Predicting Ovarian Cancer Survival Times: Performance of Parametric Methods and Random Survival Forests

by

Rachel Lipson
B.A., University of Michigan, 2010

a Project submitted in partial fulfillment
of the requirements for the degree of

Master of Science

in the
Department of Statistics & Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2014

APPROVAL

Name: Rachel Lipson

Degree: Master of Science

Title of Project: Predicting Ovarian Cancer Survival Times: Performance of Parametric Methods and Random Survival Forests

Examining Committee: **Dr. Tim Swartz, Professor**
Chair

Dr. Rachel Altman,
Associate Professor
Senior Supervisor

Dr. Thomas Loughin,
Professor
Co-Supervisor

Dr. Michelle Zhou,
Assistant Professor
Internal Examiner

Date Defended: January 3rd, 2014

Partial Copyright Licence

Abstract

This project is an exploration of the performance of parametric and nonparametric methods in predicting time to recurrence (progression of cancer) and time to death in late stage ovarian cancer patients. The Weibull survival model is a common parametric method and is fit to the data for both death and recurrence, while Ishwaran et al's method of fitting random survival forests (2008) is employed as a nonparametric method. Performance of these models is evaluated using Harrell's C-index and Lawless & Yuan's cross-validation estimator (2010).

Acknowledgments

This work would not have been possible without the tremendous support and academic insight of Dr. Rachel Altman and Dr. Tom Loughin. Alon Altman and Alex Campesato have my gratitude for their patience in answering questions and working with the data.

Roberta Sterner, Lorene Sterner, Eric Lipson, and the rest of my family provided limitless love and support, unfailingly patient listening, and financial assistance. No piece of writing is complete without a great group of people to listen to ideas : Megan McCorquodale, Elena Szefer, Catherine Peng, Camille Bonham, Darby Thompson, Brian Mann, and others too many to name. Shirin Golchi, your patience and wry sense of humor helped keep me sane. Special thanks to Jack Davis, Andrew Henrey, Fabian Moya, and Kasra Yousefi: you made the first year of graduate school bearable, and without your academic and logistical assistance, I don't know if I could have done it. The generosity of Joan Hu and Carl Schwarz in sharing their experience and time is greatly appreciated. Many thanks to Kelly Jay and Charlene Bradbury who continue to make administrative miracles happen, and whose sympathy and kindness are second nature. I would like to also thank Jason Owen-Smith, Jeff Mackie-Mason, Kerby Shedden, and Derek Bingham for putting me on this track. Your guidance and advice has been invaluable.

This would not have been possible without you all. I am truly grateful.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Ovarian cancer is one of the most deadly gynecological cancers to afflict women. It is "ranked fifth in cancer deaths among women" despite making up only 3% of cancers in women (National Cancer Institute, NIH, 2013). Once diagnosed, an ovarian cancer patient will usually undergo treatment. Treatment can last from several months up to a year after diagnosis. Treatment typically includes surgery (also referred to as debulking) and chemotherapy. These methods are invasive, painful, and expensive, and vary in their success. The physical, emotional, and financial demands of cancer treatment can have a tremendous impact on the quality of life for patients. Having a more accurate method of predicting individual patient prognosis would be beneficial to doctors and patients in planning treatment and managing quality of life.

This project is intended to explore the predictive ability of parametric and nonparametric methods for time to recurrence (or progression of cancer) and time to death after recurrence (DAR) for ovarian cancer patients based on a wide array of predictor variables. We describe these events in more detail in Section 2.1. We evaluate the performance of such methods given missing and misconfigured data as well as longitudinal measurements without associated times of observation.

Our approach differs from previous work (Rocconi et al, [2009, 2007], Asher et al [2011], Altman [2012]) in that we predict events using models that incorporate several predictors simultaneously. Other techniques such as nomograms have been used to generate prognoses for patients (Barlin et al, [2011]) that have improved prediction of mortality beyond simply classifying patients via the Federation of Gynecology and Obstetrics staging system. However, the nomogram method utilizes the results of univariate analysis of each predictor and applies a competing risks framework to account for other causes of death.

To create our predictive models, we use two methods of survival analysis, also known as the analysis of time-to-event data. We fit a variant of the Weibull model in which a separate Weibull model is fit for recurrence and DAR times. We also fit a predictive model for recurrence and DAR times using the nonparametric method of random survival forests (RSFs).

The remainder of this thesis is organized as follows. Chapter 2 contains a description of treatment of ovarian cancer and the events of interest in this project, followed by an overview of measured variables. An in-depth discussion of the statistical methods used to fit each predictive model can be found in Chapter 3, along with an overview of the methods we select to quantify prediction error, Harrell's C-index and an estimator used by Lawless & Yuan (2010). Approaches to data cleaning and manipulation of variables are described in Chapter 4. Chapter 5 provides the results of our work. We conclude with a discussion and our recommendations for future research in Chapter 6.

Chapter 2

Ovarian Cancer Data

The data for this project were collected over the course of four years on 218 late-stage ovarian cancer patients from the Tom Baker Cancer Centre in Calgary, Canada. Surgical, clinical, and hematologic (blood marker) measurements were collected during the treatment period. After diagnosis the typical patient[1] is assessed for surgery. A patient's eligibility for surgery is decided by a gynecological oncologist, who takes into account how ill the patient is (very ill patients may be deemed too fragile for surgery) and how pervasive the cancer is. If the cancer is widespread, the doctor may decide to have the patient undergo chemotherapy to shrink the cancerous tissue prior to operating. If a patient is considered a good candidate for immediate surgery, she will undergo debulking surgery to remove as much of the tumor as possible.

Approximately 4-6 weeks after surgery, the patient will start chemotherapy. The usual treatment is a platinum-containing agent, carbo-taxol, which is administered every three weeks. If the patient experiences an adverse reaction to the standard chemotherapy agent, she is switched to a different type of chemotherapeutic agent. If the patient becomes too ill to continue chemotherapy, either due to cancer or other illness, chemotherapy is paused until she recovers.

2.1 Responses of Interest

Once a patient has completed her course of treatment, the events of interest are time until recurrence of cancer and time until death following such a recurrence. These events are

[1] At this cancer center. Other facilities may have differing protocols for treatment.

similar to those considered by Rocconi et al (2009), who split survival into two categories: progression free survival and overall survival. In our work, we refer to disease progression as recurrence of cancer (defined by an increase in cancer antigen levels in the blood, increased cancerous mass, or other cancerous activity) as diagnosed by a clinician. In contrast to Rocconi et al (2009), we focus on survival of patients who experienced disease progression rather than on overall survival. In particular, we are interested in the trajectory of patients for whom treatment response was initially successful and whose deaths are likely attributable to cancer.

A brief discussion of the predictors and their implications for event times are as follows.

2.2 Hematologic Predictors

Hematologic predictors were measured repeatedly over the course of patients' treatment. The following is a list of variables that were measured.

