Predicting Ovarian Cancer Survival Times: Feature Selection and Performance of Parametric, Semi-Parametric, and Random Survival Forest Methods

by

Vinnie Liu

B.Sc., Simon Fraser University, 2017

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2019

Approval

Name:	**Vinnie Liu**
Degree:	**Master of Science (Statistics)**
Title:	**Predicting Ovarian Cancer Survival Times: Feature Selection and Performance of Parametric, Semi-Parametric, and Random Survival Forest Methods**
Examining Committee:	**Chair:** Jinko Graham Professor
	Rachel Altman Senior Supervisor Associate Professor
	Thomas M. Loughin Supervisor Professor
	Jiguo Cao Internal Examiner Professor
Date Defended:	**April 23, 2019**

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

Survival time predictions have far-reaching implications. For example, such predictions can be influential in constructing a personalized treatment plan that is of benefit to both physicians and patients. Advantages include planning the best course of treatment considering the allocation of health care services and resources, as well as the patient's overall health or personal wishes. Predictions also play an important role in providing realistic expectations and subsequently managing quality of life for the patient's residual lifetime. Unfortunately, survival data can be highly variable, making precise predictions difficult or impossible.

This project explores methods of predicting time to death for ovarian cancer patients. The dataset consists of a multitude of predictors, including some that may be unimportant. The performances of various prediction methods that allow for feature selection (the Weibull model, the Cox proportional hazards model, and the random survival forest) are evaluated. Prediction errors are assessed using Harrell's concordance index and a version of the expected integrated Brier score.

We find that the Weibull and Cox models provide the best predictions of survival distributions in this context. Moreover, we are able to identify subsets of predictors that lead to reduced prediction error and are clinically meaningful.

Keywords: censored data, survival analysis, feature selection, prediction

Dedication

In loving memory of my mom.

"'Tis a fearful thing
to love what death can touch.
A fearful thing
to love, to hope, to dream, to be -
to be,
And oh, to lose.
A thing for fools, this,
And a holy thing, a holy thing
to love.
For your life has lived in me,
your laugh once lifted me,
your word was gift to me.
To remember this brings painful joy.
'Tis a human thing, love,
a holy thing, to love
what death has touched."

-Yehuda HaLevi

Acknowledgements

First and foremost, I would like to thank Dr. Rachel Altman and Dr. Tom Loughin. I would not be here in this program without them, nor would this work be possible without their invaluable academic insight and guidance.

Rachel, I am so grateful for you. It has truly been a privilege to be your student. Thank you for your patience, kindness, understanding, and support. This process has been extremely difficult, but I cannot begin to imagine how much harder it would have been without you. Your belief in me has meant so much. I genuinely would not have made it without you.

Tom, I am thankful for your guidance and for the opportunity to have learned from you over the last few years. Not only were your classes a highlight of my academic career at SFU, they proved to be essential in completing this project.

I would like to extend a special thank you to Dr. Alon Altman for providing the opportunity to be involved in a project that is so deeply important to me and my family. I would also like to thank my classmates for their friendship and for enriching this learning experience. It has been a pleasure learning with, and from, you all.

Lastly, I would like to thank my family and friends for their continued love and support. To my sister, thank you for always being there for me and for providing an immense amount of emotional support. To my dad, thank you for trying so hard to help me through this process, even when you didn't know how. To Paula, Tim, and Celeste, thank you for providing extra love through these especially tough last few years and for opening your homes when I needed to escape. To Erin, Jacquie, Marissa, and Tara, thank you for your lasting friendship and continued encouragement.

Many thanks from the bottom of my very full heart. I truly am so grateful for you all.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Ovarian cancer is the fifth most common cancer among women and is considered the most serious cancer of the female reproductive system (Ovarian Cancer Canada, 2017). While ovarian cancer develops in only 1.45% of women and accounts for only 3% of cancers in women, it is ranked as the fifth deadliest cancer to afflict women (Canadian Cancer Society, 2017). Once the cancer has been diagnosed, a patient will typically undergo one or more types of treatment, provided she is healthy enough to receive it. In most cases, treatment consists of surgery, chemotherapy, or both, and takes factors such as cancer stage and tumour grade into consideration. An accurate method of prognosis would be highly beneficial for both physicians and patients in planning the best course of treatment, managing expectations, and optimizing quality of life.

This project builds on the work of Lipson (2014), who investigates the Weibull model and random survival forests (RSFs) as methods of predicting time to recurrence (defined as progression of disease) and time to death after recurrence (DAR) on 204 ovarian cancer patients from the Tom Baker Cancer Centre in Calgary. We use a larger data set consisting of records of not only the 204 Calgary patients but also of 255 patients from CancerCare Manitoba in Winnipeg.

This project differs from that of Lipson (2014) in a number of other ways. Most importantly, we explore feature selection for each predictive method. Feature selection is critical from a clinical perspective (predictions on future patients will be simpler if values of only a few predictors are required), a medical research perspective (identifying key predictors is of medical interest and will reduce the cost of future data collection), and a predictive perspective (including unimportant variables may reduce the predictive ability of the methods under consideration).

In addition to the (parametric) Weibull model and (non-parametric) RSF studied by Lipson (2014), we also consider the semi-parametric Cox proportional hazards model (Cox, 1972) as a predictive method. Although many methods of predicting survival times are avail-

able, these three methods are of particular interest because they produce predicted survival distributions, not simply point predictions. Point predictions can be highly imprecise and thus of limited value, as pointed out by Henderson and Keiding (2005). Moreover, from a clinical perspective, a survival distribution is more useful than a point prediction. Given our specific goals, feature selection methods investigated in this project include iterative backward selection, grouped LASSO, and iterative random forest procedures. To compare prediction error across methods, we use not only Harrell's C-index (Harrell *et al.*, 1982), the measure used by Lipson (2014), but also a novel version of the expected integrated Brier score.

We emphasize that our goal is to identify a model and feature selection method that yield the most accurate predictions for the population of ovarian cancer patients represented in our dataset (i.e., the population of patients who are subject to the treatment protocol followed at the Winnipeg and Calgary centres, among other possible criteria). We are not attempting to determine a best method for use more generally.

The remainder of this report is organized as follows. In Chapter 2, we describe the ovarian cancer data in detail. We outline the statistical methods used in Chapter 3. We provide our results and recommendations in Chapter 4 and conclude with a discussion and some comments on possible future work in Chapter 5.

Chapter 2

Ovarian Cancer Data

The individuals in this study include the 204 patients from Calgary described in the Lipson (2014) study and an additional 255 ovarian cancer patients from CancerCare Manitoba in Winnipeg. The Calgary data were collected from January 2003 to December 2007 and the Winnipeg data were collected from January 2007 to December 2010, both by retrospective chart review. The resulting dataset consists of various surgical, clinical, and hematologic (blood marker) measurements collected over the course of treatment on a total of 459 patients. In this section, we describe the treatment protocol followed at both centres. We then describe the response and predictor variables in detail.

2.1 Overview of Treatment

After diagnosis, most women undergo surgery. The patient's health care team decides the best course of treatment considering relevant factors such as the patient's age, stage, general health, and overall ability to endure the demands and side effects of treatment. Depending on how advanced or widespread the cancer is, the oncologist may prescribe chemotherapy to shrink cancerous tissues prior to surgery (neoadjuvant chemotherapy). Otherwise, the patient will undergo surgical debulking to remove as much of the tumour as possible (Canadian Cancer Society, 2019d).

Following surgery, most women are treated with chemotherapy (known as primary adjuvant chemotherapy) to destroy any residual cancer cells and to reduce the risk of recurrence. Chemotherapy generally includes a combination of two or more drugs, administered intravenously every 3-4 weeks. Common treatments include a combination of platinum-containing drugs, such as carboplatin or cisplatin, and taxane-containing drugs, such as paclitaxel or docetaxel. Different drugs may be prescribed if the patient has adverse reactions or if the cancer does not respond to the standard chemotherapy agent. The type of drugs used may depend on whether the patient develops resistance to the platinum agent, in which case other non-platinum drugs may be used. If the patient becomes too frail (for reasons related or unrelated to the cancer) to continue chemotherapy or tolerate the side effects,

the oncologist may adjust dosage or postpone chemotherapy until she recovers (Canadian Cancer Society, 2019d).

2.2 Responses of Interest

Once the patient has completed her treatment, the events of interest are the time to recurrence and the time to death. We consider time to death from the end of treatment (defined as time to death from diagnosis date for untreated patients). Cause of death is not specified in the dataset so death may or may not be a result of ovarian cancer.

2.3 Hematologic Predictors

Chemotherapy is intended to target cancerous cells; however, other cells (e.g., blood cells, which divide rapidly) may also be affected. An important aspect of treatment is monitoring blood cells, especially since the patient may already have a compromised immune system. Normal numbers in the complete blood cell count (which includes red blood cells, hemoglobin, white blood cells, and platelets) can indicate that the patient is healthy enough to continue chemotherapy, while small changes in the blood cell counts may indicate that chemotherapy is effective. Conversely, if blood cell counts are too low, the patient may be at risk for infections and be required to postpone treatment until she is stronger (Canadian Cancer Society, 2019b).

As such, each patient received a series of regular blood tests to monitor her blood cell counts throughout the course of her treatment. The following section includes a description of each hematologic predictor measured.

Albumin

Albumin is a protein found in blood's plasma that helps to maintain blood volume. It can be considered to be a measure of nutritional status, which can greatly influence a patient's overall health. Tests are used to assess protein stores or identify deficiencies. Furthermore, albumin is produced by the liver and can serve as an indicator of liver function. Low levels suggest liver damage, which may compromise the patient's ability to process medications, resist infections, turn food into energy, and recover in general (Canadian Liver Foundation, 2017). As low levels suggest a negative effect on prognosis, we include minimum albumin measurements as a predictor.

