©Copyright 2019
Marlena Bannick

Estimating time to intermediate endpoints using population-level survival data and deconvolution methods, with application to cancer progression and recurrence

Marlena Bannick

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2019

Committee:

Ruth Etzioni

Megan Othus

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Estimating time to intermediate endpoints using population-level survival data and deconvolution methods, with application to cancer progression and recurrence

Marlena Bannick

Chair of the Supervisory Committee:
Ruth Etzioni
Department of Biostatistics

Individuals diagnosed with cancer progress through disease stages with transition rates that are often unobserved but potentially estimable. We focus on deconvolution as a method to partition population-level survival data into two distinct components: (1) time from diagnosis to an intermediate endpoint and (2) time from the intermediate endpoint to death. Using data on overall survival from diagnosis and survival from the intermediate endpoint to death we propose a novel deconvolution method to estimate the distribution of the time from diagnosis to the intermediate endpoint. The method allows for an individual-level frailty to influence the correlation between time to the intermediate endpoint and time to death.

We focus on two main applications to cancer. In our first application, we validate the deconvolution method using data on individuals with prostate cancer. We estimate the time to metastasis from diagnosis and compare this with the observed time to metastasis from the SPCG-4 clinical trial. In our second application, we use deconvolution methods to estimate the time to distant metastatic recurrence of melanoma using data from the Surveillance, Epidemiology, and End Results program.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY AND ABBREVIATIONS

AJCC: American Joint Committee on Cancer

HR: Hazard Ratio

ICD-O-3: International Classification of Diseases for Oncology, Third Edition

KM: Kaplan-Meier

NCI: National Cancer Institute

PSA: Prostate-Specific Antigen

RP: Radical Prostectomy

SEER: Surveillance, Epidemiology, and End Results Program

SPGC-4: Scandinavian Prostate Cancer Group Study Number 4

WW: Watchful Waiting

ACKNOWLEDGMENTS

First, I would like to thank the chair of my thesis committee, Dr. Ruth Etzioni. Throughout the thesis process, she has given me countless opportunities to grow as a biostatistician and a researcher. I would also like to thank my second reader, Dr. Megan Othus, for her insights and helpful feedback.

I would like to acknowledge the support that I've received while writing this thesis from my current and former colleagues at the Institute for Health Metrics and Evaluation. They have been a sounding board while I worked on the derivations, they recommended software packages to use, and they gave feedback on my drafts.

Lastly, I would like to thank my family and friends for their love and support as I've pursued this degree and a full-time career simultaneously. Most importantly, I thank my husband, Jonathan Bannick, for his unwavering encouragement and support for the last three years.

DEDICATION

to my husband, Jonathan

Chapter 1

INTRODUCTION

1.1 Background

Accurate estimates of cancer progression are vital to researchers, healthcare providers, and patients. With longitudinal cohort studies where the same individuals are followed up for multiple time points, estimating the time to defined intermediate endpoints, like metastatic recurrence, is straightforward.[1] Estimating the parameters with which individuals diagnosed with cancer progress through disease stages with readily available population-level data is more difficult. Furthermore, estimating the time to an intermediate endpoint like progression to a later stage may not be immediately observable in data available to a healthcare provider. Understanding the length of time that it takes for individuals diagnosed with, for example, Stage I or II, to progress to a Stage III or Stage IV cancer has the potential to illuminate the natural history of specific types of cancer at the population level and inform early detection screening policies.

We focus on deconvolution as a method to partition population-level survival data into two distinct components: (1) time from diagnosis to an intermediate endpoint and (2) time from the intermediate endpoint to death. Recent work by Mariotto and colleagues focused on developing a deconvolution method for cancer recurrence using population-level SEER data, applied to recurrence of metastatic breast cancer. [2] A fundamental assumption of the deconvolution method from Mariotto and colleagues is that the time to an intermediate endpoint is independent of the time to the ultimate endpoint, e.g. time from diagnosis with localized cancer to distant progression is independent of time from distant progression to death. For some cancers this may be a reasonable assumption, but others might violate it. Using data on overall survival from diagnosis and survival from the intermediate endpoint to

death we propose a novel deconvolution method to estimate the distribution of the time from diagnosis to the intermediate endpoint. The method allows for an individual-level frailty to influence the correlation between time to the intermediate endpoint and time to death.

We first validate the deconvolution methods using data on individuals with prostate cancer from a clinical trial where data are available on both time from diagnosis to prostate cancer death as well as time from diagnosis to the intermediate endpoint of metastasis. In our application, we use deconvolution methods to estimate the time to distant metastatic recurrence of melanoma.