CA125

CA 125 stands for cancer antigen 125 and is a tumor marker for ovarian cancer. Persistently high levels of CA 125 throughout treatment have previously been associated with poor prognosis (Rocconi et al, 2009), while lower levels of CA 125 or achieving normalization is a positive sign. The earlier a patient's CA 125 levels normalize the better the prognosis appears (Rocconi et al, 2009).

Hemoglobin

Hemoglobin is a component of red blood cells and is responsible for carrying oxygen through the blood and regulating iron levels. Anemia, or low hemoglobin, can be a common side effect of chemotherapy and surgery. Anemic patients may face more health problems than those patients with sufficient hemoglobin. If a patient's hemoglobin drops below certain levels, transfusions may be given and if surgery is scheduled, it may be delayed.

White Blood Cell Count and Neutrophils

White blood cells are responsible for responding to infection in the body and regulating immune response. There are several different types of white blood cells and the cells responsible for immune function are called neutrophils. Chemotherapy can inhibit the body's ability to produce neutrophils, inducing a state known as neutropenia. There is some evidence that patients who experience neutropenia during treatment may have some survival advantages over those whose neutrophil levels remain high, especially among suboptimally debulked patients (Rocconi, 2007). This is attributed to the sensitivity of the neutrophil-producing stem cells to chemotherapy: similar sensitivity affects cancer producing cells. Patients experiencing very low white blood cell counts may be prescribed medication to boost white blood cell production.

Platelets

Platelets are another cell type found in the blood. They are responsible for clotting in healthy individuals. High levels can lead to blood clots and low levels can cause uncontrolled bleeding. Platelet levels at either end of the healthy range can severely affect a patient's health as well as delaying or complicating surgery.

Albumin

Albumin is a protein found in the blood and is used as a measure of nutritional status. Patients with low albumin tend to be in poorer health and may even be considered malnourished. Nutritional status has a far reaching impact on the general health and immunological status of an individual and may play a role in the effectiveness of treatment and patient response.

2.3 Surgical Predictors

One of the first assessments a patient undergoes after being diagnosed is whether she is eligible for immediate surgery. If the cancer is extremely widespread or particularly difficult to excise, the oncologist may recommend chemotherapy first followed by what is referred to as interval surgery. There is ongoing debate within gynocological oncology as to whether the type of surgery has an effect on the success of debulking and subsequent survival. (Vergote et al, 2010, Altman, A., personal communication). Optimal debulking refers to the amount of residual disease left after surgery and a patient is considered optimally debulked when residual disease is < 1 cm. Type of surgery, debulking, and blood loss during surgery were also recorded in the dataset.

2.4 Clinical Predictors

One of the primary treatments of ovarian cancer is chemotherapy. Chemotherapy is typically made up of some combination of cytotoxic-anti-neoplastic drugs, which stop rapid cell division. Some types of chemotherapy tend to be more effective than others, but individual patients may not be able to tolerate certain classes of drugs. Chemotherapy regimens that include platinum have shown some benefits compared to other chemotherapy options (Du et al, 2013) and platinum sensitivity has also been linked to faster normalization of CA 125 levels (Rocconi et al, 2009). The dataset includes the following chemotherapy-related predictors: number of primary and neoadjuvant cycles, type of chemotheraputic agent used for each cycle, and whether or not the drug contained platinum.

All patients in this dataset were in the later stages of disease progression, most with stage 3 or 4 cancer and with grade[2] ranging from 1 to 4. Patients' stage, grade, age at diagnosis, and amount of ascites (fluid in the abdomen, a possible byproduct of ovarian cancer) were also listed in the dataset.

2.5 Missing Data

Missing data are also an issue with this dataset as some of the repeatedly measured hematologic variables have only one measurement or are missing entirely for the duration of a patient's treatment. Even those variables that are measured only once (by design) can be missing. It appears that patient location (either in Calgary or the surrounding areas) may be a lurking confounder for quality of care as well as a source of missing data (Altman, A., personal communication). This issue is discussed further in Chapter 6.

[2] Tumor grade is a measure of how abnormal the cells are, ranging from barely differentiable from normal cells (grade 1) to the highly abnormal and typically fast-growing grade 4 cancer cells.

Chapter 3

Statistical Methods for Survival Data

One feature that distinguishes survival data from other types of data is the fact that not all subjects necessarily experience the event of interest while under observation, in which case those subjects' event times are considered incomplete ("censored"). There are several different types of censoring, but in this project we consider only the case in which an event did not occur during the given period of observation but is assumed to occur at some (possibly hypothetical) time in the future, after observation ends ("right censoring"). For individual i, we observe time $Y_i = \min(T_i, C_i)$, where T_i is the time of the event and C_i is the censoring time. We define the indicator variable δ_i as $\delta_i = 1$ if $Y_i = T_i$ and $\delta_i = 0$ if $Y_i = C_i$.

As mentioned previously, the events of interest for the purposes of this particular analysis are recurrence of cancer after treatment and death upon recurrence of cancer. Recurrence is measured from time of treatment completion; DAR is measured from the time of recurrence. A patient's recurrence time can be censored if she drops out of the study, dies, or reaches the end of the study without experiencing recurrence. Censoring of DAR times can be due to dropping out or surviving past the end of the study. Cause of death was not specified in the dataset but we treat death prior to recurrence as different than death after recurrence, the latter of which is assumed to be more likely due to cancer.

Let the recurrence and DAR time for subject i be T_{i1} and T_{i2}, respectively. Since we expect these event times to be correlated, it seems naive to attempt to model time until death without taking recurrence into account. Instead we specify a joint distribution for T_{i1} and T_{i2} by writing it as

$$F_{T_{i1},T_{i2}}(t_{i1}, t_{i2}) = F_{T_{i1}}(t_{i1})F_{T_{i2}|T_{i1}}(t_{i2}|t_{i1}) \ ,$$

where $F_{T_{i1},T_{i2}}$ is the joint cumulative distribution function (cdf) of the i^{th} patient's recurrence and DAR times, $F_{T_{i1}}$ is the cdf of the i^{th} patient's recurrence time, and $F_{T_{i2}|T_{i1}}$ is the cdf of the i^{th} patient's DAR time conditional on the recurrence time. We then specify one model for T_{i1} and another for $T_{i2}|T_{i1}$. Our model is a very simple example of a recurrent events model where each type of event can happen at most once.

For both our parametric and nonparametric analyses assumptions regarding censoring times are required. Specifically, since end-of-study occurs at a fixed time and drop-out presumably occurs at random, we assume that DAR times are independent of censoring times. In addition, we assume that recurrence and censoring times are independent *given the observed predictors* (since in this case, death time – which is likely related to the predictors – is treated as a form of censoring). We discuss the validity of these assumptions in more detail in Chapter 6.

