the right choice of model is less straight-forward. It depends on the research question.

- The MH model is a simple model for the hazard rates for a single outcome without considering any other competing outcome. If the goal is to assess risk factors for a specific outcome, it is acceptable to fit each transition separately. This is usually the model of choice when the research question involves the biological mechanisms relating risk factors to occurrence of a particular disease. The absolute risk, however, will depend on the rate models for all transitions, since the *number* of subjects added to any one outcome at time t depends on both the rate and the number still at risk; and the latter is influenced by all of the outcomes.
- The FG model focuses on the actual rate of occurrence of an outcome, incorporating the competing risks of other outcomes. In essence, this model focuses on the *recruitment* rate, which is the number of new additions to each outcome, as a fraction of the initial population. When assessing risk factors, the FG model does not provide the effect of the risk factor on the underlying biologic mechanism of the disease, but rather it provides the effect of the risk factor on the demand for services for this disease (e.g., how many patients with this disease will be in the waiting room). This can make interpretation of the risk factor difficult as noted in the MGUS example where the direction of the age coefficient differed for the FG model compared to the MH model. For illustration, imagine two salesmen of a new product, each starting with a territory of n potential clients, and we want to compare their relative performance after adjusting for various factors. If there is a strong competing risk, however, which is differential, for instance an epidemic of fatal disease which seriously depletes the potential clientele in only the region for salesman 2, then recruitment rates and hazard rates will differ. The second salesman will need to significantly increase their hazard rate, in fact, just to keep up.
- The pseudovalue approach can be used to mimic either the MH or FG approaches depending on the choice of transformation.

The MH and FG models will yield similar results when there is little differential effect of the competing risk, even if that risk is high, or when censoring is such that 1/3 or more always remain at risk, even if covariate effects on the competing risk are strong. neither of these conditions hold in the MGUS dataset.

However, in practice, these authors rarely use the FG model because it does not address the research question they are most interested in, which is typically related to the biological effect of the risk factors. The fact that the PH assumption cannot hold for both the MH and FG models along with the possibility that the cumulative total failure probability could exceed 1 in datasets involving elderly subjects, also make the FG model more worrisome

in practice[11]. The MH model is often the best approach, and it can extend to more complex situations, as we will see in the next chapter.

Chapter 10

Multistate data summaries

In the most general multistate processes, there are multiple transition types, and a given subject can experience multiple transitions. (Competing risks have the first characteristic but not the second, and repeated events have the second characteristic but not the first.) Four examples are shown in Figure 10.1. There are often multiple choices for the state and transition diagram, and for some datasets it is revealing to look at a problem from multiple views. Deciding which state-space diagram best matches a given research question is often the most critical step in the analysis.

Both absolute risk estimates and hazard rates are essential parts of the solution; each gives insight into the data. This chapter will discuss data set-up and absolute risk estimates. The following chapter will discuss modeling the rates. The three absolute risks that we will use most often are:

- $p_k(t;x)$: the probability of being in state k at time t, for covariate vector x.
- $E(N_k(t;x))$: the expected number of visits to state k by time t.
- $\int_0^t p_k(s;s)ds$: the expected amount of time spent in state k, up to time t. This is known as the sojourn time, or the restricted mean time in state.

As a simple starting example, consider the MGUS dataset used in the competing risks chapter. Figure 10.2 shows two state space diagrams along with the corresponding Aalen-Johansen curves. There is a large change in the curve for plasma-cell malignancy (PCM); in the left panel it represent the fraction who *ever* had a PCM diagnosis; in the right panel it is the fraction *currently* in the PCM state. At the time period represented by the study (1960-1994) there was no effective treatment for PCM, and the transition from PCM to death was rapid; in this dataset the median time from PCM to death is 20 months. The lower right hand curve rises when someone progresses (enters the PCM state) and falls when they leave.

Figure 10.3 illustrates yet another way to display this data. It combines the results from competing risk analysis along with a second fit that treats death after PCM as a separate state from death before progression, as shown in the right-hand panel of the figure. In this plot, the fraction of subjects currently in the PCM state is shown by the distance between the two curves. In this particular dataset, 860 subjects died prior to developing PCM; 103 of the 115

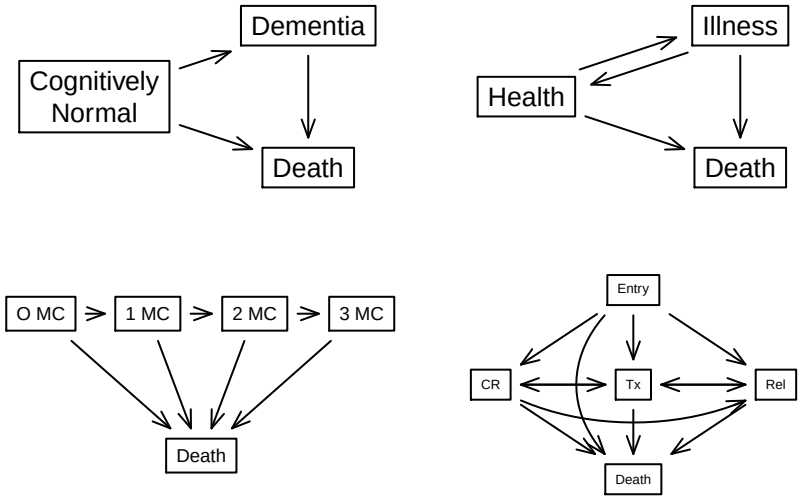

Figure 10.1 *Four multistate processes. The upper left panel is a variant of the illness-death model and will be explored in the aging example. The upper right shows an illness-death model with recovery. The lower left model will be explored in the NAFLD example: subjects have an increasing number of metabolic comorbidities (MC), with death as a competing risk. The final model represents an expansion of the MC model where each combination of comorbidities is treated as a different state.*

subjects with PCM were followed to death and of those, 90% were dead 70 days after diagnosis with PCM.

10.1 Building multistate datasets

The data setup for a multistate process has some key differences from the simpler, single event process. Building the dataset demands the most attention to detail, and quite often consumes the majority of analysis time. Analysis of multistate data using the survival package uses the following set-up.

- Counting process data with multiple rows (or observations) per subject is required.
- The status indicator is a multi-level factor. The first level must correspond to "no transition (new state) at this time".
- An id variable identifies which rows correspond to a subject.
- An optional variable specifies the current state at the beginning of the interval.
- Each subject will have:

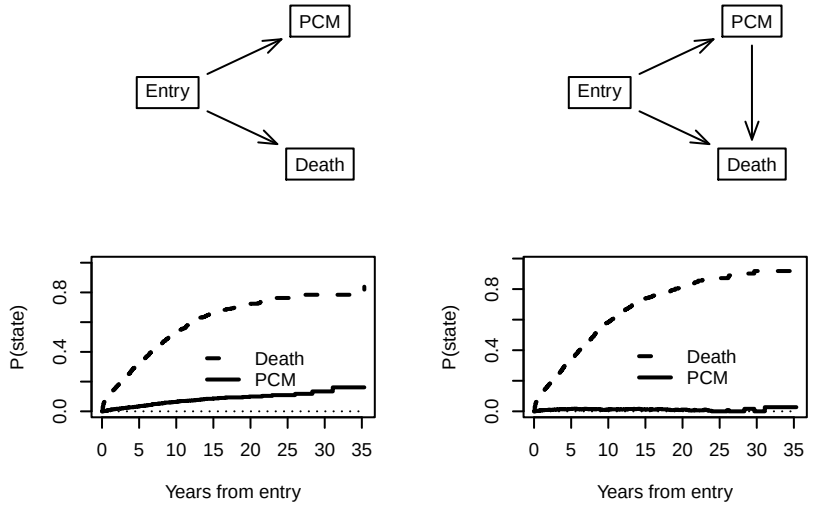

Figure 10.2 *Two state space choices for the MGUS dataset. The left is a competing risks form, the right a more general multistate. In each, the top panel is the state space, the bottom panel the resulting Aalen-Johansen estimate.*

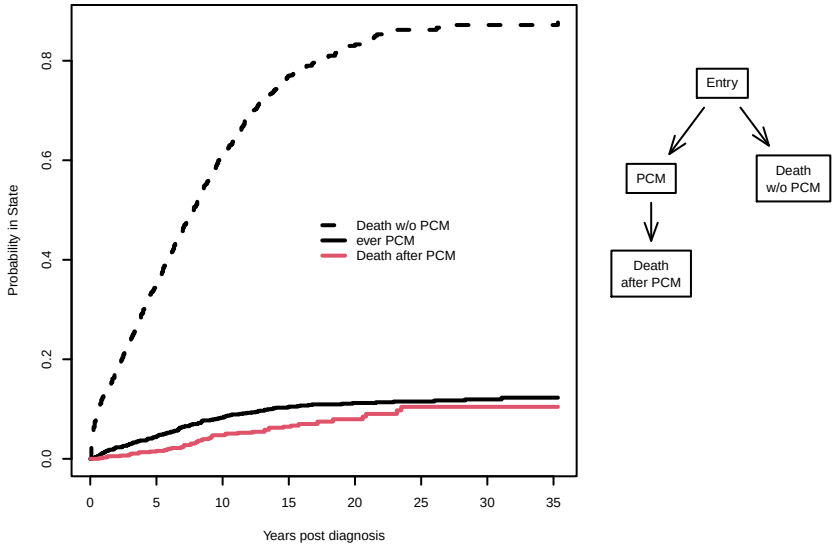

Figure 10.3 *Modified plot that shows the cumulative number of subjects ever to have had PCM (thicker solid line, competing risk model), along with death with and without prior PCM, per the state space figure in the right panel.*

- A consistent partition of time; no overlapping intervals (which would imply two copies of the subject alive at the same time), and no gaps in the time scale (where did they go?).
- A consistent pattern of states. If a subject transitions to state B at the end of interval $(t_1, t_2]$, they must begin the next interval $(t_2, t_3]$ in state B.
- The time between entry and exit of a state must be > 0.

Counting process data, first introduced in Chapter 4, was first conceived as a mechanism to deal with time-dependent covariates — and only for that purpose — in the first version of the S survival package, released in 1986. It was fairly quickly recognized that counting process data had other uses. Use of this form for general multistate data is yet another extension.

The primary changes from a simpler time-dependent covariate usage is that an id variable is now required and the transition variable is no longer a simple 0/1 indicator, but instead indicates the new state that is attained. When there are both time-dependent covariates and multiple transitions, any given subject may have a large number of rows. Unchanging variables (e.g., sex), simply repeat from row to row, while time-dependent covariates change in the expected way. Variables that encode a transition or "next state" are different than other variables, however. This is one of the more confusing aspects of counting process datasets and a common source of error.

This can be particularly confusing when one concept can play two roles. For instance, the occurrence of diabetes may be treated as an outcome in one analysis, and as a covariate in another analysis. This would require two different diabetes variables: one for *prevalent* diabetes and one for *incident* diabetes. The first is a baseline variable that repeats and the second is an event variable that does not repeat. This duality is rooted both in the history of how the counting process data form developed, and in a primary statistical fact: any covariate used to predict an outcome at some time t must be known strictly before t. A diagnosis of diabetes at time 45 days, used as a predictor (time-dependent covariate), applies only to intervals after 45 days. Failure to properly encode time-dependent covariates is one of the more common forms of immortal time bias, discussed more fully in Section 4.2.

Tools to check these aspects are an important part of the data creation process and are available in the survival package. In particular, the `survcheck` function is useful for assessing the transition between states and for checking for contiguous non-overlapping intervals. In fact, the `survfit` and `coxph` functions first call `survcheck` to verify the data format so it is well worth the time to check the data first before running additional analyses.

Interval censoring

The AJ method used by `survfit` does not account for interval censoring, also known as panel data, where a subject's current state is recorded at some fixed time, such as a medical center visit, but the actual times of transitions are

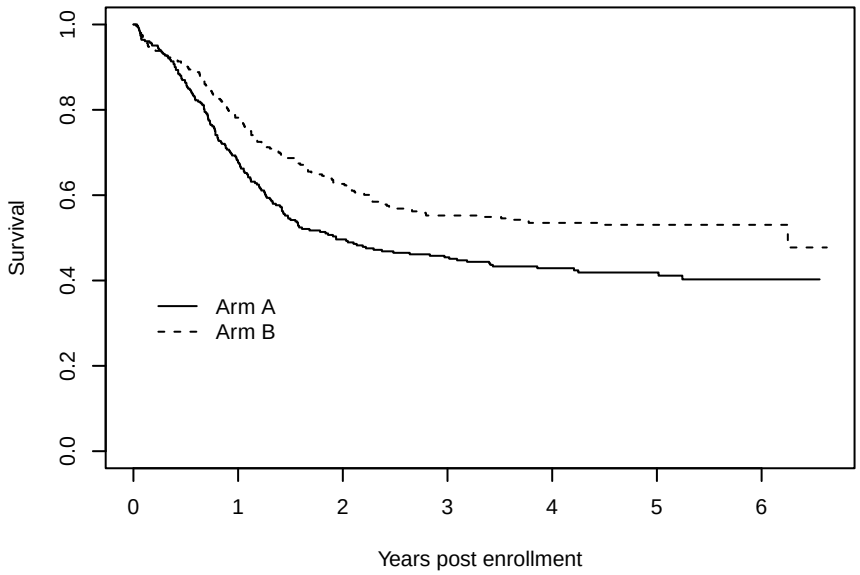

Figure 10.4 *Overall survival curves for the two study arms.*

unknown. Such data requires further assumptions about the transition process in order to model the outcomes and have a more complex likelihood; the `msm` package, for instance, deals with this type of data. If subjects reliably come in at regular intervals, then the difference between the two results can be small, e.g., the `msm` routine estimates time until progression *occurred* whereas `survfit` estimates time until progression was *observed*.

10.2 Multiple summaries

The `myeloid` dataset contains data from a clinical trial in subjects with acute myeloid leukemia. After receiving one of two treatments, subjects ideally enter the complete response (CR) state. This sets the patient up for a more aggressive regimen of hematologic stem cell transplant (SCT), then a sustained remission, followed eventually by relapse or death. That is: initial therapy $\rightarrow$ CR $\rightarrow$ SCT, then either relapse or death. Not everyone follows this ideal path, of course, as shown in Figure 10.5. This deviance from the ideal is often a reality when working with multistate data, and decisions need to be made about how to treat these exceptions.

The summaries in this section of the book uses the `myeloid` dataset to demonstrate how the same multistate data can be used to answer different research questions.

Overall survival curves for the data are shown in Figure 10.4. The difference

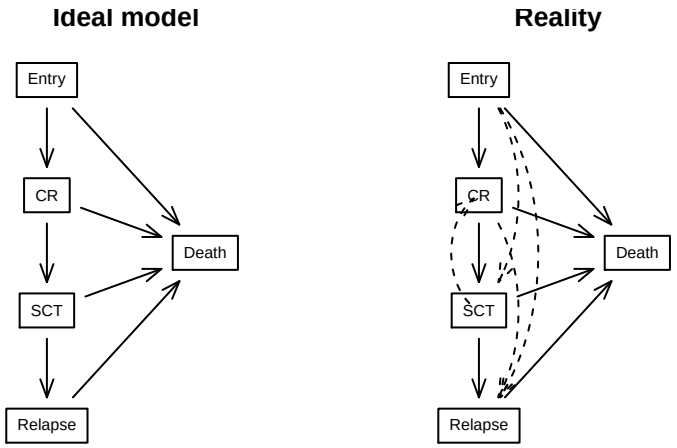

Figure 10.5 *Diagram illustrating the full multistate process for the myeloid data. The first panel shows the ideal paths and the second panel shows all the paths followed by subjects in the dataset.*

between the treatment arms A and B is substantial. A goal of this analysis is to better understand this difference.

Data check

```
  id trt sex futime death txtime crtime rltime
1  1   B   f    235     1     NA     44    113
2  2   A   m    286     1    200     NA     NA
3  3   A   f   1983     0     NA     38     NA
4  4   B   f   2137     0    245     25     NA
5  5   B   f    326     1    112     56    200
```

The first few rows of data are shown above. The dataset contains the follow-up time and alive/dead status at last follow-up for each subject, along with the time to transplant (txtime), complete response (crtime) or relapse after CR (rltime). Subject 1 did not receive a transplant, as shown by the NA value, and subject 2 did not achieve CR.

```
         to
from       CR SCT relapse death (censored)
  (s0)    443 106      13    55         29
  CR        0 159     168    17        110
  SCT      11   0      45   149        158
```

```
relapse   0  99       0    99         28
death     0   0       0     0          0
```

The table above is referred to as a transitions table and is produced by the `survcheck` function. It shows 55 direct transitions from entry to death, i.e., subjects who die without experiencing any of the other intermediate points, 159 who go from CR to transplant (as expected), 11 who go from transplant to CR, etc. No one was observed to go from relapse to CR in the dataset. This serves as a data check since it should not be possible per the data entry plan. This table helped guide the additional lines added to the right panel of Figure 10.5.

Different questions often require a different data set-up and it is generally a good idea to check each version of the analysis data to make sure it matches your diagram.

10.2.1 Competing risks

For investigating the data we would like to start off with a set of alternate endpoints, using the competing risk approach discussed in Chapter 9. Diagrams for these scenarios are shown in Figure 10.6, along with the overall probability of death.

1. The competing risk of CR and death, ignoring other states. This is used to estimate the fraction who ever achieved a complete response.
2. The competing risk of SCT and death, ignoring other states. This is used to estimate the fraction who ever achieved a transplant.

All three analyses shown in Figure 10.6 include one observation per subject however the sets of endpoints are all different. One of the checks applied to this data is whether time to complete response and time to transplant are different. In this data, there is one subject where these times are identical so some data clean-up is necessary because subjects cannot be in different states at the same time. For this subject, the decision was made to change time-to-CR to be 1 day shorter than time-to-SCT.

The right panel of Figure 10.6 overlays three separate `survfit` calls: standard survival until death, complete response with death as a competing risk, and transplant with death as a competing risk. For each fit we have shown one selected state: the fraction who have died (solid lines), the fraction ever in CR (dotted lines), and the fraction ever to receive a transplant (dashed lines), respectively. Most of the CR events happen before 2 months and nearly all the additional CRs conferred by treatment B occur between months 2 and 8. Most transplants happen after 2 months, which is consistent with the clinical guide of transplant after CR. The survival advantage for treatment B begins between 4 and 6 months, which argues that it could be at least partially a consequence of the additional CR events.

The association between a particular curve and its corresponding state

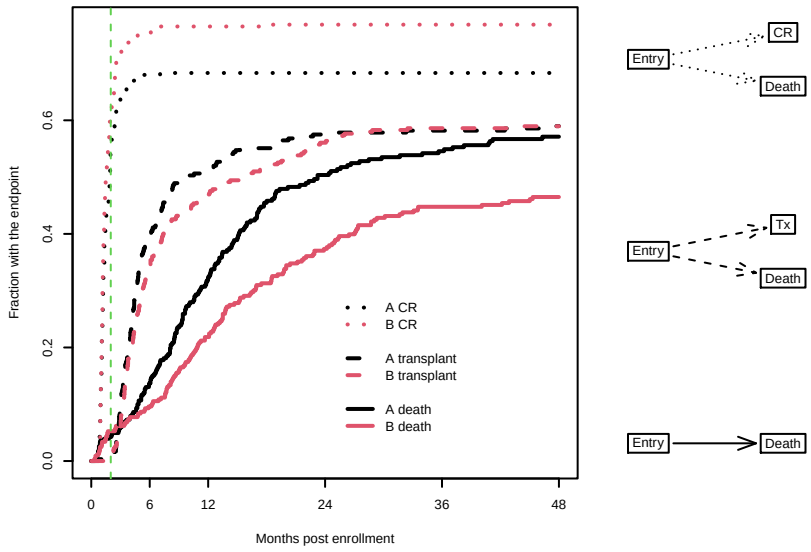

Figure 10.6 *Overall survival curves: time to death, to transplant (Tx), and to complete response (CR). Each shows the estimated fraction of subjects who have ever reached the given state. The curves were limited to the first 48 months to more clearly show early events. The right hand panel shows the state-space model for each pair of curves.*

space diagram is critical. As we will see below, many different models are possible and it is easy to get confused. Attachment of a diagram directly to each curve, as was done above, will not necessarily be day-to-day practice, but the state space should always be foremost. If nothing else, draw it on a scrap of paper and tape it to the side of the terminal when creating a dataset and plots.

10.2.2 Ever versus currently in state

Complete response is a goal of the initial therapy; Figure 10.7 looks more closely at this. As was noted before, arm B has an increased number of late responses. The duration of response is also increased: the solid curves show the number of subjects still in response, and we see that they spread farther apart than the dotted "ever in response" curves. The figure shows only the first eight months in order to better visualize the details, but continuing the curves out to 48 months reveals a similar pattern. The dotted lines are the same competing risk curves shown in Figure 10.6 whereas the solid lines use an endpoint with possible states of CR, relapse, or death.

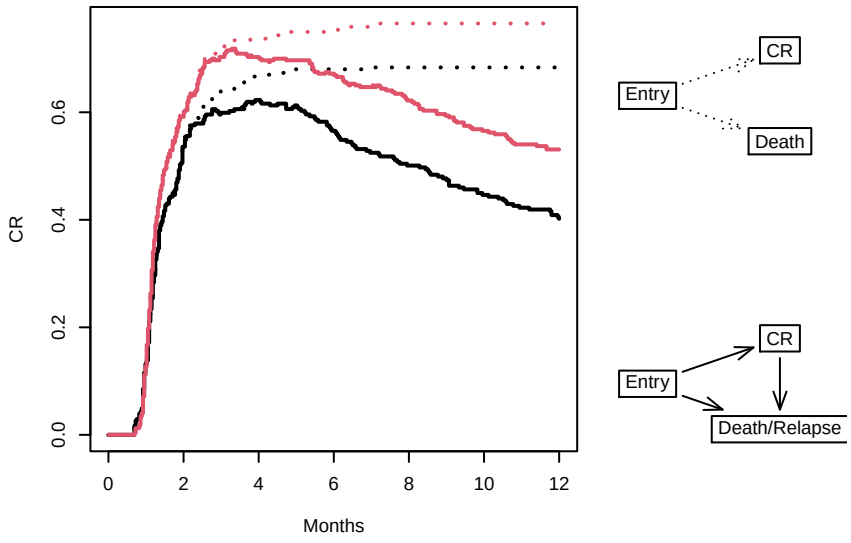

Figure 10.7 *AJ estimates for 'ever in CR' (dotted lines) and 'currently in CR' (solid lines); the only difference is an additional transition from CR to death or relapse. Both scenarios ignore transplant.*

	n	nevent	rmean	se(rmean)
trt=A, (s0)	693.0	0.0	7.1	0.8
trt=B, (s0)	739.0	0.0	5.6	0.7
trt=A, CR	693.0	206.0	16.3	1.1
trt=B, CR	739.0	248.0	21.2	1.1
trt=A, relapse	693.0	109.0	4.3	0.6
trt=B, relapse	739.0	117.0	5.5	0.6
trt=A, death	693.0	171.0	20.2	1.1
trt=B, death	739.0	149.0	15.6	1.0

Table 10.1 *Using the diagram focusing on 'currently in CR' ignoring transplants, show the restricted mean time in state by treatment within the first 48 months*

Restricted time in state

The mean time in the CR state (Table 10.1) for the first 48 months was obtained using the fit corresponding to the second diagram in Figure 10.7 since the mean time in state for competing risk events are not meaningful. The time in the CR state is extended by 21.2 - 16.3 = 4.9 months comparing arm B with arm A. A question which immediately gets asked is whether this difference is "significant", to which there are two answers. The first and more important is to ask whether 5 months is an important gain from either a

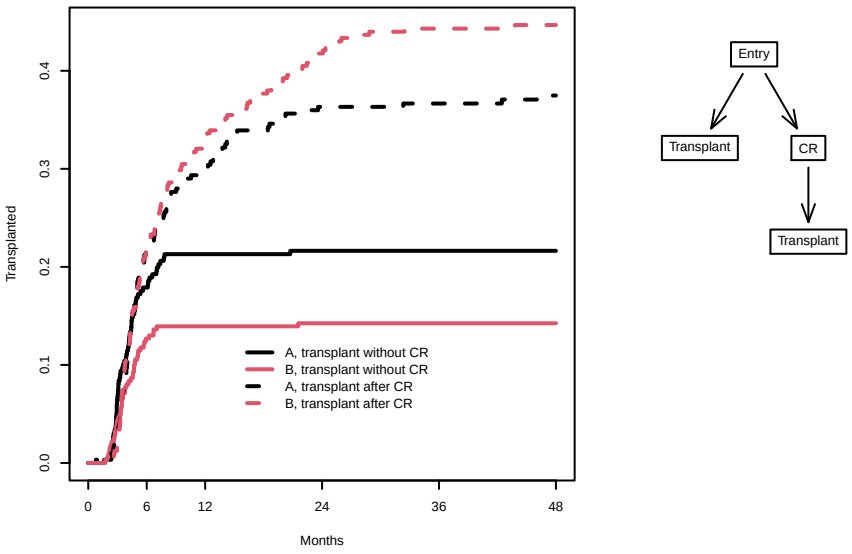

Figure 10.8 *Transplant status of the subjects, broken down by whether it occurred before (solid lines) or after (dashed lines) CR.*

clinical or patient perspective. The overall restricted mean survival for the study is approximately 30 of the first 48 months post entry; on this backdrop, an extra 5 months in CR might or might not be an meaningful advantage from a patient's point of view. The less important answer is to test whether the apparent gain is sufficiently rare from a mathematical point of view, i.e., "statistical" significance. The standard errors of the two values are 1.1 and 1.1, and since they are based on disjoint subjects the values are independent, leading to a standard error for the difference of $\sqrt{1.1^2 + 1.1^2} = 1.6$. The 5 month difference is more than 3 standard errors, so highly significant.

In summary:

- Arm B adds late complete responses (about 4%); there are 206/317 CR in arm A vs. 248/329 in arm B.
- The difference in 4 year survival is about 6%.
- There is approximately 2 months longer average duration of CR (of 48).

10.2.3 Event with and without intermediate state

CR $\rightarrow$ transplant is the target treatment path for a patient. Given the improvements listed above, why does Figure 10.6 show no change in the number transplanted? Figure 10.8 shows the transplants broken down by whether this happened before or after complete response. Most of the non-CR transplants happen by 10 months. One possible explanation is that once it is apparent

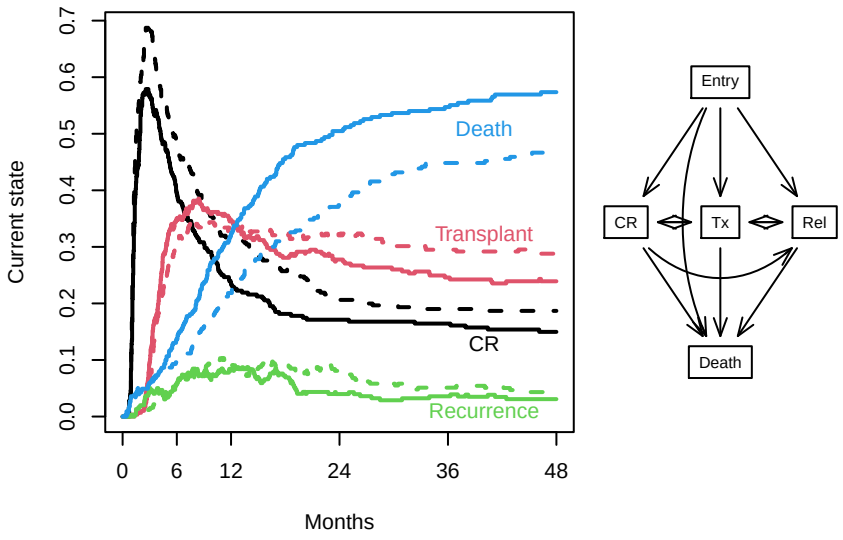

Figure 10.9 *The full multistate curves for the two treatment arms (solid line = A, dashed line = B).*

to the patient/physician pair that CR is not going to occur, they proceed forward with other treatment options. The extra CR events on arm B, which occur between 2 and 8 months, lead to a consequent increase in transplant as well, but at a later time of 12–24 months: for a subject in CR we can perhaps afford to defer the transplant date.