CA 125

Cancer antigen 125 (CA 125) is a protein found on most ovarian cancer cells. A CA 125 test may be performed to carry out a potential diagnosis, to determine if prescribed treatment is effective, or to investigate if the cancer has recurred. However, caution should be exercised when measuring CA 125 since it can be found on normal cells in patients

in non-cancerous conditions such as menstruation and pregnancy. Generally, a decrease in CA 125 levels during treatment is an indication that the cancer is responding positively to treatment, whereas unchanged or increased levels may indicate resistance. Similarly, high CA 125 levels after treatment may indicate recurrence (Canadian Cancer Society, 2019a). We therefore include minimum and maximum CA 125 levels as predictors.

Additionally, we created a predictor that reflects the difference between CA 125 measurements. Rocconi *et al.* (2009) found that lower levels, specifically at the end of treatment, lead to better prognosis and longer survival. Thus, in addition to the minimum and maximum CA 125 levels, we include the difference between first and last measurements to summarize changes in CA 125 levels over the course of treatment.

Hemoglobin

Hemoglobin is the iron-containing component of red blood cells that carry oxygen from the lungs to the rest of the body. When levels fall below 100g/L, the patient is considered anemic and may present with additional health problems. Anemia is especially a concern for those with weakened immune functions, including cancer patients. Cancer and its corresponding treatments can affect bone marrow, resulting in low levels of healthy red blood cells and consequently, hemoglobin levels. Patients may receive blood transfusions to relieve symptoms of anemia; however, multiple transfusions may be required before the bone marrow can replenish its own healthy red blood cells and return hemoglobin to normal levels (Canadian Cancer Society, 2019c). Since hemoglobin levels may reflect overall patient health, with persistently low levels specifically suggesting illness, we consider both minimum and mean hemoglobin levels as predictors.

Platelets

Platelets are a component of the blood produced in the bone marrow. The function of platelets is to react to damaged blood vessels by clumping, thereby initiating clots to stop further bleeding. High levels of platelets can lead to spontaneous blood clots and subsequent heart attacks or strokes. Conversely, low levels can lead to uncontrollable bleeding, which may complicate or delay scheduled surgeries (Johns Hopkins Medicine, 2019). Thrombocytopenia, resulting from low levels, may develop if the bone marrow does not create enough platelets, which can be a side effect of cancer and corresponding treatments. If the treatment itself is responsible for the low counts, the chemotherapy dosage may need to be adjusted or the recovery time between chemotherapy cycles may need to be re-evaluated (Canadian Cancer Society, 2019c). As platelet levels at either extreme may indicate poor health, we consider the minimum, mean, and maximum platelet levels as predictors.

White Blood Cell Count

White blood cells (WBCs) are the cells of the immune system responsible for protecting

the body and assisting in fighting off infection, disease, and other foreign attackers. The body experiences leukopenia when the total WBC count decreases. Leukopenia can be caused by chemotherapy or deficiencies in the bone marrow. If the WBC count is too low, the body may not be physically able to endure treatment and may be at greater risk of infection. Chemotherapy may need to be postponed or administered at a lower dose (Canadian Cancer Society, 2019c). We consider mean WBC count as a predictor.

2.4 Surgical Predictors

Surgery is the main avenue of treatment for most ovarian cancers. It plays a vital role in not only treating cancer but also in diagnosis and prognosis, as it is used to assess how widely the cancer has spread in the body.

Depending on the type of ovarian cancer, the goal of surgery may be to stage the cancer and debulk, or simply to debulk. Staging, i.e., identifying how widely the cancer has spread, is necessary for prescribing appropriate treatment. Debulking, on the other hand, refers to removing as much of the disease or tumour as possible. Debulking is especially important if the cancer has spread throughout the abdomen. The goal of debulking is to minimize the residual disease left after the surgery. If the thickness of residual disease is <1 cm, the patient is considered to be "optimally debulked". Patients who have been optimally debulked typically have a better prognosis than those who have been sub-optimally debulked (American Cancer Society, 2019).

Other surgical predictors in this dataset include blood loss (mL) and the amount of ascites (mL) measured during surgery. Ascites, the build-up of fluid in the abdomen, can result from high pressure in the blood vessels of the liver and low levels of albumin. It is believed that patients with certain cancers located in the abdomen, such as in the ovaries, are especially susceptible to developing ascites (Canadian Cancer Society, 2019c). Importantly, debulking, blood loss, and ascites are measured only if the patient has had surgery.

2.5 Clinical Predictors

The other primary treatment of ovarian cancer is chemotherapy, as described above. The dataset contains information on each patient's chemotherapy schedule, including treatment date and type of chemotherapeutic agent used for each cycle. From the treatment dates and surgery dates, we extracted the number of neoadjuvant and primary adjuvant chemotherapy cycles for each patient, where applicable. We also created variables to indicate whether the neoadjuvant drugs contained platinum or taxane agents.

Other clinical predictors included in the dataset are patient age at time of diagnosis and tumour grade (the degree to which the cancer cells are differentiable from normal cells). We

were also able to create the variables total treatment length and time to death/censoring using the diagnosis dates, last treatment dates, death dates, and last contact dates provided in the dataset.

2.6 Missing Data

For simplicity, we treat the missing predictor values as missing at random. For continuous variables, we replace missing values with the median of the available values. For categorical variables, we replace missing values with the mode of the available values. We acknowledge that this form of data imputation can be risky since it may not preserve the relationship between the variables and the response. However, the number of missing records for most predictors was relatively small: <5% per covariate in the Calgary dataset and approximately 12% per covariate in the Winnipeg dataset (except for albumin, which was missing 21% of the values).

For this project, we exclude all measures of neutrophils (a component of WBCs) as predictors. Lipson (2014) includes mean and minimum neutrophil levels, as well as the difference between mean WBC count and mean neutrophil count. However, approximately 80% of Winnipeg patients were missing neutrophil measurements; imputing so many missing values could distort the apparent relationship between survival time and neutrophil levels and thus seemed inadvisable.

Chapter 3

Statistical Methods for Survival Data

Survival data require special considerations for analysis. Subjects included in a study may not necessarily experience the event of interest during the observation period, resulting in "censored", or incomplete, observation times. Censoring can occur in a variety of settings; the most common type in survival data is "right censoring". Patients' survival times are right-censored if they survive beyond the end of study date or if they drop out (i.e., are "lost to follow-up"). In these cases, the exact survival times are unknown; we know only that they are greater than the censoring times (Lawless, 2003, p. 52–55).

In this project, we need to handle two types of censoring: Type I and independent random censoring. Type I censoring may occur when subjects are observed for a fixed period of time (that concludes before some of the subjects have died), while independent random censoring may occur when censoring time is random and independent of failure time. The Winnipeg data are subject to Type I censoring since the end of the study is fixed and common to all patients. No patients dropped out prior to the end of the study. The Calgary data, on the other hand, are subject to independent random censoring. In particular, Calgary patients were followed for varying lengths of time and some drop-out occurred (presumably at random).

The exact survival time of individual i will be observed if $T_i < C_i$, where T_i is the time of death and C_i is the censoring time. As such, for individual i, we observe time $Y_i = \min(T_i, C_i)$ and the censoring indicator, $\delta_i = 1$ if $Y_i = T_i$ and $\delta_i = 0$ if $Y_i = C_i$.

3.1 Models

In this section, we discuss the survival time models that we use for analyzing the ovarian cancer data. We consider parametric, semi-parametric, and nonparametric models.

3.1.1 Weibull Model

For this project, we use a Weibull model (a fully parametric model) to describe the responses of ovarian cancer patients. The Weibull model is commonly used in many applications of lifetime data due to its flexibility (relative to the exponential model) and straightforward form (Lawless, 2003, p. 18–19). The Weibull probability density function is

$$f(t) = \frac{\lambda}{\theta}\left(\frac{t}{\theta}\right)^{\lambda-1} \exp\left[-\left(\frac{t}{\theta}\right)^{\lambda}\right], \tag{3.1}$$

the survivor function is

$$S(t) = \exp\left[-\left(\frac{t}{\theta}\right)^{\lambda}\right], \tag{3.2}$$

and the cumulative hazard function is

$$H(t) = -\log[S(t)],$$

for $t > 0$, $\lambda > 0$, and $\theta > 0$. We assume that the shape parameter, λ, is constant and the scale parameter, θ, varies across patients according to their covariates, $\boldsymbol{x_i}$, with $\log(\theta_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$.

The parameters of the model, $\boldsymbol{\psi}$, can be estimated using the method of maximum likelihood. The likelihood can be expressed as

$$L(\boldsymbol{\psi}; \boldsymbol{y}, \boldsymbol{\delta}) = \prod_{i=1}^{N} f_i(y_i; \boldsymbol{\psi})^{\delta_i} S_i(y_i; \boldsymbol{\psi})^{1-\delta_i}, \tag{3.3}$$

where $f_i(\cdot; \boldsymbol{\psi})$ is the probability density function and $S_i(\cdot; \boldsymbol{\psi})$ is the survival function for the i^{th} patient, $i = 1, \ldots, N$.

Estimating the effects of the surgery-specific variables requires special consideration. As mentioned in Section 2.4, debulking, blood loss, and ascites are measured only if the patient has had surgery. We provide the details of how we specified these columns in the design matrix in Appendix A.