1.2 Objectives

In this thesis, we introduce a novel deconvolution method that simultaneously estimates the correlation at the population level between the two time to event variables, and the parameters governing those two time to event variables. We first detail the statistical theory behind our method and show its utility with a simulation study. We then use the deconvolution method in two applications.

In our first application, we use data from the Scandinavian Prostate Cancer Group Study Number 4 (SPCG-4) as a validation for our method. In the SPCG-4 trial, metastatic recurrences were recorded at the individual level and thus we can compare the empirical time to metastatic recurrence to the estimate from deconvolution using time from diagnosis to death and time from metastatic recurrence to death. For our second application, we use data extracted from the Surveillance, Epidemiology, and End Results program on survival for melanoma to estimate the time to recurrence of metastatic disease.[3] We extract data on time from diagnosis of localized disease to death and time from diagnosis of de novo metastatic disease to death. Our method is an extension of the method presented by Mariotto and colleagues for recurrence of metastatic breast cancer.[2]

Chapter 2

STATISTICAL METHODS

2.1 Methods Introduction

2.1.1 Notation and Basic Concepts

Define T_1 as the time from diagnosis to an intermediate endpoint, T_2 as the time from the intermediate endpoint to death, and $T^* = T_1 + T_2$ as the total time from diagnosis to death. Let $f_{T_1}(t_1)$, $f_{T_2}(t_2)$, and $f_{T^*}(t^*)$ be their respective probability density functions, and $S_{T_1}(t)$, $S_{T_2}(t)$, and $S_{T^*}(t^*)$ be their respective survival functions.

We use the following definition of the survival $S(t)$ and hazard $h(t)$ functions for t:

$$\begin{aligned} f(t) &:= \lim_{\Delta t \to 0} \frac{P[t \le T < t + \Delta t]}{\Delta t} = \frac{\mathrm{d}}{\mathrm{d}t} S(t) \\ S(t) &:= P[T > t] = \int_t^\infty f(u)\,\mathrm{d}u = 1 - F(t) \\ h(t) &:= \lim_{\Delta t \to 0} \frac{P[t \le T < t + \Delta t \,|\, T \ge t]}{\Delta t} = \frac{f(t)}{S(t)}. \end{aligned}$$

We use Kaplan-Meier (KM) [4], or product-limit, estimates of S(t), given by:

$$\tilde{S}(t) := \prod_{i:u_i \le t} \left(1 - \frac{d_i}{n_i}\right) \tag{2.1}$$

where u_i is the i^{th} observed event time, d_i is the number of events observed at event time u_i, and n_i is the size of the risk set at event time u_i. This construction of the Kaplan-Meier estimator assumes independent censoring times. We use Greenwood's formula for the standard error of $\hat{S}(t)$, given by:

$$\tilde{SE}(t) := \tilde{S}(t) \sqrt{\sum_{i:u_i \le t} \frac{d_i}{n_i(n_i - d_i)}}. \tag{2.2}$$

We use the following definitions of the probability density functions for exponential, gamma, and Weibull random variables (using rate parameterizations):

$$\begin{aligned}\text{Exponential: } & f(x|\delta) = \delta e^{-\delta x} \\ \text{Gamma: } & f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-x\beta} \\ \text{Weibull: } & f(x|\lambda,p) = p\lambda(\lambda x)^{p-1} e^{-(\lambda x)^p}.\end{aligned}$$

As noted above, when referring to the estimates of $S(t)$ for any random variable T, we denote $\tilde{S}(t)$ as the KM estimates of $S(t)$. We denote $\hat{S}(t)$ as the *deconvolution* estimates of $S(t)$.

We use non-linear least squares to fit parametric survival functions to the KM estimates, $\tilde{S}(t)$. Non-linear least squares minimizes the sum of squared residuals with respect to unknown parameter(s) θ. Consider KM estimats at n time points. Define the residual for the i^{th} time point t_i from the estimated survival function to the empirical survival function as

$$r(t_i, \theta) := \tilde{S}_T(t_i) - S_T(t_i, \theta) \tag{2.3}$$

and the (weighted) sum of squared residuals as

$$R_T(w, \theta) := \sum_{i=1}^{n} r(t_i, \theta)^2 \cdot w_i \tag{2.4}$$

where w_i is a weight for the i^{th} residual. We may choose not to weight the survival for each time t_i (i.e. $w_i = 1 \; \forall \; i$), or to precision weight each time t_i using the standard error of the Kaplan-Meier estimate at t_i from equation 2.2. Furthermore, define $R_T(\theta, w)$ as the objective function to be minimized with respect to θ for observed data $\tilde{S}_T(t_i)$, $i \in (0, ..., n)$.