Computation is again based on a manipulation of the event variable: in this case dividing the transplant state into two sub-states based on the presence of a prior CR. (Because of scheduling constraints within a hospital it is unlikely that a CR that is within a few days prior to transplant could have effected the decision to schedule a transplant, however. An alternate breakdown that might be useful would be "transplant without CR or within 7 days after CR" versus those that are more than a week later. There are many sensible questions that can be asked.)

10.2.4 Full multistate process

Figure 10.9 shows the full set of state occupancy probabilities for the cohort over the first 4 years. At each point in time the curves estimate the fraction of subjects currently in that state. The total who are in the transplant state peaks at about 9 months and then decreases as subjects relapse or die; the curve rises whenever someone receives a transplant and goes down whenever someone leaves the state. At 36 months treatment arm B (dashed) has a lower

fraction who have died, the survivors are about evenly split between those who have received a transplant and those whose last state is a complete response (only a few of the latter are post transplant). The fraction currently in relapse – a transient state – is about 5% for each arm. The figure omits the curve for "still in the entry state". The reason is that at any point in time the sum of the 5 possible states is 1 — everyone has to be somewhere. Thus one of the curves is redundant, and the fraction still in the entry state is the least interesting of them. Table 10.2 provides time in state estimates using the full model. Note the differences in the estimates from Table 10.1; using the full process, the treatment difference for time in the CR state is now only 2.5 months.

	n	nevent	rmean
trt=A, (s0)	807.00	0.00	2.65
trt=B, (s0)	882.00	0.00	2.92
trt=A, CR	807.00	206.00	10.41
trt=B, CR	882.00	248.00	12.87
trt=A, SCT	807.00	175.00	12.57
trt=B, SCT	882.00	189.00	13.73
trt=A, relapse	807.00	109.00	2.17
trt=B, relapse	882.00	117.00	2.92
trt=A, death	807.00	171.00	20.20
trt=B, death	882.00	149.00	15.56

Table 10.2 *Using the full multistate process, show the restricted mean time in state by treatment within the first 48 months*

Influence matrix

it isn't clear to me what we are trying to show here and whether it is useful for this audience.

For one of the curves above we returned the influence array. For each value in the matrix P = probability in state and each subject i in the dataset, this contains the effect of that subject on each value in P. Formally,

$$D_{ij}(t) = \left.\frac{\partial p_j(t)}{\partial w_i}\right|_w$$

where $D_{ij}(t)$ is the influence of subject i on $p_j(t)$, and $p_j(t)$ is the estimated probability for state j at time t. This is known as the infinitesimal jackknife (among other labels).

For treatment arm A there are 317 subjects and 426 time points in the P matrix. The influence array has subject as the first dimension, and for each subject it has an image of the P matrix containing that subject's influence on each value in P, i.e., `influence[1, ,]` is the influence of subject 1 on P. For this dataset everyone starts in the entry state, so $p(0)$ = the first row of

`pstate` will be (1, 0, 0, 0, 0) and the influence of each subject on this row is 0; this does not hold if not all subjects start in the same state.

do we need this next bit or should it go in the examples document? Yes, it should go in the examples document instead As an exercise we will calculate the mean time in state out to 48 weeks. This is the area under the individual curves from time 0 to 48. Since the curves are step functions this is simple sum of rectangles, treating any intervals after 48 months as having 0 width.

```
         [,1]  [,2] [,3] [,4]  [,5]
rmean    7.10 16.34    0 4.31 20.24
se.rmean 0.78  1.13    0 0.56  1.10
Call: survfit(formula = Surv(tstart, tstop, cr2) ~ trt, data = mdata,
    id = id, influence = TRUE)

          n nevent rmean se(rmean)*
(s0)    693      0   7.1       0.78
CR      693    206  16.3       1.13
SCT     693      0   0.0       0.00
relapse 693    109   4.3       0.56
death   693    171  20.2       1.10
   *restricted mean time in state (max time = 48 )
```

The last lines verify that this is exactly the calculation done by the `print.survfitms` function; the results can also be found in the `table` component returned by `summary.survfitms`.

In general, let U_i be the influence of subject i. For some function $f(P)$ of the probability in state matrix `pstate`, the influence of subject i will be $\delta_i = f(P+U_i) - f(P)$ and the infinitesimal jackknife estimate of variance will be $\sum_i \delta^2$. For the simple case of adding up rectangles $f(P+U_i) - f(P) = f(U_i)$ leading to particularly simple code, but this will not always be the case.

Summary

As shown by the example using the myeloid data, summaries of multistate data can get to be quite complex with muliple lines. Breaking down the analysis into smaller questions can be useful to more fully understand the data. Examination of the transition table is important to identify transitions with only a few subjects; this will be even more important when fitting models (represented by the arrows) using these subjects.

Chapter 11

Multistate models

In the previous chapter, probability in state estimates were obtained using AJ curves. The next analysis step is to understand how covariates influence these transitions (arrows), and to how they influence model predictions providing estimates of the time in state (boxes). When modeling multistate data it is still important to draw the state space diagrams and building the appropriate dataset. However it now also important to think through the different transition rates (arrows). With these models it is possible to use different covariates for different transitions or to force some of the coefficients or underlying baseline hazards to be the same across transitions.

1. Which covariates should be attached to each rate? Sometimes a covariate is important for one transition, but not for another.
2. For which transitions should one or more of the covariates be constrained to have the same coefficient? Sometimes there will be a biologic rationale for this. For other studies an equivalence is forced simply because we have too many unknowns and cannot accommodate them all. (This is the same reason that linear models often contain very few interaction terms).
3. Which, if any, of the transitions should share the same baseline hazard? Most of the time the baseline rates are all assumed to be different.

For simple two-state survival, the Cox model leads to three relationships

$$\lambda(t) = \lambda_0(t)e^{X\beta} \tag{11.1}$$

$$\Lambda(t) = \Lambda_0(t)e^{X\beta} \tag{11.2}$$

$$S(t) = \exp(-\Lambda(t)) \tag{11.3}$$

where λ, Λ and S are the hazard, cumulative hazard and survival functions, respectively. There is a single linear predictor which governs both the rate λ (the arrow in Figure 10.1) and the probability of residing in the left hand box of the figure. For multistate models this simplicity no longer holds; proportional hazards does not lead to proportional $p(t)$ curves. There is a fundamental dichotomy in the analysis namely that hazards can be computed one at a time and the probability in state must be done for all states at once.

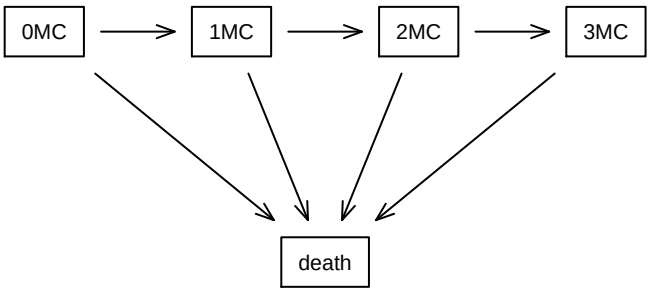

Figure 11.1 *State space for the NAFLD study.*

Non-alcoholic fatty liver disease (NAFLD) data

The `nafld` datasets come from an observational study of the incidence of non-alcoholic fatty liver disease (NAFLD), which is essentially the presence of excess fat in the liver, and parallels the ongoing obesity epidemic (See Section 1.5.4 for more details).

We will model the onset of three important components of the metabolic syndrome: diabetes, hypertension, and dyslipidemia, using the model shown in Figure 11.1. Subjects may have either 0, 1, 2, or all 3 of these metabolic comorbidities at enrollment.

The NAFLD data is represented as 3 datasets, `nafld1` has one observation per subject containing baseline information (age, sex, etc.), nafld2 has information on repeated laboratory tests, e.g. blood pressure, and nafld3 has information on yes/no endpoints.

The first step is to build a single counting process dataset from these three inputs. The steps to do this in R are detailed in the companion document, along with various checks on the process. For this analysis we think of the three conditions as one-time outcomes (you can't get diabetes twice). The final step of the process is to generate the transitions table, shown below.

```
Transitions table
       to
from     1mc  2mc  3mc death (censored)
  0mc   1829   70    4   263       5705
```

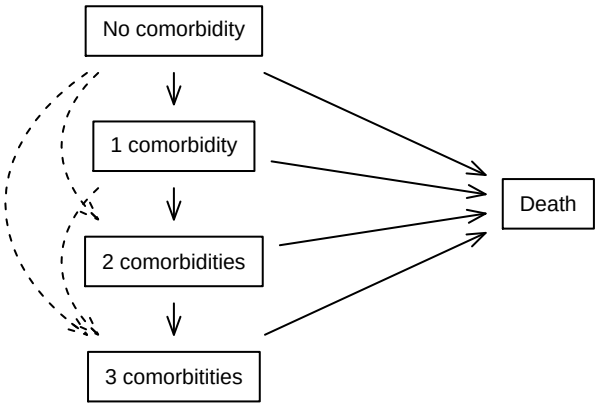

Figure 11.2 *Modified state space diagram for the NAFLD data. Dashed lines are 'jump' transitions that are observed due to incomplete monitoring of each subject.*

```
1mc       0 1843   28   243        4567
2mc       0    0 1048   417        3687
3mc       0    0    0   441        2220
death     0    0    0     0           0
```

This is a rich dataset with a large number of transitions: over a quarter of the participants have at least one event, and there are 22 subjects who transition through all 5 possible states (4 transitions). Of the 7871 subjects who enter the study with 0 of the metabolic comorbidities (MC), over half (5705) end their followup in the same state, 263 die before accumulating an MC, and the remainder transition to a higher MC state.

We see a number of subjects who "jump" states, e.g., directly from 0 to 2 comorbidities. This serves to remind us that this is actually a model of time until *detected* comorbidity; which will often have such jumps even if the underlying biology is continuous. A modified state space is shown in Figure 11.2, where the dotted lines are transformations that we observe, but would not be present if the subjects were monitored continuously.

11.1 Shared coefficients and shared hazards

As noted in the above transitions table, there are some transitions with very few subjects which will create numerical instability when fitting the model. This provides motivation for shared coefficients and shared hazards, and by

	Baseline hazard		
	Separate	Proportional	Identical
Separate coefficients	1	2	3
Shared coefficients	4	5	6

Table 11.1 *Choices for any pair of transitions in a multistate model.*

fitting all the transitions simultaneously, applying these constraints is possible. For instance, in a model with both health:death and illness:death hazards, one might wish to assume that the coefficient for male sex is identical for those two transitions. We have found in our own work that for populations over the age of 50 years, the hazard ratio for male sex is very often between 1.2 and 1.4 for the death transition, for a wide variety of studies and starting states. Another assumption would be that the two transitions have proportional baseline hazards, i.e., their baseline survival curves have the same shape.

In general, for any given pair of transitions a:b and c:d, we have 6 possibilities shown in Table 11.1. Here we make some general comments, mostly without proof.

- Options 1 and 4 in the table are the most useful: they are easy to set up in the code and straight forward to interpret.
  - Option 1 corresponds to a separate Cox model for each transition, and is the default approach in the R survival package. However, the total number of coefficients may become large.
  - Option 4 is analogous to analysis of covariance in a linear model, e.g., a common slope with separate intercepts per group, and is a common strategy to reduce the degrees of freedom for a model. The baseline hazard plays the role of an intercept.
- Option 3, separate coefficients but a shared baseline hazard, is sometimes suggested but the resulting model will not be sensible. The underlying issue is identical to one that arises when fitting a linear model with per-group slopes but only a single intercept. For instance assume the model $y = \beta_0 + \beta_M x + \beta_F x + \epsilon$. Here x is a covariate with a separate coefficient for males and for females. If x is replaced by a linear transform of x, say $z = 1 - x$, and the model is re-fit using z, the new fit will not be the same. There will be different predicted values and test statistics, not simply recoded coefficients. Users are quite cavalier about about which coding a binary variable takes, 0/1 or 1/0, which means that this model choice will result in confusion and incorrect interpretations.
  Table 11.2 lays the equivalent expressions out in detail for a Cox model, two groups and one binary covariate, where that covariate is coded first as 0/1 for male/female, and then as 1/0 for male/female. The result is a common baseline for females in the first instance and a common baseline for the males in the second; these are not exchangeable models.
- Option 6 is equivalent to collapsing the two transitions into a single end-

	F/M = 0/1		F/M = 1/0	
	Female	Male	Female	Male
Transition 1	$\lambda_a(t)$	$\lambda_a(t)e^{\alpha_1}$	$\lambda_b(t)e^{\beta_1}$	$\lambda_b(t)$
Transition 2	$\lambda_a(t)$	$\lambda_a(t)e^{\alpha_2}$	$\lambda_b(t)e^{\beta_2}$	$\lambda_b(t)$

Table 11.2 *The issue with a common baseline hazard for two transitions, but separate coefficients for a single binary covariate for sex. The left two columns show the four predicted hazards using a 0/1 coding, with fitted coefficients of* $(\lambda_a(t), \alpha_1, \alpha 2)$*, and the right two the predicted hazards under a 1/0 coding, with fitted solution of* $\lambda_b(t), \beta_1, \beta_2)$*.*

point of "either occurs". Instead of imposing constraints, it is easier to directly fit the simpler model.

- Options 2 and 5, which assume proportional baseline hazards for the two transitions, are the second most rational choices behind models 1 and 4. It is easier to justify the assumption of the "same shape" for two hazards than the assumption that they are identical, e.g., death rates might go up with age in the same pattern for two conditions, but not be identical. The use of proportional rather than identical hazards also alleviates the issue found in model 3; predicted values will be invariant to how dummy variables are coded.

The phrase "assume a common baseline hazard" appears moderately often in the literature, often without sufficient clarification. The above shows that assuming that common = "proportional" is okay, but common = "identical" is not.

11.2 Fitting multistate hazards models

Basic model

Since age is the dominant driver of the transitions in the NAFLD data, we have chosen to fit models directly on age scale rather than model the age effect. We force common coefficients for the transitions from 0 comorbidities to 1, 2 or 3, and for transitions from 1 comorbidity to 2 or 3. This is essentially a model of "any progression" from a given state. We also force the effect of male sex to be the same for any transition to death.

Results are shown in Table 11.3. Male sex increases the death rate by 1.4 fold, and the progression rate from 0 to 1 MC, 1 to 2, and 2 to 3 by 1.2, 1.28 and 1.16 fold, respectively. Given all that we know about aging, none of this is a surprise. The exposure of interest, NAFLD, has much larger effect, increasing progression rates by 1.6–2.5 fold, and death rates by 1.1-1.9 fold. The decreasing NAFLD effect on death rates with an increasing MC count, 1.8, 1.7, 1.7, 1.1, may be partly a ceiling effect. That is, when a subject has 3 serious conditions, the death rate is already high and an extra risk factor can

	NAFLD		Male	
	HR	p	HR	p
0:1-3	2.50	¡0.001	1.20	¡0.001
1:2:3	1.68	¡0.001	1.28	¡0.001
2:3	1.62	¡0.001	1.16	0.022
0:death	1.88	0.006	1.39	¡0.001
1:death	1.71	¡0.001	1.39	¡0.001
2:death	1.74	¡0.001	1.39	¡0.001
3:death	1.07	0.466	1.39	¡0.001

Table 11.3 *Estimated hazard ratio and p-values for the multistate model.*

only increase it by a limited amount. Indeed, the overall death rate is 6/1000 per year in those with 0MC and 30/1000 per year in those with 3.

In the above we forced the 0:1, 0:2, and 0:3 transitions to all have the same coefficients and baseline hazard. The first was necessary: the jump from 0 to 3 MC only has 4 observed events, which is far too few to assess 2 coefficients. The second could be relaxed.

A second available keyword is `shared`, which indicates that the baseline hazards for transitions share a common shape. Here is an example:

```
        1:2 1:3 2:3 1:4 2:4 3:4 1:5 2:5 3:5 4:5
nafld     1   1   3   1   3   5   7   9  10  11
male      2   2   4   2   4   6   8   8   8   8
ph(1:5)   0   0   0   0   0   0   0  12  13  14
```

A more complex model

It can be argued that the prior analysis' simplification into 0, 1, 2, or 3 metabolic cormorbidities is too great a simplification, i.e., diabetes is more serious than hypertension. Figure 11.3 shows an expanded model where 1MC and 2MC states of Figure 11.1 have each been expanded into its three components. This increases the total number of transitions in the figure from 3+4 =7 to 12+8 = 20. The transitions table below identifies a few more, the same double endpoints that we discussed earlier where both diabetes and hypertension onset, say, were coded at the same visit.

```
       to
from    diab  htn lipid   dh   dl   hl  dhl death (censored)
  none   106  357   818    4   14   21    2   103       3141
  diab     0    0     0   38   87    0    0    15        103
  htn      0    0     0   46    0  254    4    78        551
  lipid    0    0     0    0  212  772    9    88       2940
  dh       0    0     0    0    0    0   69    16         83
  dl       0    0     0    0    0    0  334    35        399
  hl       0    0     0    0    0    0  418   307       2490
```

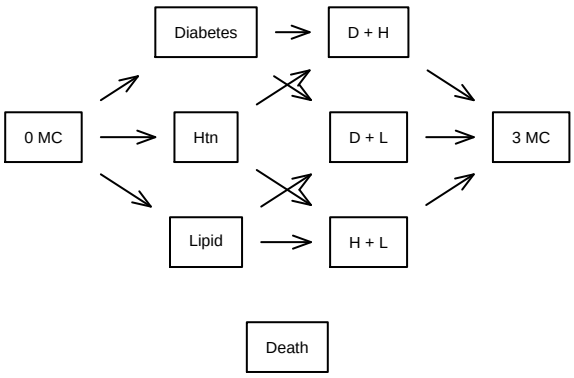

Figure 11.3 *Extended model for the NAFLD data, with intermediate states of diabetes, hypertension, dyslipidemia, and combinations of the three. The 8 arrows from each state to death were omitted to lessen overcrowding.*

```
dhl      0    0     0    0    0    0    0   370       1842
death    0    0     0    0    0    0    0     0          0
```

As a first fit, impose similar constraints as were used before. The first, that the effect of male sex is constant for all transitions to death, translates directly. The decision to use common coefficients for transitions from 1 to 2 and 1 to 3 comorbidities requires modification. We will arbitrarily order the comorbities' severity as diabetes, hypertension, lipidemia, so D:DHL is lumped with D:DH, H:DHL with H:DH, L:DHL with L:DL, none:DHL with none:D, etc.

Table 11.4 and Figure 11.4 show the results of this model.

The primary problem with this model is the complexity: how does one easily summarize 33 coefficients? An alternate model which is intermediate in size is one that assumes parallel affects. That is, a single effect for the addition of diabetes, whatever the substrate, another for the addition of hyperlipemia, and another for addition of hypertension. For the figure, we have also collapsed the transitions to death into 0–1, 2 or 3 metabolic comorbidities. It more or less corresponds to Figure 11.5, where we think of each comorbidity happening without regard to the others.

Figure 11.6 shows the result of this process. For simplicity, the transitions to death have also been collapsed. Once subjects have all three comorbidities, both the NAFLD and male sex effects on death rates are seriously attenuated, but the effect up to that point is largely constant.

	NAFLD		Male	
	HR	p	HR	p
none:D	7.67	¡0.001	1.61	0.0022
none:H	2.88	¡0.001	1.11	0.2376
none:L	1.79	¡0.001	1.19	0.0043
D:DH	0.69	0.2210	1.53	0.1339
D:DL	2.37	0.0012	2.25	0.0026
H:DH	1.72	0.0081	1.12	0.5835
H:HL	3.00	¡0.001	1.16	0.2325
L:DL	0.96	0.7988	1.22	0.0711
L:HL	1.58	¡0.001	1.33	¡0.001
DH:DHL	1.36	0.1784	0.83	0.4001
DL:DHL	1.07	0.5403	1.05	0.6120
HL:HDL	1.65	¡0.001	1.23	0.0198
0:death	1.88	0.0058	1.37	¡0.001
D:death	1.64	0.2941	1.37	¡0.001
H:death	1.87	0.0085	1.37	¡0.001
L:death	1.39	0.1346	1.37	¡0.001
DH:death	1.81	0.2567	1.37	¡0.001
DL:death	1.27	0.3979	1.37	¡0.001
HL:death	1.77	¡0.001	1.37	¡0.001
DHL:death	1.07	0.4695	1.37	¡0.001

Table 11.4 *Estimated hazard ratio and p-values for the extended model.*

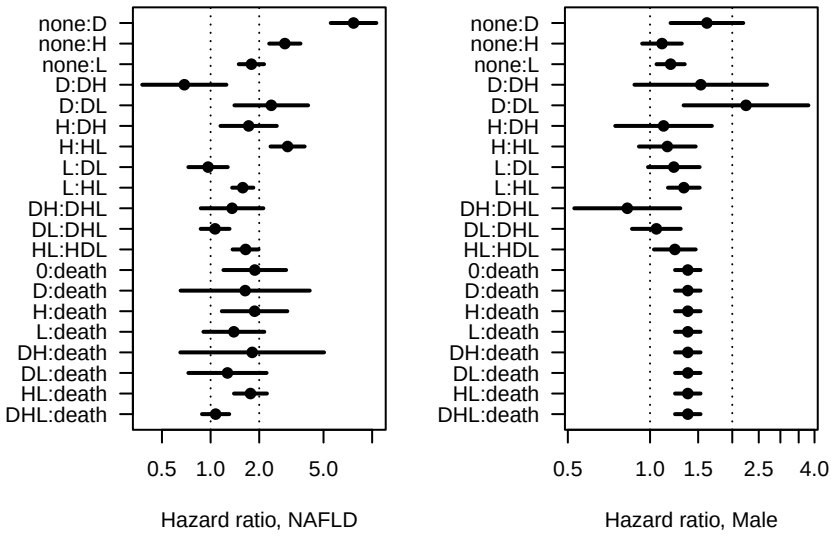

Figure 11.4 *Hazard ratios for the extended model.*

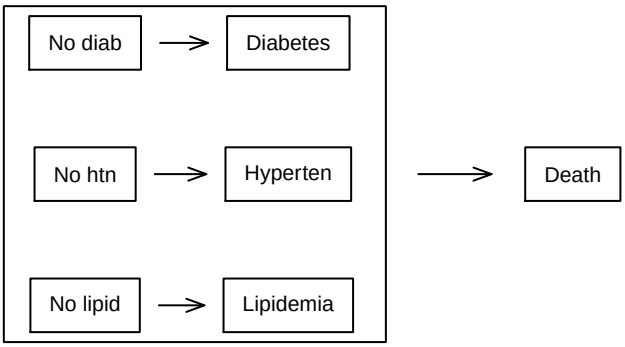

Figure 11.5 *State space where the three metabolic comorbidities acts independently of the others.*

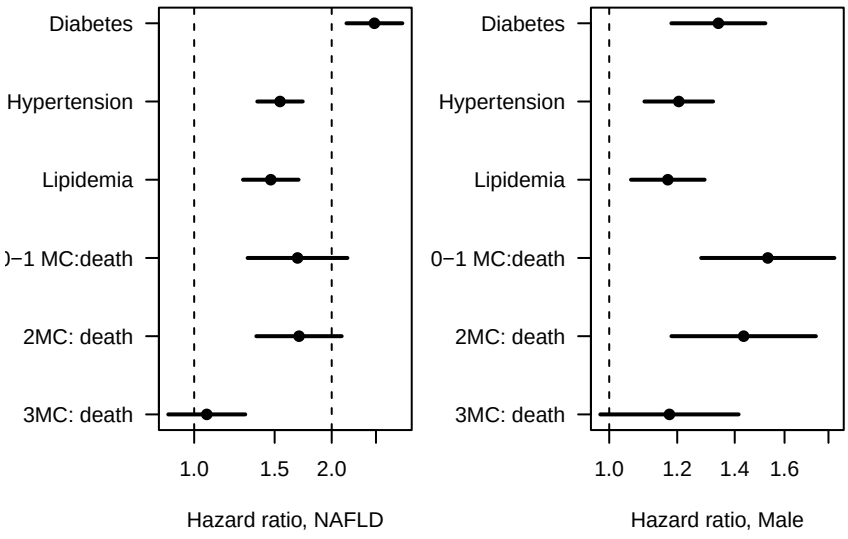

Figure 11.6 *A fit treating progression to diabetes, to hypertension, and to lipidemia each as a separate additive coefficient.*

From a higher level view, the two approaches are different ways of collapsing the 12 non-death transitions. The 0MC, 1MC, 2MC, 3MC variant groups none:d, none:h and none:l as 0-1 MC, dh:dhl, dl:dhl, and hl:dhl as 2-3 MC, and the other 6 as 1-2 MC. The parallel approach group none:d, h:dh, l:dl, and hl:dhl as "addition of diabetes", none:h, d:dh, l:lh and dl:dhl as addition of hypertension, and the remaining 3 as addition of dyslipidemia. The user will need to decide which clustering of risk is the most sensible from a scientific point of view.

Proportional baseline hazards

Each of the transitions is realized in the underlying computer code as a separate hazard model. This raises the idea of a hybrid approach, where prior hypertension and prior lipidemia are treated as covariates in a model for diabetes onset, and likewise diabetes and hypertension when considering the onset of lipidemia.

In the parallel multi-state model, there will be 4 separate baseline hazards for the none:diabetes, H:DH, L:DL and HL:DHL transitions, all four sharing the same NAFLD and male sex coefficients. The assumption is that the overall rate of progression to diabetes may be different for those with 0, 1 or 2 of the other conditions, but that the addition of NAFLD has the same proportional effect for all four cases.

We could make a stronger assumption that these four baseline hazards are proportional, which replaces the 4 non-parametric baselines with a single one, along with 3 coefficients. If this assumption of proportionality is true, then the estimate with shared baseline will be more efficient. Figure 11.7 shows the estimates for the non-parametric and shared hazard models side by side. A numeric check shows that there is a bare hint of efficiency gain, the standard error of the diabetes coefficient is 0.072 vs.0.071 but overall there is essentially no change.

As shown above, these multistate models can become very complex, which begs the question: if you are interested in death, would not the simple time-dependent covariate model be a lot easier than a multistate model? If all you are interested in are the transitions then the answer would be yes, however one of the strengths of this approach is the ability to obtain estimates of absolute risk.

