Summer 2020

Causal Inference and Prediction on Observational Data with Survival Outcomes

Xiaofei Chen
Southern Methodist University, xiaofeic@smu.edu

CAUSAL INFERENCE AND PREDICTION ON OBSERVATIONAL DATA WITH SURVIVAL OUTCOMES

Approved by:

Dr. Daniel F. Heitjan
Professor,
Department of Statistical Science, SMU
Population & Data Sciences, UTSW

Dr. Haekyung Jeon-Slaughter
Assistant Professor,
Internal Medicine, UTSW

Dr. Xinlei (Sherry) Wang
Professor,
Department of Statistical Science, SMU

Dr. Hong Zhu
Associate Professor,
Population & Data Sciences, UTSW

CAUSAL INFERENCE AND PREDICTION ON OBSERVATIONAL DATA WITH SURVIVAL OUTCOMES

A Dissertation Presented to the Graduate Faculty of the

Dedman College

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

with a

Major in Biostatistics

by

Xiaofei Chen

B.S., Economics, Shandong University
M.S., Biostatistics, Georgetown University

August 04, 2020

ACKNOWLEDGMENTS

I would like to express my special appreciation and thanks to my Ph.D advisor Dr. Daniel F. Heitjan and co-advisor Dr. Haekyung Jeon-Slaughter for the support of my Ph.D research. Their guidance and immense knowledge helped me in working on research and writing of this dissertation. I appreciate all their contributions to make me become a statistician. I also would like to express my sincere gratitude to my dissertation committee members Dr. Xinlei (Sherry) Wang and Dr. Hong Zhu for their time, interests, and helpful comments on my dissertation, and also excellent teaching during my PhD study. Last but certainly not least, I would like to thank my family and friends for their support and encouragement.

Chen, Xiaofei

B.S., Economics, Shandong University
M.S., Biostatistics, Georgetown University

Causal Inference and Prediction on Observational Data with Survival Outcomes

Advisor: Dr. Daniel F. Heitjan

Co-advisor: Dr. Haekyung Jeon-Slaughter

Doctor of Philosophy degree conferred August 04, 2020

Dissertation completed July 01, 2020

Infants with hypoplastic left heart syndrome require an initial Norwood operation, followed some months later by a stage 2 palliation (S2P). The timing of S2P is critical for the operation's success and the infant's survival, but the optimal timing, if one exists, is unknown. We attempt to estimate the optimal timing of S2P by analyzing data from the Single Ventricle Reconstruction Trial (SVRT), which randomized patients between two different types of Norwood procedure. In the SVRT, the timing of the S2P was chosen by the medical team; thus with respect to this exposure, the trial constitutes an observational study, and the analysis must adjust for potential confounding. In Chapter 1, we propose an extended propensity score analysis that describes the time to surgery as a function of confounders in a discrete competing-risk model. We then apply inverse probability weighting to estimate a spline hazard model for predicting survival from the time of S2P. In Chapter 2, we address same question by multiply imputing the potential post-S2P survival outcomes with a lognormal model under the Rubin Causal Model framework. With this approach, it is straightforward to estimate the causal effect of S2P timing on post-S2P survival by directly comparing the imputed potential outcomes. We examine the sensitivity of these results by applying a more flexible model that assumes proportional hazards as a function of S2P time, with a restricted cubic spline (RCS) for the baseline hazard. Our analysis suggests that S2P conducted at 6 months after the Norwood gives the patient

the best post-S2P survival.

In Chapter 3, we build a new 10-year ASCVD (atherosclerotic cardiovascular disease) risk prediction model for Veterans Affairs (VA) women based on data from the VA national EHR (Electronic Health Records) database.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I dedicate this dissertation to my family and friends.

CHAPTER 1

Estimating the Optimal Timing of Surgery from Observational Data

1.1. Introduction

Hypoplastic left heart syndrome (HLHS) is a congenital condition that afflicts roughly 1,000 US newborns each year [31]. Because children with this syndrome have severely compromised blood oxygenation and circulation, without prompt treatment it is uniformly fatal. In the 1970s, children with HLHS received only supportive care, and all died as neonates. The advent of the Norwood procedure followed by stage 2 palliation (S2P), introduced in the early 1980s, has led to vastly improved survival, with many patients now living to adulthood [56].

Nevertheless, HLHS patients continue to face substantial mortality risk, particularly in the interim between the Norwood and S2P operations. For example, in the Single Ventricle Reconstruction Trial, which compared two variants of the Norwood procedure, mortality before post-Norwood discharge was 16%, and a further 12% died between the Norwood discharge and S2P [12, 19, 44]. There is reason to believe that appropriately timing the S2P can further improve patient outcomes. Although no single time may be optimal in all cases, it is desirable to identify a default time that gives superior results.

The choice of the optimal time to conduct the S2P is a causal question that refers to selecting the best from a range of potential outcomes. As there has been no randomized trial of S2P timing, bias from confounding can affect any data we have on this question. Previous authors have described methods to investigate the causal effect of a

time-dependent treatment on the distribution of an outcome. Robins et al [35] proposed the *marginal structural model*, which allows for improved adjustment for confounding with time-dependent treatment or covariates. Hu et al [22] developed a structural proportional hazards model to characterize the effect of treatment initiation time on survival, using it to select the timing of antiretroviral therapy initiation for patients who are co-infected with HIV and tuberculosis.

Meza et al [28] estimated the optimal S2P timing without applying techniques from causal modeling. They used parametric hazard analysis to model transplant-free survival (TFS) from (1) Norwood procedure to S2P and (2) S2P to 3 years. They determined the optimal timing of S2P by generating nomograms of risk-adjusted TFS *vs.* the interval from the Norwood procedure to S2P. Their method does not address confounding bias; for further discussion see Section 1.4.5 below.

Some currently available methods [22, 35] address the situation where there is one type of treatment with no competing risks. By contrast, HLHS patients after the initial Norwood surgery face four possible mutually exclusive events: elective S2P, non-elective S2P, cardiac transplantation, or death before S2P. An S2P is "elective" if regular monitoring of the patient suggests that the team can safely schedule the procedure after a brief time, often a week or two. It is "non-elective" if monitoring suggests that the operation must take place urgently. Because only one of these four events can occur, we identify the pre-S2P/post-Norwood outcomes as competing risks.

To estimate without bias the effect of S2P timing on post-S2P survival, one must adjust for potential confounding effects of the choice of treatment. Within the framework of the Rubin Causal Model (RCM), one accomplishes this by modeling the selection of timing as a function of potential confounders in a propensity score analysis. In the HLHS setting we cannot create propensity scores from a logistic regression for treatment assignment because i) there are several possible times of S2P surgery, and ii) the possible events constitute competing risks. We therefore propose to use an extension of propensity score

analysis that assumes multiple treatments [23, 24, 36], basing our scores on a discrete-time competing risk analysis of the possible post-Norwood but pre-S2P outcomes. This model produces a generalized propensity score, from which we compute inverse probability weights that we apply in a proportional hazards regression [7] to describe the effect of S2P timing on post-S2P survival.

1.2. The Single Ventricle Reconstruction Trial

Our data are from the US National Heart, Lung, & Blood Institute Pediatric Heart Network Single Ventricle Reconstruction Trial (SVRT), conducted at 15 US centers between 2005 and 2008 [30]. Its primary aim was to compare one-year transplant-free survival of newborns who were randomly assigned to have their Norwood procedure either with a modified Blalock-Taussig shunt (MBTS) or a right ventricle-to-pulmonary artery (RVPA) shunt. Among 548 randomized HLHS patients who had operative data and follow-up, 139 died prior to S2P, and 9 received a heart transplant. The remaining 400 underwent S2P (Figure 1.1). The medical team determined the timing based on clinical assessment, transcutaneous oxygen saturation measurements, and echocardiographic findings.

The SVRT data set includes information on pre-Norwood medical history, Norwood procedure and post-operative course, S2P operation and post-operative course, postnatal echocardiograms (pre- and post-Norwood and pre-S2P), pre-S2P angiography, and cardiac catheterization [28]. We have divided the S2P timing (measured in months from the date of the Norwood procedure) into six categories: 1–3 months, 4 months, 5 months, 6 months, 7 months, and $\geq$ 8 months. This gives adequate numbers of patients in each category to render the generalized propensity score model estimable.

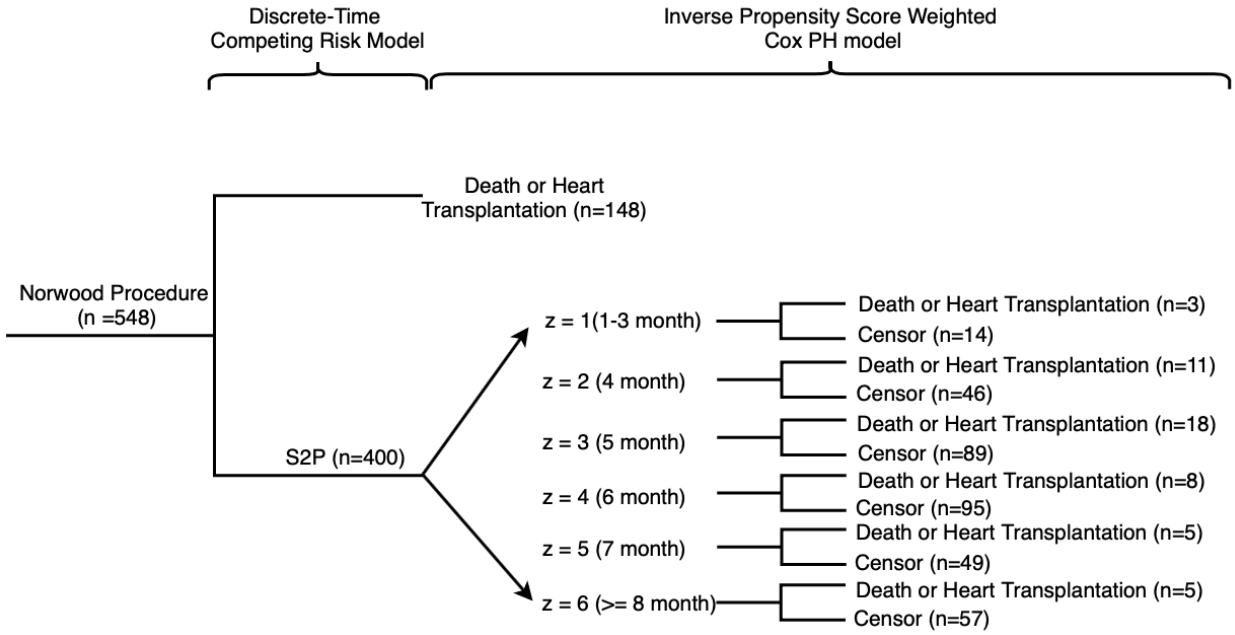

Figure 1.1: Treatment and outcome patterns in the SVRT data

1.3. A Model-Based Analysis Method

Our modeling strategy involves three elements: An extended propensity score analysis, a competing-risks model that permits estimation of propensity scores, and a hazard model relating S2P time to post-S2P survival.

1.3.1. The generalized propensity score

Assume a collection of subjects $i = 1, \ldots, n$, for each of whom we observe a baseline covariate vector X_i and a treatment $Z_i \in \{1, \ldots, M\}$. We moreover suppose that each has a notional vector $\{T_i(z)\}_{z=1}^{M}$, where $T_i(z)$ represents the post-S2P survival outcome should patient i receive treatment z. Only $T_i(Z_i)$ is observed; the other potential outcomes are counterfactuals. In this setting, Imai et al [23] define the *generalized propensity score*

$$e_i(z,x) =: \mathsf{Pr}(Z_i = z|X_i = x) = \mathbb{E}[D_i(z)|X_i = x],$$

where

$$D_i(z) = \begin{cases} 1, & \text{if } \ Z_i = z, \\ 0, & \text{otherwise} \end{cases}$$

is an indicator function for the treatment for subject i. The generalized propensity score has similar properties to the conventional propensity score [36]:

1. It is a *balancing score:*

$$D_i(z) \perp X_i \mid e_i(z, X_i) \quad \forall z;$$

   that is, given the generalized propensity score, the actual treatment and the covariate vector are independent.

2. It renders treatment assignment *unconfounded (ignorable)*:

$$D_i(z) \perp T_i(z) \mid e_i(z, X_i) \quad \forall z;$$

   that is, the potential outcomes and the actual treatment assignment are conditionally independent given the generalized propensity score.

1.3.2. A competing-risks model for post-Norwood disposition

After undergoing the Norwood procedure, patients go on to experience either S2P or death/heart transplantation. We denote Z to be the discrete time since Norwood of the first occurrence of one of these events, and we denote R to be the type of event that occurs: $R = 1$ for S2P and $R = 2$ for death/heart transplantation. This event R is distinct from the post-S2P survival outcome. Data permitting, we can re-define R as $R = 1$ for elective S2P, $R = 2$ for non-elective S2P, and $R = 3$ for death/heart transplantation to achieve a finer subgroup analysis.

As only one of these post-Norwood/pre-S2P events R can occur, we specify for Z a competing-risks model that constitutes the basis of our propensity score analysis. Let $h_r(z|x)$ denote the cause-specific hazard function for discrete time $Z \in \{1, \ldots, M\}$ of event $R \in \{1, \ldots, J\}$; that is,

$$h_r(z|x) = \mathsf{Pr}(Z = z, R = r | Z \geq z, X = x),$$

where x is a vector of covariates. The overall event hazard is $h(z|x) = \sum_{r=1}^{J} h_r(z|x)$. The probability that an event of type r takes place at time z is

$$\mathsf{Pr}(Z = z, R = r | X = x) = h_r(z|x) S(z|x), \tag{1.1}$$

where $S(z|x)$ is the survival function $S(z|X = x) = \mathsf{Pr}(Z \geq z | X = x) = \prod_{v=1}^{z-1} (1 - h(v|x))$.

Tutz et al [49] proposed a multinomial logistic model for cause-specific hazards with discrete Z:

$$h_r(z|x) = \frac{\exp(\beta_{0zr} + x\beta_r)}{1 + \sum_{j=1}^{J} \exp(\beta_{0zj} + x\beta_j)}, \tag{1.2}$$

where β_{0zr} stands for the cause-specific baseline hazard function at time z for event r, and β_r is the vector of cause-specific coefficients. This model assumes a time-constant effect of covariates on each cause-specific hazard. The probability that no event occurs at time z, given survival to that point, is

$$h_0(z|x) = \mathsf{Pr}(Z > z | Z \geq z, X = x) = 1 - \sum_{r=1}^{J} h_r(z|x) = \frac{1}{1 + \sum_{j=1}^{J} \exp(\beta_{0zj} + x\beta_j)}$$

Alternatively, one can express the model as

$$\ln\left(\frac{h_r(z|x)}{h_0(z|x)}\right) = \beta_{0zr} + x\beta_r. \tag{1.3}$$

We identify the generalized propensity score as the probability of undergoing S2P at time z given the covariate vector x, $\mathsf{Pr}(Z = z, R = 1|X = x)$. The discrete S2P time is regarded as the treatment Z in the generalized propensity score described above. This makes the generalized propensity score for outcome r in our analysis

$$e_i^{(r)}(z, x) = \mathsf{Pr}(Z_i = z, R_i = r|X_i = x).$$

(As the r is understood in subsequent analyses, we henceforth drop the superscript on e_i.) In a weighted propensity-score analysis, the weight function is the inverse of the generalized propensity score for subject i given the observed S2P time and the covariates:

$$w_i = \frac{1}{e_i(Z_i, X_i)}.$$

We denote this the *raw weight* for subject i. Because analyses using small raw weights can be unstable, it is often preferable to use the stabilized weight [35]

$$w_i^* = \frac{\mathsf{Pr}(Z_i = z, R_i = r)}{e_i(Z_i, X_i)}.$$

Here, $\mathsf{Pr}(Z_i = z, R_i = 1)$ is the marginal probability that subject i received S2P at time $z \in \{1, 2, \ldots, 6\}$. We estimate this probability by averaging the estimated values of the terms in Equation (1.1) for $R = 1$ at each time z over the S2P patients.

1.3.3. A spline model for post-S2P survival

We describe the effect of S2P timing Z on post-S2P survival T by a proportional hazards model, adjusting for confounding by inverse generalized propensity score weighting [6, 35, 55]. The hazard at time t is

$$\lambda(t|R=1;z,\alpha) = \lambda_0(t|R=1)\exp[g(z;\alpha)], \tag{1.4}$$

where $\lambda_0(t|R=1)$ is the post-S2P baseline hazard at time t for patients having S2P, and $g(z;\alpha)$ is a linear spline function in z (the S2P treatment time) with coefficient vector α:

$$g(z;\alpha) = \alpha_1 z + \sum_{q=2}^{Q+1} \alpha_q (z - K_{q-1})_+.$$

Here, $K_1, ..., K_Q$ are preselected knots and $(v)_+ = \max(0, v)$. In the linear spline model, Z is still discrete with possible values 1 to 6 corresponding to the six time categories introduced in Section 1.2. We use the Cox model with inverse generalized propensity score weighting because it is flexible and easy to estimate.

Three key points motivate our model choice:

1. We interpret Z as a discrete, ordinal representation of time, because this confers protection against potential distorting effects of small numbers of outlying S2P times.

2. The use of a regression spline to define the dependence of the hazard on Z is simply a device for parameterizing the model, analogous to the choice of a contrast matrix in an analysis of variance. Although spline regression typically presupposes a continuous predictor, it need not. By selecting Model (1.4) from among the linear splines with knots at all Z values, we obtain a parsimonious representation of the hazard function.

3. Although we consider Z here to be discrete, the model is readily applicable, without modification of notation or code, in situations where Z is continuous.

In the Appendix A we present sensitivity analyses that demonstrate robustness to the parameterization of the hazard model and to the use of continuous rather than discrete Z.

For the generalized propensity score model, we assume that there is no unmeasured confounding, effectively that the vector x contains all potential confounders. This assumption is questionable, as it would be in any analysis of observational data. For the Cox model with weighting, we assume that censoring time is independent of the potential mortality outcome given the propensity score. Because censoring of post-S2P survival in SVRT is essentially administrative, this assumption is plausible. Under these assumptions, Model (1.4) encodes the counterfactual hazard rate of $T(z)$ at t, which we will consider as our causal quantity of interest. We are interested in finding the z that minimizes this quantity.