3.1 Parametric Survival Analysis

Having outlined the underlying framework for each event, we can discuss how to apply each different method of predicting event times. Parametric survival analysis is based on the assumption that the survival times are distributed according a standard parametric distribution. The unknown parameters of this distribution, $\boldsymbol{\theta}$, are typically estimated using the method of maximum likelihood. We define the density and survival function associated with the survival time of the i^{th} individual as $f_i(t; \boldsymbol{\theta})$ and $S_i(t; \boldsymbol{\theta}) = \mathrm{P}(T_i > t; \boldsymbol{\theta})$. The likelihood is of the form

$$L(\boldsymbol{\theta}; y_1, \ldots, y_N) = \prod_{i=1}^{N} f_i(y_i; \boldsymbol{\theta})^{\delta_i} S_i(y_i; \boldsymbol{\theta})^{1-\delta_i}$$

The distribution we use to model time until recurrence and time until death (conditional on recurrence) is the Weibull distribution. The Weibull model is well established in survival analysis and is known for its flexibility in a wide variety of situations. Depending on the parameters, the Weibull distribution can be highly right-skewed, a common feature of survival data. The Weibull density is

$$f(t) = \frac{\alpha}{\phi}\left(\frac{t}{\phi}\right)^{\alpha-1} \exp\left[-\left(\frac{t}{\phi}\right)^{\alpha}\right]$$

with survival function

$$S(t) = \exp\left[-\left(\frac{\mathrm{t}}{\phi}\right)^{\alpha}\right],\ t > 0, \alpha > 0,\ \phi > 0$$

and cumulative hazard function $H(t) = -\log[S(t)]$. The shape and scale parameters for the Weibull are α and ϕ, respectively. Covariates are included in the model as $\log(\phi_{\mathrm{i}}) = \mathrm{z}'_{\boldsymbol{i}}\boldsymbol{\beta}$, where $\boldsymbol{z_i}$ is a vector of observed covariate values for individual i (Klein & Moeschberger, 2003).

A Weibull model is assumed both for T_{i1} and for $T_{i2}|T_{i1}$. We assume that there are no parameters common to $F_{T_{i1}}$ and $F_{T_{i2}|T_{i1}}$. Thus, model parameters for recurrence and DAR are estimated separately using standard maximum likelihood methods. Twenty-seven predictors are included in the Weibull model for recurrence times and the same twenty-seven predictors plus time to recurrence are used in the model for DAR times. We discuss the derivation and selection of predictor variables in Chapter 4.

Residuals for the Weibull model are formulated as described by Lawless (2003). In particular, the probability integral transform is applied to the cumulative hazard function to obtain residuals that, if the assumption of the Weibull model is appropriate, will be approximately exponentially distributed. QQ-plots can then be created by plotting the residuals against the quantiles of the standard exponential distribution. An additional discussion of these residuals can be found in Appendix A.2.

3.2 Nonparametric Survival Analysis

3.2.1 Survival Trees

A recent development in the field of survival analysis is the application of classification and regression trees (CART) to survival data (Bou-Hamad et al, 2011). Regression trees are a non-parametric method for predicting responses. The sample space is iteratively subdivided until a stopping criterion is reached, assigning each subdivision of observations a different predicted value (Loughin, 2012). In the simplest case, the sample space is divided (or “split”) based on the variable and split point that maximize the difference in responses. After we divide the observations based on this criterion, the two new subdivisions of observations are examined to determine the variable and split point that again maximizes the difference in responses. By the end of the process, observations have been grouped by similar values of the predictors in “terminal nodes”, but without a pre-selected model. For a survival tree, the observed responses are time-to-event data and the method of evaluating splits must be able to handle censored observations as well as observed event times. Ishwaran et al (2008) use the Nelson-Aalen estimate of the cumulative hazard function (CHF) to summarize the estimated survival curve (see Lawless, [2003]).

While it is possible for the tree to have so many splits that each terminal node contains only one response each, it is typical for the tree to have a stopping rule based on a minimum number of responses per node. In the case of survival trees, the minimum number of responses per node refers to the number of *observed* events per node.

With respect to defining a measure of distance between responses, survival trees typically use the estimated survival curve at each potential daughter node as a basis. The survival curves incorporate censored observations and there are many options for quantifying the distance between them (e.g. the log-rank statistic, likelihood ratio statistic, or Wilcoxon-Gehan statistic). For a more in-depth discussion of this topic, see Bou-Hamad et al (2011).

Unfortunately, using trees has some drawbacks. For example, once a variable has been selected for a split, the entire sample space is divided. Thus, small changes in one split will cascade throughout the rest of the sample space. Outliers and quirks of different sample datasets from the same population can lead to very different trees and so the predictions from a tree may be highly variable and unstable. The choice of stopping rule and the values of the various tuning parameters can also affect the final tree selection. For these and other reasons, trees are not typically used on their own.

3.2.2 Random Survival Forests

While a single tree can lead to unstable predictions, the aggregation of many trees tends to be less variable (Loughin, 2012). Random survival forests are the result of averaging over many survival trees after repeatedly resampling from the original data (Ishwaran et al, 2008). This method leads to a more robust view of the data and subsequently, better predictions. The first step in growing a random survival forest is to obtain a new dataset of size N by sampling with replacement from the original data. A survival tree is then fit to these new data. Within that tree, a subset of variables is randomly chosen at each node and the split that maximizes the difference between responses in the two daughter nodes is chosen. The stopping rule for each tree is based on the minimum number of observed deaths for each terminal node. The data are then repeatedly sampled with replacement and a tree fit to each of these new datasets. To compute the predicted outcome associated with an observation, all of the trees that were grown *without* using this particular observation (i.e. all of those trees for which this observation is out-of-bag) are considered. The observation is sent down each of these trees and the resulting estimated CHFs based on these trees are recorded. The final estimated CHF for a given observation is calculated by averaging across all these trees.

We fit RSFs to both recurrence and DAR times, using recurrence time as an additional predictor in the latter. Before the RSFs can be fitted to the data, some tuning parameters must be chosen. The number of variables randomly selected at each node, the kind of splitting rule used to determine maximal difference in survival outcomes, and, in the case of a random splitting rule, the number of random splits per variable, must also be decided. The number of variables randomly selected at each node has been extensively discussed. Both Ishwaran & Kogalur (2013) and Hastie, Tibshirani & Friedman (2009) recommend $m = \sqrt{p}$ variables selected at each node, with p being the total number of available predictors. For the purposes of this project, 27 variables including interactions were used to predict recurrence. For DAR, 28 variables were used. Thus, the number of variables chosen at each split was 5.

The splitting rule used in this case is random log-rank splitting, where the log-rank statistic is used to represent the distance between grouped observations at each node. Unlike the case in which all possible splits of the variable are considered, a random split of each variable is used instead. Conventional log-rank splitting has been shown to perform well in both proportional and non-proportional hazard settings and is easily interpretable. Random log-rank splitting retains these favorable attributes while improving computational time and has demonstrated good performance (Ishwaran et al, 2008). Random log-rank splitting rules also help to address the issue of end-cut preference (or the "favoring of uneven splits" due to outliers and extreme values [Ishwaran & Kogular, 2007]) and favoring continuous variables (Loh & Shih, 1997). "Using a reasonably small value [for the number of splits considered] mitigates bias and may not compromise accuracy" (Ishwaran & Kogalur, 2013) in predictions and too large a number of splits will diminish the benefit of using random splitting at all.