11.3 Absolute risk, based on a multistate hazards model

One aspect of simple Cox models that does not translate to the multistate case is the formula for creation of predicted survival curves. Just as a naive KM that ignores other transitions does not yield proper estimates for competing risks data, neither does the predicted survival from a single Cox model yield valid multistate estimates. An analog of the AJ estimate is required.

For a multistate hazards model, baseline hazards are estimated separately

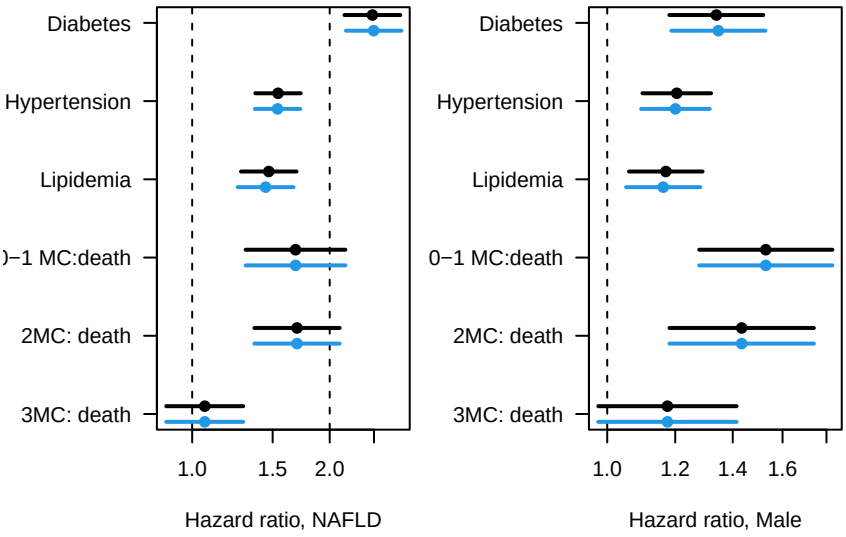

Figure 11.7 *Figure showing the estimates in a model with multiple baseline hazards (black), which matches figure 11.4, along with estimate from a model with proportional baseline hazards for addition of a new comorbidity (green).*

for each transition. That is, if $\eta_{AB} = X\beta_{AB}$ is the linear predictor for the AB transition, then

$$\hat{\lambda}_{0AB}(t) = \int_0^t \frac{\sum_i dN_{iAB}(s)}{\sum_i Y_{iA}(s)e^{\eta_{iAB}}} ds \tag{11.4}$$

$$\hat{\lambda}_{AB}(t;z) = e^{z\beta_{AB}}\hat{\lambda}_{0AB}(t) \tag{11.5}$$

Equation (11.4) is a sum over each observed A:B transition of the number of observed A:B transitions at that time, divided by the (hazard ratio) weighted number of observations at risk for the transition. The predicted A:B hazard for any particular vector of covariates z is then given by (11.5). (In computer code it is wise to subtract some centering constant c from the linear predictors, i.e., $\eta = X\beta - c$ so as to avoid large arguments in the exp function; this does not change the value of (11.5) in theory, but can lead to a large reduction in computational error.)

Let $L(t)$ be the matrix of predicted hazards at time t, below is an example for a 3 state model with states of A, B, C:

$$\hat{H}(t;z) = \begin{pmatrix} & \hat{\lambda}_{AB}(t;z) & \hat{\lambda}_{AC}(t;z) \\ \hat{\lambda}_{BA}(t;z) & & \hat{\lambda}_{BC}(t;z) \\ \hat{\lambda}_{CA}(t;z) & \hat{\lambda}_{CB}(t;z) & \end{pmatrix} \tag{11.6}$$

Each of the estimates comes from an individual proportional hazards model for the particular transition.

Depending on the state space model, some of the elements of $\hat{H}$ will be zero, e.g., there is no death:disease transition in the lower right model of Figure 8.1. Diagonal elements of the matrix are then filled in such that the row sums will be 1, satisfying the constraint that probabilities sum to 1. No matrix is created for a time point at which no transitions occur.

The simple AJ estimate of absolute risk, for this covariate vector z, is

$$\hat{p}(t; z) = p(0) \prod_{s \le t} \hat{H}(s; z) \tag{11.7}$$

over all time points s at which a transition occurred, with $p(0)$ the starting distribution of the states. This a reprise of the non-parametric AJ equation. The starting distribution $p(0)$ will most often be a single state, i.e., to display the future trajectory for a chosen (starting time, starting state) pair; the curve of interest for a particular patient.

Need to add examples of AJ curves from the above models to tie it all together

Exponential estimate

For a multistate model in continuous time, with known transition rates of $\lambda_{jk}(t)$, an alternate estimate comes from stochastic processes methods. Let A be a transition rate matrix with off diagonal elements $A_{jk} = \lambda_{jk}$, and diagonal elements such that the rows of A sum to 0. If the rates are constant over some window of time (s, t), then the transition matrix from time s to t is the matrix exponential

$$P(s, t) = e^{(t-s)A}$$

which is the solution to the corresponding differential equation.

Assuming that hazards are piecewise constant in the multistate hazards model leads to the matrix $\hat{A}$, with the same off diagonal elements as $\hat{H}$ above, but with diagonal set so that the row sums of the matrix are zero. This leads to the exponential estimate

$$\hat{p}(t; z) = p_0(t) \prod_{s \le t} e^{\hat{A}(s;z)} \tag{11.8}$$

For matrices, it is not guaranteed that $\exp(A)\exp(B) = \exp(A + B)$, so the right-hand side of (11.8) does not collapse to the exponential of a sum. Using the definition of a matrix exponential

$$\begin{aligned} e^A &= \sum_{i=0}^{\infty} A^i / i! \\ &= I + A + A^2/2! + A^3/3! + \ldots \\ e^{L(s)-I} &\approx I + (L(s) - I) = L(s) \end{aligned}$$

we see that the simple definition (11.7) can be viewed as a first order Taylor series approximation to the exponential definition.

Comparing this exponential Cox model estimator to the simple AJ extension equation (11.7):

1. If all the coefficients are 0 (or there are no covariates), then (11.7) reduces to the AJ. For a two state alive-dead model (11.7) reduces to the KM and (11.8) to the Fleming-Harrington estimator.
2. For the simple two state model with a single transition, (11.8) reduces to the standard Breslow estimate of survival for a Cox model.
3. For values of z that correspond to high transition rates, it is possible for $\sum_{k \neq j} \hat{H}_{jk}(t)$ to be greater than 1 for some time points t, which leads to a negative value on the diagonal, and in turn to negative probabilities in $\hat{p}(t)$ when using (11.7). The exponential formula, however, does not fail in this case.
4. Evaluation of the matrix exponential is numerically difficult [73, 74], and computer code that makes use of it can be surprisingly slow. Derivatives, necessary for an infinitesimal jackknife (IJ) estimate of standard error, can be doubly so.

Points 2 and 3 argue for using the exponential form, which is the default in the R survival package, while points 1 and 3 argue for the multiplicative form (the default in the R mstate package). In practice, issue 3 above tends to only arise when the number of subjects in the risk set is very small (often < 5). As a very simple example, consider the 3 state model for MGUS used in Chapter 9, and a fit with enrollment age and male sex as the covariates; coefficients for the entry:cancer and entry:death transitions are (0.013, -0.025) and (0.065, 0.393), respectively. The second to the last death is at month 321, with 9 subjects at risk, who ranged from 35 to 72 in enrollment age. The increments for the multiplicative and exponential estimates at month 321 are below; the first of these leads to a negative proportion of subjects in the entry state.

$$H(372; z) = \begin{pmatrix} -.0951 & 0 & 1.0951 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$e^{A(372;z)} = \begin{pmatrix} .3344 & 0 & .6654 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Time points where the simple method fails normally fall near the tail of the curve, a point where confidence intervals for the estimate $\hat{p}$ are wide. Since great accuracy is not really necessary at such a point, some packages instead choose to use a rough approximation at this point, e.g., set the diagonal to zero and scale the remaining elements of the row so as to add to 1. However, in a

multistate model it is quite possible to have small counts in one of the states early in the time course. I've seen this in a package, but I can't remember exactly where.

11.4 Stacked data

For the absolute risk $p(t)$, all transitions needed to be considered at once, leading to the matrix form of the AJ estimate. In the competing risks case, for instance, the probability of being in end state k is *not* equal to a KM estimate that treats other endpoints as censored. When modeling the hazards (arrows) themselves, however, the opposite is true: each transition *can* be addressed independently. That is, for any given j:k transition, we can use the following procedure:

- Starting with the full counting process dataset, select the subset of rows for which the current state = j.
- Within this subset, create a 0/1 status variable where status=1 indicates that this observation (row of the data set) ends with a transition to state k, otherwise status=0.
- Fit an ordinary proportional hazards (Cox) model to the new dataset using this 0/1 status variable.

This rather remarkable property of the hazards was shown by xxx [**?**]. Practically, it means that most of the procedures and methods found in the first part of this book immediately apply to the multistate case, including checks for proportional hazards and functional form, and issues of immortal time bias. We can also use the simple rate-based summaries of Section **??**. For models in the left hand column of Figure 8.1, and if there are no time-dependent covariates, all the time intervals will start at 0, a counting process style dataset is not required, and the individual transitions could also be modeled using accelerated failure time models. However, this is the uncommon case; more usually a counting process style of input is needed. Since accelerated failure time models do not adapt easily to (time1, time2) data, they will not be pursued further for the multistate case.

A clever programming trick for a model with k transitions (arrows) is to fit all k proportional hazards models at once using a *stacked* dataset. That is, create a new dataset with $m_1 + m_2 + \ldots + m_k$ rows; the first m_1 are the dataset we would use for the first transition's Cox model, the next m_2 are the data we would use for the second transition's model, and etc. Use the same variable name for each of the manufactured 0/1 status variables, and add a new variable `transition` which is 1 for the first m_1 observations, 2 for the next m_2 observations, etc., or use a more descriptive text label for each. If we had 2 covariates `x1, x2` then the following R code, for example, would fit all k models at once.

```
fit <- coxph(Surv(time1, time2, status) ~
                    (x1 + x2)* strata(transition),
```

```
        data = stacked)
```

Similar code applies for other packages. In each case it will turn out that 9/10 of the work will be expended setting up the data, while actually fitting the model is simple; a not uncommon occurrence in statistical analyses. There are, unfortunately, many opportunities to go wrong in creating this large dataset — some of the authors' most "interesting" results have come from exactly such errors — and as a consequence several packages have tools to aid in the process. The `mstate` package in R, for instance, has a routine `msprep` which will create a stacked dataset. The `coxph` function in the R `survival` package dispenses with creation of the stacked dataset entirely and operates directly on the counting process data (stacked data is created behind the scenes, however). Examples of both are given in the companion document.

Shared hazards

For a hazards model with separate hazards per transition, the off diagonal elements of $\hat{A}(t)$ are the simple estimates given earlier by equation (11.5). Now assume that transitions A:B and C:D share a common baseline hazard, with $\lambda_{CD0}(t) = \alpha\lambda_{AB0}(t)$, with $\eta_{iAB} = X_i\beta_{AB}$ and η_{iCD} the relevant linear predictors for subject i and the two transitions. The coefficient α is attached to the transition and β is attached to subject and their covariate values. The estimated hazards for a new subject with covariate value z, at time t will be

$$\begin{aligned}
\hat{\lambda}_0(t;z) &= \frac{\sum_i dN_{iAB}(t) + dN_{iCD}(t)}{\sum_i Y_{iA}(t)e^{\eta_{iAB}}} + \sum_i Y_{iC}(t)\alpha e^{\eta_{iCD}} \\
\hat{\lambda}_{AB}(t;z) &= e^{z\beta_{AB}}\hat{\lambda}_0(t;z) \\
\hat{\lambda}_{CD}(t;z) &= e^{z\beta_{AB}}\alpha\hat{\lambda}_0(t;z
\end{aligned}$$

This satisfies the constraint that sum(observed) = sum(expected, where the expected number of events at this time point is $\sum Y_{iA}(t)\exp(\eta_{iAB}) + \sum Y_{ic}(t)\exp(\eta_{iCD})$. The formula extends in to multiple shared curves in the obvious way; it also holds when A=C or B=D.

11.4.1 Partial likelihood

There is an important and thorny issue that arises with model 2 in Table 11.1. In short, if A, B, C, D are four possible states, a shared hazard model with $\lambda_{AB} = \lambda_{AD}\exp(\gamma)$ does not lead to a proper partial likelihood, while $\lambda_{AB} = \lambda_{CD}\exp(\gamma)$ and $\lambda_{AD} = \lambda_{CD}\exp(\gamma)$ do lead to a proper PL. That is, the partial likelihood is not well defined when two transitions, that begin in the same state, are assumed to share a common baseline. The rest of this section lays out the argument for this. The section is marked as optional for two reasons, the first is that the argument is somewhat technical and perhaps of most interest only to software authors. The second, and perhaps more important,

is that this discussion about proportional A:B and A:C hazards may only be academic. We have yet to encounter a dataset where proportional hazards for two transitions *from* a common starting state had sufficient biological plausibility to even be considered as a model. An assumption of proportional hazards for the transitions *to* a common endpoint, however, has proven useful, see the example in Section **??** for instance.

Start with the case where two arrows have the same ending state, for instance the two arrows to death for the model in the lower right of Figure 8.1, and call them A:D and B:D, D for death. If we use a stacked dataset approach, a shared baseline naively consists of simply using a manufactured strata variable, one that has the same value for those two transitions. Does this lead to an appropriate partial likelihood?

Under the assumption of proportional baseline hazards $\lambda_{A:D}(t) = \lambda_{B:D}(t)\exp(\gamma)$ for an unknown scalar $\exp(\gamma)$. When some subject i experiences an event D, say from state A, the partial likelihood

$$\begin{aligned} PL_i(t) &= \frac{\lambda_{A:D}(t)e^{X_{i\cdot}\beta_{A:B}}}{\left(\sum_{j\in A} Y_j(t)\lambda_{A:D}(t)e^{X_{j\cdot}\beta_{A:D}}\right) + \left(\sum_{j\in B} Y_j(t)\lambda_{B:D}(t)e^{\gamma+X_{j\cdot}\beta_{B:D}}\right)} \\ &= \frac{e^{Z_i\gamma+X_{i\cdot}\beta_{A:D}}}{\sum_j Y_j(t)e^{Z_j\gamma+X_{j\cdot}\beta_{A:D}+Z_jX_{j\cdot}(\beta_{B:D}-\beta_{A:D})}} \end{aligned}$$

where Z is a 0/1 time-dependent covariate whose value is 0 if the subject is in state A just before the event and 1 if in state B at that time. The denominator includes all those currently in state A who were at risk of D, and those currently in state B who are at risk of D; note separate coefficients for the A:D and B:D transitions. Per the final line, the stacked dataset can be adapted to the shared hazard case by creating a dummy variable Z for the current state, a modified strata variable that collapses the A:D and B:D transitions together, and example R code of

```
fit <- coxph(Surv(time1, time2, status) ~
                    (x1 + x2)* factor(transition) + strata(trans2),
             data = stacked)
```

An estimated baseline hazard from the modified dataset will have a jump at this point of

$$d\Lambda_{A:B}(t) = \frac{1}{\sum_j Y_j(t)e^{Z_j\gamma+X_{i\cdot}\beta_{A:B}+Z_jX_{j\cdot}(\beta_{B:D}-\beta_{A:D})}} \tag{11.9}$$

An important check on the baseline hazard is that the martingale residuals M at each event time sum to zero, where

$$M_i(t) = \int_0^t dN_i(s) - e^{X_i\beta}d\Lambda_0(s)$$

depends on both the covariates and baseline hazard for the transition. It is easy to verify that the estimate (11.9) satisfies this condition.

The case were two transitions do not overlap, A:B and C:D say, works out in the same fashion. When proportional hazard are assumed for two transitions that share the same starting state, however, the view is a bit more cloudy. As the simplest case consider a competing risks dataset with two transitions of A:B and A:C. If we use a stacked dataset approach, a shared baseline fit naively consists of simply replacing `strata(transition)` with `factor(transition)` when fitting the Cox model. But does this result in the solution to a true partial likelihood?

Under the assumption of proportional baseline hazards $\lambda_{A:C}(t) = \lambda_{A:B}(t)\exp(\gamma)$ for an unknown scalar $\exp(\gamma)$. When subject i has an event of type A:B, the most obvious partial likelihood term is

$$\begin{aligned} PL_1(t) &= \frac{\lambda_{A:B}(t)e^{X_{i.}\beta_{A:B}}}{\sum_j Y_j(t)\lambda_{A:B}(t)e^{X_{i.}\beta_{A:B}}} \\ &= \frac{e^{X_{i.}\beta_{A:B}}}{\sum_j Y_j(t)e^{X_{i.}\beta_{A:B}}} \end{aligned}$$

that is, the hazard of type B event for subject i compared to the sum of such hazards for those at risk for a type B event. Do the same in parallel for a type C event. Criticisms of this is that it yields same likelihood and solution $\hat{\beta}$ as the assumption of separate baseline hazards, and no clear path to a shared estimate of $\lambda_{A:B}$.

A second approach would be to consider the partial likelihood that subject i had an event of *either* B or C, leading to

$$PL_2(t) = \frac{\lambda_{A:B}(t)e^{X_{i.}\beta_{A:B}} + \lambda_{A:B}(t)e^{\gamma+X_{i.}\beta_{A:C}}}{\sum_j Y_j(t)\left(\lambda_{A:B}(t)e^{X_{j.}\beta_{A:B}} + \lambda_{A:B}(t)e^{\gamma+X_{j.}\beta_{A:C}}\right)} \tag{11.10}$$

The problem with this approach is that γ is not identifiable: the exact same expression occurs whether the the event was of type B or type C. In defense of (11.10), when there are identical baseline hazards and the coefficients are shared, $\gamma = 0$ and $\beta_{A:B} = \beta_{A:C}$, then this collapses to the usual PL for treating "either B or C" as a single event type.

A third approach is to multiply the above partial likelihood by the probability that the event was of type B, given that an event of type B or C did occur for subject i. This leads to

$$\begin{aligned} PL_3(t) &= PL_2(t)\frac{\lambda_{A:B}(t)e^{X_{i.}\beta_{A:B}}}{\lambda_{A:B}(t)e^{X_{i.}\beta_{A:B}} + \lambda_{A:B}(t)e^{\gamma+X_{i.}\beta_{A:C}}} \\ &= \frac{\lambda_{A:B}(t)e^{X_{i.}\beta_{A:B}}}{\sum_j Y_j(t)\left(\lambda_{A:B}(t)e^{X_{j.}\beta_{A:B}} + \lambda_{A:B}(t)e^{\gamma+X_{j.}\beta_{A:C}}\right)} \end{aligned} \tag{11.11}$$

Interestingly, this what will be computed by our "naive" application of a Cox algorithm to the stacked dataset, omitting the strata. We are not quite

comfortable calling this a partial likelihood, since it is the product of a partial likelihood for 'B or C' and a second term, but it appears to be a workable compromise. (We cannot consider it a partial likelihood directly for B, since that denominator should only contain observations that *could* experience event B. This would eliminate all of the duplicates from the denominator, duplicates that actually cannot have an event of type B by virtue of how the doubled dataset is constructed.)

The resulting estimate of $\lambda_{A:B}(t)$ will have a jump each time an event of type B or C is encountered. The printout for that estimate will look odd since the "number at risk" column will be essentially double what we expect, leading to half-size jumps. In return, there will be many more jump points than the nonparametric A:B hazard estimate.

11.5 Conclusions

When working with acute diseases, such as advanced cancer or end-stage liver disease, there is often a single dominating endpoint. Ordinary single event Kaplan-Meier curves and Cox models are then efficient and sufficient tools for much of the analysis. Such data was the primary use case for survival analysis earlier in the authors' careers. Data with multiple important endpoints is now common, and multistate methods are an important addition to the statistical toolbox. As shown above, they are now readily available and easy to use.

Grasping the big picture for a multistate dataset is always a challenge and we should make use of as many tools as possible. It is not always easy to jump between observed deaths, hazard rates, and lifetime risk. We are often reminded of the story of the gentleman on his 100th birthday who proclaimed that he was looking forward to many more years ahead, since "I read the obituaries every day, and you almost never see someone over 100 listed there".

Show cox.zph, discuss issue of combining coefficients after looking at the model fits vs making decisions based on small counts, show curves for absolute risk from model starting at different states.

Chapter 12

Multiple events of the same type

"Conducting data analysis is like drinking a fine wine. It is important to swirl and sniff the wine, to unpack the complex bouquet and to appreciate the experience. Gulping the wine doesn't work." Daniel B. Wright

Multiple events of the same type are commonly encountered in medical research. Examples include recurrent infections, hypoglycemic events among diabetics, and multiple fractures. Extra events provide additional information, but because the events within a person are correlated, they don't provide as much extra information as one might hope. In fact, these subsequent events are often worth only 1/3 to 1/10 of a first event. Multiple events per subject introduce intrasubject correlation that must be accounted for using approaches such as a robust variance or random effects. Historically, multiple events of the same type have been analyzed using the Andersen-Gill or conditional approaches using marginal or frailty models, as illustrated in Figure 12.1. However, it can also be informative to think about these analyses in the framework of multistate models which will allow us to introduce more sophisticated analyses. This chapter begins by introducing traditional analysis approaches, then revisits the examples using multistate models.

Two examples will be used to illustrate these concepts. The `cgd` dataset, first introduced in Section 1.5.2, comes from a randomized clinical trial of interferon gamma in children with chronic granulomatous disease (CGD). The primary endpoint of the study was time to the first serious infection, however, data were collected on all serious infections until the end of the follow-up. Out of the 128 subjects, 27 had one infection, nine had two infections, and eight had at least three infections.

The second example uses the `dnase` dataset, first introduced in Section 1.5.2, from a study of cystic fibrosis patients who are prone to chronic respiratory infections. The primary endpoint was the time until first pulmonary exacerbation; however, data on all exacerbations were collected for 169 days. Subjects who had an event were not considered to be at risk for another event during the course of antibiotics, nor for an additional 6 days after the end of the antibiotic treatment. If the symptoms reappeared immediately after the discontinuation of antibiotics then this would not be a new infection. Additionally, a few subjects were infected at the time of enrollment, such as subject

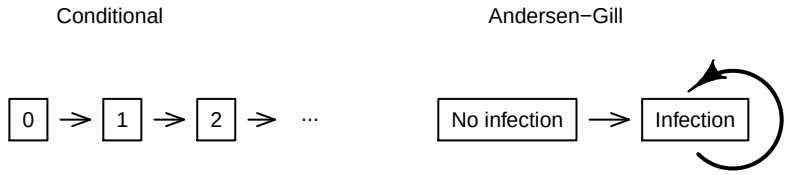

Figure 12.1 *Two possible approaches for viewing multiple events of the same type. In the left panel, the events are seen as conditional, so the states, if looking at infections as the event, are 0 infections, 1st infection, 2nd infection, etc. In the right panel, all the infections are considered to be independent, so the risk of moving from 0 infections to the 1st infection is the same as the effect of moving from the 1st infection to the 2nd infection.*

173 who had a first infection interval of -21 to 7. We do not count this first infection as an "event", and the subject first enters the risk set at day 7. In this dataset there are 644 subjects, 162 of whom experienced one infection, 53 who experienced two infections, and 27 who experienced three or more infections.

12.1 Simple estimates

For simple summaries, event rates (i.e., counts of the events divided by the person-years) may be sufficient. Table 12.1 uses the CGD data and displays the number of infections divided by the number of person-years. It suggests that there is a treatment effect and that perhaps the rates differ depending on whether the infection is a first or subsequent event. Similarly, the Kaplan-Meier and cumulative hazard plots (Figure 12.2) suggest treatment differences and are graphical approaches for summarizing multiple events. The cumulative hazard plot is the "accumulation" of the hazard over time of experiencing an infection and uses all the events. In the reliability literature the cumulative hazard is called the cumulative mean function (CMF or MCF) and is used

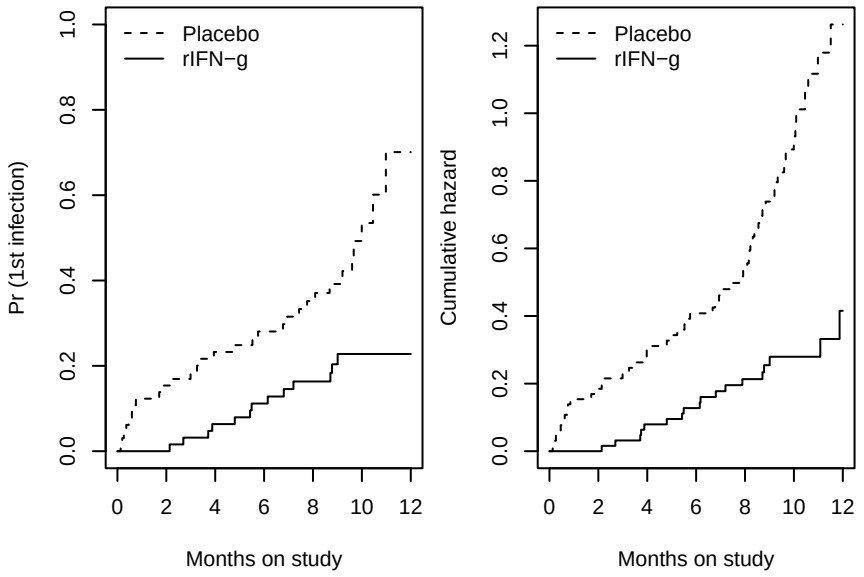

Figure 12.2 *Kaplan-Meier and cumulative hazard estimates for the two treatment arms of the CGD study counting all observed infections.*

more than the probability in state curve, reflecting the fact that counting the mean number of failures is a natural metric for that field.

treat	n.infect	n	event	pyears	rate
placebo	1st infect	65	30	37.5	0.80
rIFN-g	1st infect	63	14	47.0	0.30
placebo	2-7 infects	55	26	13.2	1.97
rIFN-g	2-7 infects	20	6	4.9	1.22

Table 12.1 *Within the CGD data, infection rates based on the first event and for subsequent events.*

12.2 Marginal regression models

Several approaches have been suggested for modeling multiple events of the same type. The most common approaches are the independent increment (Andersen-Gill, AG) and conditional (Prentice-Williams-Petersen, PWP) models. Both are considered "marginal" regression models in that $\hat{\beta}$ is determined from a fit that ignores the correlation followed by a corrected variance $\tilde{D}'\tilde{D}$. However, these approaches differ considerably in their creation of the risk sets.