We fit the model using the `survreg` function in the R package `survival`. Hazard-based residuals can then be used to check the fit of the model. As defined by Lawless (2003, p. 283–286), these residuals are calculated as the cumulative hazard function of T_i given $\boldsymbol{x_i}$ with a simple adjustment for censoring. In particular, the residuals are defined as

$$\hat{e_i} = \hat{H}(y_i|\boldsymbol{x_i}) + (1 - \delta_i). \tag{3.4}$$

If the Weibull model is correct, the residuals will behave like a random sample from the standard exponential distribution.

3.1.2 Cox Proportional Hazards Model

Parametric methods for analyzing survival data can work well if the model is specified correctly, however choosing the right distribution can be a challenge. Next, we consider an alternative method that offers more flexibility: the Cox proportional hazards model (Cox, 1972).

Letting $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$ (no intercept term included), the hazard function associated with the Cox proportional hazards model for the i^{th} patient is

$$h(t|\boldsymbol{x}_i) = h_0(t)\exp(\eta_i) \tag{3.5}$$

and the survivor function is

$$S(t|\boldsymbol{x}_i) = S_0(t)^{\exp(\eta_i)}, \tag{3.6}$$

where $h_0(t)$ is the baseline hazard function, i.e., the hazard for an individual with covariate vector $\boldsymbol{x} = \mathbf{0}$, and $S_0(t) = \exp[-H_0(t)]$.

The Cox model is referred to as semi-parametric since the covariates are entered into the model parametrically but the so-called baseline hazard function is left unspecified – the reason for its popularity. However, as the name implies, it assumes proportional hazards, which is a "relatively strong assumption" (Kalbfleisch and Prentice, 2002, p. 95). That is, the hazard ratio for two individuals with covariate vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, respectively, is

$$\frac{h(t|\boldsymbol{x}_i)}{h(t|\boldsymbol{x}_j)} = \frac{h_0(t)\exp^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{h_0(t)\exp^{\boldsymbol{x}_j'\boldsymbol{\beta}}} = \exp^{(\boldsymbol{x}_i'-\boldsymbol{x}_j')\boldsymbol{\beta}}, \tag{3.7}$$

which does not depend on t, and is thus constant over time.

For estimating $\boldsymbol{\beta}$, rather than maximizing the likelihood, Cox suggests maximizing the partial likelihood,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k}\left(\frac{\exp^{\boldsymbol{x}_{(i)}'\boldsymbol{\beta}}}{\sum_{\ell\in R_i}\exp^{\boldsymbol{x}_{(\ell)}'\boldsymbol{\beta}}}\right), \tag{3.8}$$

which does not depend on $h_0(t)$. Here, $\boldsymbol{x}_{(i)}$ is the vector of predictors corresponding to the patient with the i^{th} largest *observed* death time, $t_{(i)}$, $i = 1, \ldots, k$, and R_i is the set of individuals still at risk (i.e., alive and still under observation) just prior to time $t_{(i)}$ (Lawless, 2003, p. 342).

Using the same definitions of the surgery-specific variables outlined in Appendix A, we fit the model using the `coxph` function in the R package `survival`. Hazard-based residuals, as defined in Section 3.1.1, can also be used to check the fit of the Cox model.

3.1.3 Random Survival Forests

While the Weibull and Cox proportional hazards models are interpretable and allow for inference about the regression coefficients, the relationship between the response (via the hazard function) and the predictors must be specified before the analysis can be performed. A simpler option for prediction that does not require any distributional specifications is the random survival forest.

The fundamental building block of a random survival forest is the survival tree. The survival tree is formed in a similar way as in the classic classification and regression tree (CART) approach (Hastie *et al.*, 2009, p. 305–317). However, in the survival context, our response is subject to censoring. As such, alterations to the CART approach (namely, the stopping criterion and the method for evaluating the split selection) are required to handle the censored observations. In the survival tree, the typical stopping criterion based on the minimum number of observations in a node refers specifically to the number of *observed* events in that node (Ishwaran *et al.*, 2008). In terms of evaluating the split selection, survival trees use the difference between estimates of the survival functions at each child node. Commonly, nonparametric methods (such as the log-rank test statistic computed based on the Kaplan-Meier estimates of the two functions) are used to quantify this difference, but parametric methods (such as the likelihood ratio test statistic computed by assuming a distribution for the survival times in each node) may also be used (Bou-Hamad *et al.*, 2011). The distributions of the survival times in each terminal node are also estimated and are used for predictions.

Single trees can be appealing as they are easy to construct and to interpret. However, they are known to be highly problematic since they typically produce highly variable, unstable predictions. We do not consider them as a stand-alone prediction method in this project. However, we do use trees as the foundation of a much more effective method: the random survival forest. The random survival forest algorithm involves growing many survival trees from repeated sampling of the original data and averaging the predicted survival functions produced by each individual tree (Bou-Hamad *et al.*, 2011).

The algorithm, as described by Ishwaran *et al.* (2008), is summarized as follows:

1. Obtain a dataset of size N by sampling with replacement from the original data.
2. Fit a survival tree to the sampled dataset. Specifically, at each node, select a random subset of `mtry` variables. Next, select the variable and split that maximize the difference between the distributions of the responses in the child nodes (according to the log-rank test statistic).
3. Grow the tree until the stopping criterion (based on the minimum number of observed deaths per terminal node) is met.

4. Repeat Steps 1–3 to obtain B trees.

The predicted outcome of an individual is calculated by dropping the observation down each of the B trees that is grown without that particular observation (i.e., all trees for which this observation is "out-of-bag"). A version of the Nelson-Aalen estimates of the CHFs corresponding to the resulting terminal nodes are averaged to obtain a final estimated ensemble CHF. Specifically, for each terminal node, the Nelson-Aalen estimate of the CHF is computed up to the largest observed death time in the node. The value of the function at that time is then used as the estimated CHF for all times between the largest observed death time in the node and largest observed death time in the dataset (so that each tree produces an estimated CHF that is defined over the same domain). As a result, the ensemble CHF is, in fact, a *lower bound* on the estimated CHF. Likewise, the ensemble survivor function is an *upper bound* on the estimated survivor function.

Typically, tuning parameters should be set before fitting the model. The number of trees grown (`ntree`), the number of random splits per variable (`nsplit`), the number of variables randomly selected at each node (`mtry`), and the splitting rule used to maximize survival difference (`splitrule`) are the parameters that should be specified. For this project, we used the default values of the tuning parameters provided in the `rfsrc` function in the R package `randomForestSRC` (except for `ntree`, which we set to the high value of 5000). In particular, because of the iterative nature of the feature selection process (described in more detail in Section 3.3.3), we use the default value of `mtry`=$\sqrt{p_j}$, where p_j is the total number of predictors available at iteration j. This choice allows `mtry` to vary with p_j across iterations.

Similarly, we use the default random log-rank splitting rule (where nodes are split by maximizing the log-rank test statistic) since Ishwaran *et al.* (2008) demonstrate that the log-rank rule performs well. The random component of the log-rank splitting rule selects a random split for each of the `mtry` variables instead of considering all possible splits, resulting in increased computational speed. If any of the log-rank test statistics is significant, the node is split: the split is selected based on the variable (and random split point) that result in the largest significant test statistic value (Ishwaran *et al.*, 2008).

3.1.4 Other Prediction Methods

We conducted an extensive literature search to explore other established prediction methods that are capable both of handling censored data and of performing feature selection. We identified a number of such methods for point predictions (which we discuss briefly below for completeness) but none for predicting survival distributions, our primary goal. Therefore, we did not implement or investigate these methods further.

Alternative methods for point predictions of survival times include extensions to the support vector machine. Shivaswamy *et al.* (2007) propose a modification to standard support vector regression (SVR) by including the censoring indicator in the constraints. Van Belle *et al.* (2011) propose SVR models that include ranking constraints; comparable pairs of observations are used to predict risk ranks instead of survival times.

Gradient boosting is another method for obtaining point predictions of survival times. Chen *et al.* (2013) propose a gradient boosting algorithm that use an approximation of the concordance index, called the "smoothed concordance index", as the loss function.

3.2 Assessing Prediction Error

To compare methods of prediction, we require a way to quantify prediction error. Common ways to measure the difference between the observed and predicted values include absolute and squared error loss. However, in the presence of censoring, standard measures of prediction error cannot be evaluated since we do not observe all event times. For patients with censored event times, we know only that they did not experience recurrence or die before their censoring time. Thus, alternative measures are required for assessing error in point predictions in this context.

In addition, point predictions of survival times are known to be highly variable (Henderson and Keiding, 2005). Predicted survival distributions are more informative. Thus, we are also interested in measures of the difference between predicted and empirical distributions across patients.

The following sections discuss measures of prediction error that may be evaluated even if some observations are censored. We also provide details about the data splitting procedure that we use to estimate these measures.

3.2.1 Data Splitting

To assess training and test errors, we require an independent dataset that has not been involved in the model selection process. In scenarios where data are scarce, reserving a portion of the data for independent testing may not be practical. In such cases, options for "data re-use", such as bootstrapping and cross-validation (CV), may be preferable.

For this project, we chose a simple form of data re-use: data splitting. The basic idea of data splitting is to divide the data into two subsets, fit the model on the first subset, and evaluate the model on the second. In the case where a model selection step is desired, assuming the dataset is large enough, the best approach for model selection and assessment is to split data into three rather than two subsets: a training set to fit the model, a validation

set to evaluate prediction error for model selection, and a test set to assess overall prediction error of the final "best" model (Hastie *et al.*, 2009, p. 219-223).