1.3.4. Estimating the spline model for post-S2P survival

Suppose we have n^* subjects as above and let Δ_i be an indicator that equals 1 if subject i dies after S2P and 0 otherwise. $T^* = \min(T, C)$ is the observed post-S2P follow-up time and C is the censoring time. The risk set $\mathscr{R}_i$ is the set of subjects who are alive just prior to an observed time of death t_i. The categorical variable Z is the S2P time as above. It is reasonable to start with a saturated linear spline model and select a subset of influential knots by the LASSO [47]. We compute the coefficients $\widehat{\alpha}$ (we show in Appendix A.3 that $\widehat{\alpha}$ is consistent) by solving the estimating equations with the preselected knots:

$$\frac{\partial \ln L(\alpha)}{\partial \alpha_1} = \sum_{i=1}^{n^*} \Delta_i w_i^* \left[z_i - \frac{\sum_{l \in \mathscr{R}_i} z_l \exp\left(g(z_l; \alpha)\right)}{\sum_{l \in \mathscr{R}_i} \exp\left(g(z_l; \alpha)\right)} \right] = 0; \tag{1.5}$$

$$\frac{\partial \ln L(\alpha)}{\partial \alpha_q} = \sum_{i=1}^{n^*} \Delta_i w_i^* \left[(z_i - K_{q-1})_+ - \frac{\sum_{l \in \mathscr{R}_i} (z_l - K_{q-1})_+ \exp\left(g(z_l; \alpha)\right)}{\sum_{l \in \mathscr{R}_i} \exp\left(g(z_l; \alpha)\right)} \right] = 0, \tag{1.6}$$
$$q = 2, 3, \ldots, Q+1.$$

Here w_i^* is the stabilized weight for subject i. One can calculate standard errors for $\{\widehat{\alpha}_s\}_{s=1}^{Q+1}$ by inverting the weighted information matrix. The weighting creates a pseudo-population consisting of w_i^* (or w_i) copies of each subject i [35]. The intuition is that if the propensity score for an observed S2P timing is small, its behavior is underrepresented in

the sample, and assigning it a large weight reduces bias.

Having estimated the model, we identify the optimal S2P timing by plotting the mortality hazard linear spline against z and finding its minimum. We create a $100(1-\alpha)$ confidence interval for the optimal z by bootstrapping the data [11] and identifying the $\alpha/2$ and $(1-\alpha/2)$ quantiles of the bootstrap distribution of optimal S2P times.

1.4. Application to the SVRT Data

1.4.1. Preliminary analyses

Descriptive statistics by Norwood outcome appear in Table 1.1. We note that children who underwent S2P were heavier at birth and slightly younger at the time of the Norwood procedure and post-Norwood discharge. Roughly 60% were male and 80% were white in all groups. Subjects who underwent RVPA were more likely to survive to S2P.

Table A.1 (see Appendix A.2) presents baseline data for the 400 S2P recipients, grouped by S2P time. Mean birth weight and mean age at the Norwood procedure are similar across S2P time groups. The mean Norwood discharge age is younger for those having the S2P 4–6 months after the Norwood procedure. Figure A.1 (see Appendix A.1) shows histograms of S2P time stratified by elective status. Most non-elective S2Ps occurred within 6 months of the Norwood, whereas elective procedures occurred on average later.

1.4.2. Generalized propensity score analysis

All patients who underwent the Norwood procedure proceeded to either S2P or death or cardiac transplantation without S2P. We used the discrete-time competing-risk model described above to estimate generalized propensity scores for time of S2P, with time

Table 1.1: Descriptive statistics by outcome of the Norwood procedure

Variable	Values	S2P ($n = 400$)	Death/Tx[a] ($n = 148$)
Discrete			
Sex	M	252(63.0%)	87(58.8%)
	F	148(37.0%)	61(41.2%)
Race	White	323(80.8%)	112(75.7%)
	Black	59(14.7%)	27(18.2%)
	Other	18(4.5%)	9(6.1%)
Norwood treatment	MBTS	177(44.3%)	91(61.5%)
	RVPA	223(55.7%)	57(38.5%)
Prenatal diagnosis of congenital heart disease	Y	312(78.0%)	107(72.3%)
	N	88(22.0%)	41 (27.7%)
Aortic atresia	Y	243(60.8%)	99(66.9%)
	N	157(39.2%)	49(33.1%)
Obstructed pulmonary venous return	Y	6(1.5%)	13(8.8%)
	N	394(98.5%)	135(91.2%)
Continuous[b]			
Birth weight (gm)		3157(518)	2961(576)
Norwood age (days)		6.7(4.0)	6.8(4.2)
Norwood discharge age (days)		40.1(32.9)	42.1(46.1)

[a]Death or heart transplantation
[b]Values given as mean(SD)

origin the date of the Norwood procedure. We used as covariates sex, race, birth weight, Norwood procedure randomization arm, prenatal diagnosis of congenital heart disease, aortic atresia, obstructed pulmonary venous return, and Norwood discharge age. These are all baseline covariates.

As described above, we obtain the generalized propensity score from Equation (1.1) with $R = 1$ (for S2P) or with $R = 2$ (for death/heart transplantation). To fit the multinomial logit model (1.3) and achieve the cause-specific hazard (1.2) in discrete time, we transformed the data into a long form, with one row per patient per time category. The estimated model coefficients from the SVRT data are those shown under the column "Truth" in the simulation settings of Table 1.2.

Figure 1.2 shows the result of the extended propensity score analysis. The average probability of having S2P by month, adjusted for covariates and averaged over all patients, rises steadily until month 5 then declines to near 0. Evidently, most patients undergo the S2P from about 4 to 7 months after the Norwood. The probability of heart transplantation or death declines over time from its modal value at months 1–3.

The stabilized weight for each subject used in the following analysis is the product of the inverse generalized propensity score and the marginal probability of having S2P. We coded the 6 possible values of S2P time Z as 1 to 6. These marginal probabilities are 0.053, 0.127, 0.173, 0.138, 0.045, 0.029. Probabilities in Figure 1.2 represent all 548 subjects.

1.4.3. Assessing covariate balance

Table A.1 shows that the covariates sex, race, Norwood arm, and Norwood discharge age are poorly balanced across timing groups. We checked covariate balance given the generalized propensity score using the method of Zhu et al [58]. This involved creating a synthetic data set by sampling 5,000 observations with replacement from the original data

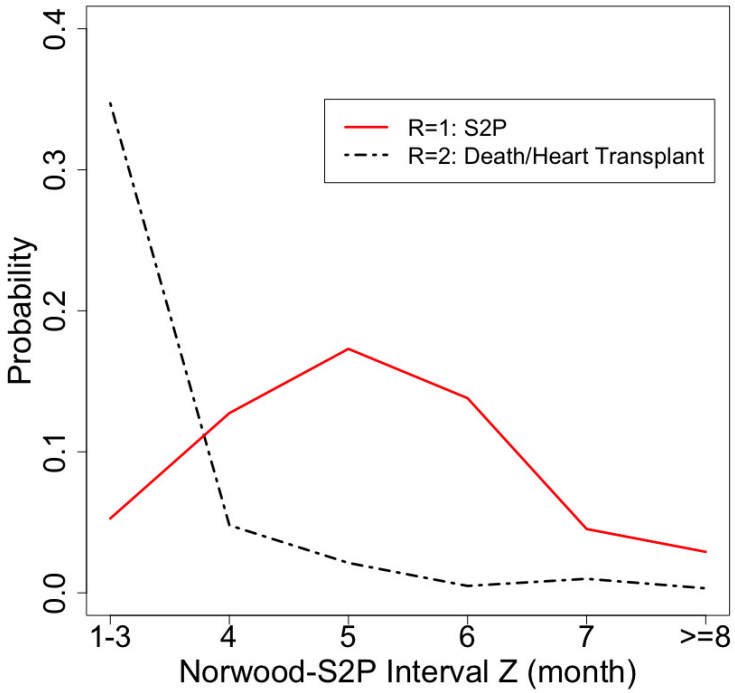

Figure 1.2: Average probability of competing events over time, estimated from the competing-risk model with outcomes S2P and death/transplantation

set with sampling probability equal to the stabilized weight. We then evaluated balance in the synthetic data by calculating Cramér's V statistic and setting the threshold at 0.1. Our analysis suggested that balance was improved and was acceptable for further analysis.

1.4.4. Optimal S2P time

Our best-fitting survival model, evaluated by AIC, placed knots at $z = 3$ and $z = 4$, corresponding to times 5 and 6 months. This gave estimated post-S2P survival hazard

$$\lambda(t|R = 1; z) = \lambda_0(t|R = 1) \exp\left[-0.19z - 0.66(z - 3)_+ + 0.99(z - 4)_+\right].$$

The standard errors for the coefficients are, respectively, 0.072, 0.125 and 0.210. The log hazard ratio over Z appears in Figure 1.3, with month 6 as the reference level. From this plot, patients receiving the S2P intervention at 6 months after the Norwood have the lowest hazard and therefore the best survival prospects. A 95% bootstrap confidence interval for the optimal S2P timing is $[5, \infty)$. The estimated 1-year post-S2P mortality risk (with 95% CI) associated with this result appears in Table A.3.

In a sensitivity analysis, we analyzed post-S2P survival including only elective or only non-elective S2P patients (Figures A.3 and A.4) by re-categorizing R: $R = 1$ for elective S2P, $R = 2$ for non-elective S2P, and $R = 3$ for death/heart transplantation. The optimal S2P timing for elective S2P patients is 6 months after the Norwood procedure with 95% bootstrap confidence interval [5, 7] months. By contrast, for non-elective S2P, the optimal S2P timing is as late as possible with confidence interval $[4, \infty)$. A possible explanation for this difference is that patients are selected for non-elective S2P because their general health condition is poor and unstable; that is, patients who are doing worse should undergo the S2P later.

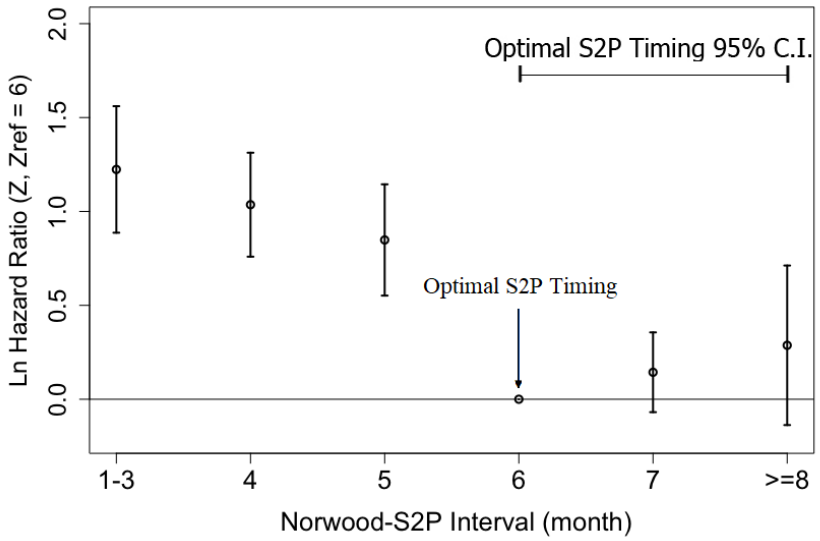

Figure 1.3: Log mortality hazard ratio as a function of S2P timing, including all S2P patients

In this relatively small clinical trial, there is little power to evaluate the interaction between Norwood procedure type (MBTS or RVPA) and S2P timing. Unsurprisingly, when we conducted this analysis we found no significant effect.

In the Appendix A, we present a sensitivity analysis that compares the discrete spline model to i) a model with indicators for the discrete Z values and ii) a model that smooths the hazard function with a linear spline on the continuous Z values. In the latter approach, although Z is taken to be continuous, we use the weights from the discrete-Z competing-risk model. Results show modest sensitivity to the hazard-modeling method when taking Z to be discrete. Estimates were similar when considering Z to be continuous, although this model exhibited greater sensitivity to the selection of knots.

1.4.5. Comparison with the analysis of Meza *et al.*

Meza et al [28] analyzed the effect of S2P timing on survival using a "parametric hazard" analysis. They began by constructing separate parametric models for transplant-free survival after the Norwood procedure and survival after S2P. They then transformed the post-S2P survival analysis results into a 3-year transplant-free survival plot over Norwood-S2P intervals to determine the optimal S2P timing, concluding that the interval 3 to 6 months is best. We have some concerns about this analysis: First, it does not explicitly adjust for covariate effects on treatment assignment, as we do in our generalized propensity score analysis; thus their estimates may be sensitive to potential model misspecification [10]. Second, when creating the nomogram to find the optimal S2P timing, they failed to consider the competing risks of death and heart transplantation. Finally, they did not distinguish elective from non-elective S2P, potentially biasing their estimate toward the earlier times often seen in non-elective S2P operations.

1.4.6. A re-analysis using the method of Hu *et al.*

Although not designed to describe sequential treatments like the HLHS program, the structural proportional hazards model of Hu et al [22] is potentially applicable to the SVRT data. Identifying HIV/TB co-infection with the Norwood procedure, and antiretroviral therapy with S2P, and censoring other Norwood outcomes (death and heart transplantation), we were able to apply their method to the SVRT data. Results appear in Figure A.2, with confidence intervals obtained by bootstrapping. This analysis gives an optimal S2P timing at month 6 (95% CI [6, 8]), later than the Meza finding.

Our method differs from that of Hu et al [22] in several ways. First, we have formulated the problem as one of maximizing post-S2P survival (T) rather than pre- plus post-S2P survival ($Z + T$). We believe this better reflects the clinical situation, where the idea is to produce the best possible long-term survival. The novelty in our method is in its use of a

competing risk framework to model treatment selection.

We note moreover that our data differ in kind from those of Hu et al., who had a large number of subjects with censored Z (we have none) and only a small fraction of deaths prior to treatment (4%, compared to 26% in our sample).

1.5. Simulations

We performed simulations to evaluate the reliability of estimation of the discrete competing risk model and the Cox model with weighting in samples of moderate size (see Appendix A.4). We ran 2,000 replicates of each model with sample sizes of 500 and 1,000, setting true parameters to their estimates from the SVRT data. The main outcome was the coverage probability of nominal 95% confidence intervals for the model parameters.

Table 1.2 shows results for the competing-risk model. With $n = 500$, coverage probabilities exceed 90%, but some are noticeably below 95%. When we increase the sample size to $n = 1,000$, all coefficients have coverage probability near the nominal 95%.

Table 1.3 shows results for simulated post-S2P survival data. Here λ^* is the scale parameter and γ^* is the shape parameter for the underlying survival distribution. We fixed the spline knots at 3 and 4; thus, there are three parameters (α_1,α_2,α_3) in total for the spline function $g(\cdot)$ discussed in Section 1.3.3. Table 1.3 shows that using the raw weight w_i leads to unacceptably low coverage probabilities, whereas using the stabilized weight w_i^* gives coverage probabilities close to 95%. Coverage probabilities are noticeably closer to the nominal level when the sample size is $n = 1,000$.

Table 1.2: Monte Carlo coverage probabilities of nominal 95% confidence intervals for parameters of the competing-risk model

				Coverage Probability (%)	
Outcome	**Variable**	**Coefficient**	**Truth**	$n = 500$	$n = 1,000$
S2P	Intercept $Z = 1$	β_{011}	-4.126	94.3	95.5
	Intercept $Z = 2$	β_{021}	2.352	93.1	95.1
	Intercept $Z = 3$	β_{031}	3.210	92.9	95.0
	Intercept $Z = 4$	β_{041}	3.914	93.4	94.7
	Intercept $Z = 5$	β_{051}	3.856	92.5	95.5
	Intercept $Z = 6$	β_{061}	4.237	93.0	96.0
	Sex:F	β_{11}	0.052	91.5	94.9
	Race:W	β_{12}	-0.037	91.7	94.9
	Race:B	β_{13}	-0.406	92.0	95.0
	Norwood treatment:RVPA	β_{14}	0.184	90.6	94.6
	Birth weight	β_{15}	0.0002	91.3	96.2
	Prenatal diagnosis of congenital heart disease:Yes	β_{16}	0.028	92.6	94.9
	Aortic atresia:Yes	β_{17}	-0.050	91.6	95.1
	Obstructed pulmonary venous return:Yes	β_{18}	0.200	93.6	94.5
	Norwood procedure discharge age	β_{19}	-0.008	92.2	95.3
Death/Heart Transplant	Intercept $Z = 1$	β_{012}	0.748	93.4	95.3
	Intercept $Z = 2$	β_{022}	-0.645	93.0	95.8
	Intercept $Z = 3$	β_{032}	-0.938	92.7	95.7
	Intercept $Z = 4$	β_{042}	-1.506	91.9	95.4
	Intercept $Z = 5$	β_{052}	0.301	92.1	94.8
	Intercept $Z = 6$	β_{062}	0.102	92.4	95.7
	Sex:F	β_{21}	0.211	90.5	94.6
	Race:W	β_{22}	-0.176	91.3	95.3
	Race:B	β_{23}	-0.071	90.5	95.7
	Norwood treatment:RVPA	β_{24}	-0.591	92.0	94.8
	Birth weight	β_{25}	-0.001	92.8	94.7
	Prenatal diagnosis of congenital heart disease:Yes	β_{26}	0.163	91.5	95.9
	Aortic atresia:Yes	β_{27}	-0.148	91.4	95.7
	Obstructed pulmonary venous return:Yes	β_{28}	1.906	91.7	95.9
	Norwood procedure discharge age	β_{29}	-0.004	91.4	95.2

Table 1.3: Monte Carlo coverage probabilities of nominal 95% confidence intervals for parameters of the post-S2P survival model

			Coverage Probability (%)			
			Raw[a]	Stb[b]	Raw	Stb
Survival distribution	**Coefficient**	**True value**	$n = 500$	$n = 500$	$n = 1,000$	$n = 1,000$
Exponential ($\lambda^* = 1$)	α_1	-0.19	51.5	94.0	54.0	95.3
	α_2	-0.66	55.8	93.9	58.6	94.2
	α_3	0.99	53.1	93.5	59.5	95.7
Weibull ($\lambda^* = 1, \gamma^* = 1.5$)	α_1	-0.19	54.4	92.5	55.3	94.4
	α_2	-0.66	57.8	93.2	60.2	96.1
	α_3	0.99	59.0	94.5	58.5	95.8
Gompertz ($\lambda^* = 1, \gamma^* = 1$)	α_1	-0.19	61.0	94.7	60.6	94.5
	α_2	-0.66	52.3	93.0	57.2	94.8
	α_3	0.99	53.8	94.3	59.8	95.5

[a]Raw weights
[b]Stabilized weights

1.6. Discussion

SVRT was effectively an observational study with respect to the choice of S2P timing; accordingly, we have proposed a generalized propensity score analysis to remove potential confounding of S2P time and post-S2P survival. The first stage involves estimating a discrete competing-risks model to create a generalized propensity score for predicting time of the S2P. The second stage involves estimating a Cox model predicting survival from S2P time, using the inverse propensity scores as weights. Our analysis of the SVRT data suggests that conducting the S2P at 6 months after the Norwood is optimal for limiting mortality risk.