2.1.2 Problem Statement

Let

$$\begin{aligned} T_1 &\sim f_{T_1}(t_1 \,|\, \theta_{T_1}) \\ T_2 &\sim f_{T_2}(t_2 \,|\, \theta_{T_2}) \\ T^* &:= T_1 + T_2 \end{aligned}$$

where θ_1 and θ_2 can be single- or multi-dimensional, and where the density of T^* is a *convolution* of the densities of T_1 and T_2, i.e.

$$f_{T^*}(t^*) := \int_0^{t^*} f_{T_1,T_2}(t^* - t_2, t_2|\theta_{T_1}, \theta_{T_2})\,\mathrm{d}t_2. \tag{2.5}$$

If convolution is distributional addition, then *deconvolution* is distributional subtraction. When $T_1 \perp\!\!\!\perp T_2$, this lends itself easily to deconvolution; deconvolution is more complicated when $T_1 \not\perp\!\!\!\perp T_2$.

Can we use deconvolution methods and information on only T_2 and T^* to infer the distribution of T_1, both in the context of independence and dependence between T_1 and T_2? More specifically, can we use data on the distributions of time from an intermediate endpoint (like metastasis, or recurrence) to death (T_2), and time from diagnosis to death (T^*), to recover the unobserved distribution of time from diagnosis to an intermediate endpoint (T_1)?

2.2 Deconvolution

2.2.1 Deconvolution with Independence

Assume that $T_1 \perp\!\!\!\perp T_2$. Then their joint distribution is given by

$$f_{T_1,T_2}(t_1, t_2|\theta_{T_1}, \theta_{T_2}) = f_{T_1}(t_1|\theta_{T_1}) f_{T_2}(t_2|\theta_{T_2}).$$

We can make the substitution $T_1 = T^* - T_2$, and then the joint distribution of T^* and T_2 is given by

$$f_{T^*,T_2}(t^*, t_2|\theta_{T_1}, \theta_{T_2}) = f_{T_1}(t^* - t_2|\theta_{T_1}) f_{T_2}(t_2|\theta_{T_2}).$$

The marginal survival functions for T_2 and T^* are then given by

$$\begin{aligned} S_{T_2}(t|\theta_{T_2}) &= \int_t^\infty f_{T_2}(t_2|\theta_{T_2})\,\mathrm{d}t_2 \\ S_{T^*}(t|\theta_{T_1}, \theta_{T_2}) &= \int_t^\infty \int_0^{t^*} f_{T_1}(t^* - t_2|\theta_{T_1}) \cdot f_{T_2}(t_2|\theta_{T_2})\,\mathrm{d}t_2\,\mathrm{d}t^*. \end{aligned}$$

2.2.2 Deconvolution with Correlation

Now assume that $T_1 \not\perp\!\!\!\perp T_2$, and that the correlation between them can be described by a shared frailty term, K, with

$$K \sim g(k|\theta_K)$$

where g is a parametric distribution with parameter(s) θ_K, which can be either single- or multi-dimensional. Thus, T_1 and T_2 are independent after conditioning on K, i.e.

$$T_1 \perp\!\!\!\perp T_2 \,|\, K$$
$$f_{T_1,T_2\,|\,K}(t_1, t_2 \,|\, \theta_{T_1}, \theta_{T_2}, k) = f_{T_1\,|\,K}(t_1 \,|\, \theta_{T_1}, k) \cdot f_{T_2\,|\,K}(t_2 \,|\, \theta_{T_2}, k).$$

Now make the substitution $T_1 = T^* - T_2$. Since the Jacobian of this transformation is 1, the joint distribution of $T^*, T_2 \,|\, K$ is given by

$$f_{T^*,T_2\,|\,K}(t^*, t_2, \,|\, \theta_{T_1}, \theta_{T_2}, k) = f_{T_1\,|\,K}(t^* - t_2 \,|\, \theta_{T_1}, k) \cdot f_{T_2\,|\,K}(t_2 \,|\, \theta_{T_2}, k).$$