We chose data splitting over bootstrapping and CV for simplicity in our iterative feature selection process (see details in Section 3.3.3). As with CV, the resulting model may produce more variable prediction errors, having only been fit on a fraction of the available data. The observations randomly selected for the training set may also affect the estimated prediction error (Loughin, 2018a). To mitigate this disadvantage, we make the data splitting method more robust by applying the data splitting process to our original dataset 25 times and averaging the resulting prediction error estimates (see details in Section 3.2.2).

Adapting the guidelines set forth by Picard and Berk (1990) and Hastie *et al.* (2009, p. 219–223), we reserved 75% of the data for the learning set and 25% for the independent test set. Within the specified learning set, we set 75% aside for the training set and 25% for the validation set. In summary, each data split yields a random selection of (approximately) 258 patients in the training set, 86 patients in the validation set, and 115 patients in the independent test set.

3.2.2 Measures of Prediction Error

For a given training, validation, and test set, we need a method for assessing prediction error (as part of both the model selection and assessment procedures). The following sections define our chosen measures of prediction error: Harrell's C-index and a version of the integrated Brier score.

C-Index

Harrell's concordance index (Harrell *et al.*, 1982) compares the relative ranking of pairs of predicted survival times to the relative ranking of the corresponding pairs of observed event times (without requiring exact event times for evaluation). Pairs of observations are comparable ("permissible") only if at least one observation is uncensored. In a given pair, if only one event time is known, this time must be shorter than the censoring time of the other observation for the pair to be permissible.

The C-index is calculated across all permissible pairs in the dataset. Consider individual i with observed survival time Y_i and event indicator δ_i, with δ_i=0 if Y_i is a censoring time and δ_i=1 if Y_i is an observed event time. Following Ishwaran et al (2008), we consider all permissible pairs (Y_i, Y_j) such that $Y_i < Y_j$ and let P_i denote the predicted mortality for the i^{th} individual, defined as the estimated CHF for the i^{th} individual summed over time points $t_1, \ldots, t_m$. Specifically,

$$P_i = \sum_{j=1}^{m} \hat{H}(t_j|\boldsymbol{x}_i). \tag{3.9}$$

In the case of the RSF, $\hat{H}(\cdot)$ is the ensemble CHF. In light of the inverse relationship between mortality and survival (higher predicted mortality implies shorter survival time, and vice versa), the C-index is defined as:

$$C = \frac{1}{R}\sum_{i,j} c_{ij}$$

where R is the total number of permissible pairs. For the $(i,j)^{th}$ pair,

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

The above definition includes an adjustment to that of Ishwaran *et al.* (2008). In particular, the original definition states that "for each permissible pair where $Y_i = Y_j$, but not both are deaths, count 1 if the death has worse predicted outcome; otherwise, count 0.5". However, we believe this statement to be inconsistent with the rest of the definition. In fact, for such a pair, we should count 0 if the death has the better predicted outcome and count 0.5 only if the predicted outcomes are tied.

C can be interpreted as the proportion of correctly predicted rankings of survival times among permissible pairs. The prediction error is the corresponding misclassification probability, $1 - C$. The C-index is not meaningful as a measure of the accuracy of individual predicted survival times. However, its simplicity – even in the presence of censoring – is advantageous, which likely accounts for its widespread use.

Average Integrated Brier Score

The C-index is a measure of the accuracy of point predictions. An alternative measure of prediction error that reflects the accuracy of predicted survival distributions (of key importance in our setting) is the average integrated Brier score (AIBS), which we define for a given data subset as

$$AIBS = \frac{1}{N}\sum_{i=1}^{N}\int_0^{\tilde{\tau}_i}\left[I(y_i > t) - \hat{S}(t|\boldsymbol{x}_i)\right]^2 dt. \tag{3.10}$$

Here, $I(y_i > t)$ and $\hat{S}(t|\boldsymbol{x}_i)$ are the empirical and predicted survivor functions, respectively, for patient i.

The standard version of the integrated Brier score has $\tilde{\tau}_i = \infty$ or $\tilde{\tau}_i = \tau$, where τ is the end time of the study (common to all individuals). However, in our setting, the empirical and predicted survivor functions may be defined only up to a finite value of t, and this value may vary across patients, data subsets, and prediction methods. In particular, patients with censored response times have empirical functions defined only up to their censoring times. Meanwhile, the predicted survivor function produced by the `randomForestSRC` package is defined only up to the maximum observed response time in the data subset. Finally, the predicted function based on the Cox model is defined only up to the maximum observation (response or censoring) time in the data subset.

In the interest of fairness when comparing prediction methods, we set $\tilde{\tau}_i$ to the maximum time at which the empirical survivor function and all predicted survivor functions (i.e., across all methods) are all defined. If patient i has a censored response, $\tilde{\tau}_i$ turns out to be her censoring time, y_i; otherwise, $\tilde{\tau}_i$ is the maximum observed response time in the data subset.

Since $\tilde{\tau}_i$ can vary by data subset, AIBS should not be compared across subsets. However, comparing AIBS (averaged across subsets) across prediction methods – our question of interest – is meaningful.

Other Measures of Prediction Error

As measures of prediction error, in addition to the C-index and AIBS, we also considered using estimated expected loss weighted by the inverse probability of censoring weights (IPCW) to adjust for censoring. In particular, we investigated the Lawless and Yuan (2010) CV estimator (used by Lipson, 2014, in her analysis of the Calgary data) and the AIBS measure suggested by Gerds and Schumacher (2007) to quantify the distance between predicted and observed values. However, the validity of the IPCW approach relies on the assumption of random censoring times, which does not hold in the case of the Winnipeg data. Therefore, we did not explore these measures further.

3.3 Feature Selection

Often datasets contain unimportant predictors. Identifying relevant variables may result in improved predictive accuracy, a simpler model, faster computation times, and greater interpretability. Similarly, from a cost perspective, the identification of a sparser set of important variables is beneficial for future data collection.

The following sections describe the methods used for feature selection in the prediction methods we investigate.

3.3.1 Backward Elimination (Weibull Model)

Methods for feature selection in the Weibull model are discussed in the literature (Das, 2017; Newcombe *et al.*, 2017). However, aside from backward elimination, implementing these methods is outside the scope of this project. Thus, we consider only this (admittedly naive) method in the Weibull context.

Because we manually code the surgery-specific categorical variables (see Section 2.4 and Appendix A) in the design matrix, the columns associated with each of these variables are not "grouped", and thus the usual functions in R cannot be used to perform backward elimination. Therefore, we created a function to perform backward elimination by iteratively conducting the likelihood ratio test. In particular, at each iteration, we manually compute a (Type II) ANOVA table using the `anova()` function. We use the likelihood ratio test as a means of testing the significance of each predictor variable by comparing the model containing all p predictors against all possible reduced models with $p-1$ predictors. We then predict the outcomes on the validation set and evaluate the error using the C-index. Based on our ANOVA table, the variable with the largest p-value is recorded and eliminated from the training set. We repeat this process until there are no more variables to eliminate. Finally, we examine all iterations and select the subset of variables that yields the smallest prediction error as our "best model".

3.3.2 Group LASSO (Cox Model)

Feature selection for the Cox proportional hazards model was performed through group LASSO (least absolute selection and shrinkage operator). Classic LASSO estimators are shrinkage estimators derived by maximizing the likelihood penalized by the L_1 penalty, $\lambda \sum_{j=1}^{p} |\beta_j|$, where λ is a tuning parameter (Hastie *et al.*, 2009, p. 90–91). This method can serve as a feature selection process since regression coefficient estimates tend to zero as λ increases (Loughin, 2018b).

Since our dataset includes categorical variables whose effects are represented by groups of regression coefficients, we require a method that sets the estimates of all regression coefficients in each group to 0 simultaneously if the corresponding variable is to be excluded from the model. The group LASSO achieves this goal by changing the penalty term to $P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{g=1}^{G} \sqrt{p_g} \left|\left|\boldsymbol{\beta}^{(g)}\right|\right|_2$, where G is the number of groups, p_g is the number of predictors in group g, and $\boldsymbol{\beta}^{(g)}$ is the vector of coefficients for group g. The key feature of this method is that estimated coefficients within a group will either all be zero or all be non-zero (Friedman *et al.*, 2010). The penalty term incorporates information on the varying group sizes through $\sqrt{p_g}$.

In the survival context, the objective function to be minimized is

$$Q(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n} D(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y}) + P_\lambda(\boldsymbol{\beta}), \tag{3.11}$$

where D is the deviance of the Cox model (-2 * partial log-likelihood) (Breheny, 2019). For this purpose, we use the `grpsurv` function in the `grpreg` package in R. The regularization parameter, λ, is chosen using 10-fold CV via the `cv.grpsurv` function.

3.3.3 Iterative Random Survival Forests

The feature selection process for the random survival forest is performed by iteratively fitting the RSF and eliminating the "unimportant" variables from the dataset at each iteration. This method uses the variable importance (VIMP) measure computed by specifying "`importance`=TRUE" in the `rfsrc` function in the R package `randomForestSRC`. The VIMP measure for a predictor, x_j, in the survival context is calculated in a similar way to the classic CART approach, namely, by computing the change in prediction error that occurs when OOB cases are dropped down in-bag trees, randomly choosing a child node each time a split on x_j is encountered. The only difference is that the C-index, not residual sum of squares, is used as the measure of prediction error. VIMP is an indication of how the predictive accuracy of our model is affected when information provided by an individual predictor is lost (Ishwaran *et al.*, 2008).