In this project, four splits per variable are considered, as this provides several options for prospective split points, but is still few enough that the prediction bias is not overwhelming. Furthermore, our preliminary exploration suggests that our results are relatively robust to the number of splits, at least when that number is in the neighborhood of four. In particular, when we re-fit the RSF trying 3, 5, and 6 splits, the resulting prediction error estimates varied only minimally.

3.3 Prediction Error

Point predictions of survival times are notoriously poor (Henderson et al, 2001). The existence of censored observations makes the assessment of prediction error (an overall measure of differences between observed and predicted values) more complicated than in other circumstances. We know that censored patients have survived or been event-free up to their censored time, but there is no information beyond that time point. For these subjects, the event of interest may never have happened, it may have occurred immediately after the censoring time, or several years later. For this reason, treating the censoring time as just another event time can lead to wildly inaccurate error estimates. We have chosen to use Harrell's C-index and the Lawless-Yuan estimator (2010), both of which allow for special treatment of the censored patient times. We discuss these estimators in detail in Sections 3.3.2-3.3.4.

We also consider plots of observed vs. predicted values. While the difference between predicted and censored values is of limited use, the distance between observed and predicted values does tell us approximately how well our model predicts for that subset of the data. The observed vs. predicted value plots can be seen in Appendix A.1.

3.3.1 Cross-Validation

In ideal circumstances, in order to estimate the prediction error associated with our models, the dataset would be randomly split into two partitions, with the models fit on one set of observations and evaluated on the other independent test set. However, with the relatively small size of the dataset and large number of predictors, this approach is not feasible. If the model is fit on the full data and the same dataset is used to evaluate the predictive ability of the model, we would expect that the model will underestimate prediction error. As an alternative, cross-validation can be used. Cross-validation splits the dataset into several groups (called folds), excludes one fold, and fits the model using the remaining observations. The observations in the omitted fold are then predicted using the fitted model. The process is repeated until the observations in each fold have been predicted based on a model fit to the non-fold data.

The number of folds is subject to a bias-variance trade-off. In particular, when a small number of folds is used, fewer data are available to train the model, creating more variable predictions and leading to an overestimate of prediction error (Loughin, 2012). However, with many folds, the observations to predict the omitted folds have greater overlap and thus the estimates of error based on each fold will be positively correlated. The overall estimated prediction error (which is computed as the average of these fold-based estimates) then tends to be more variable.

Lawless & Yuan (2010) say that n-fold cross-validation performs very well for the evaluation of prediction error of a variety of prediction methods but is computationally intensive and leads to high variance of the prediction error estimator (Hastie, Tibshirani & Friedman, 2009). Luckily, 5-fold cross-validation did not lead to excessive bias in the estimation of prediction error in the simulations done by Lawless & Yuan, while Hastie et al (2009) suggest that 5- or 10-fold cross-validation should also perform adequately in most situations. Given the relatively small size of the dataset to the large number of predictors, 10-fold cross-validation seems to be a reasonable compromise between bias and variance as well as computational time. This implies approximately 20 observations per fold in the recurrence data and approximately 15 observations per fold in the DAR data.

Since the selection of observations for each fold can affect the estimate of predicton error, we make the cross-validation method even more robust by reallocating the sample data 100 times into different folds and then averaging the resulting 100 prediction error estimates.

3.3.2 C-Index

While traditional methods of estimating prediction error yield numeric estimates of distance between predicted and observed survival times, they do not reflect the ability of a method to predict the survival times associated with censored observations. The C-index allows us to incorporate some of the information found in censored observations by ascertaining whether shorter predicted survival times correspond to actual shorter survival times, including some censored survival times.

Recall that Y_i is the observed time for subject i, and that δ_i is an indicator as to whether Y_i is the censoring time ($\delta_i = 0$) or event time ($\delta_i = 1$). The C-index is calculated by considering all pairs of observations, then removing those pairs where both subjects' times are censored or where the shorter survival time is censored. The remaining pairs are then evaluated as follows. We treat (Y_i, Y_j) as the same as (Y_j, Y_i) and present the pairs such that $Y_i \leq Y_j$. Let P_i be the predicted value for the i^{th} individual. The C-index is defined as

$$C = \frac{1}{L} \sum_{i,j} d_{ij} \ ,$$

where, for the (i, j)th pair,

$$d_{ij} = \begin{cases} 1, & \text{if } Y_i < Y_j,\ P_i < P_j,\ \delta_i = 1 \\ 0.5, & \text{if } Y_i < Y_j,\ P_i = P_j,\ \delta_i = 1 \\ 1, & \text{if } Y_i = Y_j,\ P_i = P_j,\ \delta_i = \delta_j = 1 \\ 0.5, & \text{if } Y_i = Y_j,\ P_i \neq P_j,\ \delta_i = \delta_j = 1 \\ 1, & \text{if } Y_i = Y_j,\ P_i > P_j,\ \delta_i = 0,\ \delta_j = 1 \\ 0.5, & \text{if } Y_i = Y_j,\ P_i < P_j,\ \delta_i = 0,\ \delta_j = 1 \\ 0, & \ \text{otherwise} \end{cases} .$$

C can be interpreted as the approximate percentage of correctly classified (based on length of survival time) pairs of observations. Prediction error can be calculated as $1 - C$.

3.3.3 Lawless-Yuan Estimator

The C-index provides a measure of how closely the *order* of our predicted values matches that of the observed values (including some censored observations). However, it does not reflect the *distance* between the predicted and observed values. Another method of assessing prediction error that quantifies the distance between predicted and observed survival times is the cross-validation estimator of the expected loss proposed by Lawless & Yuan (2010). The data are subdivided into V mutually exclusive groups, with the v^{th} group denoted by S^v. The predicted value for observation i in group S^v, $\hat{y}_{i(-v)}$, is derived using all observations except those in S^v. The estimator, which includes an adjustment for the probability of censoring, has the form

$$\hat{\pi}_{cv} = \frac{1}{N} \sum_{v=1}^{V} \sum_{i \in S^v} \frac{\delta_i}{\hat{S}_{c(-v)}(y_i | \boldsymbol{z}_i)} L(y_i, \hat{y}_{i(-v)}) \ , \tag{3.1}$$

where $L(y_i, \hat{y}_{i(-v)})$ is a loss function. $\hat{S}_{c(-v)}$ is the survival function for censoring times based on all observations except those in S^v.

Loss Function

For the purposes of this project, the loss function implemented in (3.1) is

$$L(y_i, \hat{y}_{i(-v)}) = \left| \log \left(\frac{y_{i(-v)}}{\hat{y}_{i(-v)}} \right) \right| ,$$

henceforth called the absolute value of the log ratio (AVLR). By using the AVLR we consider relative rather than absolute differences in survival times. The AVLR is especially appropriate for recurrence of cancer since check-ups post-treatment become less frequent the longer a patient is cancer free. More time betwen scheduled appointments may mean recorded recurrence time is less exact. Thus, we would not want to put too much weight on absolute differences observed on patients with long recurrence times. In addition, a difference in predicted and observed times of a year is less extreme in a 10 year survival time than in a 1 year survival time (Yuan, 2008). Using a loss function based on relative rather than absolute differences takes into account these considerations. Finally, the AVLR has a nice interpretation. In particular, exponentiating the AVLR is equivalent to taking a weighted geometric mean of the ratios of observed and predicted values, or more specifically, of the values $\max\left(\frac{y_i}{y_{i(-v)}}, \frac{y_{i(-v)}}{y_i}\right)$.