Because K is an *unobserved* frailty term, we want to find the marginal distribution of (T^*, T_2), and of T_2 by itself. Integrating over the support of K,

$$f_{T_2}(t_2 \,|\, \theta_{T_2}, \theta_K) = \int_K f_{T_2\,|\,K}(t_2 \,|\, \theta_{T_2}, k) \cdot g(k \,|\, \theta_K)\, \mathrm{d}k \tag{2.6}$$
$$f_{T^*,T_2}(t^*, t_2 \,|\, \theta_{T_1}, \theta_{T_2}, \theta_K) = \int_K f_{T_1,T_2\,|\,K}(t^* - t_2, t_2, \,|\, \theta_{T_1}, \theta_{T_2}, k) \cdot g(k \,|\, \theta_K)\, \mathrm{d}k. \tag{2.7}$$

To find the survival function for T_2 whose density is given by equation 2.6, we integrate the density function from t to ∞. To find the survival function for T^* whose density is given by equation 2.7 , we first integrate over T_2 from 0 to t^* to find the marginal density function for T^*, and then integrate from t to ∞. These functions are given by:

$$S_{T_2}(t \,|\, \theta_{T_2}, \theta_K) = \int_t^\infty f_{T_2}(t_2 \,|\, \theta_{T_2}, \theta_K)\, \mathrm{d}t_2 \tag{2.8}$$
$$S_{T^*}(t \,|\, \theta_{T_1}, \theta_{T_2}, \theta_K) = \int_t^\infty \int_0^{t^*} f_{T^*,T_2}(t^*, t_2, \,|\, \theta_{T_1}, \theta_{T_2}, \theta_K)\, \mathrm{d}t_2\, \mathrm{d}t^* \tag{2.9}$$

Revisiting the problem statement: for a fixed parameter governing the frailty distribution, θ_K, can we use the relationships in equations 2.8 and 2.9 and population-level data on T^* and T_2 to estimate θ_{T_1}, and θ_{T_2}, in order to learn about the distribution of the time from diagnosis to an intermediate endpoint, T_1?

2.2.3 Incorporating the Shared Frailty

We assume parametric distributions for both T_1 and T_2 that are conditional on a shared frailty term, K. Thus, the parameters for T_1 are (θ_1, θ_K), the parameters for T_2 are (θ_2, θ_K), and the distribution of the their sum, T^*, is characterized by $(\theta_1, \theta_2, \theta_K)$. We also assume some parametric distribution for K, the shared frailty term.

For mathematical tractability, we assume that this shared frailty term K modifies the *hazard* for T_1 and T_2 in a proportional manner. For example, if the hazard is $h(t|\theta)$, then the hazard modified by the frailty term is

$$h'(t|\lambda, k) = h(t|\lambda, k) \cdot k.$$

We therefore *construct* the distributions of $T_1|K$ and $T_2|K$ such that we have this proportional relationship for the hazard. This modeling strategy has precedent; it is similar to how a Cox proportional hazards model includes a function of predictor variables that proportionally scales the hazard, and how in estimating the parameters of a Poisson model with an over-dispersion parameter, there is an unobserved term that proportionally scales the mean to induce the over-dispersion.

Consider the exponential case, where the rate parameter is λ. The hazard for an exponential random variable is also λ, i.e.

$$f(t|\lambda) = \lambda e^{-\lambda t}$$
$$h(t|\lambda) = \lambda.$$

To scale the hazard proportionally by k, we just multiply the rate parameter by k, i.e.

$$f'(t|\lambda, k) = k\lambda e^{-\lambda k \cdot t} \implies h'(t|\lambda, k) = k \cdot \lambda. \tag{2.10}$$

Thus, if in the independent case T_1 and T_2 are exponential with rate parameters λ and μ, respectively, in the correlated case $T_1|K$ and $T_2|K$ have rate parameters $\lambda \cdot k$ and $\mu \cdot k$, respectively. T_1 and T_2 are still conditionally exponential on K, but their marginal distribution will depend on the distribution of K.

In the more general case, consider the variable T with distribution $f_T(t|\theta)$ and hazard function $h_T(t|\theta)$, without conditioning on K. To introduce the shared frailty in accordance with the proportional hazard setup above, we want to find $f'_{T\,|\,K}(t|\theta, k)$ such that

$$h'_{T\,|\,K}(t|\theta, k) = h_T(t|\theta) \cdot k. \tag{2.11}$$

We can now define the survival functions for T_2 and T^*, the parametric distributions $f'_{T_1\,|\,K}(t|\theta_1, k)$, $f'_{T_2\,|\,K}(t|\theta_2, k)$, and $g(k|\theta_K)$, and the equations 2.8 and 2.9, i.e.