Using the built-in VIMP measure and our own implementation of the C-index, we explore two different iterative elimination schemes inspired by Pang *et al.* (2012), hereinafter referred to as the "Pang algorithm" and the "Pang-R algorithm". A description of the Pang algorithm is as follows:

1. Fit a random survival forest on the training dataset using all available predictors. Make predictions on the validation set and quantify the prediction error via the C-index.
2. Calculate the variable importance for all predictors. Order the predictors by VIMP in descending order and store in a list called List 1.
3. Set $i = 2$.
4. Repeat the following steps until List i is empty:
   (a) Identify the variables that correspond to the lowest 20% of VIMP values in List $i-1$ and remove these variables from that list. Call the updated list List i.
   (b) Fit a random survival forest on the training dataset using only the predictors in List i. Make predictions on the validation set and quantify the prediction error via the C-index.
   (c) Set $i = i + 1$.

5. Identify the iteration, I, with the smallest C-index value and select the variables in List I for the final model.

The Pang-R algorithm uses the same approach except that it recalculates the variable importance values at each iteration, which may change the order of variables being eliminated. Díaz-Uriarte and de Andrés (2006), who also implement the idea of iterative random survival forests for classification of microarray data, suggest that recalculating VIMP may cause over-fitting. However, the authors do not provide evidence for this claim. As such, we also explore feature selection for our dataset using this alternative approach.

The goal of the Pang *et al.* (2012) paper is feature selection in the context of gene expression data, which commonly has more variables than observations. The authors use the "bottom 20%" threshold to allow for fast computation and "aggressive variable selection", which is desirable in that setting. However, our dataset contains only 22 predictors, which is manageable computationally. Thus, this "aggressive" approach may be unnecessary or, worse, lead to increased prediction error in our context.

We therefore explore not only the original but also more and less aggressive forms of the Pang and Pang-R algorithms in our work. In particular, we adapt the algorithms based on an alternative cut-point threshold (which we treat as tuning parameters), hereinafter referred to as the "Pang-K" and "Pang-K-R" algorithms. In addition to considering the bottom $h\%$ of VIMP values, we also consider the k-smallest VIMP values. The bottom percentage value (h) is tuned by searching over the values $\{0.1, 0.15, 0.2, 0.25, 0.3\}$, and the number of variables (k) is tuned by searching over the values $\{1, 2, 3, 4, 5\}$. We perform the feature selection process for each cut-point threshold, eliminating the bottom $h\%$ of variables or the k variables with the lowest VIMP values at each iteration. Then, using a 0-SE rule approach, we select the subset of variables that yields the smallest prediction error.

Chapter 4

Results and Recommendations

Having outlined the framework for our methods, we now discuss our preliminary analysis of the data and the application of each method to the prediction of ovarian cancer event times.

4.1 Preliminary Analysis

The treatment protocols at the Calgary and Winnipeg centres are close to identical and patient outcomes are, anecdotally, similar (Dr. Alon Altman, personal communication). However, theoretically, patient responses may differ across centres due, for example, to unmeasured predictors such as demographic and environmental variables. Therefore, our first step is to investigate the possibility of a "city effect" on recurrence and death time. If we are unable to detect such an effect, we will be comfortable with combining the datasets for developing prediction methods (our main goal).

We fit the Cox proportional hazards model to three different outcomes: time to recurrence (measured from end of treatment), time to death after recurrence (DAR, measured from recurrence time), and time to death (measured from end of treatment and ignoring whether the cancer recurred). For the table of parameter estimates and the QQ-plots of the residuals (which show no evidence of lack of fit), see Supplemental Material.

There is strong evidence of a city effect (after adjusting for the other predictor variables) on time to recurrence (p-value = 0.002). In particular, controlling for the other predictors, for time to recurrence, the estimated hazard for Winnipeg patients is approximately 56% of that of the Calgary patients. In contrast, there is no evidence of a difference in time to death between the two cities after adjusting for the other predictor variables (p-value = 0.6263). Given these results, we expected time to DAR to be shorter for Winnipeg patients who experienced recurrence. Although we did not find strong evidence of a city effect (after adjusting for the other predictors) on time to DAR (p-value = 0.1287), the estimated hazard for Winnipeg patients is approximately 144% of that of the Calgary patients, which

is consistent with expectations. Since time to DAR is specific to only those patients who experienced recurrence, we suspect that the substantially smaller sample size may be why we did not find statistical evidence of this effect.

These results suggest that Winnipeg patients wait longer than Calgary patients to present with symptoms of recurrence, resulting in a longer *apparent* time to recurrence. We have no evidence, however, that recurrence times are *in fact* longer in Winnipeg since these responses are, effectively, interval-censored. Our understanding of recurrence times is limited both by this interval censoring and by the inherent difficulty in determining recurrence (which is defined rather loosely as "disease progression"). Likewise, DAR, which is a function of recurrence time, is difficult to assess in a meaningful way.

In light of these findings, we do not attempt to predict recurrence or DAR times and instead focus strictly on the clearer outcome, time to death (combining the two datasets). The QQ-plots of the Weibull and Cox residuals resulting from the death time analysis (see Appendix B) show no evidence of lack of fit.

4.2 Predictions of Death Times (No Feature Selection)

We then investigate the ability of our methods to predict death times without considering feature selection. Specifically, we fit the Weibull model, Cox proportional hazards model, and RSF on the learning set. We assess each method using the C-index and AIBS on an independent test set obtained via data splitting. The results from averaging across 25 data splits are shown in Table 4.1.

Table 4.1: Estimates of prediction error: No feature selection

Model	$(1 - C)$ (SE)	AIBS (SE)
Weibull	0.269 (0.0052)	33.7 (0.54)
Cox PH	0.269 (0.0050)	33.8 (0.52)
RSF	0.231 (0.0043)	39.0 (0.61)

According to the C-index, the RSF performs best in terms of ranking patient survival. Perhaps the RSF captures non-linear or interaction effects of predictors that are not represented in the Weibull or Cox models.

In contrast, the Weibull and Cox models appear to perform similarly – and better than the RSF – in terms of predicting survival distributions (as per the AIBS). This result is perhaps unsurprising given that the predicted survival curve produced by the `randomForestSRC`

package is an upper bound, which could lead to substantial inflation of the AIBS (see Section 3.1.3).

Since the AIBS is a more meaningful measure of performance than the C-index in our context – and we give preference to simpler predictive methods – we conclude that the Weibull model is the best of the three methods for predicting survival distributions when using the full set of predictors.

4.3 Predictions of Death Times (With Feature Selection)

We implement the feature selection procedures described in Section 3.3 by fitting models on the training set and evaluating the performance on the validation set. The predictive performance of each model is measured using the C-index and the subset of variables that yields the smallest prediction error is selected as the "best" model.

4.3.1 Feature Selection Results

To assess the stability of the feature selection procedures, we apply them to each of the 25 data splits. In addition to the selected Weibull and Cox models, we present the results for the iterative RSF methods using the cut-point thresholds $h = 0.15$ in the Pang algorithm, $h = 0.10$ in the Pang-R algorithm, and $k = 1$ in both the Pang-K and Pang-K-R algorithms, as these thresholds produced the lowest prediction errors on average.

The number of times each variable was selected in each split (under each of the six feature selection schemes) is counted and the final selection proportion across the 25 data splits is obtained. The resulting proportions are shown in Table 4.2. We would expect the most important variables to be selected close to 100% of the time and unimportant variables to be selected close to 0% of the time. From these results, we see three approximate groupings: variables that appear important (e.g., 5 out of 6 of the methods select treatment length in over 95% of splits), variables that appear unimportant (e.g., no method selects minimum hemoglobin in more than 45% of splits), and variables with uncertain importance (e.g., the methods select tumour grade in 4 – 92% of splits).

4.3.2 Prediction Errors Based on Selected Models

To summarize the results of our feature selection process, we impose an inclusion threshold on selected variables. We then fit our final models on the learning set with all variables selected for each respective method that meet or exceed the specified threshold. The thresholds considered are of varying strictness, an approach adopted from Meinshausen and Bühlmann (2010). Low rigidity implies that variables selected in at least 60% of splits will be included in the final model. Similarly, medium, high, and super rigidity imply that variables selected in at least 70%, 80%, and 90% of splits, respectively, will be included in the final model.

We assess the performance of the final models on an independent test set using the C-index and AIBS measures. The average prediction errors over the 25 data splits are presented in Tables 4.3 and 4.4.