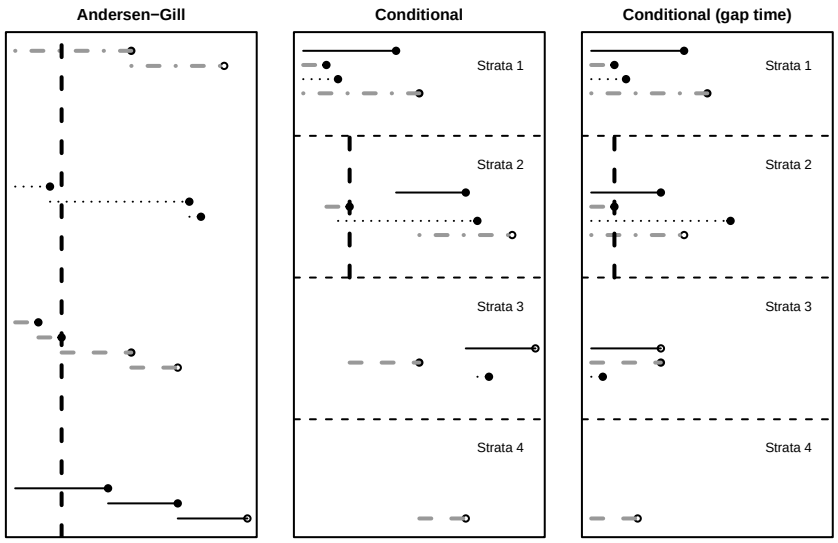

Figure 12.3 *Diagram illustrating risk sets using three different approaches for multiple events for four subjects, indicated by different line types. The thick, dashed vertical line indicates a specific event time. As indicated in the three panels, the comparison group at that event time can differ significantly. For analyses that utilize strata, only those subjects within a given strata are used for comparison purposes.*

Andersen-Gill model

The Andersen-Gill (AG) method is the simplest to visualize and set up, but makes the strongest assumptions. It is closest in spirit to Poisson regression, and can be accurately approximated with Poisson regression models in the same manner as was shown for an ordinary single-event Cox model.

Using the counting process style of data input, each subject is represented as a series of observations (rows of data) with time intervals of (entry time, first event], (first event, second event], ..., (mth event, last follow-up]. This model is ideally suited to the situation with mutually independent observations within a subject. The assumption is equivalent to each individual counting process possessing *independent increments*, i.e., the numbers of events in non-overlapping time intervals are independent, given the covariates. The variance is corrected via the infinitesimal jackknife variance estimate. In the first panel of Figure 12.3, the vertical line illustrates the comparison group at a given event time. Any subject who is under observation, regardless of how many events they may have already experienced, is included in the risk set.

```
Call:
coxph(formula = Surv(tstart, tstop, status) ~ treat + inherit +
```

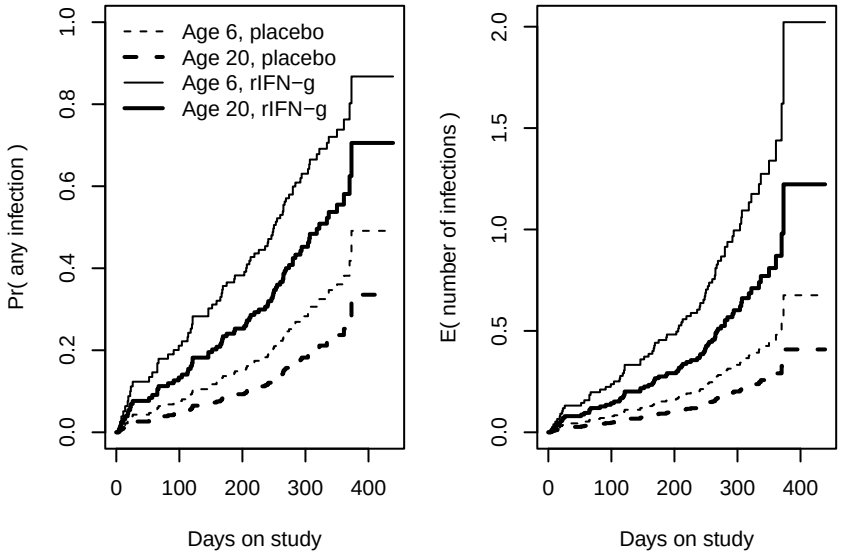

Figure 12.4 *Predicted survival and cumulative hazard curves for ages 6 and 20 years from the Andersen-Gill model of the CGD data.*

```
    age, data = cgd, id = id)

                   coef exp(coef) se(coef) robust se    z     p
treatrIFN-g      -1.096     0.334    0.262     0.305 -3.6 3e-04
inheritautosomal  0.372     1.451    0.245     0.329  1.1  0.26
age              -0.036     0.965    0.014     0.016 -2.2  0.02

Likelihood ratio test=28  on 3 df, p=3.4e-06
n= 203, number of events= 76
```

Predicted survival and/or cumulative hazard curves can then be obtained from the fitted model. In this case we will use 4 hypothetical subjects on the two treatments, at ages 6 and 20 years (near the quantiles), to create the predicted curves (Figure 12.4). We see that treatment is effective and the risk of infection decreases as age increases.

Model checks are the same as those for an ordinary Cox model, including assessment of the proportional hazards assumption, linearity of the continuous variables, and additivity by exploring interactions. For instance, there appears to be a non-linear age relationship with the endpoint, with a decrease in infection risk starting around age 25 years. Time-dependent covariates are easy to include as well, using the same approach as for a single endpoint.

Conditional model

In the conditional analysis proposed by Prentice, Williams, and Petersen (PWP), observations are stratified by the event number, i.e., the observations for the first event are in one strata and the observations for the second event are in a separate strata. Patients are not at risk for the 2nd event until the 1st event occurs. There is a separate baseline hazard (i.e., intercept / strata) for each event. This model does not assume events are independent like AG does, and it's particularly useful when the first event changes the probability or likelihood of subsequent events. Subjects may not have observations in all of the strata, which may give biased estimates due to a loss of balance in later strata.

There are two approaches that can be used for the conditional model. In the first approach, the time scale is the same as it was in the AG model which can lead to small numbers in the risk set at any given event time (see Figure 12.3, panel 2). In the CGD data, the variable `enum` indicates the start and stop times for the first event, the second event, etc. By stratifying on `enum`, only those who have experienced a first event are considered at risk for the second event, etc.

```
Call:
coxph(formula = Surv(tstart, tstop, status) ~ treat + inherit +
    age + strata(enum), data = cgd, id = id)

                   coef exp(coef) se(coef) robust se    z     p
treatrIFN-g      -0.899     0.407    0.282     0.295 -3.0 0.002
inheritautosomal  0.192     1.211    0.256     0.244  0.8 0.432
age              -0.028     0.973    0.014     0.012 -2.4 0.018

Likelihood ratio test=15  on 3 df, p=0.0021
n= 203, number of events= 76
```

Conditional model (gap time)

Often when stratifying by event, the time since last event is of primary interest. Essentially, after the subject has an event, the time to the next event starts over at time zero (Figure 12.3, panel 3). For simple datasets without time-dependent covariates, the gap time can simply be calculated as the difference between the beginning and ending of each time interval. If you have time-dependent covariates, you will need to be more careful.

```
Call:
coxph(formula = Surv(tstop - tstart, status) ~ treat + inherit +
    age + strata(enum), data = cgd, id = id)

                   coef exp(coef) se(coef) robust se    z     p
```

```
treatrIFN-g      -0.897     0.408    0.280     0.289 -3.1 0.002
inheritautosomal  0.175     1.191    0.260     0.262  0.7 0.505
age              -0.026     0.975    0.014     0.013 -2.0 0.043

Likelihood ratio test=15  on 3 df, p=0.0023
n= 203, number of events= 76
```

In this particular example there are eight strata because one subject has follow-up after their seventh event, leading to five strata (strata 4-8) with eight or fewer subjects. A modified version of the conditional analysis would be to truncate the dataset after a certain number of events. Another approach would be to group the strata into the first event and any subsequent events, which would be a hybrid of the conditional and AG models. The most important thing to remember is that within a strata, a subject should have overlapping time intervals.

The number of strata can vary considerably depending on the study. For instance, in the Exubera trial data, where the endpoint was hypoglycemic events, the number of events per subject ranged from 1 to 183, though the next largest number of events was 153. It is reasonable to wonder whether the risk of an event after the 150th event is different from the risk after the 100th event.

Table 12.2 shows the coefficients and robust standard errors using these three approaches for the CGD data. In this particular example, the results are similar.

model	variable	coef	robust_se	p.value
AG	treatment	-1.10	0.31	<0.001
	inherit	0.37	0.33	0.258
	age	-0.04	0.02	0.025
Conditional	treatment	-0.90	0.29	0.002
	inherit	0.19	0.24	0.432
	age	-0.03	0.01	0.018
Cond (gap)	treatment	-0.90	0.29	0.002
	inherit	0.17	0.26	0.505
	age	-0.03	0.01	0.043

Table 12.2 *Analysis of the CGD data using three different marginal regression models (Andersen-Gill, Conditional, Conditional using the gap time)*

12.3 Frailty models

The AG and conditional framework can also be fit using a frailty model instead of the marginal model. Instead of adjusting for multiple events per subject

after the fact with the infinitesimal jackknife variance estimate, a random subject (i.e., frailty) effect can be fit to account for within subject correlation. The outcomes are assumed to be independent conditional on the per-subject coefficient; more details can be found in Oakes [77]. Basically, frailty can be thought of as an unobserved individual random effect that acts multiplicatively on the hazard. Box-Steffensmeier and De Boef published an excellent summary of the conditional frailty model [16].

Frailty models can easily be fit using the `coxme` package, as seen with the the AG frailty model code shown below.

```
> coxme(Surv(tstart, tstop, status) ~ treat + inherit +
             age  + (1|id), data=cgd)
```

Table 12.3 shows a comparison of the marginal and frailty approaches for the AG, conditional, and conditional using gap time models.

model	variable	marginal	frailty
AG	treatrIFN-g	-1.10 (0.31)	-1.02 (0.30)
	inheritautosomal	0.37 (0.33)	0.29 (0.30)
	age	-0.04 (0.02)	-0.03 (0.02)
Conditional	treatrIFN-g	-0.90 (0.29)	-1.12 (0.33)
	inheritautosomal	0.19 (0.24)	0.28 (0.33)
	age	-0.03 (0.01)	-0.03 (0.02)
Cond (gap)	treatrIFN-g	-0.90 (0.29)	-1.00 (0.29)
	inheritautosomal	0.17 (0.26)	0.17 (0.28)
	age	-0.03 (0.01)	-0.03 (0.02)

Table 12.3 *Analysis of the CGD data using three different regression models (Andersen-Gill, Conditional, Conditional using the gap time) using marginal and frailty approaches*

12.4 Multistate models

Now reconsider the CGD data using our knowledge of multistate models. The diagram for the data looks like the left panel in Figure 12.1 where subjects move from having zero events to one to two to three, etc. With a full multistate model, the arrows to each box have a separate coefficient for movement to potentially seven infections. Table 12.4 shows the coefficients, standard error, and p-value for the treatment variable for each transition. Because there are only 8 subjects at risk moving from 3 to 4 infections, the coefficient is not estimable.

However, as shown in Section 11.1, we can easily put constraints on these estimates, provided we are willing to make the assumption that the covariate effects on different transitions are the same. Given the widely differing

transition	n.risk	coef	robust se	p.value
0:1	128	-1.140	0.347	0.001
1:2	44	0.156	0.537	0.772
2:3	16	-1.237	0.789	0.117
3:4	8		0.000	
4:5	3		0.000	
5:6	2		0.000	
6:7	1		0.000	

Table 12.4 *Unconstrained coefficients for treatment using the CGD data when transitioning between states representing the number of infections.*

coefficients for treatment across the transitions, this may not be a reasonable assumption. However, if we choose to, we can proceed with constraining the coefficients so that they are shared across all transitions (model 4 in Table 11.1).

term	estimate	std.error	robust.se	statistic	p.value
treatrIFN-g	-0.899	0.282	0.295	-3.047	0.002
inheritautosomal	0.192	0.256	0.244	0.786	0.432
age	-0.028	0.014	0.012	-2.365	0.018

Table 12.5 *Constrained coefficients for treatment using the CGD data when transitioning between states representing the number of infections.*

This provides a single coefficient for treatment based on data from all the transitions while still allowing the underlying baseline hazard to differ between transitions. If we also want to constrain both the baseline hazard and the coefficients (model 6 in Table 11.1), we can use the `\common` option. Again it may not make sense to do this, particularly if the event rates for the first and subsequent events differ. However, this constraint is very similar to the AG model.

variable	AG	Multistate
treatrIFN-g	-1.118 (0.309)	-1.096 (0.305)
inheritautosomal	0.399 (0.340)	0.372 (0.329)
age	-0.037 (0.016)	-0.036 (0.016)

Table 12.6 *Comparison of the coefficient and standard error estimates from Andersen-Gill(AG) and Multistate models*

The coefficients for the AG model and the multistate model with shared coefficients and common baseline hazards are similar, but not quite the same (Table 12.6). Why is that? It has to do the model set-up and the follow-up after the 7th event. The AG model set-up looks like the right-hand panel of Figure 12.1, where all events are the same. The multistate model set-up looks

like the left-hand panel, where each event is a separate state. In this dataset, the last event time is at day 373, but the subject with 7 events finishes their follow-up at day 350. In the AG model those last 23 days are at risk for another event, so they are included in the risk set. In the multistate model, those 23 days are at risk for event number 8, but there are no events in that stratum because no patients have an 8th infection. As no calculation are performed on a stratum without an event, those 23 days are essentially not included in the model calculations even when we specify a shared hazard across all stratum.

Another question is whether we should include those 23 days or not. We don't know with complete certainty that the study staff would have recorded an 8th event, if one had occurred. If that last time interval is removed from the data, then the coefficient from the AG matches the multistate model.

model	term	coef	std.error	robust.se	p.value
MS-shared coef	IFN.g	-0.90	0.28	0.29	0.002
	inherit	0.19	0.26	0.24	0.432
	age	-0.03	0.01	0.01	0.018
MS-shared coef/bh	IFN.g	-1.12	0.26	0.31	0.000
	inherit	0.40	0.25	0.34	0.240
	age	-0.04	0.01	0.02	0.023
AG-full	IFN.g	-1.10	0.26	0.31	0.000
	inherit	0.37	0.25	0.33	0.258
	age	-0.04	0.01	0.02	0.025
AG-trimmed	IFN.g	-1.12	0.26	0.31	0.000
	inherit	0.40	0.25	0.34	0.241
	age	-0.04	0.01	0.02	0.023

Table 12.7 *Coefficients for treatment and steroids using the CGD data using different models.*

One of the advantages for using the multistate framework is that it allows us to easily constrain the coefficients for transitions between one infection and another, but allows the transition from baseline to the first event to have a different underlying baseline hazard and coefficient. The table demonstrates that the treatment appears to have a different effect on the risk of first infection compared with subsequent infections.

	1:2	2:3	3:4	4:5	5:6	6:7	7:8
treatrIFN-g	-1.14	-0.51	-0.51	-0.51	-0.51	-0.51	-0.51
inheritautosomal	0.26	0.40	0.40	0.40	0.40	0.40	0.40
age	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02

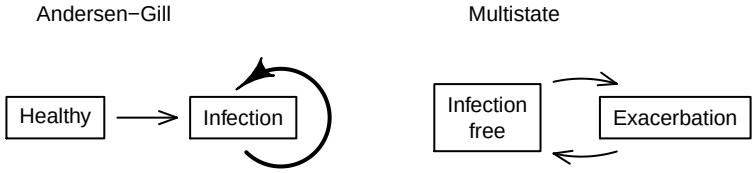

Figure 12.5 *Two possible diagrams for the rhDNase data. In the first diagram, subjects have multiple infections and the time period between when the IV starts and 6 days past when the IV ends, they are excluded from the risk set (gaps in their follow-up). In the multistate model, subjects move from being in the healthy state to being in the infection state (which includes the 6 days past when the IV ends).*

rhDNase example

The second example of the multistate framework focuses on the rhDNase trial where we noted that participants are not considered at risk for another event for six days following treatment. Figure 12.5 provides a diagram describing the states. Historically, when subjects entered the antibiotic state they were temporarily removed from the risk set. With the multistate model framework we can instead examine both the occurrence and duration of infection.

Below is the transitions table. There were 361 exacerbations, of which 316 resolved during the study period and 48 were censored.

```
         to
from      healthy  IV (censored)
  healthy       0 361        595
  IV          316   0         48
```

Figure 12.6 shows Aalen-Johansen estimates of the fraction currently in an exacerbation state, by treatment arm. The restricted mean time in state, up to day 200, is 1542 (155) and 1175 (127), 1542/200 = 7.7 and 1175/200 = 5.9, the mean heights of the placebo and rhDNase curves over the interval. This result is an amalgam of the probability of an exacerbation along with the

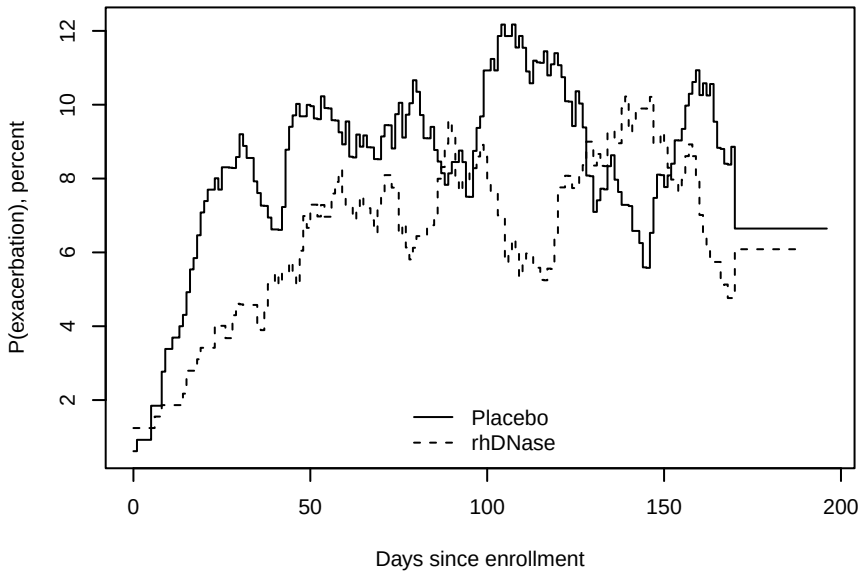

Figure 12.6 *Estimated percent of subjects currently in the exacerbation state, for those with and without using rhDNase among subjects with cystic fibrosis.*

duration of the event. If 1/100 have the event, with a duration of 100 days, or 5/100 each with a duration of 20 days, the mean time in state is 1 day for both cases. One approach is to scale: the cumulative hazard curve gives the expected number of events per subject, over time; values at 200 days are .70 and .54, giving a mean duration per exacerbation of 7.7/.70 and 5.9/.54, or 11 vs 10.9 treatment days per infection for placebo and treatment: almost the entire treatment effect appears to be due to fewer serious infections and not to a decrease in infection length.

The multistate framework allows us to investigate whether there are treatment differences for those moving from the healthy to infected state and from the infected state back to the healthy state. Using just the AG model, we would only be able to estimate healthy to infected.

```
Call:
coxph(formula = Surv(tstart, tstop, state) ~ trt + fev, data = dnase,
    id = id, istate = istate)


2:1       coef exp(coef) se(coef) robust se     z     p
  trt 0.027924  1.028318 0.115055  0.091287 0.306 0.760
  fev 0.001794  1.001796 0.002726  0.001588 1.130 0.258
```

```
1:2        coef exp(coef)  se(coef) robust se      z       p
  trt -0.283369  0.753242  0.106400  0.133326 -2.125 0.03355
  fev -0.007877  0.992154  0.002077  0.002475 -3.182 0.00146

 States:  1= healthy, 2= IV

Likelihood ratio test=22.84  on 4 df, p=0.0001361
n= 1320, number of events= 677
```

combine with above material

Table xxx [redo below as a table] shows the fit of the hazard model. The most important variable with respect to infection is forced expiatory volume (FEV), a measure of lung function. Normal values are 80–120; quantiles for this population are much lower at 38 and 83. A participant at the third quartile of FEV has a risk that is about 2.2 fold higher than someone at the first quartile (exp(.0178 * (83-38))). Treatment with rhDNase reduces the risk by 26%. The treatment has a very small apparent effect on recovery after an exacerbation: 1.03 (p= .64) fold faster transitions to normality for rhDNase. This is in line with the nonparametric results.

```
Call:
coxph(formula = Surv(tstart, tstop, state) ~ trt + fev, data = dnase,
    id = id, istate = istate)


2:1       coef exp(coef) se(coef) robust se     z     p
  trt 0.027924  1.028318 0.115055  0.091287 0.306 0.760
  fev 0.001794  1.001796 0.002726  0.001588 1.130 0.258


1:2        coef exp(coef)  se(coef) robust se      z       p
  trt -0.283369  0.753242  0.106400  0.133326 -2.125 0.03355
  fev -0.007877  0.992154  0.002077  0.002475 -3.182 0.00146

 States:  1= healthy, 2= IV

Likelihood ratio test=22.84  on 4 df, p=0.0001361
n= 1320, number of events= 677
```

Predicted probability in state curves $p(t)$ can be generated for any combination of FEV and treatment arm, as usual. A marginal estimate of the type discussed in Chapter 5 is, however, more interesting. This averages over FEV to give a single result for each of the treatment arms. The result is shown in Figure xx. To be added

12.5 Summary

Include this somewhere:

- Need coxme for multistate - software isn't there yet, complex programming task
- WLW - biased, no longer used
- discuss diabetic retinopathy data – diamond shaped multistate diagram - both ok, right eye bad, left eye bad, both eyes bad (what coefficients do we want to force to be the same, is left the same as right, ...)

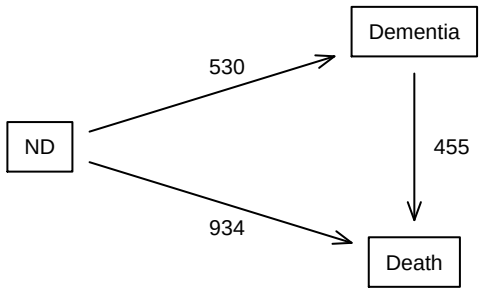

Figure 12.7 *State space for the MCSA data set, along with the number of observed transitions. All subjects start in the non-demented (ND) state.*

12.6 Dementia

The Mayo Clinic study of aging is based on an age and sex stratified random sample of subjects from Olmsted County, Minnesota (2020 population of 163 thousand) [**?**]. After initial enrollment participants are scheduled for repeated evaluation every 15 months. Subjects who have ceased active participation continue to be followed through their medical records for dementia and death. The Rochester Epidemiolgy project provides an overarching framework for the study; through it we have full enumeration of the county population, and long term follow-up of all medical care provided to county residents. Thus the initial sample is an accurate representation, subject to patient consent refer to Rosebud's paper, and post study follow-up is not subject to loss.

Jack et al [51] examined the simple 3 state model shown in Figure 12.7, with states of not demented (ND), dementia, and death. The `mcsa` data set consists of a 75% random sample of the MCSA subjects used in that analysis, which has been anonymised by the removal of all subject identifiers and dates, categorization of selected predictors, and rounding all time values to months. Figure 12.7 shows the number of observed transitions for the 3738 subjects in the shared data set.

To be filled in

- Further background on the data
- Number of subjects enrolled by age and year (full study)

- Average follow-up (9.4 years), max of 16
- Covariates
- How much of this goes in the companion? In a data section?

A primary goal of the analysis was to examine the long term utility of amyloid load in the brain, as measured by PET (positron emission tomography). Although nearly all subjects agree to the imaging part of the study, due primarily to capacity limitations only 36% of the analysis sample ever receive an amyloid PET scan (1354/Sexprsum0). Since PET imaging was not added to the MCSA until year 6 of the study, for half of those who are scanned (730/1354) the value is missing at the time of enrollment. Other covariates that play a key role in progression to dementia and death had few or no missing values, including age, sex, APOE genotype, and cardiovascular (CVD) comorbidities. As a way to retain all the samples, and increase precision for these other covariates, the decision was made to categorize amyloid level into 3 groups, using cut-points based on prior literature, with "missing" as a fourth level. This is entered into the analysis as a time-dependent covariate in order to make use of repeated scans, and as importantly, to properly account for those whose first PET evaluation was delayed. For a participant first measured at 30 months, say, an attempt to use "first PET" as a fixed covariate will be invalid, either by assigning that value at day 0, or by not entering them in the study until 30 months while those who never receive PET start at day 0. Both strategies are examples of immortal time bias, i.e., using *future* information to decided on an observation's covariate values or inclusion/exclusion at a *current* time point.

Figure 12.8 shows the overall death and dementia rates in the samples as a function of age and sex. The participant's data in this study ranges from 55 to 100 years of age, and the change in rates over this interval is over 200 fold – a huge change. The effect of age is so large that the estimated overall effect of male sex, on death rates, 1.41, looks to be of little consequence. Yet effects of 1.3 – 1.5 fold are exactly the size we often look for in clinical research. We also see that the age + sex effect, for death, is both nearly linear and additive. This is not the case for the dementia endpoint.

We choose to use age as the underlying time scale for the multistate models. The resulting partial likelihood terms, one at each event time (dementia or death), each involve an age matched sub-sample.

1. The common alternative for a Cox model is to use time since enrollment as the time scale, and add age as a covariate. If we do so, however, the functional form of the age estimate would need to be almost exactly correct. The quite common habit of adding "+ age + sex" to the code's model statement is particularly problematic in this case, due to the very large effect of age; the latter due to combination of the large age range and that age is the single most important predictor of death. Even for death, where the two lines appear to be nearly linear and additive, any pair of parallel lines added to the figure, i.e., predictions from an `age + male` model, will

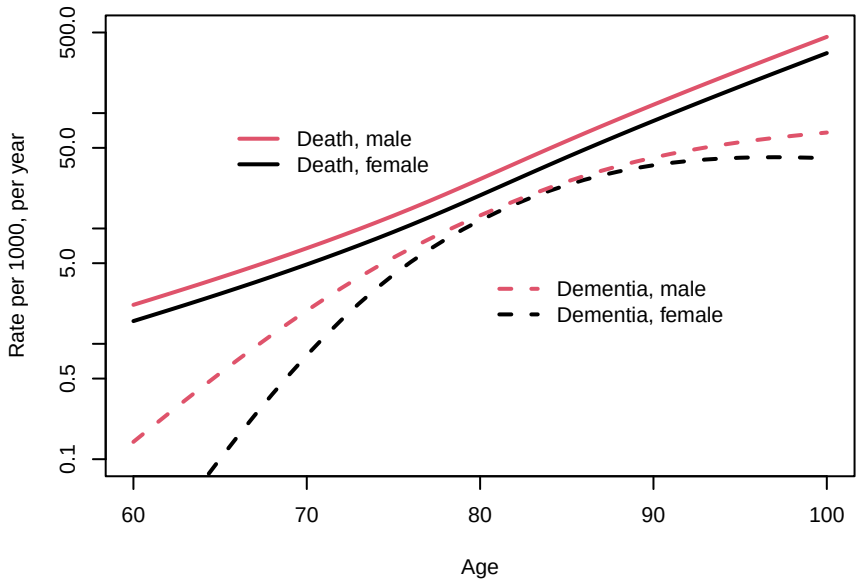

Figure 12.8 *Death and dementia rates per 1000 person years in the MCSA cohort, based on a simple Poisson model.*

be off by 2x or more in some areas. This is obviated by the matching induced by using age as a time scale.

2. As a downside, a p-value for the age effect is not produced. But, do we really need to 'prove' that death and dementia rates rise with age?
3. The standard underlying time axis of time-since-enrollment is not particularly relevant in this study. Unlike time from diagnosis or the time since a treatment intervention, the time since a subject was randomly chosen for an invitation letter, and chose to respond, does not have clear scientific interest or meaning.
4. Another consequence of this choice is that absolute risk estimates will be on an age time scale as well, rather than time since enrollment.