$$\begin{aligned} S_{T_2}(t\,|\,\theta_{T_2}, \theta_K) &= \int_t^\infty \int_K f'_{T_2\,|\,K}(t_2\,|\,\theta_{T_2}, k) \cdot g(k\,|\,\theta_K)\,\mathrm{d}k\,\mathrm{d}t_2 \\ S_{T^*}(t\,|\,\theta_{T_1}, \theta_{T_2}, \theta_K) &= \int_t^\infty \int_0^{t^*} \int_K f'_{T^*,T_2\,|\,K}(t^*, t_2, \,|\,\theta_{T_1}, \theta_{T_2}, k) \cdot g(k\,|\,\theta_K)\,\mathrm{d}k\,\mathrm{d}t_2\,\mathrm{d}t^* \end{aligned}$$

2.2.4 Incorporating Cure Fractions

For some cancers, individuals initially diagnosed with an early stage disease may never go on to develop the intermediate endpoint of interest in the deconvolution. As such, it is important to account for this "cured" fraction in the deconvolution. A mixture cure model is of the form

$$S(t) = c + (1 - c) \cdot S_{UC}(t) \tag{2.12}$$

where $S_{UC}(t)$ is the survival of the un-cured population at time t, c is the fraction of the population that is cured, and $S(t)$ is the survival for the mixed population of individuals that are both cured and un-cured. This formulation expresses the relative contribution of survival at time t for those that are cured (1) and those that are un-cured ($S_{UC}(t)$).

Let S_{T^*} be the survival function for T^* in the mixed population, $S_{T^*,UC}$ be the survival function for T^* in the un-cured population, S_{T_1} be the survival function for T_1 (time from

diagnosis to intermediate endpoint) in the mixed population, and $S_{T_1,UC}$ be the survival function for T_1 in the un-cured population. By definition, an individual that is cured does not go on to have the intermediate endpoint, so S_{T_2} only applies to those that are un-cured.

Following the implementation of Mariotto and colleagues, we fit a mixture cure model to the T^* survival data, the overall time to death for individuals diagnosed with localized, or Stage I disease, and then use the survival for the un-cured population in the deconvolution.[2] To fit the cure model, we first need to specify a parametric distribution for the survival of the un-cured fraction, $S^*_{UC}(t)$. For the purposes of the deconvolution model, we would like this to be as flexible as possible so as not to constrain the survival estimates that will go into the deconvolution method. The most flexible parametric implementation for the mixture cure model in the available packages is a generalized gamma distribution (package `flexsurvcure` in `R`)[5]. We solve for $S_{T^*,UC}$, the overall survival for the un-cured population, with the following equation:

$$S_{T^*,UC}(t) = \frac{S_{T^*}(t) - c}{1 - c}. \tag{2.13}$$

We then use the survival for the un-cured fraction of the population, $S_{T^*,UC}(t)$ in the deconvolution method in place of the typical $S_{T^*}(t)$ that is shown in equation 2.9 when we assume that no one is cured. As a result, the deconvoluted survival estimates for T_1 represent *only* the non-cured fraction of the population, $S_{T_1,UC}$. To back-calculate the deconvoluted survival function of T_1 for the mixed population is, we use the relationship from equation 2.12:

$$S_{T_1}(t) = c + (1 - c) \cdot S_{T_1,UC}(t). \tag{2.14}$$

2.2.5 Method by Mariotto and colleagues, 2018

Mariotto and colleagues proposed two methods – one analytical and one numerical – for estimating T_1 [2]. First, they fit a parametric model (e.g. Weibull or Loglogistic) to the Kaplan-Meier estimates of $S_{T^*}(t)$ to obtain estimates of both $S_{T^*}(t)$ and $f_{T^*}(t)$. For the

analytical method, they assume that T_2 is exponentially distributed, i.e.

$$f_{T_2}(t) \sim Exponential(\psi).$$

Then they show that an analytical solution for $S_{T^*}(t)$ is given by

$$S_{T_1}(t) = S_{T^*}(t) - \frac{1}{\psi} f_{T^*}(t) \tag{2.15}$$

A drawback of the analytical method is that it may produce negative estimates of S_{T_1}. Relaxing the assumption that T_2 is exponentially distributed, Mariotto and colleagues propose a numerical solution.[2]

2.3 Estimation

2.3.1 Parameter Estimation

Suppose we do not have direct data on T_2 or T^*, but we do have Kaplan-Meier (KM) estimates of $S_{T_2}(t_i)$ and $S_{T^*}(t_j)$ at fixed intervals $i \in (0, ..., n)$ and $j \in (0, ..., m)$, where n is the number of follow-up intervals for T_2 and m is the number of follow-up intervals for T^*. Our goal is to estimate the parameters that govern the distributions of T_1, T_2, and T^*, which include the parameters of the frailty distribution if we are allowing correlation between T_1 and T_2. We define an objective function that minimizes the sum of squared error for S_{T^*} and S_{T_2} combined, rather than two separate objective functions for each of S_{T^*} and S_{T_2}.