Originally, we treated death and transplant as separate competing risks. Because only a few patients experienced transplant prior to S2P, we ultimately combined these two events into a single category. Although we did this largely for convenience, in fact death and transplant both represent treatment failure; in many cases a child who died would

have instead received a transplant if a donor heart had been available at the time.

A key parameter in determining the suitability of a post-Norwood-procedure patient for S2P is the pulmonary vascular resistance (PVR), a marker of maturity of the infant's circulatory system. Surgeons are unwilling to perform S2P until the PVR has declined below a safe level. Because the SVRT was not designed to study the timing of S2P, only one PVR measurement field appears in the data set, and almost half of these observations are missing ("PVR measurement not done") [28]. Therefore, we did not include the PVR in our analysis — a potentially serious limitation of this study.

Because there are no widely accepted guidelines for checking balance of categorical covariates over a multi-level treatment, we assessed covariate balance using the Cramér's V statistic [8]. The investigation of methods for assessing balance in these complex situations is a potential topic for further study.

We conclude that the optimal timing of S2P is roughly 6 months post-Norwood, with a 95% confidence interval of 5 to infinity. This fits into the current clinical experience. The reason for this strange upper limit is that the last category of our discrete S2P time is "≥ 8". Because a large majority of HLHS patients undergo S2P within one year of the Norwood procedure, it is difficult to assess mortality for higher values of Z. Our interval is slightly later than Meza's suggestion of 3–6 months, but similar to our reinterpretation of the SVRT data using the method of Hu et al [22], which gave the interval 6–8 months. With the substantial risk of death, transplant, and unscheduled S2P, the "optimal" time is at best a target that medical teams will aspire to achieve in less complicated cases.

CHAPTER 2

Estimating the Optimal Timing of Surgery by Imputing Potential Outcomes

2.1. Introduction

Hypoplastic left heart syndrome (HLHS), a complex of congenital conditions characterized by underdevelopment or absence of the left ventricle, is uniformly fatal without prompt treatment. The combination of the Norwood procedure followed by stage 2 palliation (S2P), introduced in the early 1980s, has led to vastly improved survival, with many patients now living to adulthood [56]. It is known that the choice of S2P timing affects the post-S2P survival time, although it is not known which time, if any, is best. A key potential data source for answering this question is the Single Ventricle Reconstruction Trial (SVRT), a study that randomized HLHS patients between versions of the Norwood procedure and assigned a range of S2P times according to clinical criteria [30]. We describe the SVRT in greater detail in Section 2.3.1.

Estimating the optimal S2P time involves a causal comparison of outcomes under the range of possible times of treatment. In the parlance of the Rubin causal model [25], a *potential outcome* is a value of the outcome of interest under a particular choice of treatment. In HLHS, the outcome of interest is the post-S2P survival time, and the potential outcomes are the values of this variable under the different possible S2P times. Because in the SVRT the S2P timing was not randomly assigned, confounding can bias estimation. Current methods to investigate the causal effect of a time-dependent treatment on the distribution of an outcome mainly rely on the marginal structural model [22, 35]. These

methods use an inverse probability weighted marginal proportional hazard model to remove the confounding bias. Westreich et al [53] discuss causal inference by imputation of the potential outcomes using multiple imputation or the *g*-formula [34] based on time-fixed dichotomous treatment. They impute the potential outcomes independently, rather than simultaneously, because the original data contain no information about the joint distribution of potential outcomes. We have proposed a generalized propensity score analysis, using inverse propensity score weighting to eliminate confounding bias in estimation of the optimal S2P timing [4]. The method regards post-Norwood/pre-S2P outcomes, including death, heart transplantation and S2P, as competing risks, deriving generalized propensity scores from the estimated competing-risk model.

In this article, we propose to treat unobserved potential outcomes as missing observations, conducting a model-based analysis by multiply imputing them [38, 40]. The novelty is in using this method to estimate the causal effect of S2P timing on censored survival data. Multiple imputation is a straightforward and broadly applicable method that handles missing or otherwise incomplete observations in a principled way. We adopt it here for several reasons:

1. It is straightforward to apply in the SVRT example.
2. Unlike a propensity score analysis, it includes all subjects, even those who did not ultimately undergo S2P.
3. It simplifies analyses for sensitivity to assumptions about which the data are uninformative.
4. It correctly accounts for uncertainty about both missing observations and unobserved potential outcomes [45].

In the next section, we describe the Rubin Causal Model framework and formulate the needed assumptions and analysis strategy. Specifically, we impute all unobserved post-S2P survival potential outcomes under a lognormal model. Unlike Westreich et al [53],

who assumed independence, we can impute the potential outcomes under various assumptions about the correlation of potential outcomes (Section 2.2.2). It is straightforward to estimate the causal effect of S2P timing on post-S2P survival by directly comparing the imputed potential outcomes. We then perform a second analysis based on a restricted cubic spline (RCS) model for the baseline hazard in a proportional hazard model. This analysis relaxes the strong parametric assumption of the lognormal model. In Section 2.3 we apply our methods to the SVRT data set. In Section 2.4 we offer further discussion.

2.2. Methods

The Rubin causal model [25] takes the potential outcomes for a subject to be the outcomes that one would observe under the possible levels of a treatment or exposure. Each potential outcome is *a priori* observable, in that one could observe it if the unit were to receive the corresponding treatment, although we ultimately see only the outcome for the assigned treatment. An approach to the analysis of potential outcomes is therefore to treat the unobserved outcomes as missing observations and impute their values under an estimated statistical model. In this way, we can estimate an average causal effect by averaging the imputed causal effects across the sampled units.

A key hypothesis is the *Stable Unit Treatment Value Assumption (SUTVA)*, which posits that there is no interference between units and no hidden variation of treatments among units. “No interference” means that potential outcomes for a given unit do not vary with the treatments assigned to other units. This seems plausible in our application. “No hidden variation of treatments” means that there is a single version of each treatment, in this case the timing of the S2P; this also is plausible.

The fundamental problem of causal inference is the inevitability of missing data; for each unit, we can observe at most one potential outcome. Thus it is necessary to assume a missingness mechanism — here essentially a treatment assignment mechanism

— and elucidate its consequences for inferences [37]. For SVRT, we assume that the unobserved potential outcomes are missing at random (MAR), in the sense that the probability that a particular potential outcome is unobserved, given observed and missing variables, depends only on observed variables such as baseline predictors of outcome. That is, the probability that the subject undergoes S2P at a particular time does not depend on the survival values at the unselected times, given all observed covariate and outcome data. MAR is sufficient to justify standard likelihood-based estimation of the model for the potential outcomes. We make this assumption for both of our imputation models.

Although it is impossible to robustly test the assumption of MAR, one can evaluate the sensitivity of inferences to departures from it. Imputing data under a plausible non-ignorable model [20, 50] is a common way to evaluate robustness. If results differ under alternative assumptions, we can conclude that analyses that flow from an MAR model are unreliable.

2.2.1. Model-based multiple imputation and analysis

Suppose we have a collection of subjects $i = 1, \ldots, n$ who underwent the Norwood procedure. Among them, $n^* \leq n$ later underwent S2P, the rest either dying or receiving a heart transplant. We observe a q-dimensional baseline covariate vector X_i and an S2P treatment time $Z_i \in \{1, \ldots, \Xi\}$. For simplicity, we have divided the set of possible S2P times into six categories: 1–3 months, 4 months, 5 months, 6 months, 7 months, and ≥ 8 months, coded 1–6. With this formulation there are enough patients in each category to make parametric model estimation feasible. We moreover assume that a notional vector $T_i = (T_i(1), \ldots, T_i(6))^{\mathsf{T}}$ gives the potential outcomes for post-S2P survival under the range of possible treatments. The realized event time $T_i(Z_i)$ may be censored by an observed censoring time C_i. We let D_i be an indicator that equals 1 if subject i is observed to die after S2P and 0 otherwise, and $T_i^* = \min(T_i(Z_i), C_i)$ (Table 2.1). Because censoring in this study is administrative, C_i is independent of treatment assignment and $T_i(Z_i)$.

Table 2.1: Example observations in the SVRT data

	Potential Outcomes (days)								
Unit	$T_i(1)$	$T_i(2)$	$T_i(3)$	$T_i(4)$	$T_i(5)$	$T_i(6)$	Z_i	D_i	T_i^*
1	?	?	?	?	?	1311	6	0	1311
2	?	1331	?	?	?	?	2	0	1331
3	?	?	?	960	?	?	4	0	960
4	?	?	?	820	?	?	4	0	820
5	?	?	822	?	?	?	3	0	822
6	?	690	?	?	?	?	2	0	690
7	?	?	?	499	?	?	4	1	499
8	?	?	72	?	?	?	3	1	72
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

2.2.2. A lognormal model

We assume that the log-transformed potential outcome vector $\ln T_i$ follows the multivariate normal distribution

$$\ln T_i \,|X_i; \psi \sim \mathcal{N}(\beta X_i + \eta, \Sigma), \tag{2.1}$$

where ψ represents the model parameters. That is, the mean vector of the distribution is $\mu_i = \beta X_i + \eta$ where X_i is a q-vector of baseline predictors;

$$\beta = \begin{pmatrix} \beta(1) \\ \vdots \\ \beta(6) \end{pmatrix}$$

is a 6-by-q coefficient matrix for X_i; and $\eta = (\eta(1), \ldots, \eta(6))^{\mathsf{T}}$ is a six-vector representing the causal treatment effects on the log potential outcomes. Thus $\beta(Z_i) = (\beta_1(Z_i), \ldots, \beta_q(Z_i))$ is the q-vector of covariate effects on the outcome under the realized treatment assignment Z_i. We moreover assume that the covariance matrix has diagonal elements $\Sigma_{zz} =$

$\sigma_z^2,\ z \in \{1,\ldots,6\}$ and off-diagonal elements $\sigma_{z'z} = \rho\sigma_z\sigma_{z'}$ for $z' \neq z \in \{1,\ldots,6\}$, with $\rho \in [0,1)$. Because the likelihood function contains no information on off-diagonal diagonal elements of Σ, we set $\psi = (\beta, \eta, \sigma_1^2, \ldots, \sigma_6^2)$, effectively assuming that the potential outcomes are uncorrelated. We will use imputation to conduct sensitivity analyses under a range of assumed values of ρ to determine the impact of departures from independence on final inferences.

Another potential model simplification is to assume that the rows of β are equal — i.e., that $\beta(z) = \beta(z')$ for $z, z' \in \{1,\ldots,6\}$ [25]. This is analogous to equal covariate effects across treatment arms in the analysis of covariance. The data can speak to the validity of this assumption at the cost of estimating $5q$ additional parameters, which may be challenging if the number of observations in any observed treatment arm is modest.

With observed data (x, z, t^*, d), the log-likelihood is

$$\mathcal{L}(\psi; x, z, t^*, d) = \sum_{i=1}^{n^*} \left\{ d_i \ln \left[\frac{1}{t_i^*} \phi(\ln t_i^*; \mu_i(z_i), \sigma_{z_i}) \right] + (1 - d_i) \ln \left[1 - \Phi(\ln t_i^*; \mu_i(z_i), \sigma_{z_i}) \right] \right\},$$

where t_i^* is the observed value of T_i^*; d_i is the observed value of D_i; and $\mu_i(z_i) = \beta(z_i)x_i + \eta(z_i)$. $\phi(w; \nu, \tau)$ is the normal density and $\Phi(w; \nu, \tau)$ the normal distribution function evaluated at w with mean ν and standard deviation τ.

The first step in the imputation process is to draw M samples from the posterior distribution of the model parameters $p(\psi; x, z, t^*, d)$. We implement this by approximating $p(\psi; x, z, t^*, d)$ as $\mathcal{N}(\hat{\psi},\ I_{n^*}^{-1}(\hat{\psi}))$, where $\hat{\psi}$ is the MLE estimate and $I_{n^*}(\hat{\psi})$ is the observed information [40, 48]. The next step is to impute missing values given the observed data and the sampled parameter. There are three scenarios where potential outcomes are treated as missing:

1. A patient who does not receive S2P (pre-S2P death or heart transplant): We impute all six potential outcomes based on baseline covariates X_i.

2. A patient who underwent S2P and was observed to die thereafter: We impute the logs of the five counterfactual outcomes from their multivariate normal distribution conditional on the observed log death time.

3. A patient who underwent S2P and whose survival is censored: We sample the log event time from a truncated marginal survival distribution corresponding to observed S2P timing z_i. Conditional on this imputed event time, we sample the counterfactual log event times from their multivariate normal distribution given the sampled event time.

2.2.3. A restricted cubic spline hazard model

Inferences under Model (2.1) may be sensitive to departures from distributional assumptions that will be difficult to detect. An alternative approach is to assume a proportional hazards specification for the outcomes, modeling the baseline hazard as a flexible parametric function of time. The *restricted cubic spline (RCS)* model of Herndon et al [21] satisfies these requirements.

We specify J knots, $K_1 < K_2 < \cdots < K_J$, with K_1 and K_J designated as *boundary* knots. We define the restricted cubic spline baseline hazard $\lambda_0(u)$ at post-S2P survival time u as

$$\lambda_0(u) = \alpha_{J-1} + \sum_{j=0}^{J-2} \alpha_j \mathcal{B}_j(u),$$

where $\mathcal{B}_0(u) = u$ and, for $j = 1, \ldots, J-2$,

$$\mathcal{B}_j(u) = (u - K_j)_+^3 - \frac{(u - K_{J-1})_+^3 (K_J - K_j)}{(K_J - K_{J-1})} + \frac{(u - K_J)_+^3 (K_{J-1} - K_j)}{(K_J - K_{J-1})}.$$

This differs from the standard cubic spline in having fewer parameters and being constrained to be linear beyond the boundary knots.

We define the hazard function at post-S2P time u for subject i as

$$\lambda_i(u, z_i) = \left(\alpha_{J-1} + \sum_{j=0}^{J-2} \alpha_j \mathcal{B}_j(u) \right) \exp\left(\gamma^{\mathsf{T}} x_i + \kappa(z_i)\right), \tag{2.2}$$

where α_j is the coefficient for the restricted cubic spline; x_i is the covariate vector for subject i; z_i is the assigned treatment; γ is a vector of covariate effects; and κ is a vector of potential hazard factors. We denote the parameter of Model 2.2 as $\chi = (\alpha, \gamma, \kappa)$. With implicit conditioning on the parameter and covariate vector, the survival function $S_i(u, z_i)$ is then

$$S_i(u, z_i) = \exp\left(- \int_0^u \lambda_i(v, z_i) dv \right). \tag{2.3}$$

We impute incomplete observations under Model (2.2) by the Bayesian bootstrap [39]. To create the m-th imputed data set, we perform the following steps:

1. Sample a uniform Dirichlet deviate with the same dimension as the number of rows in the dataset.

2. Using the Dirichlet components as weights, sample a data set from the raw data with replacement.

3. Obtain $\hat{\chi}^{(m)} = (\hat{\alpha}^{(m)}, \hat{\gamma}^{(m)}, \hat{\kappa}^{(m)})$ by estimating Model (2.2) using the data from step (2).

4. Using the inverse-CDF transformation, impute the unobserved potential post-S2P survival outcomes $T_i^{(m)}(z),\ z \neq z_i$, independently from their distribution conditional on the data and the sampled parameters.

Step (4) involves creation of the completed data given observed data and the sampled parameter estimate. For completely unobserved outcomes, one samples from the survival function in Equation (2.3). For censored outcomes for the observed Z_i, one samples from the survival function restricted to lie beyond T_i. All sampling of completed observations assumes independence of potential outcomes.

2.2.4. Inference on the causal parameter

Under both statistical models, the imputed potential outcomes $T_i^{(m)}$ are the potential post-S2P survival days. For simplicity, take the S2P procedure to be successful when the imputed post-S2P survival is longer than 10 years; that is, the parameter of interest is the proportion of imputed post-S2P survival times exceeding 10 years for each treatment z. Denote this vector $\theta(z)$, and assume that its variance-covariance matrix from a complete data set is V. Denote the estimate and variance from the m-th imputed data set as $\hat{\theta}^{(m)}$ and $V^{(m)}$, respectively. Then one can create inferences about θ using the well-known Rubin combining formulas [40, 42].

We estimate the optimal S2P timing from the plot of $\hat{\theta}(z)$ against z with 95% confidence intervals. To construct a confidence interval for optimal S2P timing when using multiple imputation, we apply the method "Bootstrap+MI" of Schomaker et al [43]. This method involves creation first of a bootstrap sample of the data followed by a multiply imputed sample.

2.2.5. Computing

We created imputations in R 3.5.2 and analyzed them in SAS 9.4 using PROC MIANALYZE.

2.3. Application to the SVRT Data

2.3.1. The Single Ventricle Reconstruction Trial

The Pediatric Heart Network conducted the SVRT at 15 centers from 2005 to 2009; study data are available at http://www.pediatricheartnetwork.com. The primary goal of SVRT was to evaluate the effect of the type of Norwood procedure (stage 1 treatment) on one-year transplant-free survival of newborns with hypoplastic left heart syndrome (HLHS). The trial randomized newborns to undergo either a modified Blalock-Taussig shunt (MBTS) or a right ventricle-to-pulmonary artery (RVPA) shunt. Survivors would undergo S2P at a time determined by the surgical team, typically several months after the Norwood operation. Ohye et al [30] give more details on the study design and the data.