IPCW Weights

Each summand in (3.1) is weighted by the estimated probability that an individual's survival time will be censored after her observed survival time, called the inverse probability censoring weight (IPCW) by Lawless & Yuan (2010). In the case where censoring time is independent of the predictors, i.e., $S_{c(-v)}(y_i \mid \mathbf{z}_i) = S_{c(-v)}(y_i)$, the Kaplan-Meier method can be used to estimate the censoring distribution. As discussed earlier, since censoring of DAR times is due to either dropping out of the study or end-of-study, censoring can reasonably be assumed to be independent of the predictors. In the case of recurrence, it is much less likely that censoring could be considered independent of the predictors since death is counted as a form of censoring, and is presumably a function of clinical and treatment data. In these circumstances, $S_c(y_i|\mathbf{z}_i)$ is estimated using the Weibull model described in Section 3.1, reversing the roles of the event and censoring times.

In Lawless & Yuan's estimator, the IPCW weights should ideally be calculated for each new fold. This is feasible for the DAR analysis since we assume censoring times are iid. However, for the recurrence analysis, we allow censoring times to depend on all the predictors. In these circumstances, $S_c(y_i|\boldsymbol{z}_i)$ is estimated using the Weibull model. The large number of predictors and possibility of insufficient quantity of censored observations in each training set leads to difficulty in estimating the censoring distribution parametrically. In order to address this issue, the IPCW weights for censoring in recurrence are estimated only once using the full data. For limitations of this approach, see Chapter 6.

3.3.4 Confidence Intervals

Any estimate of prediction error (PE) may be affected by peculiarities of sample data. To gain an understanding of the distribution of our point estimators of PE, we resampled with replacement the original data M times and re-computed the C-index and Lawless-Yuan estimator based on the "new" data, B^*. With these bootstrapped estimates of PE, we can form confidence intervals (CIs). We define

$$\overline{\widehat{PE^*}} = \frac{1}{M}\sum \widehat{PE^*_b}$$

as the average point estimate of PE and

$$\widehat{\text{var}}(\widehat{PE^*}) = \frac{1}{M-1}\sum_{b=1}^{B}(\widehat{PE^*_b} - \overline{\widehat{PE^*}})^2$$

as the sample variance of PE. Due to the difficulty of calculating the variance of their prediction error estimator, Lawless & Yuan (2010) recommend using a simple normal approximation for the CIs. An approximate $100(1-\alpha)\%$ CI can be created as follows:

$$\widehat{PE} \pm \widehat{\text{var}}(\widehat{PE^*})^{(1/2)} z_{\frac{\alpha}{2}} \ ,$$

Where $\widehat{PE}$ is the estimate of PE based on the original dataset.

Lawless & Yuan suggest creating CIs for log(PE) if the estimates of PE are severely right-skewed. The logged estimates are then used in the above formula. The end points of the CIs are calculated for PE, then exponentiated. The CIs for DAR are calculated based on 1,000 bootstrapped estimates, while the CIs for recurrence are based on only 500 bootstrapped estimates due to associated computational intensity. To further reduce the computational burden and because each resampled dataset is similar to a permutation of the original data, we omit the additional re-folding procedure (discussed at the end of Section 3.3.1) when computing $\widehat{PE}^*_b$.

Chapter 4

Data Cleaning and Manipulation

Blood work and surgical indicators were measured during treatment. Unfortunately, the exact dates of repeated measurements taken during patient visits during treatment are unavailable, procluding analysis based on a joint longitudinal survival model. However, the chronological order of the measurements is available. We are thus able to derive clinically meaningful univariate summaries such as minimum, mean, and maximum values of each variable and incorporate these in subsequent models as predictors. Since the ultimate goal of this project is eventual clinical utility as well as predictive power, the simplicity of those derived predictors is a priority as well. In this chapter, we describe the rationale behind each of the predictor variables used in the models and how each final predictor is derived.

4.1 Hematologic Predictors

Rocconi et al (2009) found that CA 125 was an important predictor for patient prognosis. Not only were levels of CA 125 in and of themselves useful, but the trend of CA 125 over time and the speed with which a patient's levels normalized were also important predictors. In order to approximate this information without having treatment times, we include the difference between the first and last CA125 measurements, as well as the lowest observed CA 125 measurement. The difference between first and last measurements is intended to summarize how levels of CA 125 changed over the course of treatment, while the minimum CA 125 level provides an idea of whether the patient ever approached normal levels of the antigen.

We are able to summarize the rest of the hematologic predictors relatively simply based on the known effects of each variable on patient health outlined in Chapter 2. In particular, since hemoglobin levels reflect patient health and low hemoglobin may imply illness, we use minimum and average hemoglobin levels as predictors. Due to the the parallel sensitivity of neutrophil-producing stem cells and cancer producing cells, we include both the minimum and the average neutrophil level for each patient. Since the total white blood cell count includes neutrophils, the difference in the average level of neutrophils and average white blood cell count is also used as a predictor. When the level of neutrophils is too low, a prescription may be issued to artificially increase the number of white blood cells. The usage of neutrophil boosting drugs is also incorporated in our predictive methods. In addition, minimum, maximum, and average platelet levels are considered, as well as minimum albumin measurements.

4.2 Surgical and Clinical Predictors

Blood loss during surgery can weaken a patient and is included as a predictor, as well as interaction between type of surgery and blood loss. The effect of ascites on patient health may be moderated by treatment type and the effect of treatment may in turn be influenced by ascites. We treat ascites as binary (present/absent) and include both its main effect and its interaction with type of surgery in the models. Whether or not a patient recieved the standard chemotherapy drugs usually used in ovarian cancer (carboplatin-taxol), whether any of their chemotherapy drugs contained platinum, and the number of pre- and post-surgery chemotherapy courses, otherwise known as neoadjuvant and primary adjuvant cycles, are all considered as predictors in this project. Since the length of treatment may depend on the patient's health, and hence impact recurrence and DAR times, the length of treatment is included as a predictor. For predicting death after recurrence, the transformed log time from end of treatment to time of recurrence is also included. A list of final predictors can be found in Appendix B.

4.3 Missing Predictor Values and Incomplete Records

Some patients in our dataset did not appear to achieve normalization of their CA 125 levels. This may not be due to lack of response to treatment, but rather, may be a problem of missing data. It has been established that not all patient records were complete and so normalization of CA 125 levels may have been unrecorded if information for those visits was not included. Similarly, the difference between first and last CA 125 measurements may not have been the true difference if the last recorded measurement and the first recorded measurement were not from the first and final visits, respectively. This issue is not unique to CA 125; missing data or few recorded measurements may impact the utility of all the hematologic variables as predictors.