A sex specific APOE effect on dementia risk has been suggested in the literature and so is included in our model. (The LRT for addition of a sex specific amyloid effect is not signifcant with $p > .2$). Figure 12.9 shows the hazard ratios from the multi-state model.

The results paint a fascinating picture.

- The effect of both amyloid level and APOE positivity on dementia rates is profound. Comparing APOE+ to APOE- we see a larger effect for females than for males, in agreement with the prior literature.
- Males have a higher overall dementia rate than females, as found in Figure **??**. This analysis suggests that a primary cause may be a lower threshold on

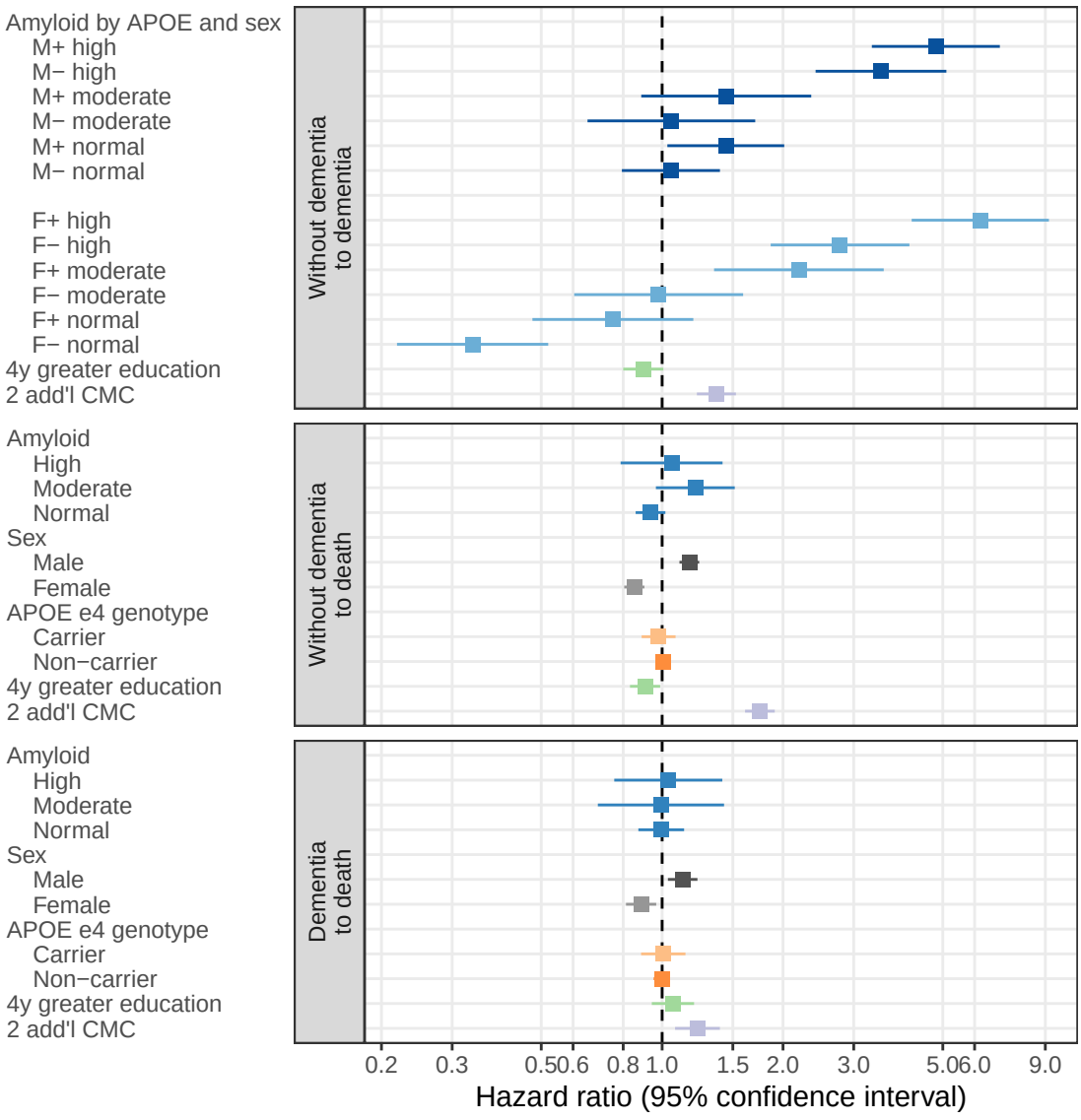

Figure 12.9 *Estimated hazard ratios for the multi-state model for the MCSA data. Categorical variables are referenced to the population mean rather than a single reference category. M+ = male, APOE e4 positive; high/medium/low = levels of amyloid burden.*

their rate: amyloid normal males do not enjoy further reductions compared to moderate amyloid, as the females do.

- Neither amyloid levels nor APOE positivity appear to have any direct effect on death rates.
- Male sex increases death rates by one third, and effect that persists with or without dementia.
- Additional cardiovascular comorbidities have a major effect on death without dementia, but also a strong effect on dementia incidence and on death after dementia.

The model also includes a term for time since enrollment, with respect to the death outcome. This term is an important addition when age is used as the underlying time scale for a Cox model. The simplest explanation is symmetry; a Cox model on the standard enrollment time scale with age as a covariate adjusts for both effects, via matching (enrollment) or modeling (age). To be comparable a model on age scale should also account for both effects, age via matching and enrollment via covariate terms. Enrollment was added as a time-dependent covariate to the MCSA fit as a time-dependent category of 0–1 years, 1–5, and 6+ (reference) based on the authors' prior experience in other studies. The fitted estimate is a death rate of .39 in the first year and .77 in years 1–5, as compared to the longer term death rate. An enrollment effect was not included for dementia, for two reasons. Primary is that one must be cautious when setting it up, e.g., follow-up visits are scheduled at 15 month intervals, and dementia is only detected at a study visit. Thus by definition the dementia rate is 0 for the first year. (Note to self, update base mcsa to not have the time effect, so that I can see the visit frequency). A full xxx% of subjects delay their first return visit, due at 15 months, by more than 30 months, making any hard threshold difficult to justify. The potential effect of incipient dementia is also debatable, will such people have a stronger incentive to enroll, or less? We did try to explore this but did not find any clarity.

Our experience has been that a study whose enrollment involves patient choice will often have a death rate in the first year that is reduced to 35–65% of the long term rate. We speculate that a primary reason is that patients who are in extremis rarely volunteer for research studies. The opposite effect can occur in a chronic disease whose course waxes and wanes: a new study might be more attractive or specifically targeted to patients who are currently experiencing a disease flare. In this case the rates in the first year will be higher than the long term average. Data sets without an enrollment effect are actually harder to find than those with an effect. One example are the control subjects in the NAFLD data set of section **??**: for each NAFLD case 4 age and sex matched controls were chosen from the underlying population, and there is no apparend time since selection effect on their mortality.

As with all multi-state models, a critical adjunct for understanding the fit are the estimates of absolute outcome. Figure **??** shows the lifetime risk of dementia, and Figure **??** the probability in state as a function of age.

Predicted curves show the future for a particular starting point: current age, current state, and a set of covariate values. Imagine advising a particular subject, currently in the physician's office, of the expected trajectory from that point. We have chosen to show a non-demented subject at either age 65 or age 80. In this particular study the covariate of primary interest was amyloid level, followed by APOE status and sex. Education and CMC were felt to be important adjusters, but not a primary target. Thus, we have integrated out these latter two by using a marginal estimate as was presented in section **??**.

The figures show a very interesting progression. Males have a higher hazard of dementia, as shown both in figure **??** and 12.9. The cumulative dementia risk in Figure **??**, however, shows very similar lifetime risks of dementia before death: higher for females than males with APOE+, somewhat lower for those without APOE. The primary reason for this shift is that females live longer, compounded by the precipitous rise in overall rate as a function of age shown in Figure **??**. Rates are higher for males, but that does not translate into a higher lifetime risk.

The estimate for "currently in the dementia state" which forms the central panel in Figure **??** amplifies this effect even more. The highest risk female curve tops out at 21% versus 13% for males: 1.6 times as large. Again, females with dementia live longer than males with dementia. The area under the curve is an estimate of E(years with dementia), values are shown in figure **??**. This value incorporates both the higher probability of getting dementia plus the longer time in state for females that experience it. We often like to think of the hazard ratios as underlying biology, with absolute risk the consequences of that biology.

12.6.1 Time dependent covariates

The prior discussion of absolute risk has ignored one key aspect, which is that the model contains time-dependent covariates. This brings us up against a harsh reality: a full understanding of the model requires that the hazard ratios be complemented by absolute risk estimates, but the standard estimators post PH-model survival curves are not valid in the presence of time dependent covariates. Working through this, for this data set, provided an important learning experience (and updates to the R survival package).

As a simple starting point, consider the

At any given time point t the increment $p(t;x) = p(t-;x)\exp(A(t;x))$ described in section zzz gives a proper update to the probability in state vector $p(t)$ for a subject with covariate vector x *at that time*. Creation of the overall curve for a subject starting in state $p(65) = (1,0,0)$ (non-demented) and baseline covariate vector x_0 requires an estimate of the the *covariate path* $x(t)$ over time, in order to compute the correct $A(t)$ term at each further time point.

This is a very interesting set of results, but now brings us up against a

harsh reality: a full understanding of the model requires that the hazard ratios be complemented by absolute risk estimates, but standard estimators are not valid in the presence of time dependent covariates. Multi-state models, however, provide a way around the issue. That is, to compute a predicted survival curve based on the model, from some time point forward, also requires prediction of the time-dependent covariate's path forward from that time point. But this can be accomplished by adding the amyloid status as a state in the model, in the same fashion as was done for the PBC data in Section **??**. The updated state space figure is shown in Figure 12.10.

```
       to
from       0    1    2    3   10   11   12   13   20
  NdA0  1970    0  107    2   40    0    0    0   92
  NdA1   621   12    0   61   28    1    0    0   49
  NdA2   357    0    0    0   61    0    0    0   37
  NdAm  6423  435  165  116  390    0    5    5  756
  DemA0   20    0    1    0    0    0    0    0   32
  DemA1   17    0    0    1    0    0    0    0   28
  DemA2   36    0    0    0    0    0    0    0   49
  DemAm  186    2    1    2    0    0    0    0  346
       to
from    NdA0 NdA1 NdA2 DemA0 DemA1 DemA2 DemAm Death (censored)
  NdAm   435  165  116     0     5     5   390   756       1242
  NdA0     0  107    2    40     0     0     0    92        678
  NdA1    12    0   61     1    28     0     0    49        223
  NdA2     0    0    0     0     0    61     0    37        129
  DemA0    0    0    0     0     1     0     0    32         10
  DemA1    0    0    0     0     0     1     0    28          6
  DemA2    0    0    0     0     0     0     0    49         19
  DemAm    0    0    0     2     1     2     0   346         39
  Death    0    0    0     0     0     0     0     0          0
```

The transition table for the expanded model reveals that demented subjects almost never change amyloid state. On reflection, this is almost guaranteed, since it is extremely rare for a demented subject to receive a PET scan. The check shows one person with two NdA0 events. A look shows that they transition from Am, A0, A1, A0. Would the example be better if we had looked ahead, and realized that subdivided dementia state is not required?

For the state2 endpoint, we want to have an additive model that matches the simple time-dependent fit with respect to coefficients. Referring to the transition matrix for the larger model, transitions within the ND sub-states or with Dementia sub-states are unconstrained, each has its own baseline hazard. Let 1-4 be the four amyloid sub-states for non-demented and 5–8 those for

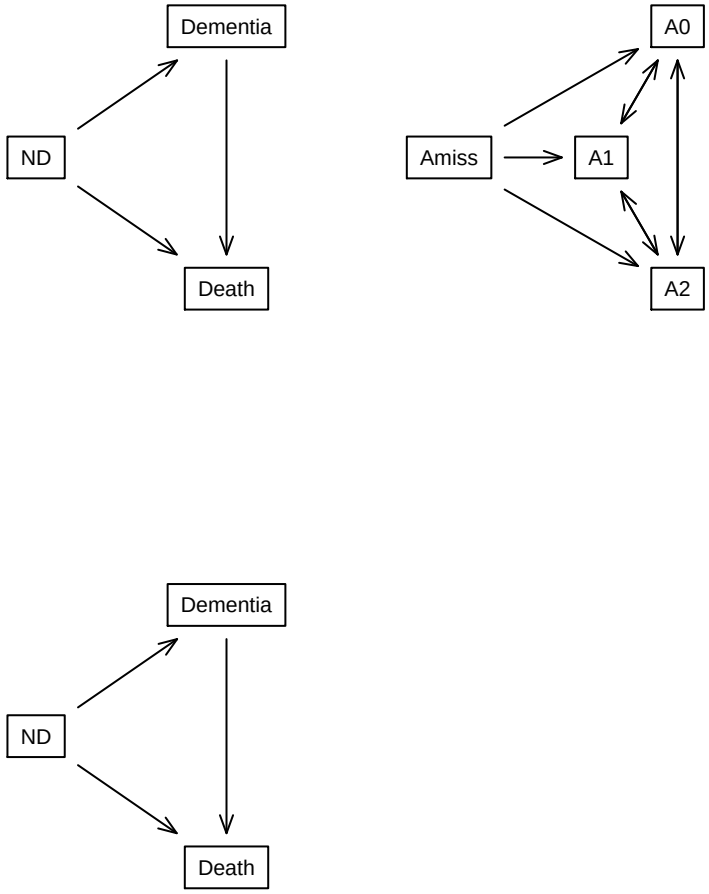

Figure 12.10 *Extended state figure for the MCSA data. The ND and dementia states each have 4 sub-states. (B&C: can we draw a dashed box around the entire left hand set, with dashed arrows to the ND and Dementia boxes?)*

the demented. We want

$$\lambda_{15}(t)\gamma_1 = \lambda_{2j}(t)\ j = 5, 6, 7, 8$$
$$\lambda_{15}(t)\gamma_2 = \lambda_{3j}(t)\ j = 5, 6, 7, 8$$
$$\lambda_{15}(t)\gamma_3 = \lambda_{4j}(t)\ j = 5, 6, 7, 8$$

Looking at the transitions matrix, many of these hazards have no representatives.

In progress:

```
temp2
   0    2    3    4    5    6    7   10   13   14   15   17   20
2024    6    9   19   14   12    3  239    3    4    1    1  590
  22   23   24   25   26   27
   1    2    9    3    1    3

   0    1    2    3    4    5    6    7
 299  812 1067  824  478  207   47    4
```

Meeting notes for TD survival curves

CMC never appeared in any plots, we always margined over it. It matters for multiple transitions. I begin to think that we could leave it as is.

It is a lot of work to recreate the same model in "state space" as we had in the usual TD. As was done with PBC and bilirubin. If you have multiple variables it will be worse. Can there be a new 'survfit' function, that is told which variables are TD, and which of those have an initial state and which are margined? Initial state only makes sense for a grouped variable.

Addition to the adjusted survival curves chapter. A discussion of strata vs covariates. Strata is more general, but you can't go too fine due to loss of a group.

Appendix A

Formulas

> Statistics is, or should be, about scientific investigation and how to do it better, but many statisticians believe it is a branch of mathematics... Now I agree that the physicist, the chemist, the engineer and the statistician can never know too much mathematics, but their objectives should be better physics, better chemistry, better engineering, and in the case of statistics better scientific investigation. Whether in any given study this implies more or less mathematics is incidental. — George Box

This section gathers all the notation and formulas for the models in one place. For the data we use counting process notation:

$N_i(t)$	cumulative number of events for subject i, up to time t
$N_{ijk}(t)$	in a multistate model, the number of transitions from state j to state k for subject i
$Y_i(t)$	0/1 indicator that subject i is at risk at time t
$Y_{ij}(t)$	in a multistate model, the indicator that subject i is at risk for a transition out of state j
X	n by p matrix of predictors where p is the number of predictors, X_i the row vector for observati
w	vector of case weights for the subjects, W a diagonal matrix of the weights
β	vector of estimates
$\eta = X\beta$	vector of linear predictors

A.1 Nelson-Aalen and Kaplan-Meier estimates

The most common estimates of the cumulative hazard and survival are the Nelson-Aalen and Kaplan-Meier estimates.

$$\hat{\Lambda}(t) = \int_0^t \frac{\sum w_i dN_i(s)}{\sum_i w_i Y_i(s)} \tag{A.1}$$

$$\text{var}(\hat{\Lambda}(t)) = \int_0^t \frac{\sum w_i dN_i(s)}{\left(\sum_i w_i Y_i(s)\right)^2} \tag{A.2}$$

$$\begin{aligned} \widehat{S}(t) &= \prod_{s \le t} \left(1 - \frac{\sum w_i dN_i(t)}{\sum_i Y_i(t)}\right) \\ &= \prod_{s \le t} \left(1 - \hat{\lambda}(s)\right) \end{aligned} \tag{A.3}$$

$$\text{var}(\widehat{S}(t)) = \widehat{S}^2(t) \int_0^t \frac{\sum w_i dN_i(s)}{\sum_i w_i Y_i(s) \left[\sum_i w_i (Y_i(s) - dN_i(s))\right]} \tag{A.4}$$

$$M_i(t) = N_i(t) - Y_i(t)\hat{\Lambda}(t)$$

The variance estimate (A.4) for the Kaplan-Meier is known as the Greenwood estimator. An alternate variance is $\widehat{S}^2(t)\text{var}(\hat{\Lambda}(t))$, but it has been found to be inferior to the Greenwood.

The quantity $M_i(t)$ is the martingale residual process for subject i, it is a running total of the observed - expected number of events for the subject. (Since the theory of martingales was motivated by cumulative winnings in games of chance, it's applicability to cumulative events is perhaps not surprising). At any time point $\sum w_i M_i(t) = 0$, and M_i without a time indicator is understood to be the terminal value.

When there are multiple events tied at a single time, an alternate estimate of the cumulative hazard can be based on the idea of *coarsened* data. That is, if the time scale had been more finely measured the ties would not have occurred. Say that there were 10 subjects at risk of which 3 had an event. Using the data, observed on a coarsened time scale, the Nelson-Aalen estimate will have a jump of 3/10, but if time had been measured continuously the jump would have been 1/10 + 1/9 + 1/8 = .336. This estimate is explored by Fleming and Harrington [33].

In the more general setting of case weights, if there were d tied events at time t we can apply the Nelson-Aalen estimate to a synthetic data set in which there are d distinct event times t, $t - \epsilon$, $t - 2\epsilon$, ..., and each of the tied observations is split into d synthetic subjects with weight w_i/d, one of which has an event at the first of the synthetic times, one at the second synthetic time, etc. The idea is that each subject could have been the first, second, third, etc of the d events, so we spread each of them out evenly over those possibilities. Computationally, we do not need to actually create the synthetic data. If there are d events at some time t, let $n_2(t) = \sum w_i dN_i(t)$

be the sum of weights for those who have an event at that time and $n_1(t) = (\sum w_i Y_i(t)) - n_2()$ be the sum for all the others. The the increment to the hazard estimate and the derivative are

$$d\hat{\Lambda}(t) = \frac{n_2(t)}{d} \sum_{m=1}^{d} 1/(n_1 + (m/d)n_2) \tag{A.5}$$

$$\frac{\partial \hat{\Lambda}(t)}{\partial w_i} = \int_0^t dN_i(s) d\hat{\Lambda}(s) - Y_i(s) \sum_{m=1}^{d} \frac{(m/d)I(i \in n_2) + 1I(i \in n_1)}{(n_1 + (m/d)n_2)^2} \tag{A.6}$$

$$\neq \int_0^t \frac{dM_i(s)}{\sum_i w_i Y_i(s)} \tag{A.7}$$

From equation (A.5), by moving the divisor d to the right we see that the method is essentially using an average denominator for the FH increment. The step from (A.6) to (A.7) that was used in the Nelson-Aalen case fails here since products do not factor — $\sum a_i b_i \neq (\sum a_i)(\sum b_i)$; one can't separately sum the dM contributions at a tied event time.

The coarsening argument does not affect the Kaplan-Meier estimate, which will be 7/10 on the coarsened scale and the product (9/10)(8/9)(7/8)= 7/10 for continuous time, i.e., the solution does not change. This is true for weighted data as well.

For a continuous survival distribution $S(t) = \exp(-\Lambda(t))$. Since $\exp(-x) \approx 1 - x$ when x is small, this is also approximately true for the non-parametric estimates: $\widehat{S}(t) \approx \exp[-\hat{\Lambda}(t)]$. In this case the exponent of the Fleming-Harrington hazard estimate will be closer to the KM than the exponent of the Nelson-Aalen [33], but the differences between all three of KM, exp(-NA) and exp(-FH) are normally very small until the number at risk becomes small, at which point the standard errors of the curves are large as well. The exponential estimate plays an important role in curves for a Cox model, but is rarely used for simple survival.

A.2 Aalen-Johansen estimate

The multistate analog to the Kaplan-Meier curve is the Aalen-Johansen estimator. It estimates $p(t)$, a vector containing the probability of being in each of the states at time t. Mathematically the estimate is simple. For each unique time that an event occurs form a transition (or hazards) matrix $H(t)$ with elements $\hat{\lambda}_{jk}(t)$ = the fraction of subjects who transition from state j to k at time t, among those in state j just prior to t. Formally

$$\hat{\lambda}_{jk}(t) = \frac{\sum_i w_i Y_{ij}(t) dN_{ijk}(t)}{\sum_i w_i Y_{ij}(t)}, \; j \neq k$$

where $Y_{ij}(t)$ is 1 if observation i is in state j at time $t-$ (just before the event), $dN_{ijk}(t)$ is 1 for a transition from state j to k, and w is an optional

case weight. The diagonal elements of H are such that the row sums of H are 1.

Then

$$p(t) = p(0) \prod_{s \le t} H(s) \tag{A.8}$$

where $p(0)$ is the initial distribution of subjects. H is equal to the identity matrix at any time point without an observed transition, so we only need to include transition times in the product.

The elements of $p(t)$ sum to 1 at each time point: everyone has to be somewhere. Likewise, the rows of each transition matrix H sum to 1: everyone has to either stay put (diagonal) or transition to a different state. For the simple alive→dead model there are only two states and H has the simple form

$$H(s) = \begin{pmatrix} \frac{\sum Y_i(s) - dN_i(s)}{\sum_i Y_i(s)} & \frac{\sum dN_i(s)}{\sum_i Y_i(s)} \\ 0 & 1 \end{pmatrix}$$

$H_{21} = 0$ since no one transitions from dead to alive; death is an *absorbing* state. Writing out the matrices for the first few transitions and multiplying them leads to

$$p_1(t) = \prod_{s \le t} \frac{\sum Y_i(s) - \sum dN_i(s)}{\sum Y_i(s)} \tag{A.9}$$

which we recognize as the Kaplan-Meier estimate of survival.

An alternate estimator is based on matrix exponentials

$$\begin{aligned} p(t) &= p_0(t) \prod_{s \le t} \exp(H(s) - I) \\ &= p_0(t) \prod_{s \le t} \exp(A(s)) \end{aligned} \tag{A.10}$$

The off diagonal elements of A are $\hat{\lambda}_{jk}(t)$, the same as H, but the diagonal elements are constrained so that rows sum to zero. We normally cannot further collapse equation (A.10), since in general $\exp(C)\exp(D) \neq \exp(C + D)$ for two matrices C and D. We refer to (A.8) and (A.10) as *direct* and *exponential* estimates of the probability in state matrix. For single transition models these correspond to the Kaplan-Meier and Fleming-Harrington estimates, respectively. The $A(s)$ notation is common in the counting process literature, in which case $A(s) + I$ usually replaces $H(s)$.

The matrix exponential is defined as

$$\exp(A) = I + A + A^2/2! + A^3/3! + \ldots$$

we can also view the direct estimate as a first order Taylor series approximation to the exponential.

A.3 Cox Model

Let $r_i(t) = \exp(\eta_i(t) - c) = \exp(X_i(t)\beta - c)$ be the risk score for each subject at time t, with c an arbitrary centering constant. Mathematically we can set $c = 0$, since it cancels out of the partial log-likelihood (LPL), but computationally it is important to choose a value that avoids extreme arguments to the exponential function. (The value c is universally left out of textbook formulas, and at the same time universally recognized as crucial by writers of widely distributed code.) A common choice is the mean risk score $(1/n)\sum \eta_i$, $\eta = X\hat{\beta}$, but the exact centering value is not critical.

We have the following quantities:

$$LPL = \sum_i \int \left(w_i \log(r_i(s)) - \log \left[\sum_j Y_j(t) w_j r_j(s) \right] \right) dN_i(s) \tag{A.11}$$

$$\hat{\lambda}(t; c) = \frac{\sum w_i dN_i(t)}{\sum_i Y_j(t) w_i r_i(t)} \tag{A.12}$$

$$\hat{\Lambda}(t; c) = \int_0^t \hat{\lambda}(s) \tag{A.13}$$

$$M_i(t) = N_i(t) - Y_i(t) r_i \hat{\Lambda}(t)$$

$$\overline{x}(t) = \frac{\sum_i Y_i(t) w_i r_i(t) X_i(t)}{\sum_i Y_j(j) w_i r_i(t)}$$

$$U = \sum_i \int [x_i - \overline{x}(s)] dN_i(s) \tag{A.14}$$

$$H = \sum_i \int dN_i(t) \frac{\sum_{j=1}^n w_j r_j(s) [X_j(s) - \overline{x}(s)]^2}{\sum_j Y_j(s) w_j r_j(s)} dN_i(s) \tag{A.15}$$

$$\tag{A.16}$$

Since $N(t)$ is a step function, integrals with respect to dN are equivalent to a sum that contains a term at each death time. Each term of the log partial-likelihood (LPL) compares the risk score r_i of a subject who had an event to the sum over all those at risk. The baseline hazard estimate (A.13) is for a hypothetical subject with risk score $r = 0$ or equivalently $X\hat{\beta} = c$. Textbooks commonly use the formula with $c = 0$ and refer to the result as *the* baseline hazard $\hat{\lambda}_0$, but again, this is a very unwise choice in computer code due to potential numeric errors.

When $\beta = 0$ then $\hat{\Lambda}$ is equal to a Nelson-Aalen estimate, and the the estimated survival $\exp(-\Lambda(t))$ is referred to as the Breslow estimate. $\overline{x}$ and H are the weighted mean and variance of the covariate vectors at each event time. The total of $H(t)$ over the death times is the second derivative of the LPL, also known as the Hessian or information matrix. U is the contribution to the first derivative of the LPL.

The martingale residual is the observed - expected number of events for

each subject i. M_i without an argument is taken to be the value at the maximum follow-up for the subject.

Three important identities are

$$\sum_i w_i M_i(t) = 0$$

$$M_i = \frac{\partial LPL}{\partial \eta_i} \tag{A.17}$$

$$\begin{aligned} U &= \frac{\partial LPL}{\partial \beta} \\ &= M'WX \end{aligned} \tag{A.18}$$

where W is the diagonal matrix of weights. Equation (A.18) plays an important role in MCMC or machine learning approaches which make use of first derivative information, and for which X may be both large and sparse. Since the marginale residual M can be computed in $O(n)$ steps this allows for fast sparse matrix calculation of the first derivative.