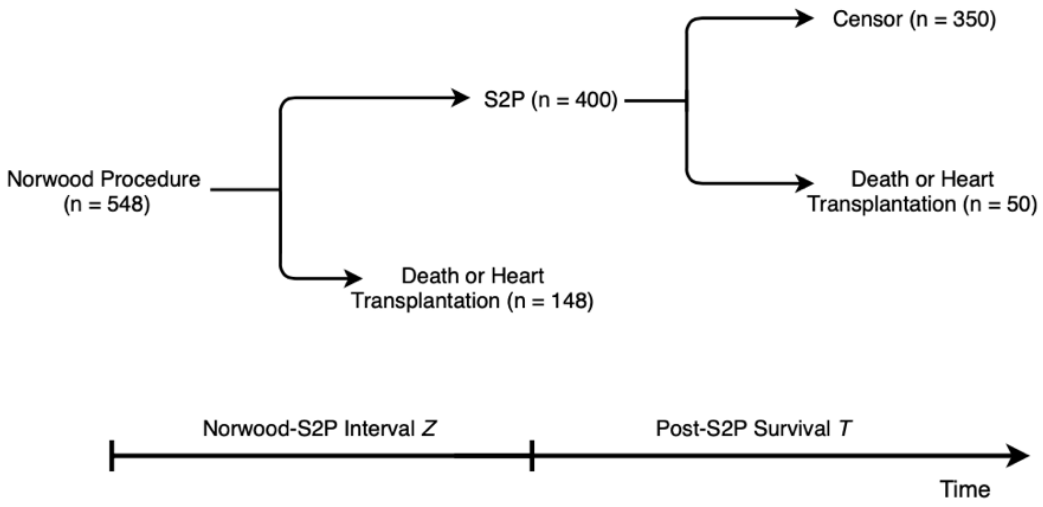

Figure 2.1: Treatment and outcome patterns in the SRVT data

The data set contains baseline variables, information on the Norwood and S2P operations, and survival time. Among 548 randomized HLHS patients who had operative data and follow-up, 139 died prior to S2P, and 9 received a heart transplant. The remaining 400 underwent S2P (Figure 2.1).

Roughly 60% of the infants were male, and 80% were white. Newborns in the RVPA arm were more likely to undergo S2P. Subjects who underwent S2P were heavier at birth and slightly younger at the time of the Norwood procedure and post-Norwood discharge, compared with the subgroup who died or underwent cardiac transplantation without S2P.

2.3.2. Analysis under the lognormal model

Using the Akaike information criterion (AIC), we selected a panel of predictive covariates including sex, race, randomized Norwood procedure (MBTS *vs.* RVPA), prenatal diagnosis of congenital heart disease, aortic atresia, obstructed pulmonary venous return, birth weight, age at Norwood, and age at Norwood discharge.

We conducted a sensitivity analysis to determine whether assumptions about the correlation of the potential outcomes would influence final causal inferences. Assuming a homogeneous compound symmetry model ($\sigma_1^2 = \ldots = \sigma_6^2$), we imputed data under the lognormal model separately for common correlation values $\rho \in \{0, 0.3, 0.6, 0.9\}$. Figure 2.2 shows that there is minimal sensitivity to the assumed value of ρ.

Figure 2.3 shows results from $M = 200$ imputations under the lognormal model with varying common correlation coefficient. The peak fraction surviving occurs at the S2P time 6 months after Norwood under all four models. The proportion of imputed post-S2P survival times greater than 10 years ranges from 0.68 to 0.84 (Table 2.2).

Table 2.2: Parameter estimates using the lognormal model ($M = 200$)

Correlation Coefficient	Parameter	Estimate	Standard Error	95% Confidence Interval Lower Limit	95% Confidence Interval Upper Limit
ρ = 0	$\theta(1)$	0.683	0.025	0.634	0.731
	$\theta(2)$	0.702	0.024	0.654	0.749
	$\theta(3)$	0.792	0.020	0.753	0.830
	$\theta(4)$	0.832	0.018	0.798	0.867
	$\theta(5)$	0.815	0.020	0.775	0.855
	$\theta(6)$	0.784	0.021	0.743	0.825
ρ = 0.3	$\theta(1)$	0.676	0.024	0.627	0.724
	$\theta(2)$	0.707	0.024	0.660	0.755
	$\theta(3)$	0.794	0.019	0.755	0.832
	$\theta(4)$	0.838	0.017	0.803	0.873
	$\theta(5)$	0.816	0.020	0.776	0.857
	$\theta(6)$	0.778	0.020	0.737	0.818
ρ = 0.6	$\theta(1)$	0.667	0.024	0.619	0.716
	$\theta(2)$	0.697	0.023	0.650	0.744
	$\theta(3)$	0.784	0.019	0.745	0.823
	$\theta(4)$	0.831	0.017	0.797	0.866
	$\theta(5)$	0.814	0.020	0.774	0.854
	$\theta(6)$	0.781	0.020	0.740	0.821
ρ = 0.9	$\theta(1)$	0.675	0.024	0.628	0.723
	$\theta(2)$	0.690	0.023	0.642	0.737
	$\theta(3)$	0.784	0.019	0.745	0.822
	$\theta(4)$	0.835	0.017	0.801	0.869
	$\theta(5)$	0.808	0.020	0.768	0.849
	$\theta(6)$	0.781	0.020	0.741	0.822

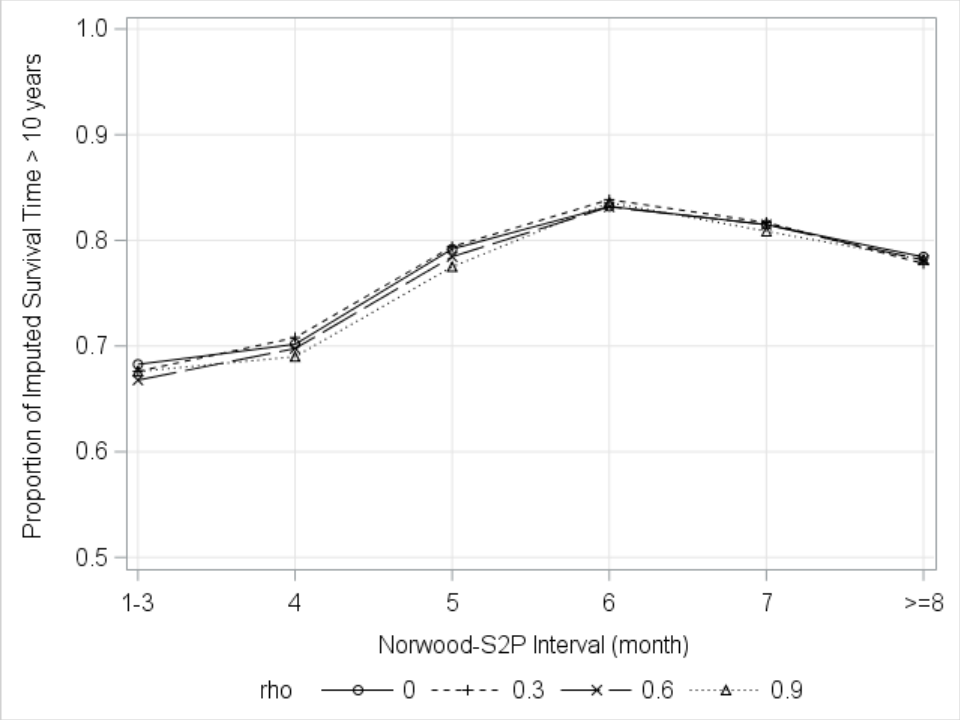

Figure 2.2: Inference on the causal parameter as a function of the common correlation ρ in the lognormal model ($M = 200$)

The results above assume equal covariate effects across treatment arms; that is, $\beta(z) = \beta(z')$ for $z, z' \in \{1, \ldots, 6\}$. A larger data set would allow us to estimate a separate covariate effect in each arm. In addition, a homogeneous compound symmetry model ($\sigma_1^2 = \ldots = \sigma_6^2$) simplifies the likelihood function, rendering the optimization more stable and likely to converge. We conducted an analysis assuming non-homogeneous error variance (not shown), which yielded the same overall results.

2.3.3. Analysis under the RCS model

Estimates of θ imputed under the RCS model appear in Table 2.3 and Figure 2.4. After a grid search using AIC as fit criterion, we set knots at 700 and 1400 days post-S2P survival. We then applied multiple imputation with $M = 20$, 100, 200 and 500. Table 2.3 shows that the proportion of imputed survival times greater than 10 years for $z = 1, \ldots, 6$ are roughly 0.77, 0.81, 0.87, 0.91, 0.88 and 0.86, regardless of M, with

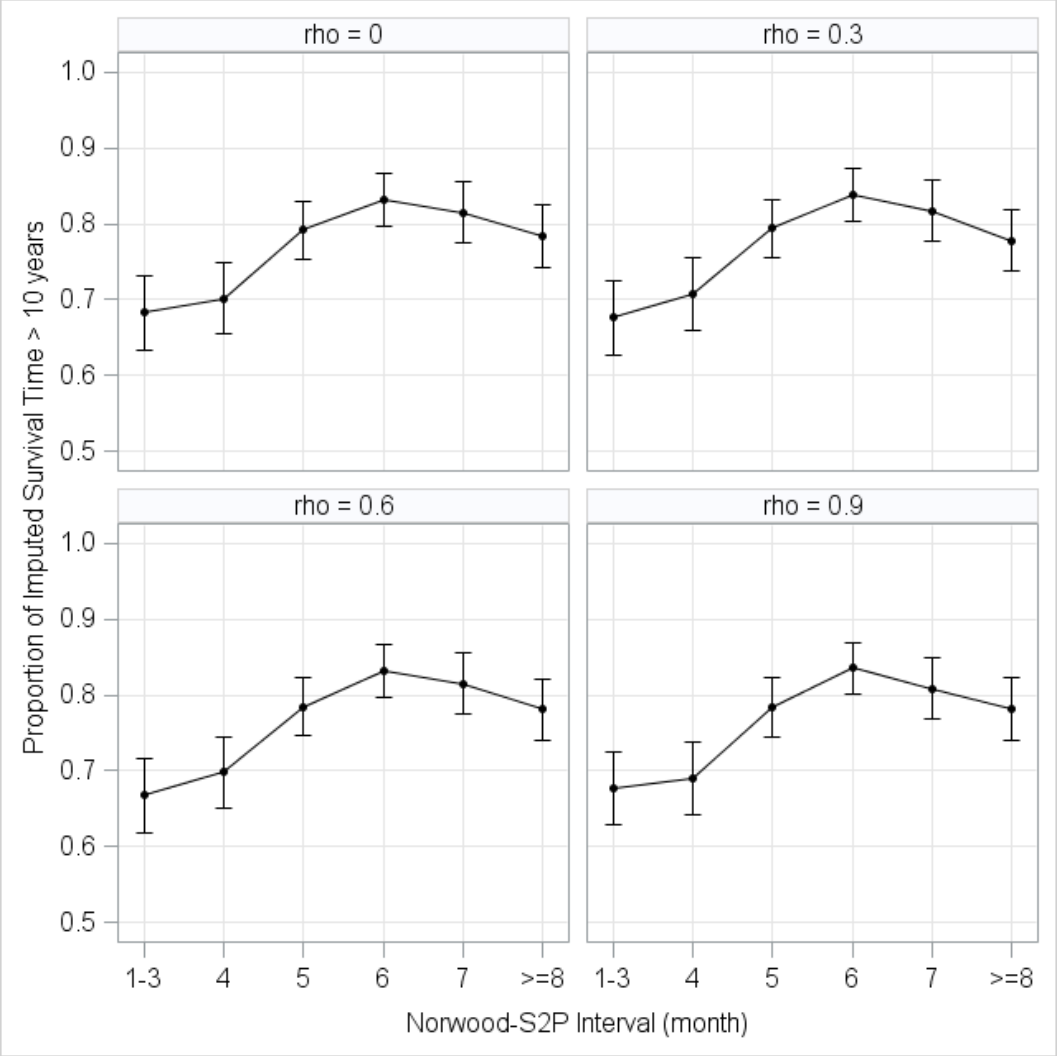

Figure 2.3: Confidence intervals on the causal parameter as a function of the common correlation ρ in the lognormal model ($M = 200$)

estimates stabilizing by $M = 200$. Although θ values are generally higher than under the lognormal model, the ranking of $\theta(z)$ is the same. We identify the optimal S2P timing as 6 months after the Norwood procedure.

To obtain a 95% confidence interval for the optimal S2P timing, we drew 100 bootstrap samples, performing the MI procedure $M = 20$ times within each. In the 100 bootstrap samples, the best S2P time was 6 months 82 times, 7 months 8 times, and ≥ 8 months 6 times (Figure 2.5). Thus, a bootstrap 95% confidence interval for optimal S2P timing is 6

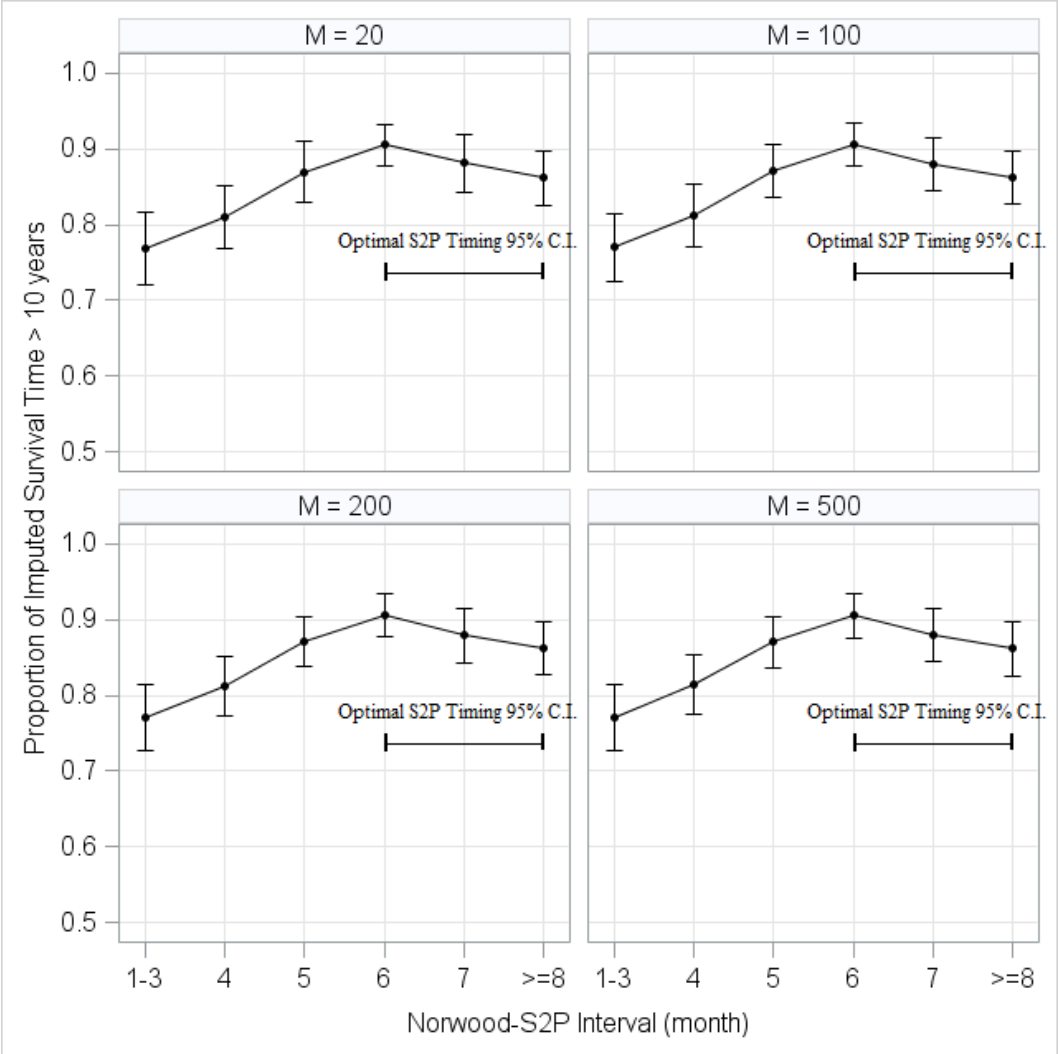

Figure 2.4: Inference on the causal parameter as a function of the number of imputations in the restricted cubic spline model

Table 2.3: Parameter estimates using the RCS model

M	Parameter	Estimate	Standard Error	95% Confidence Interval Lower Limit	95% Confidence Interval Upper Limit
20	$\theta(1)$	0.768	0.024	0.721	0.815
	$\theta(2)$	0.811	0.021	0.768	0.852
	$\theta(3)$	0.869	0.020	0.829	0.909
	$\theta(4)$	0.906	0.014	0.878	0.932
	$\theta(5)$	0.881	0.019	0.842	0.919
	$\theta(6)$	0.862	0.018	0.825	0.897
100	$\theta(1)$	0.770	0.026	0.726	0.815
	$\theta(2)$	0.813	0.021	0.771	0.854
	$\theta(3)$	0.871	0.018	0.836	0.905
	$\theta(4)$	0.906	0.014	0.878	0.934
	$\theta(5)$	0.880	0.018	0.844	0.915
	$\theta(6)$	0.862	0.018	0.827	0.897
200	$\theta(1)$	0.771	0.022	0.727	0.815
	$\theta(2)$	0.813	0.020	0.774	0.853
	$\theta(3)$	0.870	0.017	0.838	0.903
	$\theta(4)$	0.906	0.014	0.878	0.934
	$\theta(5)$	0.879	0.018	0.844	0.915
	$\theta(6)$	0.862	0.018	0.827	0.897
500	$\theta(1)$	0.771	0.022	0.727	0.814
	$\theta(2)$	0.814	0.020	0.775	0.853
	$\theta(3)$	0.870	0.017	0.837	0.905
	$\theta(4)$	0.905	0.015	0.876	0.935
	$\theta(5)$	0.879	0.018	0.845	0.914
	$\theta(6)$	0.861	0.018	0.826	0.879

to infinity.

In the SVRT data, most patients have pre-S2P events (death or heart transplantation) or are censored within 3 years of S2P. One patient, however, is purported to have a post-S2P censoring time of 2,704 days, or 7.4 years. This is a possibly influential outlier. The non-parametric RCS model is more responsive to this point, potentially explaining the generally higher values of θ under this model. The parametric lognormal model gives a slightly lower fitted survival probability in the tail of the survival curve.