We treat the missing predictor measurements as missing completely at random. When a patient is missing all measurements on a given covariate, the median value for the covariate measurement is substituted in the case of continuous variables and the mode in the case of categorical variables. While this is a very rudimentary method of incomplete data imputation, the number of patients missing all values of a covariate is reasonably small (< 10 per covariate). One patient who is missing all covariate values is excluded from the analysis.

For the six patients who are missing last contact dates and were alive at their last check-up, the end of study date is used as their last follow-up date. Three patients are missing recurrence dates, four patients experienced significant disease progression prior to finishing treatment, two patients are recorded as having died before treatment ended, four patients are documented as experiencing recurrence on the same day they completed treatment, and one patient's records indicate recurrence after death. All 15 of the aforementioned subjects are excluded from analysis for time until recurrence, leaving 204 patient records for analysis. Of these 204 patients, 153 patients experienced recurrence of their disease. Of these 153 patients, 112 patients died after recurrence.

Chapter 5

Results

Once the dataset is processed as described in Chapter 4, the models presented in Chapter 3 are fit and their fit evaluated (in the case of our parametric analysis).

For the Weibull models, the QQ-plots of the residuals (see Appendix A.2) show no evidence of lack of fit for either recurrence or DAR times. Residuals are also plotted against all continuous covariates and no evidence of lack of fit is apparent (plots not shown). Both observed recurrence and DAR times are plotted against the predicted values based on the Weibull and RSF methods and can be seen in Appendix A.1.

CIs are created for prediction error for recurrence and DAR times, for both Weibull and RSFs. The results of assessing each method using the C-index (where $1 - C$ can be interpreted as the proportion of misclassified observations) can be seen in Table 5.1. The bootstrapped estimates of $1 - C$ appeared to be approximately normally distributed, lending support for our choice to use the normal quantiles for constructing CIs.

Table 5.1: Estimates of Prediction Error ($1 - C$)

Event	Weibull	RSF
Recurrence	0.42 (0.35, 0.49)	0.40 (0.35, 0.45)
DAR	0.41 (0.34, 0.49)	0.43 (0.36, 0.50)

From these results, it appears that the Weibull and RSF methods have similar error rates. Given the presumable noise in these estimates, it is unclear whether either method is superior according to this measure of prediction error. Neither method appears to be doing very well: an error rate of approximately 40% is substantial and means that 40% of the time, our predictions of longer or shorter survival times did not correspond to observed longer or shorter survival times. That said, if we naively predicted Y_i by choosing a smaller or larger value with 50% probability, a C-index of 0.5 would result. Both prediction methods appear to perform slightly better than this random allocation of outcomes.

Next, we compute the Lawless-Yuan estimator. Since the initial distribution of bootstrapped estimates of the Lawless-Yuan error appears to be somewhat skewed, we take the log of the estimates. After this adjustment, the distribution of the Lawless-Yuan estimator appears approximately normal. We thus determine point estimates and CI endpoints based on the logged estimates. The point estimates and CI endpoints in Table 5.2 are the twice-exponentiated values of these quantities. They can be interpreted as the IPWC-weighted geometric average distance between predicted and observed values. In terms of implications for individual prediction, these estimates represent the factor by which, on average (in the geometric sense) predicted observations will differ from their true values. In particular, our methods over- or under-estimate survival times by a factor of approximately 2.

Table 5.2: Estimates of Prediction Error (Lawless-Yuan)

Event	Weibull	RSF
Recurrence	1.99 (1.79, 2.24)	1.82 (1.64, 2.08)
DAR	2.24 (1.78, 3.08)	1.87 (1.5, 2.67)

Neither RSFs nor the Weibull model differ egregiously in terms of the width of each CI within an event type. DAR does have greater estimated PE and wider CIs, which is unsurprising given that the sample size for the DAR analysis is much smaller than that for the recurrence analysis.

Where the C-index does not demonstrate great differences between the predictive abilities of RSFs and the Weibull model, the Lawless-Yuan estimates are slightly lower for RSFs. Since the CIs based on these estimates overlap, we cannot make definitive conclusions about the relative performance of the two methods, but all else being equal, the slightly smaller observed prediction error would seem to justify use of RSFs. Thus, we would tentatively recommend using RSFs in lieu of parametric methods for prediction of ovarian cancer survival times.

Chapter 6

Discussion

While preliminary results indicate that the RSFs may have better performance, there are some caveats. Firstly, peculiarities of this particular dataset may greatly influence our results and the missing covariate and measurement time data may obscure the true relationship between predictors and the events of interest.

All of our models were trained on subsets of the same data and so the estimates of prediction error were also based on the same dataset, despite all attempts to conduct pseudo-independent error evaluation through cross-validation. Cross-validation itself is not without faults: it can lead to overestimates of error (Hastie et al, 2009) and the number of folds will impact the bias-variance trade-off mentioned in Section 3.3.1. However, the comparisons between methods should still be valid if both utilize cross-validation.

One of the other issues that arose with having a small dataset is the re-estimation of the censoring distribution for each fold. By re-estimating the censoring distribution for each fold and each new dataset B^* (as we did for DAR), the estimates of PE incorporate the uncertainty about the censoring distribution. In contrast, as a result of estimating the IPCW weights only once (as we did for recurrence), the estimates of PE will not. In other words, we are implicitly assuming that the IPCW weights for recurrence are from the true censoring distribution, which is not likely the case.

Furthermore, the choice of loss function will have an effect on the bias of PE (Hastie et al, 2009). Efron (2004) suggests several bias adjustment penalties for different loss functions, although the absolute loss is not among them. Further work is warranted on the behavior and performance of potential loss functions in a survival setting.

Our ability to gauge the performance of a given method by using the C-index would be augmented by CIs; however, creation of such intervals is not addressed by either Lawless &

Yuan (2010) or Ishwaran et al (2008). We have proposed approximately normal confidence intervals for a bootstrap-based method but its performance has not yet been explored.

The predictive abilities of our methods may be improved by the addition of new predictors that perform better than those available to us – or the subtraction of unimportant variables. While it is beyond the scope of this project, assessing the impact of geographic location of patients should be undertaken as well. It is possible that the area in which patients lived and their distance from Calgary are important predictors of survival times and missingness.

A larger dataset will allow for independent training and test sets. We are expecting such a dataset from Winnipeg to be available within the next one or two years. We are optimistic that more conclusive results concerning the performance of our prediction methods will follow. In addition, our current results are confined to those patients for whom treatment was successfully completed and the analysis of patient survival limited to those who experienced recurrence of their disease. A larger dataset would permit the analysis of death times prior to recurrence and the application of more sophisticated methods for handling missing data.

Finally, our choice of tuning parameters in RSFs may also impact our estimated prediction error.