A.3.1 Derivation using the chain rule

One approach to derive U and H is to use the chain rule, starting with the view of $LPL = f(\eta_1, \eta_2, \ldots)$. First, remind ourseleves of the rules for a simple function of two derived variables. I'll use dx/dt for the inner nesting and ∂ for the outer, to help readability.

$$\begin{aligned} g(t) &= f(a(t), b(t)) \\ g'(t) &= \frac{\partial f}{\partial a}\frac{da}{dt} + \frac{\partial f}{\partial b}\frac{db}{dt} \\ \frac{d}{dt}&\left(\frac{\partial f}{\partial a}\frac{da}{dt}\right) \\ &= \left(\frac{\partial^2 f}{\partial a^2}\frac{da}{dt} + (\frac{\partial^2 f}{\partial a \partial b}\frac{db}{dt}\right) + \frac{\partial f}{\partial a}\frac{d^2a}{dt^2} \\ g''(t) &= \frac{\partial^2 f}{\partial a^2}\left(\frac{da}{dt}\right)^2 + \frac{\partial^2 f}{\partial b^2}\left(\frac{db}{dt}\right)^2 \\ &+ \left(\frac{\partial f}{\partial a}\right)^2 \frac{d^2a}{dt^2} + \left(\frac{\partial f}{\partial b}\right)^2 \frac{d^2b}{dt^2} \\ &+ \left(\frac{\partial^2 f}{\partial a \partial b} + \frac{\partial^2 f}{\partial b \partial a}\right)\frac{da}{dt}\frac{db}{dt} \end{aligned}$$

Our interior (a, b, ...) functions are $\eta_i = X_i\beta$. The first derivative with respect to β_k is X_{ik} and second derivatives are 0; this causes the second line of g'' above to disappear, giving

$$\frac{\partial LPL}{\partial \beta_k} = \sum_i \left[\frac{\partial L}{\partial \eta_i} \frac{\partial \eta_i}{\beta_k} \right]$$
$$= \sum_i \frac{\partial L}{\partial \eta_i} X_{ik} \tag{A.19}$$
$$\frac{\partial L}{\partial \beta_k \beta_m} \sum_i \sum_j \left[\frac{\partial L}{\partial \eta_i \partial \eta_j} X_{ik} X_{jm} \right] \tag{A.20}$$

$$LPL(\eta) = \sum_i \int \left(w_i \eta_i(s) - \log \left[\sum_j Y_j(t) w_j e^{\eta_j(s)} \right] \right) dN_i(s) \tag{A.21}$$
$$\frac{\partial LPL}{\partial \eta_i} = \int \left[w_i dN_i(t) - \frac{Y_i(t) w_i e^{\eta_i(t)}}{\sum_j Y_j(t) w_j e^{\eta_j(t)}} dN(t) \right]$$
$$= w_i m_i \tag{A.22}$$
$$\frac{\partial LPL}{\partial \eta_i \partial \eta_j} = \frac{\partial w_i m_i}{\partial \eta_j} \tag{A.23}$$
$$= -w_i \int \left[I_{i=j} \frac{Y_i(t) e^{\eta_i(t)}}{\sum_l Y_l(t) w_l e^{\eta_l(t)}} - \frac{Y_i(t) Y_j(t) w_j e^{\eta_i} e^{\eta_j}}{\left(\sum_l w_l Y_l(t) e^{\eta_l(t)} \right)^2} \right] dN(t) \tag{A.24}$$

Equations (A.19) and (A.22) give the identity $U = wm'X$ where m is the vector of martingale residuals, which can be very handy for compuation. This is paticularly true when X is a sparse matrix, with efficient library routines available to do the sum. The martingale residuals sum to zero.

The martingale residual is an observed- expected quantity, where $e_i = \sum e_i(t)$, the right hand side has one term for each event in the data. More formally

$$e_i = \int \frac{Y_i(t) e^{\eta_i}}{\sum_j Y_j(t) e^{\eta_j}} dN(t)$$

which is a discrete sum since $N(t)$ is a counting process. Each of summands is the prior probability, under the model, that observation i will contribute the next event.

Equations (A.20) and (A.24) collapse to $H = X'WMWX$, where M is the n by n matrix of second derivatives defined by (A.24) and W is the diagonal matrix of weights. The matrix M has row and column sums of 0. This is not an efficient compuation, however, even when X is sparse. Another way to view M is as a sum with one term per event, each term a multinomial variance matrice with diagonal elements $e_i(t)(1 - p_i(t))$ and off-diagonal $-e_i(t)e_j(t)$.

A grouped sum of the martingale residuals is the log-rank statistic, which can be written as a vector with element $O_g - E_g$ for each group, and variance

matrix equal to the collapsed version of M (sum up within blocks). An approximation to the log-rank test is $\sum_g (O_g - E_g)^2/E_g$, a parallel to the classic chisquare statistic of a 2x2 table. Using the diagonal of M for the denominator refines this slightly; note that if risk sets are large $\text{diag}(M) \approx e$, but smaller than e for highly stratified Cox models. The coordinatewise descent algorithm found in glmnet uses the martingal residual m directly to update the linear predictor, with the diagonal of A used as a scaling factor for the update.

A.3.2 Tied event times

The theory underlying the Cox model is derived for the case of continuous t, but in real data there often are tied event times. There are four primary approaches to handle ties. The simplest is simply to ignore them and continue to use the formulas just above. The upshot is that if there are d events at some time t, each of these d subjects is considered to be at risk for all d of the events. This is known as the Breslow estimate of $\hat{\beta}$; the corresponding estimate of cumulative hazard $\hat{\Lambda}$ is called the Breslow estimate of the baseline hazard.

When dealing with a terminal event such as death the above calculation is flawed by the fact that subjects must by definition leave the risk set when the event occurs. That is, if we assume the true data is continuous then there is *some* true order in which the deaths occurred; whomever died first cannot have been at risk for the later events. There are four common approaches to resolve the issue, of which the first is to ignore it, i.e., the Breslow estimate. The Efron approximation is based on the same coarsened data logic as the Fleming-Harrington estimate of the cumulative hazard: if there are d tied deaths, imagine d separate times separated by a small increment ϵ; subject 1 has $1/d$ chance of being the first death, $1/d$ of being the second death, ..., and similarly for the other ties. Mechanically, one can create d distinct synthetic times clustered at the tied event time, and divide each of the tied events into d synthetic subjects with weights w_i/d, one portion of the first subject perishes at the first synthetic time, one at the second synthetic time, etc. Consequently all d of the synthetic clones of a subject are at risk at the first event time, $d-1$ of them will be at risk at the second giving a total weight of $w_i(d-1)/d$, ... Computationally, divide the LPL denominator sum into two portions where $s_2(t)$ is the sum over the tied events and $s_1(t)$ the sum over the remaining subjects at risk, then the first and second terms of the increment to the LPL at time t are

$$\text{first term} = \sum_i w_i dN_i(t)(\eta_i - c)$$

$$\text{second term} = \sum_{k=1}^{d} -\log[n_1(t) + (k/d)n_2(t)]$$

The first term is unchanged from the Breslow approach while the values in the the second term are decreased.

The estimate of the cumulative hazard will also be changed, essentially becoming a Fleming-Harrington increment rather than a Breslow increment at each time. Although Efron [31] did not discuss estimation of the cumulative hazard, that paper being focused on properties of $\hat{\beta}$, for consistency we refer to this as the Efron estimate of the hazard function. Using this approach the martingale identities (A.17) and (A.18) still hold. Though it is more work to create the computer code for the Efron approach than it is for the Breslow approximation, the compute time for the Efron and Breslow approaches is essentially identical.

Some software packages use an Efron approach to compute the likelihood and $\hat{\beta}$, but then switch to the Breslow estimate of the baseline hazard when computing the residual, e.g. SAS. For this hybrid estimate equations (A.17) and (A.18) will no longer hold; this also impacts the robust variance estimate. The hybrid residuals still sum to zero.

A third approach to ties is to compute the exact partial likelihood (EPL), as found in Cox's original paper. This approach views the time scale as discrete. The first term in the LPL is unchanged from before, but if there are d events out of n subjects at risk the second term is now a sum over all $\binom{n}{d}$ ways of choosing a subset of size d. For even modest values of n and d this can be a *very* substantial computation. A clever nested algorithm [38] speeds this up considerably, but it is not clear that the EPL is actually worth the all the effort. Many software packages implement this approximation, but none (that the authors are aware of) go through the extra work of computing a matching hazard estimate, martingale residual, or robust variance. In practice the solution using an Efron approximation is often very close to the EPL estimate.

A fourth approach due to Prentice takes advantage of an algebraic identity, namely that the EPL summation for a given term of the partial likelihood is precisely the analytical solution to a particular integral after d nested applications of an integration by parts formula. The code can then evaluate said integral using numerical integration. This last is implemented in only a few packages.

For multi-state models the argument is a bit more complex. Consider the case of 2 tied events.

- If both involve the same transition, then all the above considerations hold as is.
- If both involve the same starting state, say the transitions were A:B and A:C, then one can carry forward the prior arguments. Assume that the transitions were A:B for subject 1 and A:C for subject 2, then the risk set for the A:C transition might have had both subjects (A:C came first) or only one of them while A:B has the opposite. However, implimenting this in computer code will be significantly more difficult since the feature mentioned in section 11 will no longer hold. That is the code can no longer treat each transition as a disjoint computation.

	Few ties		Moderate ties	
	coef	se	coef	se
Breslow	0.48	0.11	0.46	0.11
Efron	0.48	0.11	0.48	0.11
Exact	0.48	0.11	0.5	0.12

	Heavy ties	
	coef	se
Breslow	0.41	0.11
Efron	0.46	0.11
Exact	0.53	0.13

Table A.1 *Approximations for ties using the lung data set from the survival package, with ph.ecog as the single covariate. Few ties: original data; there are 115 unique death times, 22 with 2 deaths, and 2 with 3 tied deaths. Moderate ties: replace time with floor(time/30); there are 1–15 deaths at each unique death time. Heavy ties: replace time with floor(time/100); there are only 10 unique death times.*

- If there are multiple starting states for the tied events, A:B and B:C, say, then the question is harder. Even if time were known exactly, and the A:B transition happened first, would that subject actually be at risk for a near immediate transtion to state C? In human subject research at least, this seems improbable.

The authors' opinion at the current time is that

1. When ties are infrequent the approximation for ties has no impact of consequence. Use whichever computation is convenient, e.g., the default for whatever software you favor.
2. When there are a moderate number of ties or if one wants to be more "pure" about the underlying assumption of continuous time, then use a formal correction. Even then, the numeric difference will often be slight, i.e., less than one fifth the standard error of the estimate. Table A.1 below gives a concrete example.
3. When time is actually discreet and there are a large number of ties a good argument can be made for the exact partial likelihood (EPL). However, the calculation quickly becomes intractable, and the Efron approximation is normally sufficiently close in value.
4. Once an estimate is chosen, be consistent. An argument that a hybrid algorithm which mixes an Efron or EPL partial likelihood along with a Breslow hazard is "numerically close enough" immediately begs the question of why one would not use the simpler Breslow approach throughout.
5. The entire discussion becomes much more complex for a multistate model, and it is no longer as clear what is actually accomplished by each of the approximations. In this case simplicity may be the best guide, i.e., use the Breslow estimate.
6. Our overall advice is not to worry about ties. Though at one time this topic generated substantial thought and interest, it simply isn't very important.

A.3.3 Absolute risk

Predicted survival curves $S(t)$ from a Cox model, or equivalently probability in state estimates $p(t)$ from a multistate model, are most often based on an analog of the exponential estimate (A.26) rather than the direct estimate (A.25). Predicted curves are for a hypothetical subject with chosen covariates z. Let $A(s;z)$ have off diagonal elements $\hat{\lambda}(s;z)$, and then fill in the diagonal element such that row sums of A are zero. This is essentially the $\hat{\lambda}$ used earlier, but with centering constant $c = z\beta$. The direct and exponential estimates of the probability in state are

$$\hat{\lambda}(t;z) = \frac{\sum_i dN_i(t)}{\sum_i Y_i(t) e^{(X_i - z)\beta}}$$

$$\hat{p}(t;z) = p(0) \prod_{s \le t} (I + A(s;z)) \tag{A.25}$$

$$\hat{p}(t;z) = p(0) \prod_{s \le t} e^{A(s;z)} \tag{A.26}$$

(A.27)

Unless the number of subjects at risk is small ($< 10 - -20$) the two estimates will often be nearly equivalent, numerically. One issue with the direct estimate occurs for predictions of high risk subjects: the diagonal of $I + A$ may be negative, which is an invalid transition matrix (more than 100% of the subjects are predicted to make a transition). The exponential form avoids this error. The issue normally occurs only when the number at risk is very small (often ≤ 5). Most often this at the far right tail of the curve, a point where the standard deviation of the estimate is large. Because of the large se a precise answer for further time points may not matter and some software will truncate the estimate at this point as a way to go forward.

For a simple alive-dead model the equations simplify to

$$\hat{\Lambda}(t;z) = \int_0^t \hat{\lambda}(s;z) ds \tag{A.28}$$

$$= e^{z\beta} \int_0^t \frac{\sum_i dN_i(s)}{\sum_i Y_i(s) e^{X_i \beta}} \tag{A.29}$$

$$S(t;z) = \exp(-\hat{\Lambda}(t;z)) \tag{A.30}$$

$$S(t;z) = \prod_{s \le z} \left(1 - \hat{\lambda}(s;z)\right) \tag{A.31}$$

Equation (A.29) often appears in print, with the right hand term (for $z = 0$) labeled as the baseline hazard, but in computation it is wiser to use (A.28). The exponential estimate (A.30) is known as the Breslow estimate, and is uniformly used. The direct estimate A.31 can generate a negative survival estimate for z values corresponding to a high risk subject ($\hat{\lambda}(t;z) > 1$). An

alternate estimate due to Kalbfleisch and Prentice avoids this issue but is only applicable to the simple survival case. It has the form

$$\begin{aligned} \hat{S}(t;z) &= \prod_{s \le t} \alpha_s \\ \sum \frac{dN_i(s) r_i}{1 - \alpha_s^{r_i}} &= \sum Y_i(s) r_i \\ r_i &= e^{(X_i - z)\beta} \end{aligned}$$

where the individual terms $\alpha(s)$ satisfy the second equation. If $\beta = 0$ it agrees with the Kaplan-Meier.

The fact that the exponential form is the most prominent for single state Cox models is at first surprising, since the exponential form is almost never used for non-parametric curves. This is, we think, more due to the fact that it was easier to implement in code than any statistical argument. For multistate curves, we will see below that the variance is somewhat harder to compute for the exponential form, which has led to a higher prevalence of the direct estimate in early multistate software.

The standard variance for the cumulative hazard for a subject with covariate vector z has two terms $A + B$. The first is an analog of the Nelson-Aalen variance and the second accounts for the variance in $\hat{\beta}$.

$$\begin{aligned} A(t) &= r_z^2 \int_0^t \frac{\sum w_i dN_i(s)}{\left(\sum_i w_i r_i(s) Y_i(s)\right)^2} \\ B(t) &= d(t)' \mathcal{I}^{-1} d(t) \\ d(t) &= \frac{\partial \hat{\Lambda}_z(t)}{\partial \beta} \\ &= r_z \int_0^t (\overline{x}(s) - z) d\hat{\Lambda}(s) \end{aligned}$$

The integral in the last line can be recognized as -1 times the score residual process for an observation with covariate z, which measures the effect of each subject on the first derivative of $\hat{\beta}$. The risk score $r_z = \exp(z'\beta - c)$ is centered, while $\overline{x}(s) - z$ is independent of centering.

A.3.4 Matrix exponential and multistate models

Absolute risk estimates require the matrix exponential of the transition matrix to be computed, sometimes hundreds of times. For matrices A and B, $\exp(A)\exp(B) \neq \exp(A+B)$, so formula (A.26) does not have a shortcut, i.e., you cannot use the matrix exponential of the cumulative hazard. The matrix exponential is formally defined as

$$\exp(A) = I + \sum_{j=1}^{\infty} A^i / i! \tag{A.32}$$

The computation is nicely solved by the expm package in R *if* we didn't need derivatives and/or high speed. We want both.

We can break this down into 4 main cases.

1. When there is only one event at a particular time point, say from state j to state k, then A is zero for all but the jj and jk elements, $\exp(A)$ will equal the identity matrix for all but element jj, which will be $\exp(-\hat{\lambda}_{ij}(t))$, and element jk which will be 1 minus this. That is, the computation is no different than a simple exponential, and the derivative is likewise simple. For many data sets this case will hold for almost all time points.
2. If there are multiple events at the time point, but all all share the same initial state then the formula is likewise simple, and is shown below. This will be the case for competing risk models, for instance.
3. If $A = BDB^{-1}$ where D is diagonal, then $\exp(A) = B\exp(D)B^{-1}$, the exponential of a diagonal matrix is simply a diagonal matrix of the elementwise exponentials. The derivative also has a simple form [52]. If a model is acyclic (no loops) it can be arranged so that A is upper triangular, and the solution found using a generalized cholesky decomposition.
4. In the general case we use a Pade-Laplace algorithm, which is the standard method for the matrix exponential, along with additions to compute the derivative.

The fact that A is a rate matrix, i.e., each row sums to zero and the off diagonal elements are non-negative, implies that $\exp(A)$ will be a transition matrix: all elements are non-negative and each row sums to 1. It also means that many of the edge cases which plague a general matrix exponential algorithm will not arise.

If there is only a single departure state j, then all rows of A except row j will be zero. It is easy to verify that the matrix power A^i also has zeros in all rows but the jth. Equation (A.32) can then be applied element by element to work out the result. In the below $k \neq j$.

$$\begin{aligned}
\left(e^A\right)_{jj} &= e^{A_{jj}} \\
\left(e^A\right)_{jk} &= (e^{A_{jj}} - 1)A_{jk}/A_{jj} \\
\left(e^A\right)_{kk} &= 1 \\
\left(e^A\right)_{kl} &= 0
\end{aligned}$$

As a specific case consider a competing risks model with the states labeled as 0–2 where 0 is the initial state. Only the first row of A will be non-zero, and rows 2–3 of both $(A + I)$ and $\exp(A)$ match an identity matrix. In closed

form, the estimated fraction in state 1 for the direct estimate is

$$S(t) = \prod_{s \le t} 1 - [\hat{\lambda}_{01}(s) + \hat{\lambda}_{02}(s)] \tag{A.33}$$

$$p_1(t) = \int_0^t \hat{\lambda}_{01}(s) S(s-) \tag{A.34}$$

and for the exponential estimate

$$\hat{\Lambda}(t) = \sum_{s \le t} \hat{\lambda}_{01}(s) + \hat{\lambda}_{01}(s) \tag{A.35}$$

$$p_1(t) = \int_0^t \hat{\lambda}_{01}(s) e^{-\hat{\Lambda}(s-)} \left[\frac{e^{A_{00}(s)} - 1}{A_{00}(s)} \right] ds \tag{A.36}$$

Both approaches lead to consistent estimates of $p(t)$, that is, $0 \le p_k(t) \le 1$ and $\sum_k p_k(t) = 1$ for all timepoints t. A commonly quoted formula for the cumulative incidence is, interestingly, not consistent, i.e., equation (A.34) using $\exp(-\hat{\Lambda})$ rather than S.

A.4 Robust variance

The survival package allows for a robust variance for most of the estimates; for multistate models or data sets where a subject can have more than one event the robust variance is essential. Computations are based on the infinitesimal jackknife (IJ) estimate [32]. Let $\hat{\beta}$ be an estimate, which could be the coefficient of a regression model, the value of a survival curve, etc., and let w_i be a case weight or sampling weight for each subject. The leverage matrix D is then defined to have elements

$$D_{ij} = \left. \frac{\partial \hat{\beta}_j}{\partial w_i} \right|_w$$

One interesting property of D is that column sums are 0, which can act as a useful check on computations.

The IJ variance matrix is then

$$V_{IJ} = D' W^2 D$$

where W is a diagonal matrix of observation weights; most commonly the weights are 1. The simple grouped jackknife replaces the central W^2 term with $WBB'W$ where B is an n by g 0/1 grouping matrix. B is essentially the design matrix for a linear model which had the grouping variable as a single prediction factor. More complex sampling designs can be realized with other replacements for B [14] but this topic is outside of our scope. (Many of the package routines will return the weighted influence matrix WD, however, so users can wrap their own.)

The IJ estimate is familiar in multiple statistical contexts, under multiple

names. In a generalized estimating equations models the simple grouping matrix B leads to the *working independence* estimate of variance. It also arises in survey sampling as the Horvitz-Thompson estimate. For a linear regression model, the infinitesimal jackknife approach leads to the estimate

$$D'D = (X'X)^{-1}X'RX(X'X)^{-1}$$

of White [113, 114], where R is a diagonal matrix containing the squared residuals. White recommends its use when the data are heteroscedastic. If one believed the data to be homoscedastic, then a natural step would be to replace R with $\hat{\sigma}^2 I$, the "average" squared residual times an identity matrix. The estimator then collapses down to the usual linear model variance estimate. We can rewrite $D'D$ in the sandwich form as well,

$$\hat{\sigma}^2(X'X)^{-1}\ [\hat{\sigma}^{-2}X'RX\hat{\sigma}^{-2}]\ \hat{\sigma}^2(X'X)^{-1},$$

a nonparametric variance estimator sandwiched between two copies of the usual variance matrix $\hat{\sigma}^2(X'X)^{-1}$. For more general models this is called the Huber-White estimate.

A.4.1 Nelson-Aalen and Kaplan-Meier estimates

$$\begin{aligned}
\hat{\Lambda}(t) &= \int_0^t \frac{\sum w_i dN_i(s)}{\sum_i w_i Y_i(s)} \\
\mathrm{var}(\hat{\Lambda}(t)) &= \int_0^t \frac{\sum w_i dN_i(s)}{\left(\sum_i w_i Y_i(s)\right)^2} && \text{(A.37)} \\
\frac{\partial \hat{\Lambda}(t)}{\partial w_j} &= \int_0^t \frac{dN_j(s) - Y_j(s)/\sum_i w_i Y_i(s)}{\sum_i w_i Y_i(s)} && \text{(A.38)} \\
&= dM_j(s)/\sum_i w_i Y_i(s) && \text{(A.39)} \\
\widehat{S}(t) &= \prod_{s\le t}\left(1-\hat{\lambda}(s)\right) && \text{(A.40)} \\
\mathrm{var}(\widehat{S}(t)) &= \widehat{S}^2(t)\int_0^t \frac{\sum w_i dN_i(s)}{\sum_i w_i Y_i(s)\left[\sum_i w_i(Y_i(s)-dN_i(s))\right]} && \text{(A.41)} \\
\frac{\partial \widehat{S}(t)}{\partial w_j} &= \frac{\partial \widehat{S}(t-)}{\partial w_j}[1-\hat{\lambda}(t)] - \widehat{S}(t)\frac{\partial \hat{\lambda}(t)}{\partial w_j} && \text{(A.42)}
\end{aligned}$$

The usual variance estimates for the Nelson-Aalen estimate of the cumulative hazard and for the Kaplan-Meier are given in (A.37) and (A.41). Robust variance estimates based on (A.39) and (A.42) will be appropriate when there are observation weights or grouping structure. The full set of influence estimates J would have a row for each subject and a column for each event time,

and $J'WJ$ would then be an estimate of the full variance-covariance matrix of all the time points. This is impractical and unnecessary as we normally only want variances for each time point separately. For this is suffices to retain a vector of per-observation influence values, at the current time, updating as we move forward in time.

The Fleming-Harrington estimate of hazard and its derivative are below. In this case the first derivative does not admit of the final simplification of equation (A.39) above.

$$\begin{aligned}
\hat{\lambda}(t) &= \frac{n_2(t)}{d} \sum_{m=1}^{d} 1/(n_1 + (m/d)n_2) \\
\frac{\partial \hat{\Lambda}(t)}{\partial w_i} &= \int_0^t dN_i(s) d\hat{\Lambda}(s) - Y_i(s) \sum_{m=1}^{d} \frac{(m/d)I(i \in n_2) + I(i \in n_1)}{(n_1 + (m/d)n_2)^2} \\
&\neq \int_0^t \frac{dM_i(s)}{\sum_i w_i Y_i(s)}
\end{aligned}$$

A.4.2 Aalen-Johansen

The multistate analog to the Nelson-Aalen hazard estimator is no different than the the cumulative hazard for a single state: each transition from a state j to a state k is computed separately, independent of the other transitions. The derivatives of the multistate hazard and cumulative hazard thus follow directly. The absolute risk estimate $p(t)$ involves matrix products so is a bit more work.

Assume s states and let $p(t)$ be a vector of length s with $p_j(t) =$ probability that the process is in state j at time t. Then for the direct estimate

$$\begin{aligned}
p(t) &= p(0) \prod_{s \le t} (A(s) + I) \\
\frac{\partial p(t)}{\partial w} &= U(t) \qquad \text{(A.43)} \\
&= U(t-)(A(t) + I) + p(t-)\frac{\partial A(t)}{\partial w} \qquad \text{(A.44)}
\end{aligned}$$

U is a matrix with one row per observation and s columns.

The jk element of $A(t)$ is the hazard increment found in equation (A.1) above, for the j to k transition, the derivative will be that found in (A.39). That is

$$\frac{\partial A_{jk}(t)}{\partial w_i} = Y_{ij}(t) \frac{dN_{ijk}(t) - A_{jk}(t)}{\sum_i Y_{ij}(t) r_i}$$

Any observation which is at risk contributes to only one row of the matrix derivative, the row corresponding to their starting state j, and $p_j(t-)$ times that row is added to that row of U. This "select my row" operation does not

have a compact matrix formula, but is very simple computationally. Since each row of A must sum to 0, each row of the derivative sums to 0.

For the exponential form we have similarly

$$p(t) = p(t-)e^{A(t)}$$
$$U(t) = \frac{\partial p(t)}{w_i} = U(t-)e^{A(t)} + p(t-)\frac{\partial e^{A(t)}}{\partial w_i}$$

Derivatives of matrix exponential are a bit more involved, but follow directly from the definition:

$$\exp(A) = I + \sum_{i=1}^{\infty} A^i/i!$$

If we had a matrix B containing element-wise derivatives of A, then the derivative of the $A^3/3!$ term above would be $(BAA+ABA+AAB)/3!$. (The routines for the matrix exponential use more efficient algorithms for both the exp and its derivatives. They accept A and one or more B matrices, returning the exp and derivative matrices.) Each derivative is a significant computation, and we would like to avoid separate computations of ∂w_i for all n observations in a study. Since A is a transition matrix each element of $\exp(A)$ is ≥ 0, each row sums to 1, and if there are no transitions from state j to k then the jk element is 0. With these and other considerations, no more than $d(d-1)$ different B matrices are ever needed, where d is the number of events at that time.

A.4.3 Cox model

Coefficients: The derivative of the score statistic and $\hat{\beta}$ with respect to case weights is

$$\frac{\partial U}{\partial w_i} = \int_0^{\infty} (x_i - \overline{x}(s))dM_i(s) \tag{A.45}$$
$$D = \frac{\partial \hat{\beta}}{\partial w_i}$$
$$= \frac{\partial(\mathcal{I}^{-1}U)}{\partial w_i}$$
$$= \mathcal{I}^{-1}\frac{\partial U}{\partial w_i} + \frac{\partial \mathcal{I}^{-1}}{\partial w_i}U \tag{A.46}$$
$$\approx \mathcal{I}^{-1}\frac{\partial U}{\partial w_i} \tag{A.47}$$

The right hand side of (A.45) is called the score residual. The inverse of the information matrix, $\mathcal{I}^{-1}$, is the usual asymptotic variance estimate of $\hat{\beta}$. The n by p matrix of derivatives of $\hat{\beta}$ is called the *dfbeta* matrix D, with ith row defined by (A.47). A robust variance for $\hat{\beta}$ is $D'W^2D$. The second term of (A.46) turns out to be small; it is a major nuisance to compute and so is universally ignored.