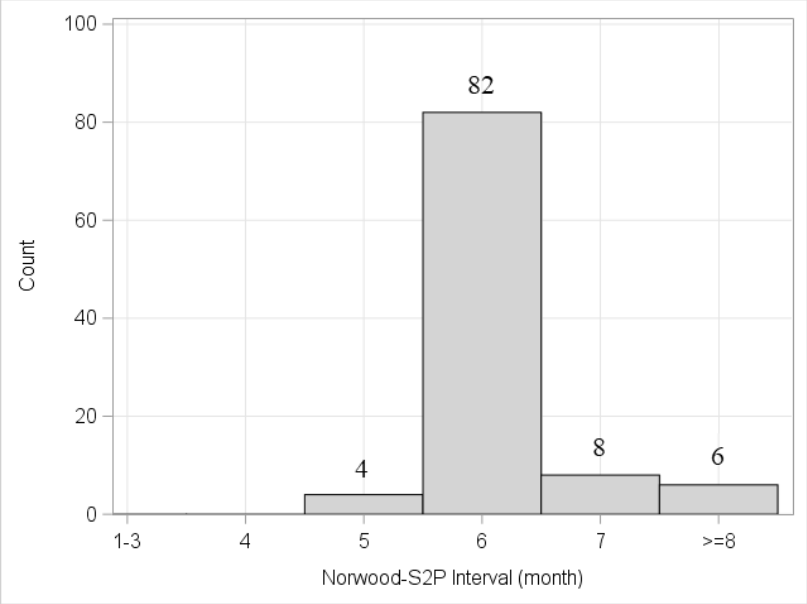

Figure 2.5: Count of optimal S2P timing from 100 times bootstrap ($M = 20$)

2.3.4. Testing linear hypotheses for the parameters

To determine the causal effect of S2P timing on θ, we tested linear contrasts of the $\theta(z)$ terms. Results under the RCS model appear in Table 2.4. The only significant contrast is $\theta(3) - \theta(2)$ with estimate 0.06 ($M = 200$) and 95% confidence interval 0.01–0.11. This means having S2P at 5 months after Norwood procedure is significantly better than having S2P at 4 months. The conclusion is insensitive to the number of imputations. Under the lognormal model, we find same single significant contrast, and the conclusion is insensitive to the correlation coefficient ρ.

Table 2.4: Causal contrasts of θ terms

			95% Confidence Interval	
M	**Parameter**	**Estimate**	**Lower Limit**	**Upper Limit**
20	$\theta(2)-\theta(1)$	0.042	−0.020	0.104
	$\theta(3)-\theta(2)$	0.059	0.009	0.109
	$\theta(4)-\theta(3)$	0.036	−0.013	0.086
	$\theta(5)-\theta(4)$	−0.025	−0.073	0.024
	$\theta(6)-\theta(5)$	−0.019	−0.074	0.035
100	$\theta(2)-\theta(1)$	0.042	−0.017	0.101
	$\theta(3)-\theta(2)$	0.058	0.008	0.108
	$\theta(4)-\theta(3)$	0.035	−0.009	0.081
	$\theta(5)-\theta(4)$	−0.026	−0.072	0.019
	$\theta(6)-\theta(5)$	−0.018	−0.068	0.032
200	$\theta(2)-\theta(1)$	0.042	−0.016	0.100
	$\theta(3)-\theta(2)$	0.057	0.008	0.107
	$\theta(4)-\theta(3)$	0.035	−0.008	0.080
	$\theta(5)-\theta(4)$	−0.027	−0.071	0.018
	$\theta(6)-\theta(5)$	−0.018	−0.068	0.032
500	$\theta(2)-\theta(1)$	0.042	−0.015	0.099
	$\theta(3)-\theta(2)$	0.057	0.007	0.107
	$\theta(4)-\theta(3)$	0.035	−0.008	0.080
	$\theta(5)-\theta(4)$	−0.027	−0.071	0.018
	$\theta(6)-\theta(5)$	−0.018	−0.068	0.032

2.4. Discussion

We have proposed a multiple imputation strategy to create causal inferences for the effect of S2P timing on survival outcomes in infants with HLHS. Imputation under a lognormal model showed that inferences are insensitive to the assumed, and inestimable, correlation of potential outcomes. A more flexible cubic spline hazard model gave estimates that were generally larger for the fraction surviving 10 years, although causal contrasts between z values were similar to those from the lognormal. Both sets of analyses suggest that the optimal S2P time occurs at 6 months.

Imbens and Rubin [25], working in the context of continuous, non-censored outcomes, observed that causal inferences are generally insensitive to the assumed value of the inestimable correlation coefficient of the potential outcomes. Our sensitivity analysis under the lognormal model gave a similar result. Our RCS model assumes independence of the potential outcomes given the covariates; it is unclear whether these inferences are also insensitive.

In medical practice, the dynamic variable pulmonary vascular resistance (PVR), a biomarker of the maturity of the infant's circulatory system, is a key determinant of the timing of S2P. A limitation of this study is that the data contain only a baseline PVR measurement, which is absent in roughly half of the data. Therefore, we were unable to include PVR in our analysis.

In a previous study [4], we used inverse propensity score weighting to eliminate confounding bias in estimation of the causal effects of S2P timing on survival in the SVRT data. The key ignorability assumption in that analysis was weak unconfoundedness [23, 57]. Our analyses necessarily excluded the 148 newborns who died or underwent heart transplantation (and therefore did not undergo S2P). The analysis in this article includes those subjects, as they are assumed to have the full vector of potential outcomes

even if we do not know their Z_i values.

By constructing M completed datasets and properly combining inferences based on them, one can eliminate bias due to systematic differences between the observed and unobserved data [2]. Multiple imputation also offers a straightforward path to sensitivity analysis through the incorporation of non-ignorable features into imputation models [41].

In conclusion, the optimal timing of S2P is roughly 6 months after Norwood procedure with a 95% confidence interval of 6 to infinity. This result is consistent with the previous analysis that we conducted that created propensity scores using a competing-risks analysis and estimated hazard ratios using inverse propensity score weighting [4].

CHAPTER 3

Developing a VA Women's Cardiovascular Disease Risk Score

3.1. Introduction

Atherosclerotic cardiovascular disease (ASCVD) is the leading cause of death and disability in North America and the third-leading cause of death in women veterans [52]. Women military service members and veterans have significantly higher numbers of cardiovascular risk factors and poorer health status than their civilian counterparts [16, 27]. Previous studies [3, 51] report that women service members have almost twice the burden of traditional cardiovascular disease (CVD) risk factors, such as hypertension, at younger ages (< 40 years) than their civilian counterparts. Currently, women enlistees are significantly younger than male enlistees in the military, and military exposure earlier in life may alter the aging trajectory on ASCVD risk factors [54]. Therefore, aging-related changes in ASCVD risk factors among Veterans Affairs (VA) women may differ from those experienced by their civilian counterparts.

The current DoD (Department of Defense)/VA guidelines for cardiovascular disease screening and treatment of hypertension and hypercholesterolemia rely on the American College of Cardiology/American Heart Association (ACC/AHA) ASCVD risk assessment model. The ACC/AHA model was developed from Goff et al [13] based on a pooled epidemiologic dataset containing civilian men and women aged 40 to 79 years old. Consistent with this, the current DoD/VA guideline recommends screening women for ASCVD risk starting at age 45 [33]. But because the traumatic stress associated with military

service earlier in life may alter the aging effect on ASCVD risk factors [9], it is unclear that the current screening strategy is adequate for VA women.

VA women differ from women in the general population not only in prevalence of traditional and non-traditional CVD risk factors [16], but also in racial and ethnic group representation. To examine whether the current ACC/AHA ASCVD model is applicable to VA women, we re-estimated VA women's ASCVD risk using the VA national Electronic Health Records (EHR) database, which includes younger women service members aged 30–39, similarly to the current ACC/AHA models [5]. This database is the largest longitudinal health record in the U.S. health care system and the best data available on the VA women population. Our analysis identified a curvilinear association of aging with ASCVD risk in VA women starting at ages as young as 30 years across all races. This differs substantially from the previous consensus, which was that women face minimal ASCVD risk until age 45. Our observation suggests a need for cardiovascular risk screening of VA women at ages less than 45, and the development of a new validated CVD risk model adequately assessing CVD risk for VA women.

We aim to develop a new CVD risk model for VA women using the VA national EHR database. To improve prediction accuracy (or calibration) and precision (or discrimination) of the new CVD model, the study incorporated all available visit data and aging-related changes in CVD risk factors. Using information from multiple visits, we have developed the *VA Women CVD Risk Score*, based on a parsimonious predictive model and calculated using time-dependent covariate Cox regression [7].

3.2. Study Cohort and Data Preparation

The study cohort is defined by women veterans aged were between 30 and 79 years and alive on January 1, 2007 with at least one outpatient or inpatient visit diagnosis. 92,682 VA women with records in the VA Corporate Data Warehouse (CDW) met these

criteria. Of these, we excluded 14,126 women who were missing data on baseline covariates, e.g., blood pressure and lipids. Because we focused on inference for white, African American, and Hispanic patients, we excluded a further 8,982 VA women of other races. We were left with data on 69,574 women service members and veterans: 36,172 non-Hispanic whites (52%), 29,231 non-Hispanic African Americans (42%), and 4,171 Hispanics (6%). See Figure 3.1.

We extracted data using the Structured Query Language (SQL) Server Management Studio (SSMS, Version 2017, Microsoft Corp., Redmond, WA) and conducted statistical and graphical analyses using SAS Enterprise Guide (version 7.1, SAS Institute, Cary, NC) and R (Version 3.5.3, cran.r-project.org). We performed all data extraction, preparation, and analyses within the domain of the VA Informatics and Computing Infrastructure (VINCI). Death event and cause of death data are obtained from the VINCI Vital Status File, which compiles data from the BIRLS (Beneficiary Identification Records Locator Subsystem), the VA Medicare Vital Status File, and the National Death Index (NDI) for Veterans, which is a part of the VA Suicide Data Repository (SDR). All variables of interest in the current study are extracted from VA national EHR data, situated in the VA national Corporate Data Warehouse (CDW). The CDW data contain health records of all patients treated in the VA Health Care System.

We transformed the data to represent values in six-month equal-length windows, with data considered to be missing when there are no visit records in the window. Specifically, the first interval (denoted the baseline visit) runs from January 1, 2007 to June 30, 2007. Subsequent intervals run from July 1, 2007 to December 31 2007, January 1, 2008 to June 30, 2008, etc. There are 22 intervals in total. If there is no visit within the first interval, the first following available visit is set as the baseline. We consider each interval to contain one record per patient; if a patient has multiple visits in a window, we represent risk factor data as the average across the visits for continuous variables (e.g. HDL) or maximum for binary variables (e.g. diabetes, current smoking status).

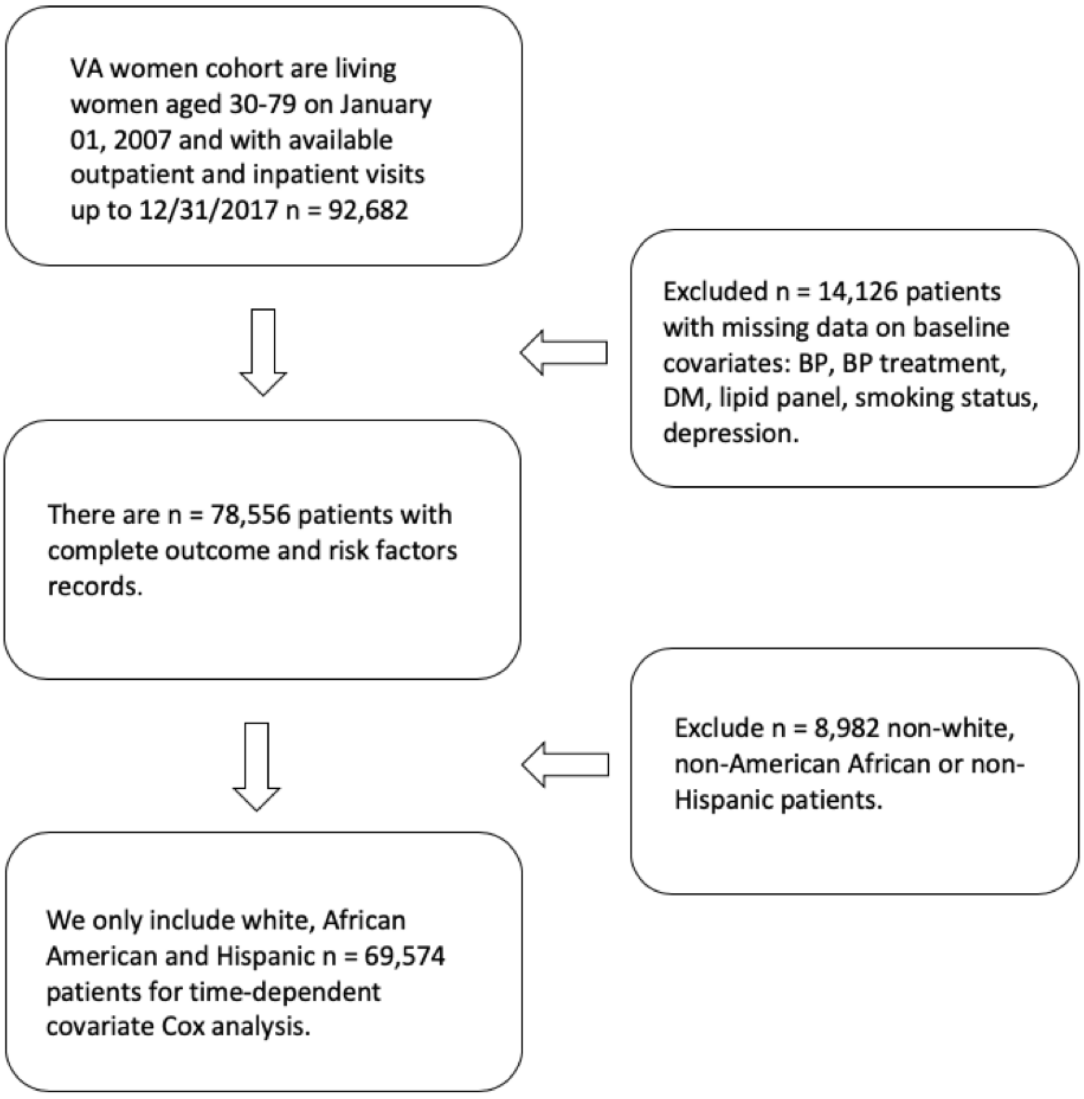

Figure 3.1: Data extraction process

Because of the sensitive nature of the Veterans Affairs (VA) data collected for this study, access is restricted to VA-affiliated researchers trained in human subject confidentiality protocols. Requests may be sent to the VA North Texas Health Care System Institutional Review Board (IRB) at NTXIRBAdmin@va.gov.

3.3. Method

3.3.1. Re–estimating the ACC/AHA model

The ACC and AHA seek to improve the prevention, detection, and treatment of cardiovascular disease (CVD) through the promulgation of guidelines. The Goff [13] CVD risk assessment tool, based on the Cox proportional hazards model, is the basis of the current ACC/AHA risk assessment guideline.

As mentioned above, the Goff model applies to civilians aged 40–79 years with no previous history of nonfatal myocardial infarction (recognized or unrecognized), stroke, heart failure, percutaneous coronary intervention, coronary artery bypass surgery, or atrial fibrillation.

The Goff model evaluates risk separately for whites and American Africans. The model for whites includes age, age squared, untreated systolic blood pressure (SBP), treated SBP, current smoking status, diabetes, total cholesterol, high-density lipoprotein (HDL), and several interaction terms: current smoking status*age, total cholesterol*age, and HDL*age. The African American model omits the age squared term and smoking*age and total cholesterol*age interactions, while including the SBP*age interaction. All continuous predictors in these models appear on the log scale.

We re-estimated the Goff model using our VA EHR data and evaluated it by C-statistics and calibration plots. Because this model cannot handle time-varying covariates, we used

only predictors measured at the baseline visit. The ACC/AHA model estimates separate coefficients for white women and African American women, and excludes Hispanic women. We applied both the white and African American ACC/AHA models to Hispanic women to estimate their ten-year ASCVD risk.

We assume that each of n subjects, labelled $i = 1, \ldots, n$, has a p-dimensional predictor X_i of biomarkers and risk factors. Let T_i stand for the CVD event time for subject i, and assume that it is potentially censored at time C_i. The observed event/censoring time is $T_i^* = \min(T_i, C_i)$, with the indicator Δ_i equal to 1 if T_i^* is an event time and 0 if T_i^* is a censoring time.

We let $\lambda_i^*(t|X)$ refer to the ASCVD event hazard function for subject i. The Cox proportional hazards model takes the following form:

$$\lambda_i^*(t|X_i) = \lambda_0^*(t)\exp(\beta X_i) = \lambda_0^*(t)\exp\left(\sum_{j=1}^{p}\beta_j X_{ij}\right),$$

where $\lambda_0^*(t)$ is the baseline hazard at time t, and β_j is the coefficient for the j^{th} risk factor. We interpret $\exp(\beta_j)$ as the hazard ratio (HR) resulting from a one-unit increase in predictor j. We note here two facts about the model:

1. The baseline hazard $\lambda_0^*(t)$ depends on t, but not on X_i.
2. The hazard ratio $\exp(\beta X_i)$ depends on X_i but not on time t.

3.3.2. VA Women Cardiovascular Disease Risk Score

In our VA EHR database, patients have multiple visit records, with measured values of the risk factors potentially varying across visits. Thus we can represent the risk factor vector X_i as a function of time: $X_i(t)$ is the vector of risk factors for subject i, measured at time t. With this representation we can apply a more general form of the Cox model that

allows the hazard to vary over time in response to changes in the predictors. We define the time-dependent covariates Cox proportional hazard model as

$$\lambda_i(t|X_i(t)) = \lambda_0(t)\exp(\beta X_i(t)) = \lambda_0(t)\exp\left(\sum_{j=1}^{p}\beta_j X_{ij}(t)\right),$$

where $\lambda_0(t)$ is the baseline hazard at time t, and the hazard $\lambda_i(t|X_i(t))$ for subject i at time t depends on the value of the risk factor vector at time t.