Table 4.2: Proportion of splits in which variables are selected by prediction method (in decreasing order, by average proportion across methods)

Variable	Weibull	Cox	Pang	Pang-R	Pang-K	Pang-K-R
primary adjuv. chemo	0.88	0.96	1.00	1.00	1.00	0.92
debulking	0.68	0.84	0.96	1.00	0.96	0.96
treatment length	0.36	1.00	1.00	1.00	1.00	0.96
surgery	0.84	1.00	0.80	0.84	0.88	0.80
min. CA125	0.40	0.68	1.00	0.96	1.00	1.00
min. albumin	0.96	1.00	0.76	0.88	0.80	0.52
max. CA125	0.32	0.68	0.80	0.88	0.80	0.72
neoadjuv. chemo	0.80	0.32	0.56	0.76	0.64	0.76
taxane	0.72	0.88	0.48	0.56	0.40	0.52
mean hemoglobin	0.64	0.80	0.48	0.60	0.48	0.48
age at diagnosis	0.52	0.68	0.44	0.56	0.28	0.36
mean WBC	0.52	0.88	0.24	0.56	0.16	0.36
max. platelets	0.24	0.24	0.52	0.76	0.36	0.48
stage	0.32	0.76	0.20	0.52	0.24	0.24
mean platelets	0.36	0.56	0.20	0.64	0.08	0.28
tumour grade	0.36	0.92	0.12	0.48	0.04	0.16
platinum	0.52	0.32	0.32	0.44	0.16	0.32
blood loss	0.24	0.52	0.24	0.56	0.20	0.24
min. hemoglobin	0.44	0.24	0.24	0.44	0.12	0.40
CA125 diff.	0.28	0.20	0.28	0.64	0.16	0.20
min. platelets	0.32	0.24	0.24	0.32	0.24	0.36
ascites	0.24	0.40	0.12	0.24	0.04	0.48

The C-index values across the four inclusion thresholds are generally consistent with expectations. Specifically, the performances of all the methods appear to improve with the introduction of feature selection (i.e., the low threshold) and worsen with the high and super thresholds. The AIBS values show a similar trend across inclusion thresholds for the Weibull and Cox models. We illustrate this trend, driven by the bias-variance trade-off, with the performance of the Weibull model with backward elimination in Figure 4.1. When using the full set of predictors to fit our model, we expect lower bias and higher variance in our predicted outcomes. Alternatively, when using the highly sparse set of predictors in the super threshold, we expect higher bias and lower variance. The box plots for the other methods, which illustrate the performance of both the C-index and AIBS values across thresholds, as well as the variation across the data splits, can be found in the Supplementary Material.

Table 4.3: Average prediction error $(1 - C)$ for varying degrees of strictness when selecting features

		Strictness Level				
Model	FS Method	None (SE)	Low (SE)	Medium (SE)	High (SE)	Super (SE)
Weibull	BackElim	0.269 (0.0052)	0.248 (0.0044)	0.259 (0.0042)	0.283 (0.0051)	0.403 (0.0056)
Cox PH	GrpLASSO	0.269 (0.0050)	0.259 (0.0042)	0.258 (0.0043)	0.262 (0.0054)	0.272 (0.0052)
RSF	Pang ($h = 0.15$)	0.231 (0.0043)	0.226 (0.0035)	0.226 (0.0035)	0.228 (0.0046)	0.236 (0.0046)
	Pang-R ($h = 0.10$)	0.231 (0.0043)	0.228 (0.0034)	0.226 (0.0035)	0.229 (0.0046)	0.236 (0.0046)
	Pang-K ($k = 1$)	0.231 (0.0043)	0.225 (0.0033)	0.226 (0.0036)	0.228 (0.0046)	0.236 (0.0046)
	Pang-K-R ($k = 1$)	0.231 (0.0043)	0.229 (0.0036)	0.229 (0.0036)	0.237 (0.0046)	0.236 (0.0046)

However, the AIBS values for the RSFs seem to *increase* with feature selection. A possible explanation is the use of the C-index to evaluate the subsets in the iterative feature

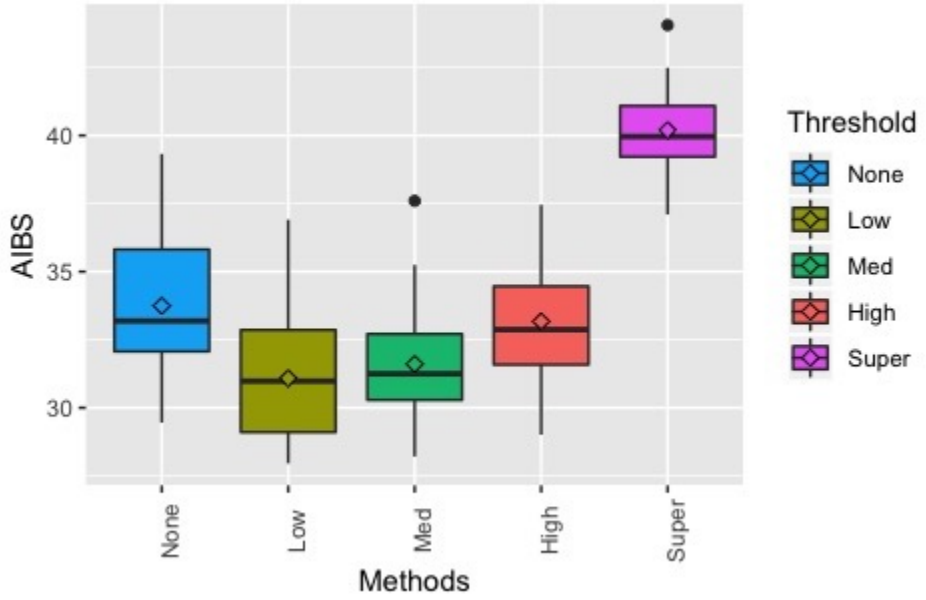

Figure 4.1: Prediction error (AIBS) of the Weibull model with backwards elimination by varying degrees of strictness when selecting features

selection process but the AIBS to evaluate the final model. Interestingly, mixing the two error measures does not have this effect on the performance of the Weibull model (which also used an iterative feature selection process based on the C-index).

Table 4.4: Average prediction error (AIBS) for varying degrees of strictness when selecting features

		Strictness Level				
Model	FS Method	None (SE)	Low (SE)	Medium (SE)	High (SE)	Super (SE)
Weibull	BackElim	33.7 (0.54)	31.1 (0.49)	31.6 (0.45)	33.2 (0.46)	40.2 (0.31)
Cox PH	GrpLASSO	33.8 (0.52)	32.0 (0.47)	31.9 (0.47)	32.3 (0.49)	33.1 (0.47)
RSF	Pang ($h = 0.15$)	39.0 (0.52)	40.1 (0.70)	40.1 (0.70)	39.6 (0.71)	41.7 (0.73)
	Pang-R ($h = 0.10$)	39.0 (0.52)	39.6 (0.66)	40.1 (0.69)	39.6 (0.71)	41.7 (0.73)
	Pang-K ($k = 1$)	39.0 (0.52)	39.4 (0.68)	40.1 (0.69)	39.6 (0.70)	41.7 (0.73)
	Pang-K-R ($k = 1$)	39.0 (0.52)	39.7 (0.70)	39.7 (0.70)	41.9 (0.74)	41.7 (0.73)

As we are interested primarily in the accuracy of the predicted survival distributions, we base recommendations on the AIBS values. In particular, we recommend using the Weibull or Cox model for predicting time to death. While a corrected implementation of the RSF method may be a good option in the future, at present, the Weibull and Cox models fit our data adequately and provide better predictions of survival distributions. The Weibull and Cox methods perform comparably (with the exception of the Weibull model at the super threshold). Comparing across the four inclusion thresholds, we recommend implementing a low to medium threshold, which seems to yield the lowest prediction errors. However, the Weibull and Cox models still show improvement at the high threshold and the Cox model still shows improvement at the super threshold. Therefore, we present the best subsets of predictors associated with all thresholds that lead to reasonable AIBS values in Table 4.5.

Since the Weibull backward elimination method and Cox group LASSO method yield different subsets of predictors, we fit the Cox model using the variables selected with backward elimination and fit the Weibull model using the variables selected with group LASSO. The comparison of the average AIBS values over the 25 data splits relative to the original final models are presented in Tables 4.6 and 4.7. See Appendix C for the corresponding

Table 4.5: Proposed models that yield reasonable AIBS values

Model	Threshold	# of Vars	Proposed Subset
Weibull	Low	7	min. albumin, adjuv. chemo, neoadjuv. chemo, surgery, taxane, min. hemoglobin, debulking
	Medium	5	min. albumin, adjuv. chemo, neoadjuv. chemo, surgery, taxane
	High	4	min. albumin, adjuv. chemo, neoadjuv. chemo, surgery
Cox	Low	13	treatment length, min. albumin, tumour grade, adjuv. chemo, surgery, mean hemoglobin, mean WBC, taxane, debulking, stage, diagnosis age, max. CA125, min. CA125
	Medium	10	treatment length, min. albumin, tumour grade, adjuv. chemo, surgery, mean hemoglobin, mean WBC, taxane, debulking, stage
	High	9	treatment length, min. albumin, tumour grade, adjuv. chemo, surgery, mean hemoglobin, mean WBC, taxane, debulking
	Super	5	treatment length, min. albumin, tumour grade, adjuv. chemo, surgery

comparison of C-index values. Given the similar performance of the alternative final models and the original final models – and given that preference is given to simpler predictive methods – we conclude that the Weibull model is the best of the three methods for predicting survival distributions. Moreover, the AIBS values associated with the Weibull model fit to all proposed subsets of predictors (with the exception of those selected by the Weibull backward elimination method in the super threshold) are similar.

Table 4.6: Comparison of the Weibull and Cox models fit using variables selected by the Weibull backward elimination method: Average prediction error (AIBS) for varying degrees of strictness when selecting features

		Strictness Level				
Model	FS Method	None (SE)	Low (SE)	Medium (SE)	High (SE)	Super (SE)
Weibull	BackElim	33.7 (0.54)	31.1 (0.49)	31.6 (0.45)	33.2 (0.46)	40.2 (0.31)
Cox PH	Weibull BE Variables	33.8 (0.52)	31.5 (0.48)	32.0 (0.42)	32.7 (0.47)	40.3 (0.25)

Table 4.7: Comparison of the Weibull and Cox models fit using the variables selected by the Cox group LASSO method: Average prediction error (AIBS) for varying degrees of strictness when selecting features

		Strictness Level				
Model	FS Method	None (SE)	Low (SE)	Medium (SE)	High (SE)	Super (SE)
Weibull	Cox GL Variables	33.7 (0.54)	32.8 (0.56)	32.4 (0.52)	32.1 (0.53)	32.6 (0.51)
Cox PH	GrpLASSO	33.8 (0.52)	32.0 (0.47)	31.9 (0.47)	32.3 (0.49)	33.1 (0.47)

Minimum albumin levels and the number of primary adjuvant chemotherapy cycles are selected in every proposed model. For each method, the directions of the estimated effects of these predictors suggest that higher values of albumin levels and more primary adjuvant chemotherapy cycles are associated with better prognosis. Debulking is included in a number of the proposed models; its estimated effect implies that optimal debulking is associated with better prognosis. Likewise, tumour grade appears in some models; its estimated effect suggest that higher values (i.e., poorly differentiable tumour cells) are associated with worse prognosis. In general, the selected predictors and their estimated effects are consistent with anecdotal evidence (Dr. Alon Altman, personal communication).