Cumulative hazard for a Cox model: The predicted hazard and/or survival is always for a hypothetical subject; let z be the covariates for that subject. Recenter the risk scores $\eta_{iz}(t) = (X_{i.} - z)\beta$ at z, leading to the simple cumulative hazard formula

$$\hat{\lambda}(t;z) = \frac{\sum_i w_i dN_i(t)}{\sum_i w_i Y_i(t) \exp(\eta_{iz})}$$
$$\hat{\Lambda}(t;z) = \int_0^t \lambda(s;z)$$

The derivative is then

$$d(t;z) \equiv \sum_{i=1}^{n} Y_i(t) e^{\eta_{iz}} \tag{A.48}$$

$$\frac{\partial \hat{\lambda}(t;z))}{\partial w_k} = \left(\frac{dN_k(t)}{d(t;z)} - \frac{Y_k(t) e^{\eta_{kz}} \hat{\lambda}(t;z)}{d(t;z)} \right) - (\hat{\lambda}(t;z)/d(t;z)) \sum_{p=1}^{m} \frac{\partial d(t;z)}{\partial \beta_p} \frac{\partial \beta_p}{\partial w_k} \tag{A.49}$$

$$(1/d(t;z)) \frac{d(t;z)}{\partial \beta_p} = \frac{\sum_{i=1}^{n} Y_i(t)(X_{ip} - z_p) e^{\eta_{iz}}}{\sum_{i=1}^{n} Y_i(t) e^{\eta_{iz}}} = \overline{x}_p(t) - z_p \tag{A.50}$$

where m is the length of the coefficient vector $\hat{\beta}$. Gathering these together, and noting that $d\beta_p/dw_i$ is the k,p element of the dfbeta matrix D in equation (A.47), we have

$$\frac{\partial \hat{\lambda}(t;z))}{\partial w_k} = \frac{dM_k(t)}{d(t;z)} - \hat{\lambda}(t;z) D_{k.} (\overline{x}(t) - z)'$$

where $D_{k.}$ is the kth row of the dfbeta matrix.

The first term is a direct analog of equation (A.39), and the second accounts for the impact of each observation on $\hat{\beta}$. Notice that M, D and $\overline{x}$ do not involve z.

For a multistate model the estimates of the hazard for each transition are disjoint, so the above derivation still holds, one transition at a time. The hazard, martingale process M and dfbeta matrix D will be hazard specific, while $\overline{x}$ and $d(t)$ are based on all those at risk in the starting state of the transition (foot of the arrow).

The estimated probability in state $p(t)$ is a transformation of the hazard of the same form as the Aalen-Johansen, so the same formulas hold, albeit

with matrix exponentials. That is

$$p(t;z) = p(t-;z)e^{A(t;z)}$$
$$\frac{\partial p(t)}{\partial w_k} = \frac{\partial p(t-)}{\partial w_k}e^{A(t)} + p(t-)\frac{\partial e^{A(t)}}{\partial w_k}$$

with $A_{jk}(t) = \lambda_{jk}(t)$ for $j \neq k$.

A time-dependent covariate x carries through in the above derivation: at each jump in $\hat{\Lambda}$ there is a new covariate value driving the risk. Computer code is more difficult though. It is also not clear what such a curve would *mean*, so a restriction to time fixed covariates is no great loss.

A.5 The IJ estimate of variance for the Kaplan-Meier

An interesting fact is that for the Kaplan-Meier, the variance based on the infinitesimal jackknife (IJ) is identical to the usual variance based on the Greenwood estimate. To make the notation in this section more compact, let $n(t) = \sum Y_i(t)w_i$ be the weighted number of subjects at risk at time t, $e(t) = \sum w_i dN_i(t)$ be the number of events at time t, and $d_1, d_2, \ldots$ be unique event times. The Greenwood and IJ variance estimates are

$$\begin{aligned} V_G(t) &= S^2(t) \sum_{d_j \le t} \frac{e(d_j)}{n(d_j)[n(d_j) - e(d_j)]} \\ &= S^2(t)g(t) \\ V_{IJ}(t) &= \sum_{i=1}^{n} u_i^2(t) \\ &= \sum_{i=1}^{n} \left(\frac{\partial S(t)}{\partial w_i}\right)^2 \end{aligned}$$

This equivalence is a surprise since formulas look nothing at all alike, but it was noticed in data examples that the two were identical. The following proof by induction is due to to Anne Eaton [30].

First:

$$\frac{\partial[1 - e(t)/n(t)]}{\partial w_i} = \frac{Y_i(t)e(t)}{n^2(t)} - \frac{dN_i(t)}{n(t)}$$

$$\begin{aligned} \sum_i w_i \left(\frac{\partial[1 - e(t)/n(t)]}{\partial w_i}\right)^2 &= \sum_i w_i Y_i(t) e^2(t)/n^4(t) + \sum_i w_i dN_i(t)/n^2(t) \\ &\quad - 2\sum_i w_i Y_i(t) dN_i(t) e(t)/n^3(t) \qquad \text{(A.51)} \\ &= e^2(t)/n^3(t) + e(t)/n^2(t) - 2e^2(t)/n^3(t) \\ &= \frac{e(t)}{n^2(t)} \frac{n(t) - e(t)}{n(t)} \qquad \text{(A.52)} \end{aligned}$$

At the first event time d_1, $S(d_1) = 1 - e(d_1)/n(d_1)$, and from (A.52)

$$\begin{aligned}\sum_i u_i^2(d_1) &= \frac{e(d_1)}{n^2(d_1)} \frac{n(t) - e(t)}{n(t)} \frac{n(d_1) - e(d_1)}{n(d_1) - e(d_1)} \\ &= \left(\frac{n(d_1) - e(d_1)}{n(d_1)}\right)^2 \frac{e(d_1)}{n(d_1)[n(d_1) - e(d_1)]} \\ &= S^2(d_1) g(d_1) \end{aligned} \tag{A.53}$$

which shows that the theorem is true at the first event time d_1.

Second: Assume that the theorem holds for event times $t < d_j$. Then

$$S(d_j) = S(d_{j-1}) \frac{n(d_j) - e(d_j)}{n(d_j)} \tag{A.54}$$

$$u_i(d_j) = u_i(d_{j-1}) \frac{n(d_j) - e(d_j)}{n(d_j)} + S(d_{j-1}) \frac{\partial [1 - e(d_j)/n(d_j)]}{\partial w_i} \tag{A.55}$$

$$\sum_i u_i^2(d_j) \equiv \sum_i (a_i + b_i)^2 \tag{A.56}$$

$$\begin{aligned}\sum_i a_i^2 &= \sum_i u_i^2(d_{j-1}) \left(\frac{n(d_j) - e(d_j)}{n(d_j)}\right)^2 \\ &= S^2(d_{j-1}) g(d_{j-1}) \left(\frac{n(d_j) - e(d_j)}{n(d_j)}\right)^2 \\ &= S^2(d_j) g(d_{j-1}) \end{aligned} \tag{A.57}$$

$$\sum_i b_i^2 = S^2(d_{j-1}) \left(\frac{n(d_j) - e(d_j)}{n(d_j)}\right)^2 \frac{e(d_j)}{n(d_j) - e(d_j)} \tag{A.58}$$

$$= S^2(d_j) \left(g(d_j) - g(d_{j-1})\right) \tag{A.59}$$

$$\sum_i a_i b_i = 0 \tag{A.60}$$

Equation (A.54) is the definition of the Kaplan-Meier, equation (A.55) follows from the product rule for derivatives, and (A.56) is shorthand to make the remaining lines fit on the page. Step (A.57) follows from the induction hypothesis, and (A.58) is a repeat of the steps for (A.53). Equation (A.60) follows from 3 observations

- $\frac{\partial 1 - e(d_j)/n(d_j)}{\partial w_i} = 0$ for all subjects i who are not at risk at time d_j
- $\sum_i \frac{\partial 1 - e(t)/n(t)}{\partial w_i} = 0$
- All subjects i who are at risk at time d_j share the same value for $u_i(d_{j-1})$.

The last statement reveals that this proof will not extend to a data set with delayed entry.

A.6 IPC weights

Inverse probability of censoring weights appear in several contexts. An important aspect of these, often overlooked, is the handling of tied times. Consider a simple study with death as the endpoint and a particular death time at day 100, say. The analysis data set will contain a death at day 100, and may also contain one or more censored observations at $t = 100$, as well as changes in one or more time-dependent covariates at $t = 100$. To sort this out we assume the following order: the deaths, then any censoring, then any covariate changes. The first of these is a reflection of the observation process itself: "censored on day 100" is the coding used for a patient who was last observed alive on day 100, i.e., their death time, whatever it is, must be > 100. Thus in the code censors happen after deaths, something that is baked in to every Kaplan-Meier or Cox model routine. The second condition, that covariate changes happen after everything else, is a consequence of the underlying mathematics, namely that the models are valid only if the covariates form what is known as a *predictable process*. In gambling parlance, you must place your bet before the roulette wheel is spun.

A.6.1 Inverse probability weight

The IPC weight is $1/G(t)$ where G is the censoring distribution. A common way to estimate this is with a Kaplan-Meier, with the 0/1 status variable reversed, and this approach is *almost* correct. The issue comes about with tied time values. Say we have a data set with the following 4 observations for (time, status) of (31, 0), (52,0), (52,1), and (85,1). For computation of the ordinary KM of the event (death say), observations 2,3 and 4 are each considered to have been at risk for the event on day 52: censoring happens after the event. This agrees with how clinical data is gathered: a subject last observed to be alive on day 52 is coded as being censored on day 52; and we know that their death time is > 52. For computation of a Kaplan-Meier estimate $\hat{G}$ of the censoring distribution, however, the death at time 52 is not at risk for being censored at time 52: you cannot be "lost to followup" if you have already died. The risk set at time 31 will have all four obs, the risk set at time 52 should have only 2. One way to cause this to occur is to subtract a small value ϵ from each of the event times (but only those) before computing the Kaplan-Meier. An ϵ value of 1/2 the smallest difference between unique event times is a good default.

Asking whether this fix makes any practical difference is a fair question, and the answer is that normally the effect is small to negligible, and the naive reversed KM will be just fine. We are less charitable to those who write general software to be used by others; then the computation should be done correctly, for professional pride if no other reason.

A.6.2 Redistribute to the right

In the redistribute to the right algorithm the above ordering is integral: when weights are redistributed, a subject who has an event at the concurrent time is not included in the recipient list. An alternate algorithm for the RTR process, and the one that is much more commonly used, is to give, at time t, each observation a weight of 0 if censored or the inverse probability of censoring (IPC) weight of $1/G(t)$ if uncensored, where G is a Kaplan-Meier estimate of the censoring distribution.

A common description of G is to "use the KM, with the status variable reversed", but this approach is correct only if there are no time points that share a censoring and an event time. The correct algorithm has two important differences.

1. When computing $G(t)$ ensure that events are addressed before censoring times. One way to do this is to replace all event times t with $t - \epsilon$; then a computation with reversed status will retain the correct behavior.
2. When assigning weights at a given time t, those censored *before* t get a weight of zero, and all others a weight of $1/G(t + \epsilon)$. That is, use a left-continuous form of G rather than the default right-continuous form.

A reasonable value for ϵ is half the minimum spacing between any two unique time values in the data set.

One feature of the RTR algorithm is that the sum of weights is constant over time, they are simply re-distributed. Use of the incorrect algorithm results in a sum the grows each time there is a shared event/censoring time.

Computing the Brier score

Create a "reversed status" survival curve $G(t)$ using a response of $(t-\epsilon, \delta = 0)$ for the events and $(t, 1)$ for the censored subjects. The small increment ϵ causes the correct accounting at time points with both an event and a censoring. At time 0 all subjects start with a weight of $1/n$, then going forward in time, at each event time s, in order,

1. Compute the weighted Brier score at time s
   $B(s) = \sum_{i=1}^{n} w_i(s)(y_i(s) - \hat{y}_i(s))^2$
2. Reset the weights of each subject who is still at risk ($t_i > s$) to $1/nG(s)$, and the weight of any subject censored at s to 0.

Doing the two steps in the proper order leads to left-continuity for G, which is important for the underlying theory. Using this approach the sum of the weights at any given time point will always be 1.

Many programs for the Brier score overlook the issue with ties, leading to a sum of weights that grows with time, just as for the RTR and IPC. The effect of this depends on the number of tied event/censoring values in the data set, which is, thankfully often quite small. List some values?

A.7 Approximating a Cox model

The Cox model can be approximated using Poisson regression, a trick that was well known when Cox models were not yet common in the major software packages [115, 60]. The coefficients and standard errors for the Poisson regression, which uses the number of events for each person as the y variable, are usually quite close to those of the Cox model, which is focused on a censored time value as the response. In fact, if the baseline hazard of the Cox model $\lambda_0(t)$ is assumed to be constant over time, the Cox model is equivalent to Poisson regression.

One non-obvious feature of a Poisson fit is the use of an `offset` term. This is based on a clever "sleight of hand", which has its roots in the fact that a Poisson likelihood is based on the number of events (y), but that we normally want to model not the number but rather the *rate* of events (λ). Then

$$\begin{aligned} E(y_i) &= \lambda_i t_i \\ &= \left(e^{X_i\beta}\right) t_i \\ &= e^{X_i\beta + \log(t_i)} \end{aligned} \tag{A.61}$$

We see that treating the log of time as another covariate, with a known coefficient of 1, correctly transforms from the hazard scale to the total number of events scale. An `offset` in glm models is exactly this, a covariate with a fixed coefficient of 1.

The hazard rate in a Poisson model is traditionally modeled as $\exp(X\beta)$ (i.e. the inverse link $f(\eta) = e^{\eta}$) rather than the linear form $X\beta$, for essentially the same reason that it is modeled that way in the Cox model: it guarantees that the hazard rate is positive.

Example

In order to better understand the underlying hazard rates, start by plotting the cumulative hazard for the dataset, and approximate it with a set of connected line segments. The kidney cancer data, for instance, is moderately well approximated using cutpoints at 45 and 500 days, as shown in Figure A.1.

The Poisson regression approximation can then be obtained by breaking the time scale into intervals based on these cutpoints and fitting a Poisson model with one intercept per interval. We see that the coefficients and standard errors are very similar (Table A.2, row 2).

If each unique death time is made into its own interval the Cox result can be duplicated exactly. One more correction is needed for perfect agreement, which is to toss away any partial contributions. If there were unique death times at 100 and 110 days, for instance, and a subject were censored at time 103, those 3 days are not counted in the in the 100–110 interval. If there is someone with follow-up after the last event, that is removed as well. After removal, everyone in the same interval will have the same number of days at risk, which means that an offset correction is no longer needed.

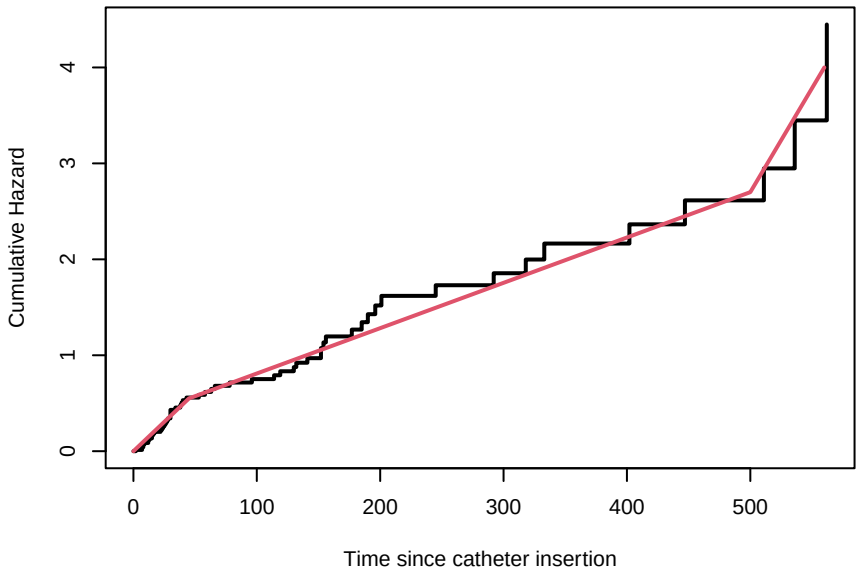

Figure A.1 *Cumulative hazard rates and approximations of the rate in three time periods using the kidney dataset with cutpoints at 45 and 500 days.*

	age	sex
Cox	0.0022 (0.0092)	-0.8210 (0.2987)
poisson1	0.0032 (0.0093)	-0.7512 (0.2943)
poisson2	0.0022 (0.0092)	-0.8210 (0.2987)
binomial	0.0028 (0.0095)	-0.8993 (0.3138)

Table A.2 *Comparison of Cox, poisson and binomial models of the kidney data predicting survival using the covariates age and sex. The poisson1 results are based on breaking follow-up into three intervals (cutpoints at 45 and 500 days) and the poisson2 results use intervals based on each unique death time. The binomial model uses the same data as was used for poisson2.*

The Poisson coefficients now exactly match those of the Cox model (Table A.2, row 3). Almost all of the intervals have only a single event, i.e., both the event count and the event rate are low, which is the case in which the binomial and Poisson distributions closely approximate each other. Consequently, the last model in Table A.2 shows that binomial fits are also effective. This computational trick can be particularly useful in contexts where there is readily available code for binomial outcomes but time-to-event models are lacking, e.g., machine learning.

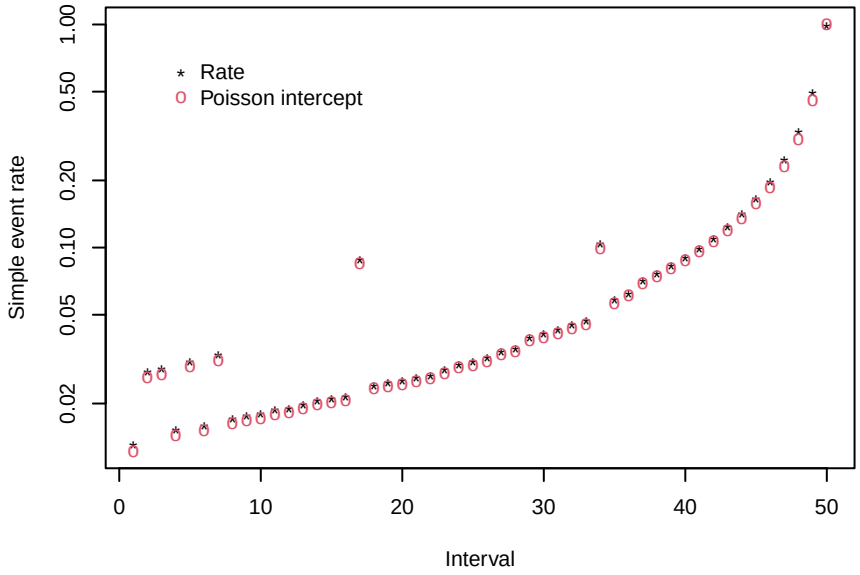

Figure A.2 *Poisson intercepts and event rates separated by intervals where events occurred using the kidney dataset.*

Pre-centering data

We can make the connection between the Cox, poisson, and binomial models easier to exploit by pre-centering the data. The plot shows that when this is done, the intercepts are very close to the simple event rate for each interval of (number of events) / (number of observations). We can add this variable to the model as an offset.

	age	sex
Cox	0.00218 (0.00922)	-0.82100 (0.29872)
Poisson, multiple intercepts	0.00218 (0.00922)	-0.82100 (0.29872)
Poisson, offset interval rate	0.00217 (0.00920)	-0.81845 (0.29757)
Poisson, no offset or intercept	0.00381 (0.00903)	-0.78852 (0.29063)
Binomial, offset interval rate	0.00278 (0.00950)	-0.89619 (0.31210)
Binomial, no offset	0.00403 (0.00928)	-0.83705 (0.30074)

Table A.3 *Comparison of Cox, poisson and binomial models of the kidney data predicting survival using the covariates age and sex. The data has been split into intervals based on event times and the data has been pre-centered. Coefficients and standard errors are shown for the different models.*

The coefficients from the different models, as shown in Table A.3, illustrate that adding an approximate per-interval intercept via the offset term gives a very close approximation to the `coxph` fit with the Poisson and a reasonable

one with the binomial. Using no intercept at all produces some bias, the size of which will be related to the correlation between interval rate `phat` and the covariate in question. The advantage of the offset models is a smaller number of coefficients, particularly for large datasets where the number of intercepts would be excessive.

A.8 Concordance

> "mille vie ducunt hominem per secula Romam" (a thousand roads lead men forever to Rome) Alain de Lille (1175)

For continuous data the concordance C between two variables x and y is defined as $P(y_i > y_j | x_i > x_j)$, i.e., the probability that the two variables share the same ordering for a randomly chosen pair of subjects i and j. For time-to-event data the concordance statistic has also become popular, modified to deal with censored data [45]. What is particularly interesting, however, is that there are close connections between concordance, two sample tests such as the Gehan-Wilcoxon statistic, and the Cox model score statistic. As such, there also turn out to be connections between suggestions for improvements, e.g., Uno's 2011 [109] modification of C is an echo of Schemper's 1992 [91] proposal for modifications of the Cox score statistic, which is itself a continuation of a 1972 Peto [82] modifcation of the Gehan-Wilcoxon test. This set of interconnections is not widely appreciated. (Several of them became clear in programming the survival package, i.e., deja vu moments we we realized that "I have seen this computation before".)

Though of interest, these historical connections are not central to message and purpose of the book, and recounting them in the body of the text would be a distraction. They do, however, inform our suggestions with respect to how the concordance statistic is best used and computed, and so are presented here in the appendix. We have largely omitted proofs as they can be found elsewhere.

A.8.1 Measures

We will focus on the concordance of a measured outcome y and the prediction $\hat{y}$ of that outcome. For continuous data, four common measurements of the concordance are Kendal's τ_a and τ_b, the Goodman-Kruska γ statistic, and Somers' d [1]; they differ only in how ties are handled. The four measures each lie between -1 and 1, similar to R^2.

Let A, D, T_x, T_y and T_{xy} be a count of the pairs that are concordant (or agree), discordant (disagree), tied on $\hat{y}$ (but not y), tied on y (but not $\hat{y}$), and

[1] or D_{xy}, depending on the reference

tied on both. $A + D + T_x + T_y + T_{xy} = n(n-1)/2$, the total number of pairs.

$$\tau_a = \frac{A - D}{n(n-1)/2} \tag{A.62}$$

$$\tau_b = \frac{A - D}{\sqrt{(A + D + T_x)(A + D + T_y)}} \tag{A.63}$$

$$\gamma = \frac{A - D}{A + D} \tag{A.64}$$

$$d = \frac{A - D}{A + D + T_x} \tag{A.65}$$

- Kendall's tau-a (A.62) is the most conservative; ties shrink the value of the statistic towards 0.
- The Goodman-Kruskal γ statistic (A.64) ignores ties in either y or $\hat{y}$.
- Somers' d (A.65) treats ties in y as incomparable and they are removed from the denominator.
- Kendall's tau-b (A.63) is similar to Somers' d, but treats y and $\hat{y}$ symmetrically.

Let $C = P(\hat{y}_i > \hat{y}_j | y_i > y_j)$, the fraction of time that a prediction turns out to be correct; $0 \leq C \leq 1$. Consider the following simple experiment

- Present pairs of subjects to an oracle — a statistical prediction rule, a domain expert such as an M.D., a local seer, whatever — and count the number of times that the oracle correctly predicts the order of the outcome.
- Pairs for which the outcome $y_i = y_j$ are not presented, as they would be uninformative with respect to the oracle's accuracy.
- Pairs for which the oracle cannot decide count as 1/2, e.g. tied values of the prediction rule.
- Define C = fraction correct.

For this approach of handling ties, it turns out that $C = (d+1)/2$, a rescaled version of Somer's d.

This definition has two other equivalencies:

- If y is a 0/1 variable, then C = the area under the receiver operating curve (AUROC). (This is one of those identities that looks like it would have a 2 line proof, but takes a bit more work than that. Nevertheless it is well known.)
- When y is a survival time, then pairs for which the ordering of y_i and y_j cannot be ascertained are not used; this leads to Harrell's C_H.

For survival data, pairs of observations y that cannot be unabiguously ordered are treated as tied values. This would include the pair $(10+, 20)$, the first censored at time 10 and the second with an event at time 20: we do not know if observation 1 will have it's eventual event before or after time 20. The pair $(20+, 20)$ *is* ordered, however, since we know that the first observation's event happens sometime after $t = 20$.

A.8.2 Rank tests and the Cox model

For computation, start by sorting the data set by the observed survival response y. Then the numerator of Somers' d is

$$d = \sum_{i=1}^{n} \sum_{y_j > y_i} \text{sign}(\hat{y}_j - \hat{y}_i) \tag{A.66}$$

$$= \sum_{i=1}^{n} N_i \sum_{j \in R_i} \text{sign}(\hat{y}_j - \hat{y}_i) \tag{A.67}$$

$$= \sum_{i=1}^{n} \int \left[\sum_j Y_j(t) \, \text{sign}(\hat{y}_j(t) - \hat{y}_i(t)) \right] dN_i(t) \tag{A.68}$$

where N_i is the 0/1 indicator of censored/event and R_i is the risk set for the event (if any) for subject i. The transition from (A.66) to (A.67) follows first from our definition: if observation y_i is censored then any observations with a larger time are not comparable; the first sum need only include the events. The second insight is that within a set of tied times, the sum of the sign terms will be zero. Equation (A.68) writes this more carefully using counting process notation.

Antolini et al [7] made use of this identity to extend the C statistic to a model with crossing hazards. In this case one cannot make the simplifying assumption that $\eta_i > \eta_j$ ensures that the predicted survival $S_j(t)$ for subject j will be greater than that for subject i, for any time t. Equation (A.68) only needs the ordering of predictions for i and y at the time of i's event. They apply this to a fit with a time-dependent coefficient.

Therneau and Watson [105] carried this one step further to rewrite (A.68) as the score statistic from a Cox model. This provides a direct connection between the concordance and many standard tests. Let D be the indices of the events (deaths) in the data set. Consider three intersecting facts

A. Let E be the set of data indices for the events (or deaths). For a single covariate x the score statistic for a time-weighted Cox model is

$$\sum_{i \in E} v(t_i)[x_i - \overline{x}(t_i)] \tag{A.69}$$

where t_i is the event time for observation i, $v(t)$ is a fixed time dependent weight function, and $\overline{x}(t)$ is the mean over all those at risk at time t. Normally $v(t) = 1$ for a Cox model. The variable x may also be time dependent.

B. When x is a 0/1 variable, then equation (A.69) is the Gehan-Wilcoxon test statistic when $v(t) = n(t) =$ the number at risk at time t. Other choices include the log-rank with $v(t) = 1$, the Peto-Wilcoxon with $v(t) = S(t)$, and several more.