We hope that our work will provide a starting point for creating a tool for patients and physicians to generate accurate predictions of time until recurrence and post-recurrence survival. We also recommend the development of methods for obtaining prediction intervals for a given patient. Naively, we might assume that the estimated survival function is the true underlying survival function and build prediction intervals based on the quantiles of this function. However, such intervals would not incorporate uncertainty regarding the estimated survival function. Lawless (2003) explores more nuanced methods in the case where a pivotal function not relying on estimated parameters can be used to create exact prediction intervals for new data. Lawless also briefly addresses how to build prediction intervals based on non-parametric simulation under different types of censoring and using Bayesian methods. While there is much to be done on this topic, we hope that our preliminary analysis will provide a solid foundation for subsequent research.

Bibliography

Altman, A., Nelson, G., Chu, P. Nation, J., and P. Ghatage (2012). “Optimal Debulking Targets in Women With Advanced Stage Ovarian Cancer: A Retrospective Study of Immediate Versus Interval Debulking Surgery”, *Journal of Obstetric Gynaecology*, 34 (6); 558-566.

Asher, V., Lee, J., and A. Bali (2011). “Preoperative Serum Albumin is an Independent Prognostic Predictor of Survival in Ovarian Cancer”, *Medical Oncology*, 29 (3); 45-97.

Barlin, J.N., Yu, C., Hill, E.K, Zivanovic, O., Kolev, V., Levine, D.A., Sonoda, Y., Abu-Rustum, N.R., Huh, J., Barakat, R.R., Kattan, M., and D.S. Chi (2012). “Nomogram for Predicting 5-year Disease-Specific Mortality After Primary Surgery for Epithelial Ovarian Cancer”, *Gynecologic Oncology*, 125 (1); 25-30.

Bou-Hamad, I., Larocque, D.,and H. Ben-Ameur (2011). “A Review of Survival Trees”, *Statistics Surveys*, 5 (1); 44-71.

Du XL, Parikh RC, Lairson DR, Giordano SH, and P. Cen (2013). “Comparative Effectiveness of Platinum-Based Chemotherapy Versus Taxane and Other Regimens for Ovarian Cancer”, *Medical Oncology*, 30 (1); 440.

Efron, B. (2004) “How Biased is the Apparent Error Rate of a Prediction Rule?”, *Journal of the American Statistical Association*, 81 (394); 461-476.

Hastie, T., Tibshirani, R.,and J. Friedman (2009). *The Elements of Statistical Learning, Second Edition*, Springer. 587-604.

Henderson, R., Jones, M., and J. Stare (2001). “Accuracy of Point Predictions in Survival Analysis”, *Statistics in Medicine*, 20 (1); 3083-3096.

Ishwaran, H. and U.B. Kogalur (2013). “Package ’randomSurvivalForest’ ”, *R Documentation*, CRAN Repository. Dec. 23rd, 2013.

Ishwaran, H. and U.B. Kogalur (2007). “Random Survival Forests for R”, *R News*, 7 (2); 25-31.

Ishwaran, H. Kogalur, U.B., Blackstone, E.U., and M.S. Lauer (2008). “Random Survival Forests”, *The Annals of Applied Statistics*, 2 (3); 841-860.

Klein J.P., and M.L. Moeschberger (2003). *Survival Analysis: Techniques for Censored and Truncated Data*, 22-61.

Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data, 2nd Edition*, 269-329.

Lawless, J.F., and Y. Yuan (2010). “Estimation of Prediction Error for Survival Models”, *Statistics in Medicine*, 29 (2); 262-274.

Loh, W. and Y. Shih (1997). “Split Selection Methods for Classification Trees”, *Statistica Sinica*, 7; 815-840.

Loughin, T. (2012). “Model Assessment and Multi-Model Inference”, *Statistics 890: Modern Applied Statistics*, Lecture on September 11, 2012.

Loughin, T. (2012). “Regression Trees”, *Statistics 890: Modern Applied Statistics*, Lecture on October 15, 2012.

National Cancer Institute at the National Instutites of Health. (2013) “A Snapshot of Ovarian Cancer”, *http://www.cancer.gov/researchandfunding/snapshots/ovarian*, Accessed 12/06/2013.

Rocconi, R.P, Matthews, K.S., Kemper, M.K., Hoskins, K.E., Huh, W.K., and J.M. Straughn Jr.(2009). “The Timing of Normalization of CA-125 Levels During Primay Chemotherapy is Predictive of Survival in Patients with Epithelial Ovarian Cancer”, *Gynecologic Oncology,* 114 (2); 242-245.

Rocconi, R.P, Matthews, K.S., Kemper, M.K., Hoskins, K.E., and M.N. Barnes (2008). “Chemotherapy-related Myelosuppresision as a Marker of Survial in Epithelial Ovarian Cancer Patients”, *Gynecologic Oncology,* 108 (2); 336-341.

Yuan, Y. (2008). “Prediction Performance of Survival Models”, *PhD Thesis, University of Waterloo.*

Vergote I., Trop C.G., Amant F., Kristensen G.B., Ehlen T., Johnson N., Verheijen R.H., van der Burg M.E., Lacave A.J., Panici P.B., Kenter G.G., Casado A., Mendiola C., Coens C., Verleye L., Stuart G.C., Pecorelli S., and N.A. Reed (2010). "Neoadjuvant Chemotherapy or Primary Surgery in Stage IIIC or IV Ovarian Cancer", *New England Journal of Medicine*, 2; 363 (10); 943-953.