Recall the definition from equation 2.3 for the sum of squared errors between the KM estimates of a survival function and the estimated survival function from the deconvolution, r. Now, let r_{T_2} be the residuals for S_{T_2} and r_{T^*} be the residuals for S_{T^*}:

$$r_{T_2}(t_i, \theta_{T_2}, \theta_{T_K}) = \tilde{S}_T(t_i) - S_T(t_i|\theta_{T_2}, \theta_K)$$
$$r_{T^*}(t_j, \theta_{T_1}, \theta_{T_2}, \theta_K) = \tilde{S}_{T^*}(t_j) - S_{T^*}(t_j|\theta_{T_1}, \theta_{T_2}, \theta_{T_K})$$

where θ_K is included only when including a shared frailty. We define an objective function that combines both of these residuals, and with the option for weighting observations with

w:

$$R_T(\theta_{T_1}, \theta_{T_2}, \theta_K, w) = \sum_{i=1}^{n} r_{T_2}(t_i, \theta_{T_2}, \theta_K) w_i + \sum_{j=1}^{m} r_{T^*}(t_j, \theta_{T_1}, \theta_{T_2}, \theta_K) w_j$$

Estimates of the parameters θ_1, θ_2, and θ_K are given by the values that minimize the objective function, i.e.

$$\min_{\theta_1, \theta_2, \theta_K} \quad R_T(w, \theta_{T_1}, \theta_{T_2}, \theta_K) \tag{2.16}$$

The estimates of $S_{T_1}(t)$, $S_{T_2}(t)$, and $S_{T^*}(t)$ are then given by following:

$$\hat{S}_{T_1}(t) = S_{T_1}(t|\hat{\theta}_{T_1}, \hat{\theta}_K) \tag{2.17}$$

$$\hat{S}_{T_2}(t) = S_{T_2}(t|\hat{\theta}_{T_2}, \hat{\theta}_K) \tag{2.18}$$

$$\hat{S}_{T^*}(t) = S_{T^*}(t|\hat{\theta}_{T_1}, \hat{\theta}_{T_2}, \hat{\theta_K}) \tag{2.19}$$

where $\hat{S}_{T_1}(t)$ is the primary quantity of interest.

2.3.2 Uncertainty Estimation

In order to obtain uncertainty intervals for $\hat{S}_{T_1}(t)$, $\hat{S}_{T_2}(t)$, and $\hat{S}_{T^*}(t)$, we use the bootstrap technique. Bootstrapping allows us to estimate uncertainty by re-sampling with replacement from the data and re-running the deconvolution on those new data resamples.

With individual level data for T_2 and T^*, we can re-sample rows from the data set B times, separately for T_2 and T^*, and calculate the Kaplan-Meier estimates on the bootstrapped data. Done once, this gives us one random realization of the Kaplan-Meier estimates. We can then repeat B times to create many bootstrapped Kaplan-Meier estimates, and perform the deconvolution on all of them. This gives us B parameter estimates and B predicted survival estimates for each time t. A $(1-a)\%$ confidence interval for the survival estimates at each time t is then given by the $(a/2)^{th}$ and $(1-a/2)^{th}$ quantiles of the survival estimates at each time t. If we choose to incorporate a cure fraction, each time that we perform a bootstrap resample, we re-estimate the cure fraction, c_b for $b \in 1:B$ and run the deconvolution method bootstrap for resample b with the cure fraction c_b.

2.4 Model Choices

Now that we have set up the framework for both independent and correlated deconvolution models, we choose parametric distributions for both T_1 and T_2. The simplest distribution to choose for survival data is the exponential distribution. A slightly more complex model is the Weibull distribution. For both of these choices, we will consider what the distribution of T^* looks like under independence and correlation. For the correlated scenarios, we use a Gamma distribution for the frailty term, K.

Recall that the aim is to estimate S_{T_1}. To do so, we want to find functions $S_{T_1}(t|\theta_{T_1})$, $S_{T_2}(t|\theta_{T_2})$, $S_{T^*}(t|\theta_{T_1,T_2})$ (with the additional parameters θ_K when we allow correlation between T_1 and T_2) that are determined by the parameters θ_{T_1}, θ_{T_2} (and θ_K) and that we can optimize with non-linear least squares to fit to the Kaplan-Meier curves for S_{T_2} and S_{T^*}. We will now walk through each of these four model types in sequence – exponential-exponential, exponential-exponential-Gamma, Weibull-Weibull, and Weibull-Weibull-Gamma – providing the closed form survival functions where possible. We have included the derivations of these closed form survival functions in appendix A. In the model choices where we include a Gamma frailty, we fix the shape parameter $\alpha = 1$.