We constructed ASCVD risk factors from VA EHR data following Sussman et al [46], and ASCVD events (non-fatal myocardial infarction, non-fatal stroke, and cardiac death) using ICD-9 and ICD-10 diagnostic and procedural codes from VA national EHR data and the NDI data. We validated myocardial infarction and stroke events by searching for words such as "MI", "myocardial infarction", and "stroke", respectively, embedded in health providers' narratives and notes extracted from VA EHR.

After building the model, we used the Harrell C-statistic [17, 18] to evaluate its precision and ability to discriminate ASCVD events. We also used calibration plots to evaluate the quality of the prediction model.

The calibration plot compares the predicted probability with the observed probability. Under an ideal model, pairs of observed and predicted probabilities would lie on the 45-degree line, meaning that they agree exactly. We describe here the process of obtaining the calibration plot for the time-dependent covariate Cox model:

1. Randomly split the original data into training and test data sets in the ratio 4:1. That is, we sample 80% of unique patient IDs without replacement and use them as the training set, with the remaining 20% representing the test set.

2. Estimate the proposed time-varying Cox model using the training set; obtain the predicted 10-year ASCVD risk by plugging in the covariate values from the test set; the final predicted risk is the average risk over all patients in the test set.

3. Calculate the observed risk as one minus the Kaplan-Meier (KM) survival estimate at 10 years.

We repeat steps (1)–(3) 100 times to get 100 pairs of predicted and observed risk. The final calibration plot, with a 45–degree straight line as the reference, is the line plot connecting the above 100 dots (x-axis: predicted risk; y-axis: observed risk) after sorting the data.

3.4. Results

3.4.1. The VA EHR data

Table 3.1 shows that the study cohort consisted of 69,574 women: 52% white with mean age 46; 42% African American with mean age 44; and 6% Hispanic with mean age 43. Patients under 40 constitute 16% of the sample. African American women have the highest mean SBP and HDL cholesterol, but lowest mean total cholesterol. They also have the highest prevalence of diabetes and the lowest of current smoking.

The enrolled women experienced 2,176 deaths from any cause (3.1%); of the women who died 1,321 were white (1.9%); 781 were African American (1.1%); and 74 were Hispanic (0.1%). Table 3.2 describes ASCVD events by race and type. Non-fatal myocardial infarction is the most common ASCVD event, followed by non-fatal stroke and cardiac death. The rate of stroke is significantly higher in African American women (2.0%) than in white women (1.5%).

Table 3.1: Baseline risk factors stratified by race and ethnic group.

	White $n = 36,172(52\%)$	**African American** $n = 29,231(42\%)$	**Hispanic** $n = 4,171(6\%)$
Continuous[a]			
Age (year)	45.9(8.7)	44.2(7.8)	43.1(8.4)
Systolic blood pressure (mmHg)	123.8(14.8)	127.1(15.8)	122.2(14.6)
Total cholesterol (mg/dL)	200.0(40.8)	192.4(38.9)	195.6(38.3)
High-density lipoprotein (mg/dL)	53.8(16.7)	56.9(17.4)	53.83(15.6)
Discrete			
Diabetes	8,405(23.2%)	9,569(32.7%)	1,056(25.3%)
Current smoking	10,864(30.3%)	5,111(17.5%)	994(23.8%)

[a]Values given as mean(SD)

Table 3.2: ASCVD events stratified by race and ethnic group

ASCVD type	**White**	**African American**	**Hispanic**
Non-fatal myocardial infarction (MI)	1,515	1,148	148
Non-fatal stroke	538	592	61
Cardiac death	245	144	16
Heart failure	175	151	18
Cardiac arrest	9	6	2

3.4.2. Re–estimating the ACC/AHA model

We re-estimated the ACC/AHA model using the VA EHR data with patients younger than 40 years old. Figure 3.2 plots the covariate-adjusted aging effect. We can see that the estimated 10-year ASCVD risk for VA women increases curvilinearly with older age, starting at age 30.

The Figure 3.2 (a) displays the aging effect on 10-year risk of ASCVD event for whites (solid line) and African Americans (dotted line). Both of the predicted 10-year risk of ASCVD event are about 5% at age 30 years old. That risk of white veteran increases slowly and surpasses that of African American at age 55. It reaches around 15% at age 79, comparing with 8% for African American veteran. Since Goff et al [13] only have the models for white or African American veteran, we apply both of the two models for the Hispanic veteran VA women in our dataset. The result shows in Figure 3.2 (b) having similar pattern as Figure 3.2 (a).

3.4.3. VA women Cardiovascular Disease Risk Score

Goff et al [13] proposed separate time-independent covariate models for white and African American civilians. To simplify prediction, we have built a single, parsimonious model that is valid for all three race/ethnic groups. The risk factors are age, untreated systolic blood pressure, treated systolic blood pressure, diabetes mellitus status (yes or no), current smoking status (yes or no), major depression (yes or no), total cholesterol, HDL, and antihypertensive treatment (yes or no). Unlike Goff's models, ours contain no interaction terms.

We have created two separate models: Model 1 predicts any of non-fatal myocardial infarction (MI), non-fatal stroke, and cardiovascular death; and Model 2 predicts these outcomes plus heart failure and cardiac arrest. Both models include all of the predictors

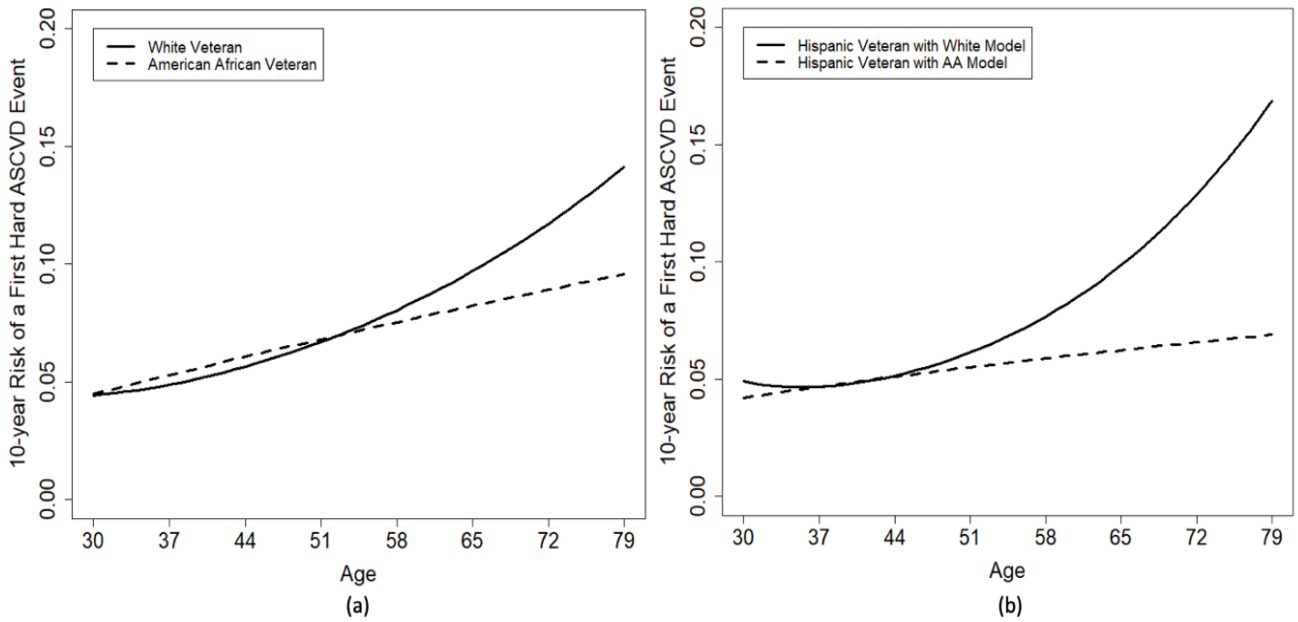

Figure 3.2: Estimated effect of aging on increased 10-year ASCVD risk, stratified by race for women military service members: (a) whites (solid line) and African Americans (dotted line); (b) Hispanic veterans by the models for white and African Americans.

indicated above. The coefficients of the ASCVD risk factors, interpreted as log hazard ratios, appear in Table 3.3, stratified by race. Table 3.3 shows that model 1 and model 2 typically have similar coefficient values. A positive (negative) coefficient means that the risk factor has a positive (negative) correlation with the ASCVD risk. In all race/ethnic groups, high ASCVD risk is associated with older age, higher SBP and total cholesterol level, lower HDL, and presence of smoking, diabetes, and major depression.

Baseline survival probabilities for the Model 1 events at 10 years, estimated from the Kaplan-Meier curve, are 0.944, 0.944, and 0.954 for whites, African Americans, and Hispanics, respectively. These probabilities are slightly lower after adding heart failure and cardiac arrest as outcome events (as in Model 2). We calculated the predicted 10-year ASCVD risks using the two outcome models for all race groups, fixeing the covariates as age 38, total cholesterol 199 mg/dL, HDL 50 mg/dL, systolic BP 138mmHg, no diabetes, no major depression, no current smoking and no antihypertensive treatment. These results appear as the final row in Table 3.3, where we see that African American women

Table 3.3: Time dependent Cox PH model estimates

	White		African American		Hispanic	
Model	1[a]	2[b]	1	2	1	2
Age	2.399	2.493	2.058	2.074	2.191	2.291
Untreated hypertension	1.008	1.018	0.411	0.526	0.653	0.812
Treated hypertension	−0.208	0.463	1.246	1.664	−3.714	−1.391
Diabetes mellitus	0.425	0.457	0.276	0.350	0.315	0.301
Current smoking	0.072	0.073	-0.020	-0.004	0.356	0.413
Major depression	0.244	0.254	0.231	0.282	0.311	0.393
Total cholesterol	0.024	0.167	0.180	0.096	0.099	0.107
High-density lipoprotein	−1.350	−1.295	−1.339	−1.276	−1.225	−1.208
Antihypertensive treatment	1.263	−1.966	−5.795	−7.698	18.290	7.061
C-statistic	0.700	0.710	0.680	0.683	0.660	0.671
$S(10)$[c]	0.944	0.939	0.944	0.939	0.954	0.949
10-year CVD risk (%)[d]	4.070	4.340	4.970	5.200	3.810	4.220

[a]Outcome is MI + stroke + cardiovascular death

[b]Outcome is MI + stroke + cardiovascular death + heart failure + cardiac arrest

[c]10-year CVD event free survival probability

[d]1-$S(10)^{\exp(\hat{\beta}X-\hat{\beta}\bar{X})}$, where X is a vector of covariates, $\bar{X}$ is the mean value of the corresponding covariates, and $\hat{\beta}$ is the vector of estimated regression coefficients (log hazard ratios). Specific values used in the table are age 38 years, total cholesterol 199 mg/dL, HDL 50 mg/dL, SBP 138 mmHg, no diabetes, no major depression and no current smoking status.

have the highest predicted risk at nearly 5%.

The performance of our VA women risk score is evaluated by C-statistics and prediction accuracy (calibration plot). All C-statistics for our models are around 0.7; thus they are uniformly better than the ACC/AHA models in Section 3.3.1, which give C-statistics slightly over 0.6. A C-statistics of 0.7 generally is interpreted to represent good prediction precision.

Figure 3.3 displays calibration plots for Model 1 by race group. For white women, Model 1 overestimates the 10-year ASCVD risk slightly when the observed risk is lower than 0.1, and underestimates risk when the observed risk greater than 0.1. As the calibration plot is roughly centered around the 45-degree line, this discrepancy is considered acceptable. The calibration plots for African American and Hispanic women reveal similar patterns. Calibration plots for the model 2 outcome (Appendix Figure C.1) have similar appearance.

To evaluate the prediction discrepancy between our risk score and the ACC/AHA model, we calculated the 10-year predicted ASCVD risk for each subject in our VA EHR database using these models (Table 3.4). For simplicity, we classify estimated ASCVD risk as either low ($< 7.5\%$), moderate (7.5–19.9%), or high ($\geq 20\%$) [15]. Because ACC/AHA models only apply to whites and African Americans, we omit Hispanics from the comparison. As Table 3.4 shows, our risk score classifies white women who are at low risk by ACC/AHA into moderate ($n = 6\,962$) or high-risk groups ($n = 3\,763$). For African Americans, our model predicts 4,531 women in the low risk category whom the ACC/AHA model marks as of moderate risk. Most of the re-classifications are nevertheless consistent.

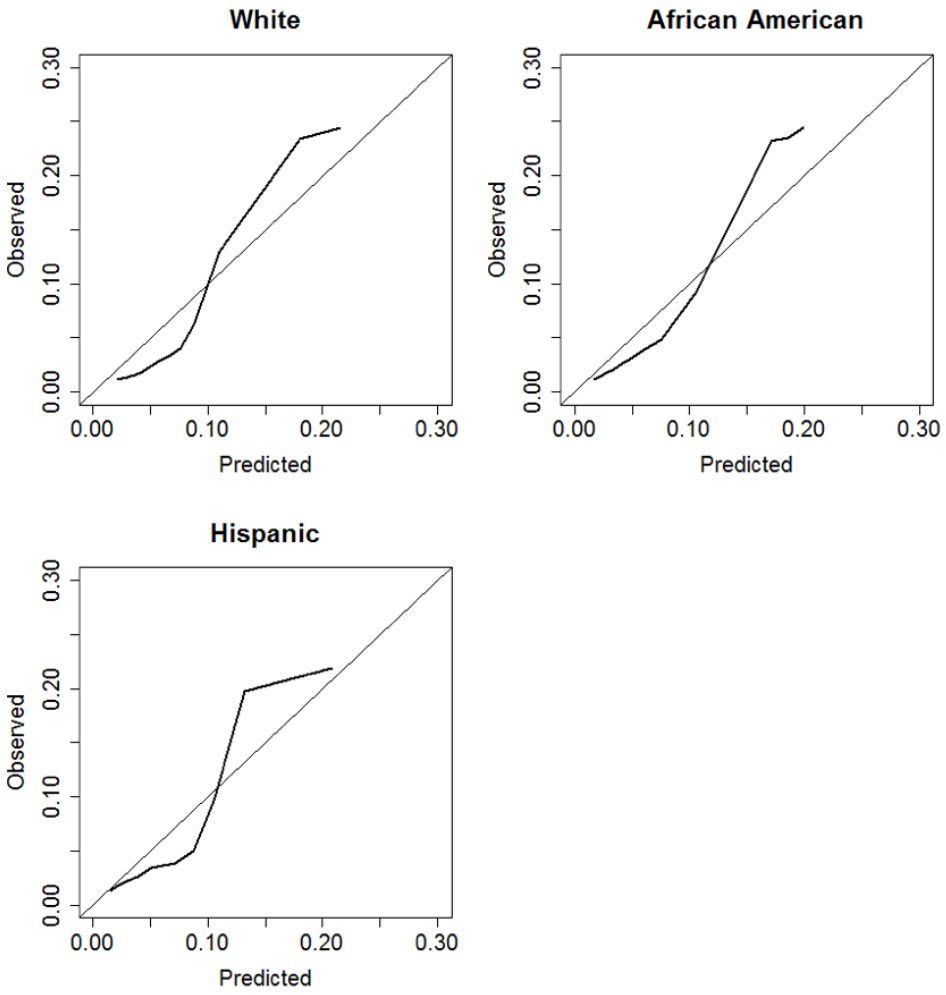

Figure 3.3: Calibration plots for VA women ASCVD risk model 1 by race

Table 3.4: Comparison of model prediction between VA women risk score and ACC/AHA model

		VA women ASCVD risk score		
		$< 7.5\%$	$7.5\% - 19.9\%$	$\geq 20\%$
ACC/AHA model	$< 7.5\%$	19,572(20,185)[a]	6,962(1,485)	3,763(0)
	$7.5\% - 19.9\%$	50(4,531)	3,108(2,828)	2,636(147)
	$\geq 20\%$	0(46)	0(0)	80(9)

[a]Values given are for whites (African Americans)

3.5. Discussion

We have proposed a new 10-year ASCVD risk prediction model for Veterans Affairs women based on data from the VA national EHR database. We evaluated the new model by C-statistics, calibration plots, and reclassification performance when compared with the current ACC/AHA CVD risk model [13]. In predicting risk for VA women, our new model is more accurate, precise, parsimonious, and stable than the ACC/AHA model. Our proposed ASCVD risk score includes the traditional CVD risk factors: age, SBP, diabetes, current smoking status, total cholesterol, and HDL. We have also identified a new risk factor — major depression — that leads to statistically significant improvement in ASCVD risk prediction.

Our study has substantial strengths:

1. It is based on the VA national EHR database with a sample size of more than 90,000; therefore it is representative of the VA women population and stable.
2. The proportion of minority women is moderate, making the model reliable for different race groups.
3. The VA national EHR data, unlike other non-VA EHR databases, is able to follow patients over substantial periods of time. This longitudinal (multiple visits) feature allows us to build a Cox PH model that has time-dependent predictors and is therefore potentially more accurate.
4. The database includes a substantial number of VA women under age 40, improving the credibility of the model for predicting ASCVD risk in this young but susceptible population.
5. Provider narrative notes (medical notes) in the database enable us to verify the outcome events, like non-fatal MI, non-fatal stroke, etc., identified through their ICD codes.

EHR data are often criticized for containing mis-classified ICD-9 and ICD-10 codes. For example, some studies have found inaccruate ICD diagnosis codes for stroke events, with concordance of only 50%–61% with providers' note [14, 32]. In our database, the concordance between ICD codes and medical note is 92.5% for non-fatal stroke and 96.9% for non-fatal MI. ASCVD events derived from ICD codes are therefore reliable in the VA EHR database.

Our analysis has several limitations:

1. Our model cannot incorporate information about menopause or history of other pregnancy-related conditions/complications. Future work could incorporate these female-specific biomarkers.

2. Our analysis assumes non-informative censoring, in the sense that loss to follow up is statistically independent of the time to the ASCVD event. Violation of this assumption could bias estimates. This would occur if, for example, patients suffering stroke or MI are more likely to visit the nearest hospital rather than their regular VA hospital, and in that case our model would underestimate ASCVD risk. Future studies could assess the sensitivity of results to violation of this assumption.

3. The current study is limited to VA women with complete data on vital signs, SBP, total cholesterol and HDL at baseline visits. If women who do not have all of these variables available differ from those who do with respect to risk factors for ASCVD, some bias may result.