Chapter 5

Discussion and Future Work

Our results indicate the Weibull and Cox models may have better performance than the RSF and that the elimination of unimportant variables may improve predictions. In particular, we recommend the Weibull model (because of its simplicity) and provide seven subsets of predictor variables that all lead to similar estimates of prediction error. However, some issues remain to be addressed.

We did not know beforehand that the `randomForestSRC` package would produce such poor predicted survivor functions. While the Weibull and Cox models may inherently outperform the RSF in this regard, we suspect that the way the `randomForestSRC` package produces the predicted survivor function – not the RSF method in general – is responsible for our results. Correcting the predicted survivor function and repeating our analysis would be a worthwhile endeavour. However, in a repeated analysis, we suggest, like Díaz-Uriarte and de Andrés (2006), that recalculating variable importance at every iteration is not necessary (see Appendix D for a brief comparison of the effects of recalculating variable importance).

Within the iterative feature selection processes, the performance of the selected subsets was evaluated only using the C-index, not the AIBS. This choice may have impacted our selected final models. Using AIBS for feature selection is another important avenue for future work.

Further work is warranted with regard to feature selection methods, particularly for the Weibull model. We expect the predictive ability of the Weibull model to improve with a more sophisticated feature selection method than backward elimination. Other extensions might also improve our feature selection processes. For example, our iterative feature selection methods employed a "0-SE rule", in which the subset with the lowest error was chosen for further analysis. Use of a "1-SE" rule might result in sparser final models with similar performance.

We also recommend applying the feature selection methods on a larger dataset and over more data splits. While the results of our feature selection methods identify subsets of important variables that are consistent with anecdotal evidence (such as minimum albumin and debulking), we would expect an analysis based on more data and more splits to lead to a clearer picture of the important variables. We would also expect more conclusive results if we instead examined the proportion of times each proposed *subset* is selected, as opposed to the proportion of times each individual *variable* is selected. This approach would require many more data splits and would be more computationally expensive than the approach we took for this project.

In addition to these potential improvements within our existing feature selection methods, we are optimistic about other predictive methods capable of feature selection, such as SVMs, for predicting survival times. Further development of methods that produce predicted survival distributions would also be a worthwhile avenue of research.

Our dataset was large enough to allow for sufficient validation sets and independent testing. However, a larger dataset would allow for more substantial validation and test sets without taking away from the training set. Additionally, the potential overlap of observations selected for the learning sets and test sets across data splits may have affected our results. Due to our modest sample size, we to chose to use the full dataset for each of the data splits. However, a larger dataset would benefit the data splitting process by allowing a fully independent test set overall, as opposed to independent test sets for each data split. More accurate estimates of prediction errors would likely result.

Our choice of default tuning parameters in the random survival forests may have impacted our results, both in the iterative feature selection process and in the performance of our final RSF models. More thoughtful specification of the tuning parameters might impact the subsets of predictors selected and could improve our prediction errors. Efforts were also made to find new splitting rules, however our literature search into alternatives to the `randomForestSRC` splitting rules did not yield any methods that were clearly superior, easy to implement, and within the scope of our project.

A more sophisticated method of data imputation would reflect the relationship between predictors and the response more accurately. In addition, the possibility of "informative missingness" of the hematological predictors should be considered. As different patients with varying prognoses have varying needs, the available hematologic and surgical predictors for each patient may also vary. More specifically, missing lab tests could be an indication of a patient's overall health.

We hope that our work will provide a useful starting point for physicians wishing to generate predicted distributions of time to death using a relatively sparse subset of informative

predictors. Our work also serves to highlight which predictors should be a priority when collecting data for future analysis. Much work on feature selection methods and predictive methods for censored survival data remains, but we hope that our analysis has made a meaningful contribution to future research.

References

American Cancer Society (2019). Surgery for ovarian cancer. https://www.cancer.org/cancer/ovarian-cancer/treating/surgery.html. accessed July 31, 2018.

Bou-Hamad, I., Larocque, D., and Ben-Ameur, H. (2011). A review of survival t rees. *Statistical Surveys*, **5**(1), 44–71.

Breheny, P. (2019). Models. http://pbreheny.github.io/ncvreg/articles/web/models.html. accessed July 31, 2018.

Canadian Cancer Society (2017). Canadian cancer statistics. http://www.cancer.ca/en/cancer-information/cancer-101/canadian-cancer-statistics-publication/past-editions-canadian-cancer-statistics/. accessed July 31,2018.

Canadian Cancer Society (2019a). Cancer antigen 125. http://www.cancer.ca/en/cancer-information/diagnosis-and-treatment/tests-and-procedures/cancer-antigen-125-ca-125/. accessed July 31, 2018.

Canadian Cancer Society (2019b). Chemotherapy. http://www.cancer.ca/en/cancer-information/diagnosis-and-treatment/chemotherapy-and-other-drug-therapies/chemotherapy/. accessed July 31, 2018.

Canadian Cancer Society (2019c). Managing symptoms and side effects. http://www.cancer.ca/en/cancer-information/diagnosis-and-treatment/managing-side-effects/. accessed July 31, 2018.

Canadian Cancer Society (2019d). Treatments for ovarian cancer. http://www.cancer.ca/en/cancer-information/cancer-type/ovarian/treatment/. accessed July 31,2018.

Canadian Liver Foundation (2017). Your liver. https://www.liver.ca/your-liver/. accessed July 31, 2018.

Chen, Y., Jia, Z., Mercola, D., and Xie, X. (2013). A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Computational and Mathematical Methods in Medicine*, **2013**.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology*, **34**(2), 187–220.

Das, U. (2017). Variable selection for survival data under weibull distribution. *Calcutta Statistical Association Bulletin*, **68**(1 & 2), 52–68.

Díaz-Uriarte, R. and de Andrés, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**(3).

Friedman, J. H., Hastie, T. J., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. In *arXiv preprint arXiv:1001.0736*.

Gerds, T. A. and Schumacher, M. (2007). Efron-type measures of prediction error for survival analysis. *Biometrics*, **63**(4), 1283–1287.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and a. Rosati, R. (1982). Evaluating the yield of medical tests. *JAMA*, **247**(18), 2543–2546.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2 edition.

Henderson, R. and Keiding, N. (2005). Individual survival time prediction using statistical models. *Journal of Medical Ethics*, **31**(12), 703–703.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, **2**(3), 841–860.

Johns Hopkins Medicine (2019). What are platelets and why are they important. `https://www.hopkinsmedicine.org/heart_vascular_institute/clinical_services/centers_excellence/womens_cardiovascular_health_center/patient_information/health_topics/platelets.html`. accessed July 31, 2018.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, New York, 2 edition.

Lawless, J. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley, New York, 2 edition.

Lawless, J. F. and Yuan, Y. (2010). Estimation of prediction error for survival models. *Statistics in Medicine*, **29**(2), 262–274.

Lipson, R. (2014). Predicting ovarian cancer survival times: Performance of parametric methods and random survival forests.

Loughin, T. (2018a). "Data re-use methods to assess error and multi-model inference". Lecture on Sept 24, 2018.

Loughin, T. (2018b). "The LASSO: A modern approach to variable selection". Lecture on Sept 17, 2018.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology*, **72**(4), 417–473.

Newcombe, P. J., Ali, H. R., Blows, F. M., Provenzano, E., Pharoah, P. D., Caldas, C., and Richardson, S. (2017). Weibull regression with bayesian variable selection to identify prognostic tumour markers of breast cancer survival. *Statistical Methods in Medical Research*, **26**(1), 414–436.

Ovarian Cancer Canada (2017). About ovarian cancer. `https://ovariancanada.org/About-Ovarian-Cancer`. accessed July 31, 2018.

Pang, H., George, S. L., Hui, K., and Tong, T. (2012). Gene selection using iterative feature elimination random forests for survival outcomes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **9**(5), 1422–1431.

Picard, R. R. and Berk, K. N. (1990). Data splitting. *The American Statistician*, **44**(2), 140–147.

Rocconi, R. P., Matthews, K. S., Kemper, M. K., Hoskins, K. E., Huh, W. K., and Jr., J. M. S. (2009). The timing of normalization of ca-125 levels during primary chemotherapy is predictive of survival in patients with epithelial ovarian cancer. *Gynecologic Oncology*, **114**(2), 242–245.

Shivaswamy, P. K., Chu, W., and Jansche, M. (2007). A support vector approach to censored targets. In *Proceedings of the Seventh IEEE Internation Conference on Data Mining (ICDM 2007)*, pages 655–660. IEEE.

Van Belle, V., Pelckmans, K., Huffel, S. V., and Suykens, J. A. (2011). Support vector methods for survival analysis: A comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, **53**(2), 107–118.