2.4.1 Exponential-Exponential Model

The simplest distributions we could assume for T_1 and T_2 are exponential distributions. The distributions of T_1 and T_2 are as follows:

$$T_1 \sim Exponential(\lambda)$$
$$T_2 \sim Exponential(\mu)$$

The survival function for T_1 and T^* are given by (derivations in section A.1:

$$S_{T_2}(t) = e^{-\mu t}$$
$$S_{T^*}(t) = \frac{\lambda e^{-\mu t} - \mu e^{-\lambda t}}{\lambda - \mu}$$

After finding $\hat{\mu}$ and $\hat{\lambda}$, as described in section 2.3.1, the deconvolution estimate of $S_{T_1}(t)$ is given by

$$\hat{S}_{T_1}(t) = e^{\hat{\lambda}t}$$

2.4.2 Exponential-Exponential-Gamma Model

Building off of the Exponential-Exponential Model, we introduce K to be a shared frailty term that influences the hazard in both T_1 and T_2. Let

$$K \sim Gamma(\alpha, \beta)$$

and let T_1 and T_2 be conditionally exponential, given K. In accordance with section 2.2.3, we want k to be a proportional multiplier for the hazard, $h_{T_1}(t)$ and $h_{T_2}(t)$. We showed in equation 2.10 that to in order to have $T_1|K$ and $T_2|K$ conditionally exponential and have k be a proportional multiplier on the hazards, that we can multiply the rate parameter for an exponential by k, i.e.

$$T_1 \,|\, K \sim Exponential(\lambda k)$$
$$T_2 \,|\, K \sim Exponential(\mu k)$$

After conditioning on the shared frailty term K, the distributions of T_1 and T_2 are no longer marginally exponential, but are conditionally exponential given K.

To define the survival functions for T_2 and T^*, we first need the marginal distribution for T_2 and the marginal distribution for T^*. We find that the marginal survival distribution for T_2 is are given by (appendix A.2.1)

$$S_{T_2}(t) = \left(\frac{\beta}{\mu t + \beta}\right)^{\alpha} \tag{2.20}$$

and the marginal survival distribution for the convolution of T_1 and T_2, T^* is given by (A.42)

$$S_{T^*}(t) = \frac{\beta^{\alpha}}{\lambda - \mu}\left[\frac{\lambda}{(\mu t + \beta)^{\alpha}} - \frac{\mu}{(\lambda t + \beta)^{\alpha}}\right]. \tag{2.21}$$

For the exponential-exponential-Gamma model, and for a fixed α, there are multiple parameter solutions for λ, μ, and β to the equations. However, all of the parameter sets that minimize the objective function produce the same predicted survival probabilities for S_{T_1}, S_{T_2}, and S_{T^*}. In our analyses, we fix $\alpha = 1$. After finding $\hat{\mu}$, $\hat{\lambda}$, $\hat{\beta}$, as described in section 2.3.1, the deconvolution estimate of $S_{T_1}(t)$ is given by

$$\hat{S}_{T_1}(t) = \left(\frac{\hat{\beta}}{\hat{\lambda}t + \hat{\beta}}\right)^{\alpha}. \tag{2.22}$$

2.4.3 Weibull-Weibull Model

Using exponential distributions for T_1 and T_2 may suffice for some cancers, but expanding our method to use Weibull models allows for more flexibility when the exponential model performs poorly. A Weibull distribution with a Gamma frailty term is a generalization of the Loglogistic distribution.[6] As the Weibull distribution is a generalization of the exponential distribution, the Weibull-Weibull and Weibull-Weibull-Gamma model are generalizations of the exponential-exponential and exponential-exponential-Gamma models, respectively (where the shape parameters p and q defined below are both set to 1). Let

$$T_1 \sim Weibull(\lambda, p)$$

$$T_2 \sim Weibull(\mu, q)$$

The closed form survival function for T_2 is given by:

$$S_{T_2}(t) = e^{-(t\mu)^q}$$

and the survival function for the sum of T_1 and T_2, T^*, is given by the following integral:

$$\begin{aligned} S_{T^*}(t) &= \int_t^{\infty} \int_0^{t^*} f_{T_1}(t^* - t_2) \cdot f_{T_2}(t_2)\, \mathrm{d}t_2\, \mathrm{d}t^* \\ &= \int_t^{\infty} \int_0^{t^*} p\lambda((t^* - t_2)\lambda)^{p-1} e^{-((t^*-t_2)\lambda)^p} \cdot q\mu(t_2\mu)^{q-1} e^{-(t_2\mu)^q}\, \mathrm{d}t_2\, \mathrm{d}t^*. \end{aligned}$$