4. VA women younger than age 30 and older than 80, and smaller minority groups such as Asians are not included in the current study cohort.

In summary, we have proposed an accurate and parsimonious 10-year ASCVD risk prediction model that performs better in VA women than the current ACC/AHA guideline model. It is a consistent risk score that fits the three major race/ethnic groups of VA women and demonstrates a new aging effect on ASCVD risk. Because the number of

women in the U.S. military has increased rapidly in recent years [1], a precise and usable ASCVD risk prediction model will be an important element in the future care of women veterans.

APPENDIX A

APPENDIX of CHAPTER 1

A.1.

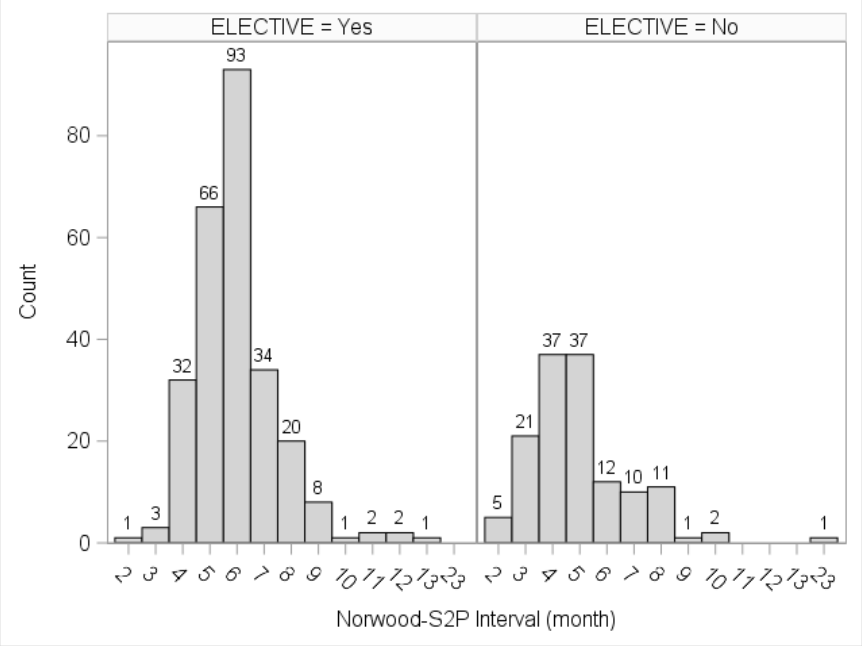

Figure A.1: Histogram of S2P times of 400 patients who underwent S2P, by elective status of the procedure

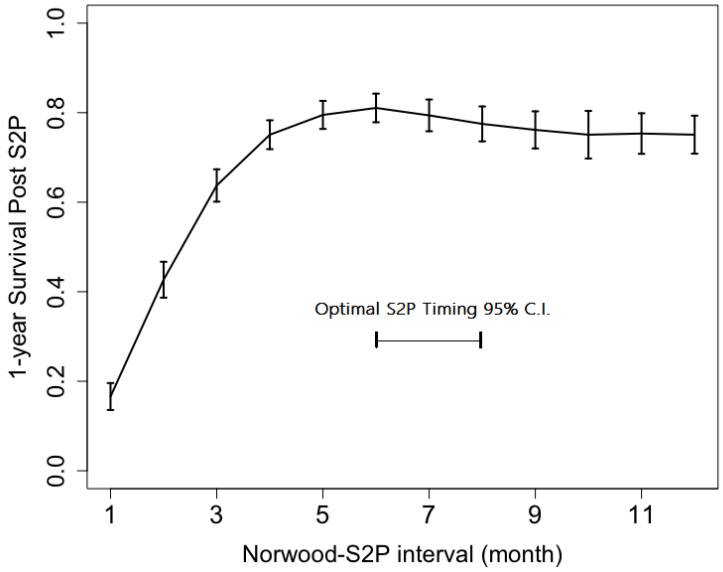

Figure A.2: Estimated one-year survival probability after S2P, estimated by the method of Hu *et al.* (2018)

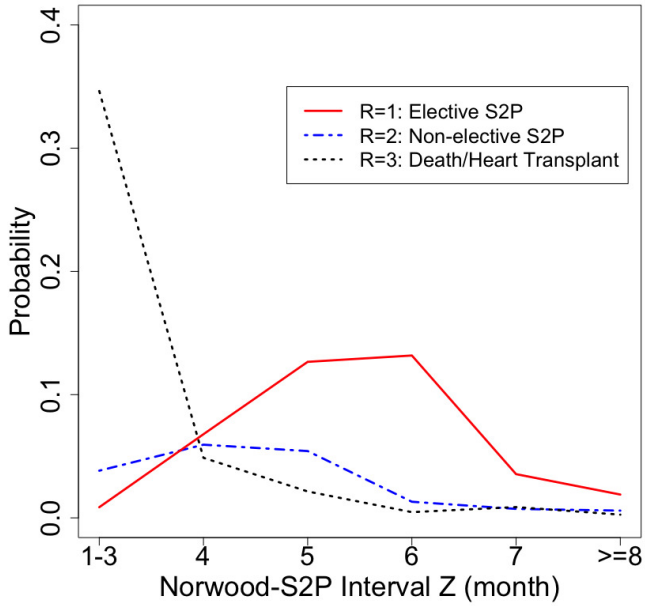

Figure A.3: Average probability of the three competing events — elective S2P, non-elective S2P, and death/cardiac transplantion — over time

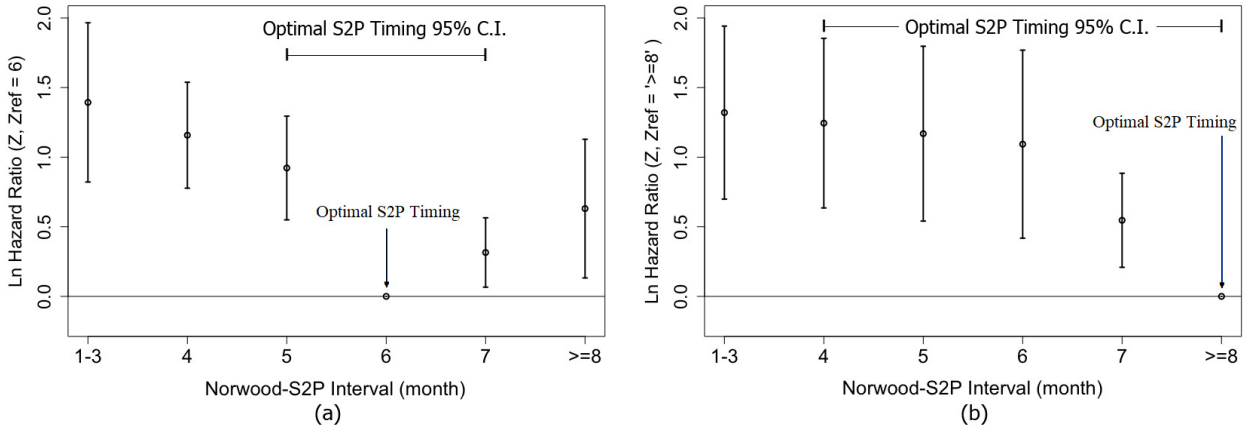

Figure A.4: Log mortality hazard ratio as a function of S2P timing: (a) Elective S2P only; (b) non-elective S2P only.

Figure A.5 presents the log hazard ratio as a function of time from Norwood to S2P (Z) estimated under three models: Unsmoothed, using discrete Z (shown in red, with a triangle at the estimate); smoothed by a linear spline, with discrete Z (black, circle); and smoothed using a linear spline with continuous Z (blue, star). Both discrete-Z methods give a minimum hazard ratio at $Z = 6$, and the continuous-Z method gives a minimum hazard at $Z = 6$ (with knots chosen at integer Z). Estimated hazard ratios are similar at all times. Standard errors for the analysis with continuous Z are often less than those with discrete Z.

Figure A.6 presents the log hazard ratio as a function of time from Norwood to S2P (Z) estimated under the three models, now stratified by elective status of the S2P. The models are coded as in Figure A.5. The discrete-Z methods give nearly identical results. The continuous-Z method gives similar results with smaller standard errors but is sensitive to the choice of knots in the spline function.

In the survival models with continuous Z we used weights estimated from the discrete-Z competing-risk model. We also removed one patient who had an outlying S2P time (23 months; see Figure A.1). We selected a final model by the Akaike information criterion, starting from a family of models with knots at the discrete Z values.

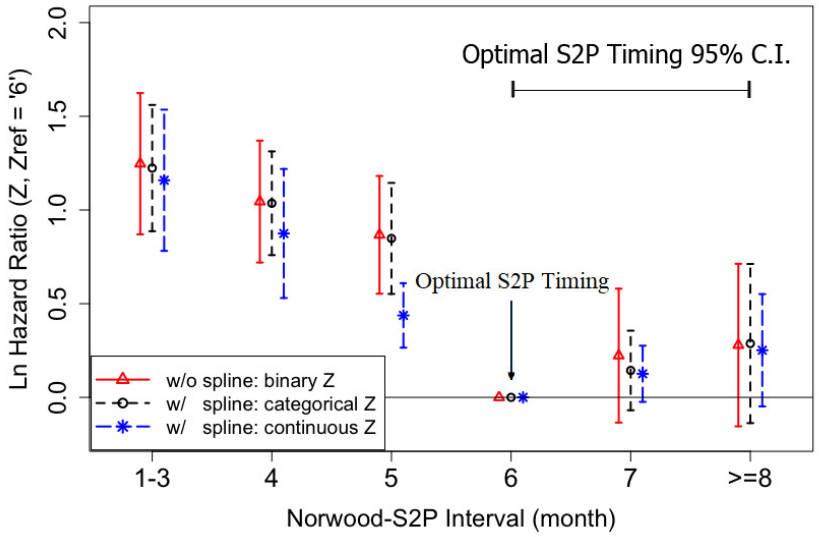

Figure A.5: Log mortality hazard ratio as a function of S2P timing, estimated under three models using all S2P patients

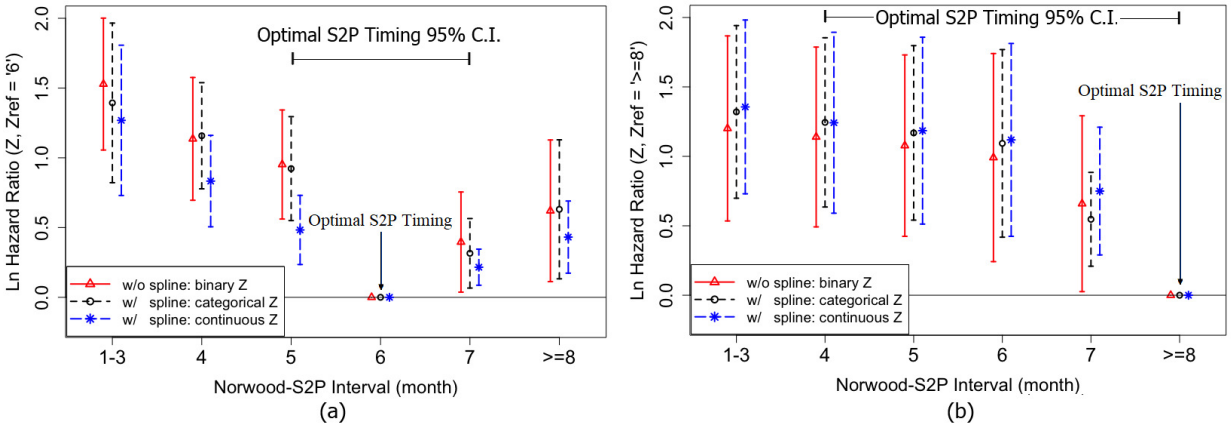

Figure A.6: Log mortality hazard ratio as a function of S2P timing, estimated under three models and stratified by elective S2P status: (a) Elective S2P; (b) non-elective S2P.

A.2.

Table A.1: Descriptive statistics by S2P time (in months)

Variables		**1–3 (n=30)** $Z=1$	**4 (n=69)** $Z=2$	**5 (n=103)** $Z=3$	**6 (n=105)** $Z=4$	**7 (n=44)** $Z=5$	**≥8 (n=49)** $Z=6$
Discrete							
Sex	M	21(70%)	45(65%)	64(62%)	68(65%)	25(57%)	29(59%)
	F	9(30%)	24(35%)	39(38%)	37(35%)	19(43%)	20(41%)
Race	White	28(94%)	61(88%)	83(80%)	83(79%)	32(72%)	36(73%)
	Black	1(3%)	5(7%)	15(15%)	16(15%)	10(23%)	12(24%)
	Other	1(3%)	3(5%)	5(5%)	6(6%)	2(5%)	1(3%)
Norwood arm	MBTS	18(60%)	17(25%)	49(48%)	54(51%)	16(36%)	23(47%)
	RVPA	12(40%)	52(75%)	54(52%)	51(49%)	28(64%)	26(53%)
Prenatal Dx[a]	Y	20(67%)	57(83%)	79(77%)	86(82%)	33(75%)	37(76%)
	N	10(33%)	12(17%)	24(23%)	19(18%)	11(25%)	12(24%)
Aortic atresia	Y	19(63%)	42(61%)	67(65%)	57(54%)	23(52%)	35(71%)
	N	11(37%)	27(39%)	36(35%)	48(46%)	21(48%)	14(29%)
Obstructed pulmonary venous return	Y	1(3%)	1(2%)	2(2%)	1(1%)	0(0%)	1(2%)
	N	29(97%)	68(98%)	101(98%)	104(99%)	44(100%)	48(98%)
Continuous[b]							
Birth weight (grams)		3160(452)	3228(455)	3141(492)	3146(561)	3159(619)	3114(514)
Norwood age (days)		7.3(3.5)	7.2(3.5)	6.2(3.8)	6.9(4.6)	6.6(4.6)	6.6(3.3)
Norwood discharge age (days)		45(22)	37(25)	38(30)	37(28)	48(43)	47(49)

[a]Prenatal diagnosis of congenital heart disease
[b]Values given as mean(SD)

Table A.2: Descriptive statistics by outcome of the Norwood procedure

Variables		**Elective S2P (n=263)**	**Non-elective S2P (n=137)**	**Death/Tx [a](n=148)**
Discrete				
Sex	M	161(61.2%)	91(66.4%)	87(58.8%)
	F	102(38.8%)	46(33.6%)	61(41.2%)
Race	White	207(78.7%)	116(84.7%)	112(75.7%)
	Black	46(17.5%)	13(9.5%)	27(18.2%)
	Other	10(3.8%)	8(5.8%)	9(6.1%)
Norwood treatment	MBTS	124(47.2%)	53(38.7%)	91(61.5%)
	RVPA	139(52.8%)	84(61.3%)	57(38.5%)
Prenatal diagnosis of congenital heart disease	Y	206(78.3%)	106(77.4%)	107(72.3%)
	N	57(21.7%)	31(22.6%)	41 (27.7%)
Aortic atresia	Y	161(61.2%)	82(59.9%)	99(66.9%)
	N	102(38.8%)	55(40.1%)	49(33.1%)
Obstructed pulmonary venous return	Y	4(1.5%)	2(1.5%)	13(8.8%)
	N	259(98.5%)	135(98.5%)	135(91.2%)
Continuous[b]				
Birth weight (grams)		3193(535)	3088(478)	2961(576)
Norwood age (days)		6.9(4.4)	6.4(3.0)	6.8(4.2)
Norwood discharge age (days)		38.8(30.2)	42.7(37.5)	42.1(46.1)

[a]Death or heart transplantation
[b]Values given as mean(SD)

Table A.3: 1-year estimated post-S2P mortality risk by S2P time

S2P time Z	1-yr mortality risk(%)	Lower 95% CI	Upper 95% CI
1 (1-3 months)	12.10	8.70	15.37
2 (4 months)	10.11	7.25	12.89
3 (5 months)	8.44	6.04	10.78
4 (6 months)	3.70	2.63	4.76
5 (7 months)	4.24	3.01	5.46
6 (≥ 8 months)	4.86	3.46	6.25

A.3.

We show that when censoring is noninformative and there are no unmeasured confounders, the estimate $\widehat{\alpha}$ obtained by solving Equations (1.5) and (1.6) is consistent for α in Model (1.4).

Define the counting process $N_i(t) = I(T_i^* \leq t, \Delta_i = 1)$, where $I(\cdot)$ is the indicator function, T_i^* is the observed post-S2P follow-up time for subject i, and Δ_i is an indicator that equals 1 if subject i dies after S2P and 0 otherwise. Denote $\xi_i(t)$ to be the indicator that subject i is at risk at time t.

Consider a hypothetical randomized study in which the treatment (S2P timing) is randomly assigned under distribution $f(z)$, where z is discrete as in this study. The score equation for α in Model (1.4) from the hypothetical randomized study is

$$\sum_{i=1}^{n^*} U_i^*(T_i^*, Z_i; \alpha) = \sum_{i=1}^{n^*} \int_0^\infty \Big\{ A(Z_i, K) - \frac{\sum_{l=1}^{n^*} A(Z_l, K)\xi_l(t)\exp(g(Z_l;\alpha))}{\sum_{l=1}^{n^*} \xi_l(t)\exp(g(Z_l;\alpha))}) \Big\} dN_i(t) = 0.$$

Here, $A(\cdot)$ is the design matrix under the linear spline model with knots K. The score term in the estimating equation $\frac{1}{n^*}\sum_{i=1}^{n^*} U_i^*(T_i^*, Z_i; \alpha) = 0$ is an unbiased estimator of $\mathbb{E}_R\{U^*(T^*, Z; \alpha)\}$, where $\mathbb{E}_R\{\cdot\}$ refers to expectation under the hypothetical randomized study [26]. Murphy et al [29] derived the Radon-Nikodym derivative for the distribution of

the data from a hypothetical randomized study with respect to the distribution of the data from an observational study under the no-unmeasured-confounder assumption.