Appendix A

Surgery-Specific Predictor Variables

The following describes how surgery-specific predictor variables are defined in the design matrix.

For patient i,

$$(a)_{2i} = \begin{cases} 1, & \text{if surgery = "Yes"} \\ 0, & \text{otherwise} \end{cases}$$

$$(ab)_{22i} = \begin{cases} 1, & \text{if debulking = "No", given surgery = "Yes"} \\ 0, & \text{otherwise} \end{cases}$$

$$(ab)_{23i} = \begin{cases} 1, & \text{if debulking = "Unknown", given surgery = "Yes"} \\ 0, & \text{otherwise} \end{cases}$$

$$(ag)_{22i} = \begin{cases} 1, & \text{if blood loss = "<1000 mL", given surgery = "Yes"} \\ 0, & \text{otherwise} \end{cases}$$

$$(ag)_{23i} = \begin{cases} 1, & \text{if blood loss = ">1000 mL", given surgery = "Yes"} \\ 0, & \text{otherwise} \end{cases}$$

$$(ag)_{24i} = \begin{cases} 1, & \text{if blood loss= "Unknown", given surgery = "Yes"} \\ 0, & \text{otherwise} \end{cases}$$

$$(ap)_{22i} = \begin{cases} 1, & \text{if ascites = "<1000 mL", given surgery = "Yes"} \\ 0, & \text{otherwise} \end{cases}$$

$$(ap)_{23i} = \begin{cases} 1, & \text{if ascites = "1000-3999 mL", given surgery = "Yes"} \\ 0, & \text{otherwise} \end{cases}$$

$$(ap)_{24i} = \begin{cases} 1, & \text{if ascites = ">4000 mL", given surgery = "Yes"} \\ 0, & \text{otherwise} \end{cases}$$

$$(ap)_{25i} = \begin{cases} 1, & \text{if ascites= "Unknown", given surgery = "Yes"} \\ 0, & \text{otherwise} \end{cases}$$

Appendix B

Model Fit Diagnostics

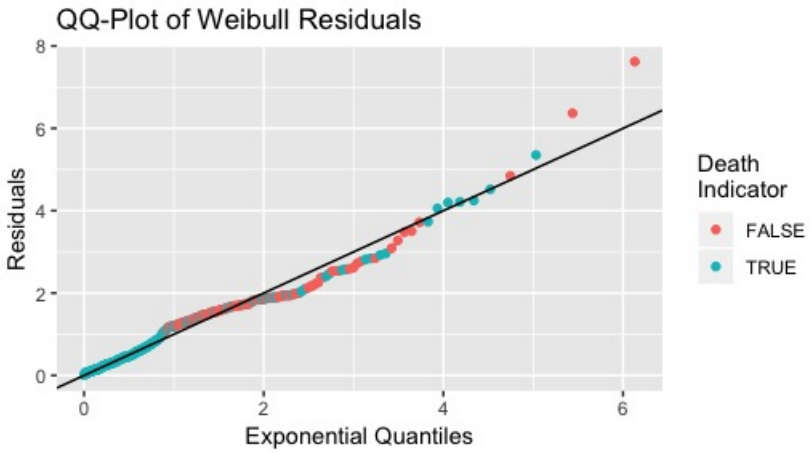

Figure B.1: Weibull Residuals: Time to death

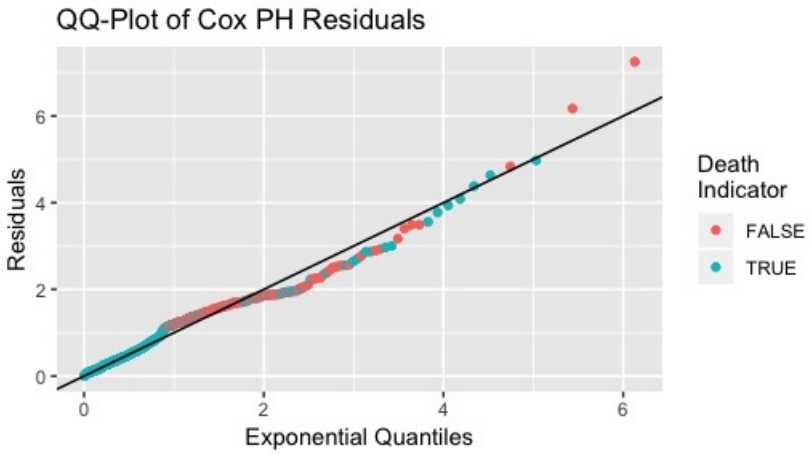

Figure B.2: Cox PH Residuals: Time to death

Appendix C

Alternative Final Models

Table C.1: Comparison of the Weibull and Cox models fit using variables selected by the Weibull backward elimination method: Average prediction error $(1 - C)$ for varying degrees of strictness when selecting features

		Strictness Level				
Model	FS Method	None (SE)	Low (SE)	Medium (SE)	High (SE)	Super (SE)
Weibull	BackElim	0.269 (0.0052)	0.248 (0.0044)	0.259 (0.0042)	0.283 (0.0051)	0.403 (0.0056)
Cox PH	Weibull BE Variables	0.269 (0.0050)	0.249 (0.0050)	0.262 (0.0049)	0.280 (0.0052)	0.404 (0.0046)

Table C.2: Comparison of the Weibull and Cox models fit using the variables selected by the Cox group LASSO method: Average prediction error $(1 - C)$ for varying degrees of strictness when selecting features

		Strictness Level				
Model	FS Method	None (SE)	Low (SE)	Medium (SE)	High (SE)	Super (SE)
Weibull	Cox GL Variables	0.269 (0.0052)	0.267 (0.0050)	0.264 (0.0041)	0.259 (0.0048)	0.271 (0.0057)
Cox PH	GrpLASSO	0.269 (0.0050)	0.259 (0.0042)	0.258 (0.0043)	0.262 (0.0054)	0.272 (0.0052)

Appendix D

Iterative RSF Variable Importance

As mentioned in Section 3.3.3, to determine the final RSFs, we performed tuning on the four variations of the Pang algorithm to determine the best thresholds for eliminating variables at each iteration (resulting in the thresholds described in Section 4.3.1). Specifically, we selected the tuning parameter that yielded the lowest C-index for each of our final proposed models. In Figure D.1, we illustrate the effects of "aggressive" cutting approaches as well as the effects of recalculating the VIMP values at each iteration. We fit the corresponding final models on the training data as chosen by each respective feature selection method across all proposed values of cut-point thresholds and evaluate the errors as described in Section 3.3.3.

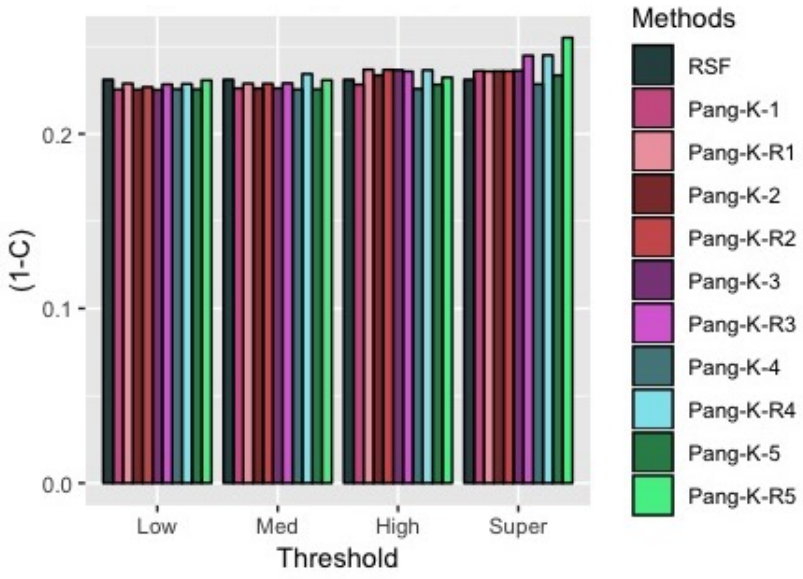

Figure D.1: Comparison of prediction error $(1 - C)$ of the RSF with Pang-K vs. Pang-K-R algorithms by varying degrees of strictness when selecting features

The comparison of the bottom h percent and k cut-points shows that differences between "gentle" and "aggressive" cutting approaches are minimal. Overall, as the cut-point thresh-

olds increase, the predicted errors either remain approximately the same or increase only slightly. The one exception is the high error resulting from applying the more aggressive methods at the "super" threshold.

The exploration into the effects of recalculating the VIMP values at each iteration yielded results consistent with expectations. The comparison of Pang-K vs. Pang-K-R (Figure D.1) illustrates a noticeably worse performance by the methods that recalculated VIMP, as Díaz-Uriarte and de Andrés (2006) suggest. The comparison of the Pang vs. Pang-R methods yields similar, though subtler, conclusions. See Supplementary Material for the Pang vs. Pang-R plots comparing $(1 - C)$ and AIBS values, as well as the Pang-K vs. Pang-K-R plot comparing AIBS values.

Appendix E

Supplementary File: Preliminary Analysis

Description:

The accompanying supplementary file contains the full report of the preliminary analysis of the three different outcomes: time to recurrence (measured from end of treatment) time to death after recurrence (DAR, measured from recurrence time), and time to death (measured from end of treatment and ignoring whether the cancer recurred).

Filename:

PreliminaryAnalysis.pdf

Appendix F

Supplementary File: Plots

Description:

The accompanying supplementary file provides the box plots of the C-index and AIBS values across all methods and inclusion thresholds, illustrating the distribution of these measures across the 25 data splits.

Filename:

SupplementalPlots.pdf