C. Let $z(t)$ be a time dependent covariate defined as

$$z_i(t) = \frac{\sum_j Y_i(t) Y_j(t) +, \text{sign}(\hat{y}_i - \hat{y}_j)}{\sum_j Y_j(t)} \tag{A.70}$$

where sign is the sign function with value -1, 0, or 1. In words, for each observation i in the risk set at time t, z is the fraction of others whose prediction

is smaller, minus the fraction whose predicted value is larger; $r = (z+1)/2$ is a time dependent rank which lies in (0, 1). Then

$$C_H = \sum_{i \in D} n(t_i)[r_i - \overline{r}(t_i)] \tag{A.71}$$

is Harrell's C statistic. That is, C_H is equivalent to a Cox model score statistic, based on the time depedent *rank* of the prediction. (If there are tied values in $\hat{y}$ it is equivalent to the Cox model score statistic using the Breslow approximation.)

Three immediate consequences are that

1. Asymptotics for C_H now follow immediately from standard counting process arguments, in the same way as other Cox model results.
2. The concordance is well defined for time-dependent covariates.
3. The decades of consideration and debate over the "best" weighting $v(t)$ for a test of survival carry forward directly to the concordance.

The last of these is the most interesting. Peto and Peto [82] for instance noted that $n(t) \approx n(0)S(t)G(t)$, where S and G are the survival functions for death and censoring, respectively. They postulated that using $S(t)$ as the weight would result in a more stable test when the two treatment arms had different censoring patterns. Prentice [83] later showed this superiority for an empirical dataset, and indeed, implementations of the Gehan-Wilcoxon in the major statistics packages most often implement the Peto variant. We might even label equation (A.71) the "Gehan-Wilcoxon concordance" rather than Harrell's C, one with $v(t) = S(t)$ the Peto-Wilcoxon concordance, and etc.

Schemper [90] proposes a weight of $v(t) = S(t)/G(t)$ in the usual Cox model. The argument is that if proportional hazards does not hold, one would like an estimate that is consistent with respect to subject follow-up, that is, if one were to analyze the same study after t years of follow-up, and again after $t + s$ years, the coefficients from the two analyses will estimate the same quantity. This can be accomplished by using a time-dependent weight of $S(t-)/G(t-)$.

Uno et al. [109] essentially reprise the Schemper reasoning, and recommend a weight of $v(t) = n(t)/G^2(t-)$. Note that $n(t)/G^2(t-) \approx n(0)S(t-)/G(t-)$ by Peto's argument. When using estimates $\hat{G}$ and $\hat{S}$ of G and S the Uno and Schemper weights actually turn out to be precisely identical (if G is calculated properly). Uno et al.s make the even stronger claim that the C statistic that will arise with complete follow-up of all the subjects is the "correct" target for estimation (we do not agree), and hence that C_H is 'biased' (we strongly disagree). Since in practice the hazard ratio almost invariably becomes closer to 1 over time in clinical studies, given sufficient follow-up, this means that the usual C_H statistic using $n(t)$ (Gehan-Wilcoxon) weighting will be larger than this asymptotic value.

A.8.3 Variance

The connection to the Cox model provides a natural measure of variance, namely the Cox model's second derivative or Hessian. This applies for any of the weighting choices, with or without time dependent covariates. The Cox model variance is correct under the null hypothesis, and so provides a valid test of concordance = 0.5. However, when C is far from 0.5 the resulting variance is too large, resulting in wider confidence intervals than needed.

An alternative is to compute a robust variance based on the infinitesimal jackknife. Rewriting equation (A.65) and adding case weights w, the numerator of Somers' d is

$$\begin{aligned} A &= \sum_{i \in E} v(t_i) w_i \sum_j w_j Y_j(t_i) I(t_j > t_i) \text{sign}(x_i - x_j) \\ U_k &= \frac{\partial A}{\partial w_k} \\ &= \sum_{i \in E} v(t_i) w_i Y_k(t_i) I(t_k > t_i) \text{sign}(x_i - x_k) \\ &\quad + I(k \in E) v(t_k) \sum_j w_j Y_j(t_k) I(t_j > t_k) \text{sign}(x_k - x_j) \\ \text{var}(A) &= \sum_k U_k^2 \end{aligned}$$

A.8.4 Truncated values

We argue in section **??** that the approriate target of validation will often be a time-limited prediction. You might for instance compute the RMST over 2 years for "standard" treatment based on a fitted model for that treatment, using the result to decide on use of an experimental therapy. In that situation, the ability of the model to discriminate between a 3 year and a 5 year survival is immaterial, and it natural to compute the concordance on truncated data. That is $C(\min(y, 2), \hat{y})$, where all survival time > 2 are replaced by the censored value 2+ before doing the compuatation. Since the compuation only sums over pairs (i, j) where one survival time is known to be larger, this simple strategm effectively removes said 3 vs 5 year comparison from the calculation. In R, the concordance function has an optional `ymax` argument to make this particularly easy. Variance of the statistic carries through as before.

A more severe constraint is to dichomize the response as (survival ≤ 2 years) versus (survival > 2 years). Since the response is now a 0/1 variable, without censoring, C is equal to the well known area under the receiver operating curve (AUROC). Our major problem with this comes from a career-long battle against the disease of "dichotomania"; the nearly relentless of our medical colleages to transform everything into black vs white. (We've sometimes joked about a new automobile made especially for researchers, with a single red light in place of the speedometer). Time dichotomy has been promoted

under the label "time dependent AUROC", starting with a paper by Heagerty and Zheng [46] and with a large subsequent literature. This is a brilliant label, making dichotomy appear to be a virtue rather than a vice. (And feels like an own goal by the statistics team).

In its favor, the notion of "predicted 2 year survival" is easy to communicate to our audience, certainly easier than C or truncated C. As with all dichotomization there is a price to pay in terms of efficiency, and the dichotomized C has larger variation.

A primary technical issue is the "without censoring" qualifier above. What is to be done with the subject censored at 1.5 years, when creating the 0/1 variable for 2 year survival? One solution is to apply the RTTR algorithm to all those censored before time 2, redistributiong their case weight, then compute a weighted concordance. Figure **??** shows the truncated and dichomized C statistic at multiple time points for a model fit to the Rotterdam data, assessed on the GBSG data set as validation, along with 1 SE error bars. For the truncated data both the Harrell and Uno weightings are shown, the dichomized statistic has dealt with censoring in another way. [This figure is in the validation chapter.]

A.8.5 Synthetic measures

Göen and Heller [41] take a different approach to the censoring issue. They note that if the proportional hazards model holds, then given two subjects with risk scores η_i and η_j, the probability that $t_i > t_j$ is $\exp(\eta_j - \eta_i)$. Looking ahead from time 0, one can compute the expected concordance of the model $E(C)$ directly using only the covariate matrix X and the estimated coefficient $\hat{\beta}$. This completely sidesteps censoring.

Sort the data in order of the estimated risk score $\hat{\eta}$, then the estimate is

$$C_G = \sum_{i>j} \frac{1}{1 + e * \eta_j - \eta_i} / [n(n-1)/2] \tag{A.72}$$

The significant downside to this approach is that we get an estimate of how well the PH model *would* predict, in some future dataset that has complete follow-up and exactly follows the model, i.e., proportional hazards over all time, perfect linearity wrt covariates, and no outliers. This is, however, not an evaluation of how well the model *actually* works for the data set at hand.

A.9 Dates and roundoff error

The raw data for studies often has dates of entry and exit from a study, which means that the natural unit of time for follow-up is days. Or we should say that is the form which is natural to the computer, since software packages like R, SAS, and STATA can all subtract dates directly. Users will often be more comfortable working in years, however, either age or years since entry

for instance. It is not widely appreciated that this can lead to some subtle round off errors.

As an example, consider two subjects, both born on 1973-3-10, the first enrolled on 1998-09-13 and followed until 12-04, the second enrolled on 1998-09-17 and followed until 12-08; both have 81 days of followup. However, if we compute age at enrollment as (entry date - birth date)/365.25, and similarly for age at last follow-up, the simple difference (follow-up age - entry age) will *not* be exactly the same for the two subjects. This is due to the fact that 365.25 does not have an exact representation in binary,

For parametric models this tiny error has no impact whatsoever, but for the Kaplan-Meier and Cox model exact ties are treated differently, since they determine which observations are or are not in a given risk set. Users can be surprised when a change from days to years leads to a different result. (Some software is aware of this and tries to avoid the problem, most is not.)

Bibliography

[1] M. Abrahamowicz and T. A. MacKenzie. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Stat. in Medicine*, 26:392–408, 2007.

[2] A. M. Allen, T. M. Therneau, J. J. Larson, A. Coward, V. K. Somers, and P. S. Kamath. Nonalcoholic fatty liver disease incidence and impact on metabolic burden and death: A 20 year community study. *Hepatology*, 67:1726–1736, 2018.

[3] P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer-Verlag, New York, 1993.

[4] P. K. Andersen and K. Liestol. Attenuation caused by infrequently updated covariates in survival analysis. *Biometrics*, 4:633–649, 2003.

[5] P. K. Andersen and M. Pohar Perme. Pseudo-observations in survival analysis. *Stat Methods Medical Research*, 19:71–99, 2010.

[6] J. R. Anderson, K. C. Cain, and R.D. Gelber. Analysis of survival by tumor response. *J Clinical Oncology*, 1:710–719, 1983.

[7] Laura Antolini, Patriza Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Stat. in Medicine*, 24:3927–3944, 2005.

[8] P. C. Austin. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*, 33:3083–3107, 2009.

[9] P. C. Austin. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experimens. *Stat Med*, 28:1242–1258, 2009.

[10] P. C. Austin, F. E. Harrell, and D. vanKlaveren. Graphical calibration curves and the integrated calibration index for survival models. *Stat Med*, 39, 2020.

[11] P. C. Austin, E. W. Steyerberg, and H. Putter. Fine-gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: Cumulative total failure probability may exceed 1. *Stat Med*, 40, 2021.

[12] Joseph Berkson and Robert P. Gage. Survival curve for cancer patients following treatment. *J. Amer. Stat. Assoc.*, 47:501–515, 1952.

[13] G. Berry. The analysis of mortality by the subject years method. *Biometrics*, 39:173–184, 1983.

[14] D. A. Binder. Fitting Cox's proportional hazards models from survey data. *Biometrika*, 79:139–147, 1992.

[15] G. Bonadonna and P. Valagussa. Dose-response effect of adjuvant chemotherapy in breast cancer. *New England J. Medicine*, 34:10–15, 1981.

[16] J. M Box-Steffensmeier and S. De Boef. Repeated events survival models: The conditional frailty model. *Stat. in Medicine*, 25:3518–33, 2006.

[17] P. Brindle, A. Beswick, T. Fahey, and S. Ebrahim. Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review. *Heart*, 92:1752–9, 2006.

[18] M. Buyse and P. Piedbois. The relationship between response to treatment and survival time. *Stat. in Medicine*, 15:2797–2812, 1996.

[19] B. Van Calster, D. Nieboer, Y. Vergouwe, B. De Cock, M.J. Pencina, and E.W. Steyerberg. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol.*, 74:167–176, 2016.

[20] Direct standardization. Accessed: 2018-12-03.

[21] Health effects of cigarette smoking. Accessed: 2018-11-30.

[22] B. Choodari-Oskooei, P. Royston, and M. K. B. Parmar. A simulation study of predictive ability measures in a survival model i: Explained variation measures. *Stat. in Medicine*, 31:2627–43, 2012.

[23] B. Choodari-Oskooei, P. Royston, and M. K. B. Parmar. A simulation study of predictive ability measures in a survival model ii: explained randomness and predictive accuracy. *Stat. in Medicine*, 31:2644–59, 2012.

[24] RM Conroy, K Pyorala, AP Fitzgerald, S Sans, A Menotti, G De Backer, D De Bacquer, P Ducimetiere, P Jousilahti, U Keil, I Njolstad, RG Oganov, T Thomsen, H Tunstall-Pedoe, A Tverdal, H Wedel, P Whincup, L Wilhelmsen, and IM Graham. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J*, 24:987–1003, 2003.

[25] R. D. Cook and Jerald F. Lawless. *Multistate models for the analysis of life history data*. CRC Press, Boca Raton, 2018.

[26] MT Cooney, AL Dudina, and IM Graham. Value and limitations of existing scores for the assessment of cardiovascular risk: a review for clinicians. *J Am Coll Cardol*, 54:1209–27, 2009.

[27] C. S. Crowson, E. J. Atkinson, and T. M. Therneau. Assessing calibration of prognostic risk scores. *Stat Methods Med Res*, 25:1692–1706, 2016.

[28] A. Dispenzieri, J. Katzmann, R. Kyle, D. Larson, T. Therneau, C. Colby,

R. Clark, G. Mead, S. Kumar, L.J. Melton III, and S.V. Rajkumar. Use of monclonal serum immunoglobulin free light chains to predict overall survival in the general population. *Mayo Clinic Proc*, 87:512–523, 2012.

[29] A. Dupuy and R. M. Simon. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J National Cancer Inst*, 99:147–57, 2007.

[30] A. Eaton. Equivalence of the robust and Greenwood variance estimates for the Kaplan-Meier estimate. personal communication, Mayo Clinic, 2018.

[31] B. Efron. The efficiency of Cox's likelihood function for censored data. *J. Amer. Stat. Assoc.*, 72:557–565, 1977.

[32] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia, 1982.

[33] T. R. Fleming and D. P. Harrington. Nonparametric estimation of the survival distribution in censored data. *Comm. Stat. Theory Methods*, 13:2469–2486, 1984.

[34] T. R. Fleming and D. P. Harrington. *Counting Processes and Survival Analysis*. Wiley, New York, 1991.

[35] Moons K. G., Altman D. G., Reitsma J. B., Ioannidis J. P., Macaskill P., Steyerberg E. W., Vickers A. J., Ransohoff D. F., and Collins G. S. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): explanation and elaboration. *Ann Intern Med*, 162:W1–73, 2015.

[36] M. H. Gail. Does cardiac transplantation prolong life? a reassessment. *Annals Internal Medicine*, 76:815–817, 1972.

[37] M. H. Gail and D. P. Byar. Variance calculations for direct adjusted survival curves, with applications to testing for no treatment effect. *Biometrical J.*, 28:587–599, 1986.

[38] M. H. Gail, J. H. Lubin, and L. V. Rubinstein. Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika*, 68:703–707, 1981.

[39] R. B. Geskus. Cause-specific cumulative incidence estimation and the Fine and Gray model under both left truncation and right censoring. *Biometrics*, 67:39–49, 2011.

[40] W. A. Ghali, H. Quan, R. Brant, G. van Melle, C.m. Norris, P.D. Faris, P.D. Galbraith, M.L. Knudtson, and APPROACH investigators. Comparison of 2 methods for calculating adjusted survival curves from proportional hazards models. *JAMA*, 286:1494–1497, 2001.

[41] M. Göen and G. Heller. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92:965–970, 2005.

[42] P. M. Grambsch and T. M. Therneau. Proportional hazards tests and

diagnostics based on weighted residuals. *Biometrika*, 81:515–526, 1994.

[43] Putter H., Fiocco M., and Geskus R. B. Tutorial in biostatistics: Competing risk and multi-state models. *Stat. in Medicine*, 26:2389–2430, 2007.

[44] Frank E. Harrell. *Regression Modeling Strategies*. Springer, 2015.

[45] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18):2543–2546, 1982.

[46] Patrick J Heagerty and Yingye Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105, 2005.

[47] H.W. Hense, E. Koesters, J. Wellman, C. Messinger, H. Volzke, and U. Keil. Evaluation of a recalibrated systematic coronary risk evaluation cardiovascular risk chart: results from systematic coronary risk evaluation germany. *Eur J Cardiovasc Prev Rehabil*, 15:409–15, 2008.

[48] M. A. Hernan. The hazards of hazard ratios. *Epidemiology*, 21:13–15, 2010.

[49] J. Hilden and T.A. Gerds. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Stat. in Medicine*, 33:3405–3414, 2014.

[50] Kent J. and O'Quigley J. Measures of dependence for censored survival data. *Biometrika*, 75:525–34, 1988.

[51] Clifford R. Jack Jr., Terry M. Therneau, Emily S. Lundt, Heather J. Wiste, Michelle M. Mielke, David S. Knopman, Jonathan Graff-Radford 3, Val J. Lowe, Prashanthi Vemuri, Christopher G. Schwarz, Matthew L. Senjem, Jeffrey L. Gunter, and Ronald C. Petersen. Long-term associations between amyloid positron emission tomography, sex, apolipoprotein e and incident dementia and mortality among individuals without dementia: hazard ratios and absolute risk. *Brain Communications*, doi:10.1093/braincomms/fcac017, 2022.

[52] J. D. Kalbfleisch and J. F. Lawless. The analysis of panel data under a Markov assumption. *J. Amer. Stat. Assoc.*, 80:863–871, 1985.

[53] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, New York, 1980.

[54] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data, second edition*. Wiley, 2002.

[55] M. W. Kattan and T. A. Gerds. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diag Prog Res*, 2, 2018.

[56] N. Keiding and D. Clayton. Standardization and control for confounding in observational studies: a historical perspective. *Stat Science*, 29:529–558, 2014.

[57] E. L. Korn and R. Simon. Measures of explained variation for survival data. *Stat. in Medicine*, 9:487–503, 1990.

[58] R. Kyle, T. Therneau, Rajkumar S., J. Offord, D. Larson, M. Plevak, and L. J. Melton (III). A long-term study of prognosis in monoclonal gammopathy of undetermined significance. *New England J Medicine*, pages 564–569, 2002.

[59] R. A. Kyle. Moncolonal gammopathy of undetermined significance and solitary plasmacytoma. Implications for progression to overt multiple myeloma. *Hematology/Oncology Clinics N. Amer.*, 11:71–87, 1997.

[60] N. Laird and D. Olivier. Covariance analysis of censored survival data using log-linear analysis techniques. *J. Amer. Stat. Assoc.*, 76:231–240, 1981.

[61] B. Langholz and L. Goldstein. Risk set sampling in epidemiologic cohort studies. *Statistical Science*, 11:35–53, 1996.

[62] F. Lawless, J. *Statistical Models and Methods for Lifetime Data*. John Wiley and Sons, New Jersey, 2003.

[63] J. G. Le-Rademacher, R. A. Peterson, and T. M. Therneau. Application of multi-state models in cancer clinical trial. *Clinical Trials*, 15:489–498, 2018.

[64] M. J. Leening, E. W. Steyerberg, B. Van Calster, Sr. R. B. D'Agostino, and M. J. Pencina. Net reclassification improvement and integrated discrimination improvement require calibrated models: relevance from a marker and model perspective. *statmed*, 33:3415–3418, 2014.

[65] D. Y. Lin, L. J. Wei, and Z. Ying. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80:557–572, 1993.

[66] J. Liu, Y. Hong, R.B. D'Agostine, Z. Wu, W. Wang, Sun J., P.W. Wilson, W.B. Kannel, and D. Zhao. Predictive value for the chinese population of the framingham chd risk assessment tool compared with the chinese multi-provincial cohort study. *Jama*, 291:2591–9, 2004.

[67] DM Lloyd-Jones. Short-term versus long-term risk for coronary artery disease: implications for lipid guidelines. *Curr Opin Lipidol*, 17:619–25, 2006.

[68] DM Lloyd-Jones, MG Larson, A Beiser, and D. Levy. Lifetime risk of developing coronary heart disease. *Lancet*, 353:89–92, 1999.

[69] C. L. Loprinzi, J. A. Laurie, H. S. Wieand, J. E. Krook, P. J. Novotny, J. W. Kugler, J. Bartel, M. Law, M. Bateman, N. E. Klatt, A. M. Dose, P. S. Etzell, R. A. Nelimark, J. A. Mailliard, and C. G. Moertel. Prospective evaluation of prognostic variables from patient-completed questionnaires. *J. Clinical Oncol.*, 12:601–607, 1994.

[70] S. Manzi, E.N. Meilahn, J.E. Rairie, C.G. Conte, Jr. Medsger, T.A., L. Jansen-McWilliams, R.B. D'Agostino, and L.H. Kuller. Age-specific

incidence rates of myocardial infarction and angina in women with systemic lupus erythematosus: comparison with the framingham study. *Am J Epidemiol*, 145:408–15, 1997.

[71] D. F. McCaffery, B. A. Griffen, D. Almirall, M. E. Slaughter, R. Ramchand, and L. F. Burgette. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med*, 32:3388–3414, 1991.

[72] W. Q. Meeker and L. A. Escobar. *Statistical Methods for Reliability Data*. Wiley-Interscience, 1998.

[73] C. Moler and C. van Loan. Nineteen dubious ways to take the exponential of a matrix. *Siam Review*, 20:801–36, 1978.

[74] C. Moler and C. van Loan. Nineteen dubious ways to take the exponential of a matrix, twenty-five years later. *Siam Review*, 45:1–46, 2003.

[75] N. J. D. Nagelkerke, J. Oosting, and A. A. M. Hart. A simple test for goodness of fit of Cox's proportional hazards model. *Biometrics*, 40:483–486, 1984.

[76] W. Nelson. *Applied life data analysis*. Wiley, New York, 1982.

[77] D. Oakes. Frailty models for multiple event times. In J. P. Klein and P. K. Goel, editors, *Survival Analysis, State of the Art*. Kluwer, Netherlands, 1992.

[78] D. Oakes, A.J. Moss, J.T. Fleiss, Jr. Bigger, J.T., T.M. Therneau, S.W. Eberly, M.P. McDermott, A. Manatunga, E. Carleen, J. Benhorin, and the Multicenter Diltiazem Post-Infarction Research Group. Use of compliance measures in an analysis of the effect of Diltiazem on mortality and reinfarction after myocardial infarction. *J. Amer. Stat. Assoc.*, 88:44–49, 1993.

[79] Digitalis Subcommittee of the Multicenter Post-Infarction Research Group, Moss A. J., J. T. Jr Bigger, E. Carleen, J. L. Fleiss, C. L. Odoroff, L. Rolnitzky, and Therneau T. The mortality risk associated with digitalis treatment after myocardial infarction. *Carivascular Drugs Ther*, 1:125–132, 1987.

[80] Puri P. and Sanyal A. J. Nonalcoholic fatty liver disease: Definitons, risk factors, and workup. *Clin Liver Dis*, 1:99–103, 2012.

[81] P. Peduzzi, J. Concato, E. Kemper, T.R. Holford, and A.R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol.*, 49:1137–9, 1996.

[82] R. Peto and J. Peto. Asymptotically efficient rank invariant test procedures (with discussion). *J. Royal Stat. Soc. A*, 135(2):185–206, 1972.

[83] Ross L Prentice and P Marek. A qualitative discrepancy between censored data rank tests. *Biometrics*, 35(4):861–867, 1979.

[84] M. S. Rahman, G. Ambler, B. Choodari-Oskooei, and R. Z. Omar. Re-

view and evaluation of performance measures for survival prediction models in external validation settings. *BMC Med Res Methodol*, 17:60, 2017.

[85] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *CoRR*, abs/1811.12808, 2018.

[86] C. Redmond, B. Fisher, and H. S. Wieand. The methodologic dilemma in retrospectively correlating the amount of chemotherapy received in adjuvant therapy protocols with disease free survival: a commentary. *Cancer Treatment Reports*, 67:519–526, 1983.

[87] P. Royston and D. G. Altman. External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology*, 13, 2013.

[88] P. Royston and W. Sauerbrei. A new measure of prognostic separation in survival data. *Stat. in Medicine*, 23:723–748, 2004.

[89] M. Schemper and R. Henderson. Predictive accuracy and explained variation in Cox regression. *Biometrics*, 56:249–255, 2000.

[90] M. Schemper, S. Wakounig, and G. Heinze. The estimation of average hazard ratios by weighted Cox regression. *Stat. in Medicine*, 28(19):2473–2489, 2009.

[91] Michael Schemper. Cox analysis of survival data with non-proportional hazard functions. *The Statistician*, 41(4):455–465, 1992.

[92] D. Schoenfeld. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, 67:145–153, 1980.

[93] D. A. Schoenfeld. Sample-size formula for the proportional-hazards regression model. *Biometrics*, 39:499–503, 1983.

[94] M. Schumacher, N. Hollander, and W. Sauerbrei. Resampling and cross-validation techniques: a tool to reduce bias caused by model building. *Stat. in Medicine*, 16:2813–2827, 1997.

[95] S. Senn. *Statistical Issues in Drug Development*. Wiley, New York, 2007.

[96] J. Stare, M. Pohar Perme, and R. Henderson. A measure of observed variation for event history data. *Biometrics*, 67:750–59, 2010.

[97] E. W. Steyerberg and Y. Vergouwe. Towards better clinical prediction models: seven steps for development and an abcd for validation. *Eur Heart J.*, 35:1925–31, 2014.

[98] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, N. Obuchowski, M. J. Pencina, and M. W. Kattan. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, 21:128–138, 2010.

[99] E.W Steyerberg. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer, New York, 2010.

[100] E.W. Steyerberg, F.E. Harrell, G.J. Borsboom, M.J. Eijkemans, and

Y. Vergouwe. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol*, 54:774–81, 2001.

[101] S. Suissa. Immortal time bias in pharmacoepidemiology. *American J Epidemiology*, 167:492–499, 2008.

[102] E. B. Tapper and Loomba R. Nonalcoholic fatty liver disease, metabolic syndrome, and the fight that will define clinical practice for a generation of hepatologists. *Hepatology*, 67:1657–1659, 2017.

[103] T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York, 2000.

[104] T. M. Therneau, P. M. Grambsch, and T. R. Fleming. Martingale based residuals for survival models. *Biometrika*, 77:147–160, 1990.

[105] T. M. Therneau and D. A. Watson. The concordance statistic and the Cox model. Technical Report 85, Department of Health Science Research, Mayo Clinic, 2015.

[106] A. C. M. Thiébaut and J. Bénichou. Choice of time-scale in cox's model analysis of epidemiologic cohort data: a simulation study. *Stat. in Medicine*, 23:3803–3820, 2004.

[107] R. Tibshirani. The lasso method for variable selection in the cox model. *Stat Med*, 16:385–95, 1997.

[108] James Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26:24–36, 1958.

[109] H. Uno, T. Cai, M. J. Pencina, R. B D'Agnostino, and L. J. Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. in Medicine*, 30(10):1105–1117, 2011.

[110] H. Uno, T. Cai, L. Tian, and L. J. Wei. Evaluating prediction rules for t-year survivors with censored regression models. *J. Amer. Stat. Assoc.*, 102:527–37, 2007.

[111] Hans C. van Houwelingen and Hein Putter. *Dynamic prediction in clincal survival analysis*. CRC Press, 2012.

[112] H. A. Verheul, E. Dekker, P. Bossuyt, A. C. Moulijn, and A. J. Dunning. Background mortality in clinical survival studies. *Lancet*, 341:872–875, 1993.

[113] H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48:817–838, 1980.

[114] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–26, 1982.

[115] J. Whitehead. Fitting Cox's regression model to survival data using GLIM. *Applied Stat.*, 29:268–275, 1980.

[116] F. Yates. The analysis of multiple classifications with unequal numbers in the different classes. *J. Amer. Stat. Assoc.*, 29:51–66, 1934.

[117] SS Yoon, CF Dillon, K Illoh, and M Carroll. Trends in the prevalence of coronary heart disease in the U.S.: National Health and Nutrition Examination Survey. *Am J Prev Med*, 51:437–45, 2016.