One can obtain unbiased estimating equations for α in the observational study by weighting the terms according to their Radon-Nikodym derivatives under the hypothetical randomized study [26]. Suppose the probability distribution of the data (T^*, Z, X, Δ) under the hypothetical distribution is $\Pr_R(\cdot)$ and under the observational study is $\Pr(\cdot)$. Murphy et al [29] have shown that the distribution of (T^*, Z, X, Δ) under $\Pr_R(\cdot)$ is absolutely continuous with respect to the distribution of (T^*, Z, X, Δ) under $\Pr(\cdot)$ with a Radon-Nikodym derivative. Because the S2P treatment censored by death or heart transplantation is specified by the discrete-time competing risk model, the Radon-Nikodym derivative reduces to $\mathbb{E}\left\{\frac{f(z)}{f(z|x)}|T^* = t, Z = z, X = x, \Delta = \delta\right\}$, where $\mathbb{E}\{\cdot\}$ is the expectation under the observational study. Now,

$$\mathbb{E}_R\{U^*(T^*, Z; \alpha)\} = \mathbb{E}\left\{U^*(T^*, Z; \alpha) \times \frac{f(z)}{f(z|x)}\right\} = \mathbb{E}\{U(T^*, Z; \alpha)\}$$

where $U(T^*, Z; \alpha) := U^*(T^*, Z; \alpha) \times \frac{f(z)}{f(z|x)} = U^*(T^*, Z; \alpha) \times w^*$ and $U(T^*, Z; \alpha)$ is the score function of Model (1.4) for the observational study. We estimate $f(z|x)$ from the discrete-time competing-risk model and use it to compute the generalized propensity score $e^{(r)}(z, x)$, and the stabilized weight w^*:

$$\sum_{i=1}^{n^*} U_i(T_i^*, Z_i; \alpha) = \sum_{i=1}^{n^*} U_i^*(T_i^*, Z_i; \alpha) \times w_i^*$$
$$= \sum_{i=1}^{n^*} w_i^* \int_0^\infty \left\{A(Z_i, K) - \frac{\sum_{l=1}^{n^*} A(Z_l, K)\xi_l(t)\exp(g(Z_l; \alpha))}{\sum_{l=1}^{n^*} \xi_l(t)\exp(g(Z_l; \alpha))}\right\} dN_i(t).$$

By solving $\sum_{i=1}^{n^*} U_i(T_i^*, Z_i; \alpha) = 0$, which is identical to the estimating equations in (1.5) and (1.6), we obtain a consistent estimate for α in Model (1.4) [26].

A.4.

To estimate frequentist properties of the parameter estimates, we repeat the following procedures 2,000 times. In each iteration, we generate a data set, then apply our method to the data set and record the estimates. Each simulation consists of five steps:

1. Consistent with our application to the SVRT data, we bootstrap the real data set with the desired sample size ($n = 500$ or $n = 1000$). We only keep the baseline covariates X and the post-Norwood event type R (signifying elective S2P, non-elective S2P, death, or heart transplantation).

2. Plugging the observed x and r into the generalized propensity score model estimated from the real data, we obtain the probability Pr$(Z, r|x)$. We sample the Norwood outcome event time Z from the distribution with probability mass function Pr$(Z, r|x)$.

3. We sample post-S2P survival from a standard parametric distribution, as indicated in Table 1.3. We simulate the post-S2P survival time T under the spline model estimated from the real data for the corresponding R.

4. The censoring of post-S2P survival is administrative. Thus we censor the simulated survival time T at a designated censoring time C.

5. We apply our proposed analyses to the above generated data.

APPENDIX B

APPENDIX of CHAPTER 2

We provide here the form of the log-likelihood function in the restricted cubic spline model.

The cumulative hazard $\Lambda_i(u, z_i)$ at post-S2P time u with treatment z_i for subject i is:

$$\Lambda_i(u, z_i) = \exp\left(\gamma^{\mathsf{T}} x_i + \kappa(z_i)\right)\Lambda_0(u),$$

where

$$\Lambda_0(u) = \alpha_{J-1}u + \sum_{j=0}^{J-2} \alpha_j \mathfrak{B}_j(u);$$

$\mathfrak{B}_0(u) = \frac{1}{2}u^2$, for $j = 1, \ldots, J-2$, and

$$\mathfrak{B}_j(u) = \frac{1}{4}(u - K_j)_+^4 - \frac{(u - K_{J-1})_+^4(K_J - K_j)}{4(K_J - K_{J-1})} + \frac{(u - K_J)_+^4(K_{J-1} - K_j)}{4(K_J - K_{J-1})}.$$

$\mathfrak{B}_j(u)$ is the integral of $\mathcal{B}_j(u)$, where $\mathcal{B}_0(u) = u$ and, for $j = 1, \ldots, J-2$,

$$\mathcal{B}_j(u) = (u - K_j)_+^3 - \frac{(u - K_{J-1})_+^3(K_J - K_j)}{(K_J - K_{J-1})} + \frac{(u - K_J)_+^3(K_{J-1} - K_j)}{(K_J - K_{J-1})}.$$

The full log-likelihood is then

$$\mathcal{L}(\chi; x, z, t^*, d) = \ln\left(\prod_{i=1}^{n^*} \lambda_i(t_i^*, z_i)^{d_i} \exp(-\Lambda_i(t_i^*, z_i))\right)$$

$$= \sum_{i=1}^{n^*}\left(d_i \ln\left\{\alpha_{J-1} + \sum_{j=0}^{J-2}\alpha_j \mathcal{B}_j(t_i^*)\right\} + d_i\left(\gamma^\mathsf{T} x_i + \kappa(z_i)\right) - \exp\left(\gamma^\mathsf{T} x_i + \kappa(z_i)\right)\Lambda_0(t_i^*)\right)$$

$$= \sum_{i=1}^{n^*}\left(d_i \ln\left\{\alpha_{J-1} + \sum_{j=0}^{J-2}\alpha_j \mathcal{B}_j(t_i^*)\right\} + d_i\left(\gamma^\mathsf{T} x_i + \kappa(z_i)\right)\right.$$

$$\left. - \exp\left(\gamma^\mathsf{T} x_i + \kappa(z_i)\right)\left\{\alpha_{J-1} t_i^* + \sum_{j=0}^{J-2}\alpha_j \mathfrak{B}_j(t_i^*)\right\}\right),$$

where n^* is the number of patients who underwent S2P; t_i^* is the observed value of T_i^*; and d_i is the observed value of D_i for subject i.

APPENDIX C

APPENDIX of CHAPTER 3

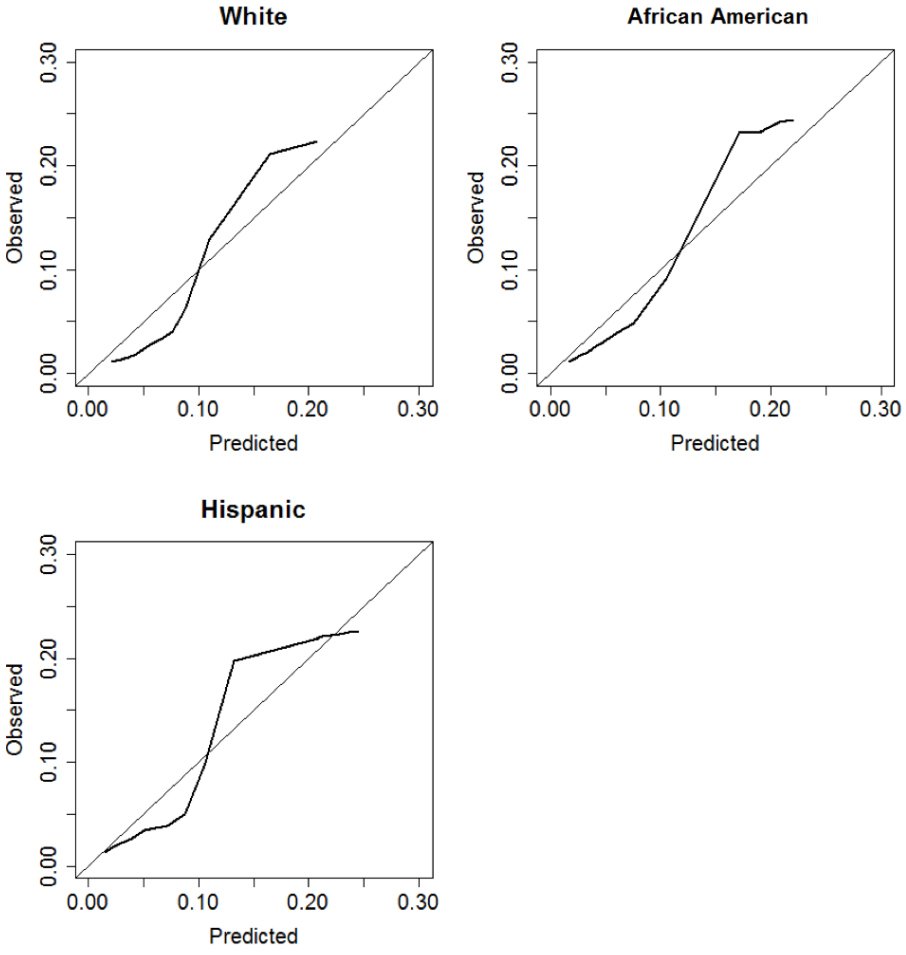

Figure C.1: Calibration plots for VA women ASCVD risk model 2 by race

BIBLIOGRAPHY

[1] Shervin Assari. Veterans and risk of heart disease in the united states: A cohort with 20 years of follow up. *International Journal of Preventive Medicine*, 5(6):703, 2014.

[2] John Barnard and Xiao-Li Meng. Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research*, 8:17–36, 1999.

[3] Brian T Bateman, Kate M Shaw, Elena V Kuklina, William M Callaghan, Ellen W Seely, and Sonia Hernández-Díaz. Hypertension in women of reproductive age in the united states: Nhanes 1999-2008. *PloS one*, 7(4), 2012.

[4] Xiaofei Chen, Daniel F Heitjan, Gerald Greil, and Haekyung Jeon-Slaughter. Estimating the optimal timing of surgery from observational data. *Biometrics*, 2020. doi: https://doi.org/10.1111/biom.13311.

[5] Xiaofei Chen, Bala Ramanan, Shirling Tsai, and Haekyung Jeon-Slaughter. Differential impact of aging on cardiovascular risk in women military service members. *Journal of the American Heart Association*, 2020. doi: https://doi.org/10.1161/JAHA.120.015087.

[6] Stephen R Cole and Miguel A Hernán. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168:656–664, 2008.

[7] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34:187–202, 1972.

[8] Harald Cramér. *Mathematical Methods of Statistics*. Princeton: Princeton University Press, 1946.

[9] Nancy F Crum-Cianflone, Melissa E Bagnell, Emma Schaller, et al. Impact of combat deployment and posttraumatic stress disorder on newly reported coronary heart disease among us active duty and reserve forces. *Circulation*, 129(18):1813–1820, 2014.

[10] Christiana Drake. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, 49:1231–1236, 1993.

[11] Bradley Efron. Bootstrap methods: Another look at the jackknife. In *Breakthroughs in Statistics*, pages 569–593. Springer, 1992.

[12] Nancy S Ghanayem, James S Tweddell, George M Hoffman, et al. Optimal timing of the second stage of palliation for hypoplastic left heart syndrome facilitated through home monitoring, and the results of early cavopulmonary anastomosis. *Cardiology in the Young*, 16:61–66, 2006.

[13] David C Goff, Donald M Lloyd-Jones, Glen Bennett, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the american college of cardiology/american heart association task force on practice guidelines. *Journal of the American College of Cardiology*, 63(25 Part B):2935–2959, 2014.

[14] Larry B Goldstein. Accuracy of icd-9-cm coding for the identification of patients with acute ischemic stroke: Effect of modifier codes. *Stroke*, 29(8):1602–1604, 1998.

[15] Scott M Grundy, Neil J Stone, Alison L Bailey, Craig Beam, et al. 2018 AHA/ACC guideline on the management of blood cholesterol: A report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Journal of the American College of Cardiology*, 73(24):e285–e350, 2019.

[16] Janet K Han, Elizabeth M Yano, Karol E Watson, and Ramin Ebrahimi. Cardiovascular care in women veterans: A call to action. *Circulation*, 139(8):1102–1109, 2019.

[17] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.

[18] Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, 1996.

[19] David A Hehir, Troy E Dominguez, Jean A Ballweg, Chitra Ravishankar, Bradley S Marino, Geoffrey L Bird, et al. Risk factors for interstage death after stage 1 reconstruction of hypoplastic left heart syndrome and variants. *The Journal of Thoracic and Cardiovascular Surgery*, 136:94–99, 2008.

[20] Daniel F Heitjan and J Richard Landis. Assessing secular trends in blood pressure: A multiple-imputation approach. *Journal of the American Statistical Association*, 89(427):750–759, 1994.

[21] James E Herndon and Frank E Harrell. The restricted cubic spline as baseline hazard in the proportional hazards model with step function time-dependent covariables. *Statistics in Medicine*, 14:2119–2129, 1995.

[22] Liangyuan Hu, Joseph W Hogan, Ann W Mwangi, and Abraham Siika. Modeling the causal effect of treatment initiation time on survival: Application to HIV/TB co-infection. *Biometrics*, 74:703–713, 2018.

[23] Kosuke Imai and David A Van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866, 2004.

[24] Guido W Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87:706–710, 2000.

[25] Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, & Biomedical Sciences*. Cambridge University Press, 2015.

[26] Brent A Johnson and Anastasios A Tsiatis. Semiparametric inference in observational duration-response studies, with duration possibly right-censored. *Biometrika*, 92:605–618, 2005.

[27] Keren Lehavot, Katherine D Hoerster, Karin M Nelson, Matthew Jakupcak, and Tracy L Simpson. Health indicators for military, veteran, and civilian women. *American Journal of Preventive Medicine*, 42(5):473–480, 2012.

[28] James M Meza, Edward J Hickey, Eugene H Blackstone, Robert DB Jaquiss, Brett R Anderson, William G Williams, et al. The optimal timing of stage 2 palliation for hypoplastic left heart syndrome: An analysis of the pediatric heart network single ventricle reconstruction trial public data set. *Circulation*, 136:1737–1748, 2017.

[29] Susan A Murphy, Mark J van der Laan, James M Robins, and CPPRG. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96:1410–1423, 2001.

[30] Richard G Ohye, Lynn A Sleeper, Lynn Mahony, Jane W Newburger, Gail D Pearson, Minmin Lu, Caren S Goldberg, Sarah Tabbutt, Peter C Frommelt, Nancy S Ghanayem, et al. Comparison of shunt types in the Norwood procedure for single-ventricle lesions. *New England Journal of Medicine*, 362:1980–1992, 2010.

[31] Samantha E Parker, Cara T Mai, Mark A Canfield, Russel Rickard, Ying Wang, Robert E Meyer, et al. Updated national birth prevalence estimates for selected birth defects in the united states, 2004–2006. *Birth Defects Research Part A: Clinical and Molecular Teratology*, 88:1008–1016, 2010.

[32] DM Reker, BB Hamilton, PW Duncan, SC Yeh, and A Rosen. Stroke: who's counting what? *Journal of Rehabilitation Research and Development*, 38(2):281—289, 2001.

[33] Report. Department of defense. va/dod clinical practice guideline for the management of dyslipidemia for cardiovascular risk reduction, 2014.

[34] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

[35] James M Robins, Miguel Angel Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560, 2000.

[36] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[37] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[38] Donald B Rubin. Multiple imputations in sample surveys — a phenomenological Bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, volume 1, pages 20–34. American Statistical Association, 1978.

[39] Donald B Rubin. The Bayesian bootstrap. *The Annals of Statistics*, 9:130–134, 1981.

[40] Donald B Rubin. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley, 1987.

[41] Donald B Rubin. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31:161–170, 2004.

[42] Joseph L Schafer. *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC, 1997.

[43] Michael Schomaker and Christian Heumann. Bootstrap inference when using multiple imputation. *Statistics in Medicine*, 37(14):2252–2266, 2018.

[44] Janet M Simsic, SM Bradley, MR Stroud, and AM Atz. Risk factors for interstage death after the norwood procedure. *Pediatric Cardiology*, 26:400–403, 2005.

[45] Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *British Medical Journal*, 338:b2393, 2009.

[46] Jeremy B Sussman, Wyndy L Wiitala, Matthew Zawistowski, Timothy P Hofer, Douglas Bentley, and Rodney A Hayward. The veterans affairs cardiac risk score: Recalibrating the ascvd score for applied use. *Medical care*, 55(9):864, 2017.

[47] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1996.

[48] Xin Ming Tu, Xiao-Li Meng, and Marcello Pagano. The AIDS epidemic: Estimating survival after AIDS diagnosis from surveillance data. *Journal of the American Statistical Association*, 88(421):26–36, 1993.

[49] Gerhard Tutz and Matthias Schmid. *Modeling Discrete Time-to-Event Data*. Springer, 2016.

[50] Stef Van Buuren, Hendriek C Boshuizen, and Dick L Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18:681–694, 1999.

[51] Varsha G Vimalananda, Donald R Miller, Cindy L Christiansen, Wenyu Wang, Patricia Tremblay, and B Graeme Fincke. Cardiovascular disease risk factors among women veterans at va medical facilities. *Journal of General Internal Medicine*, 28(2):517–523, 2013.

[52] Janet Weiner, Therese S Richmond, Joseph Conigliaro, and Douglas J Wiebe. Military veteran mortality following a survived suicide attempt. *BMC Public Health*, 11(1):374, 2011.

[53] Daniel Westreich, Jessie K Edwards, Stephen R Cole, Robert W Platt, Sunni L Mumford, and Enrique F Schisterman. Imputation approaches for potential outcomes in causal inference. *International Journal of Epidemiology*, 44(5):1731–1737, 2015.

[54] London AS Wilmoth JM. Handbook of aging and the social sciences. *Handbook of Aging and the Social Sciences*, 2016.

[55] Stanley Xu, Colleen Ross, Marsha A Raebel, Susan Shetterly, Christopher Blanchette, and David Smith. Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value in Health*, 13:273–277, 2010.

[56] Mouhammad Yabrodi and Christopher W Mastropietro. Hypoplastic left heart syndrome: From comfort care to long-term survival. *Pediatric Research*, 81:142–149, 2017.

[57] Shu Yang, Guido W Imbens, Zhanglin Cui, Douglas E Faries, and Zbigniew Kadziola. Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, 72(4):1055–1065, 2016.

[58] Yeying Zhu, Donna L Coffman, and Debashis Ghosh. A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of Causal Inference*, 3:25–40, 2015.