Appendix A

Model Fit Diagnostics

A.1 Observed vs. Predicted Value Plots

A.1.1 Recurrence

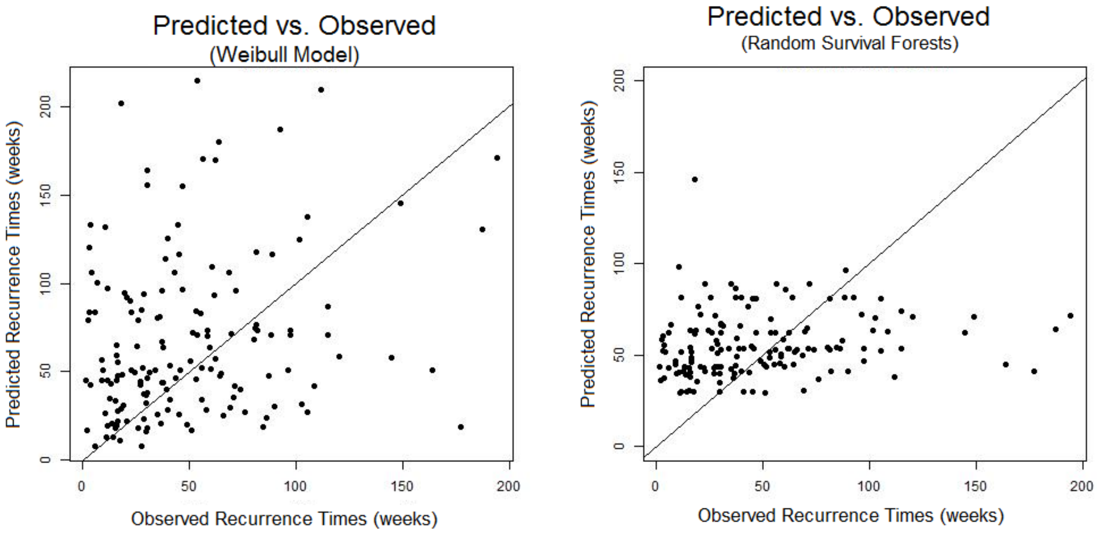


Figure A.1: Predicted vs. Observed Recurrence Times

A.1.2 DAR

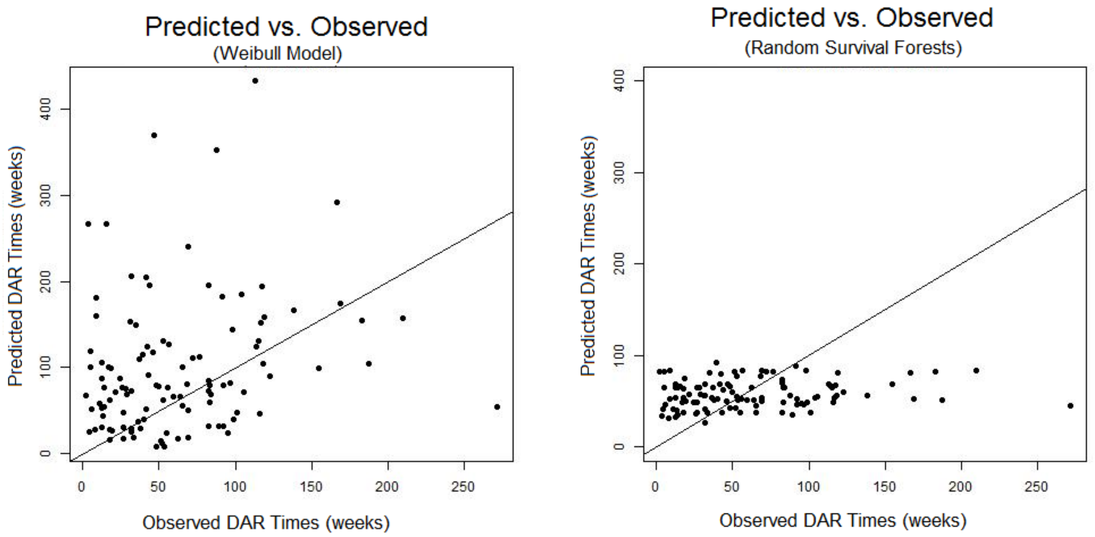


Figure A.2: Predicted vs. Observed DAR Times

A.2 Weibull Residuals

Residuals for the Weibull model were derived using the probability intergral transform on the CHF (Lawless, 2003). If our specified model is correct, the standard exponential distribution will approximately describe the distribution of these residuals. The CHF for the Weibull is

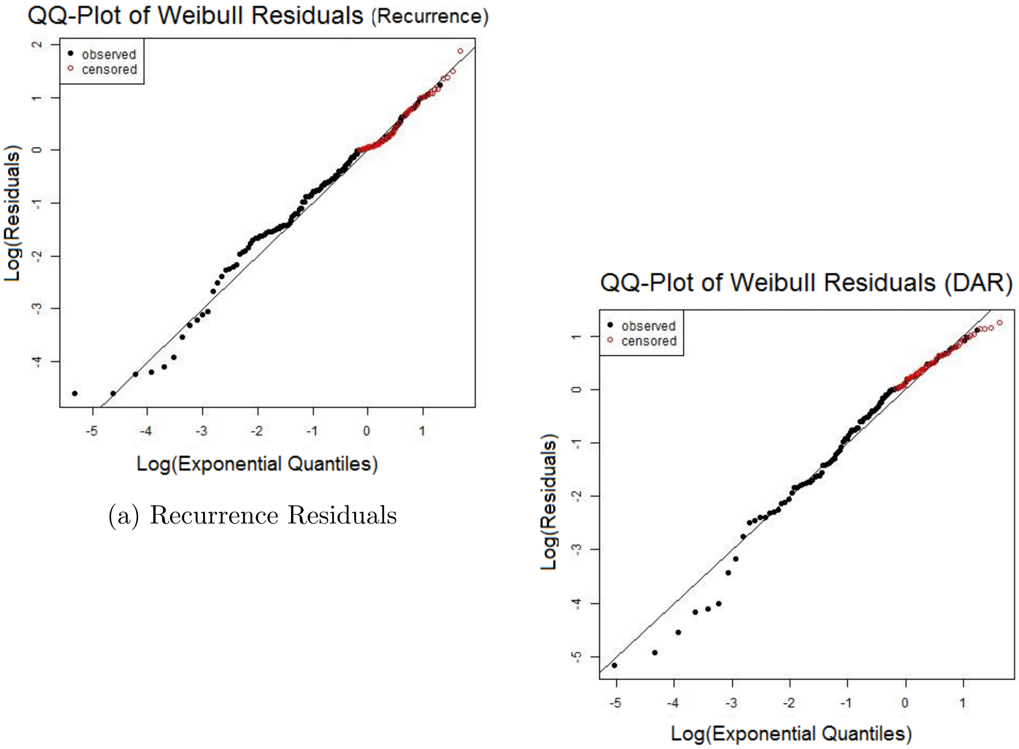
$$H_i(t) = \frac{1}{\phi_i^{\alpha}} t^{\alpha}$$

and so we formulate residuals as

$$e_i = \left(\frac{y_i}{\phi_i}\right)^{\alpha} + (1 - \delta_i) \ .$$

The residuals can be highly skewed so taking the log of both the residuals and the appropriate standard exponential quantiles is recommended. This transformation facilitates the detection of observations with relatively small associated residuals that deviate from our proposed model. For more details, see Lawless (2003).

Plots of the Weibull residuals for both recurrence times and DAR times are shown in Figure A.3.



(a) Recurrence Residuals

(b) DAR Residuals

Figure A.3: QQ-Plots

Appendix B

Final Predictors

All predictors were included in both models, with the exception of time until recurrence, which was included only in the DAR models. Table B.1 contains the full list of predictors.

Table B.1: Predictors

Hematologic Predictors	Clinical Predictors	Surgical Predictors
Min Hemoglobin	Age	Surgery Type
Avg Hemoglobin	Stage	Optimally Debulked (binary)
Min Neutrophil Count	Grade	Blood Loss
Avg Neutrophil Count	Presence of Ascites (binary)	Surgery Type x Debulking
Min Platelet Count	Platinum Containing Chemo (binary)	Surgery Type x Ascites
Max Platelet Count	# of Neoadjuvant Cycles	Surgery Type x Blood Loss
Avg Platelet Count	# of Primary Adjuvant Cycles	
Min CA 125 Value	Prescribed Growth Factor (binary)	
Difference (First, Last CA 125 Values)	Treatment Length	
Min Albumin Level	Time to Recurrence (DAR models only)	
Difference (Avg White Blood Cell, Avg Neutrophils)		