After finding $\hat{\mu}$, $\hat{\lambda}$, $\hat{p}$, and $\hat{q}$, as described in section 2.3.1, the deconvolution estimate of $S_{T_1}(t)$ is given by

$$\hat{S}_{T_1}(t) = e^{-(t\hat{\lambda})^{\hat{p}}}. \tag{2.23}$$

2.4.4 Weibull-Weibull-Gamma Model

To define the Weibull-Weibull-Gamma Model, that incorporates the shared frailty term, let

$$K \sim Gamma(\alpha, \beta)$$

and let T_1 and T_2 be conditionally Weibull, given K. In accordance with section 2.2.3, we want k to be a proportional multiplier for the hazards, $h_{T_1}(t)$ and $h_{T_2}(t)$. We showed in equation 2.10 the result for the exponential case. In the Weibull case, we want to find some parameterization of the Weibull (referred to as $f'(t)$ in 2.2.3) where the hazard is multiplied proportionally by k.

The hazard function for a Weibull(λ, p) random variable is

$$h(t) = p\lambda(\lambda t)^{p-1}$$

If we have $f'(t) \equiv Weibull(\lambda k^{\frac{1}{p}}, p)$, then

$$\begin{aligned} h'(t) &= p\lambda k^{\frac{1}{p}}(\lambda k^{\frac{1}{p}} t)^{p-1} \\ &= p\lambda(\lambda t)^{p-1} k = h(t) \cdot k \end{aligned}$$

and we have that the hazard is multiplied proportionally by k. Therefore, let

$$\begin{aligned} T_1 \,|\, K &\sim Weibull(\lambda \cdot k^{\frac{1}{p}}, p) \\ T_2 \,|\, K &\sim Weibull(\mu \cdot k^{\frac{1}{q}}, q) \end{aligned}$$

The distributions of T_1 and T_2 are not marginally Weibull, but are conditionally Weibull given K. The conditional density function for T_2 simplifies to:

$$f_{T_2|K}(t_2|k) = k\mu^q q t_2^{q-1} e^{-k(\mu t_2)^q}$$

and the joint distribution between T_2 and K is:

$$f_{T_2,K}(t_2, k) = k^\alpha \mu^q q t_2^{q-1} e^{-k(\mu t_2)^q} \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-k\beta}. \tag{2.24}$$

Integrating 2.24 over the support of K, we find that the marginal density for T_2 is

$$f_{T_2}(t_2) = \mu^q q t_2^{q-1} \alpha \frac{\beta^\alpha}{(\beta + (\mu t_2)^q)^{\alpha+1}} \tag{2.25}$$

Again we need to find the marginal survival functions for T_2 and T^*. There is no closed form solution for their survival functions, and so we use numerical integration techniques to solve the following integrals:

$$S_{T_2} = \int_t^\infty f_{T_2}(t_2)\,\mathrm{d}t_2 \tag{2.26}$$

$$S_{T^*} = \int_t^\infty \int_0^{t^*} \int_0^\infty f_{T_1|K}(t^* - t_2|k) \cdot f_{T_2|K}(t_2|k) \cdot g_K(k)\,\mathrm{d}k\,\mathrm{d}t_2\,\mathrm{d}t^* \tag{2.27}$$

where $g_K(k)$ is the distribution of the frailty term, and in this case is $Gamma(\alpha, \beta)$ with $\alpha = 1$. After finding $\hat{\mu}$, $\hat{\lambda}$, $\hat{p}$, $\hat{q}$, and $\hat{\beta}$ as described in section 2.3.1, the deconvolution estimate of $S_{T_1}(t)$ is given by (using equations 2.25 and 2.26):

$$\hat{S}_{T_1}(t) = \int_t^\infty \hat{\lambda}^{\hat{p}} \hat{p} t_1^{\hat{p}-1} \alpha \frac{\hat{\beta}^\alpha}{(\hat{\beta} + (\hat{\lambda} t_1)^{\hat{p}})^{\alpha+1}}\,\mathrm{d}t_1. \tag{2.28}$$

2.5 Software Implementation

The code used to perform all analyses in this thesis is available on GitHub.[7]

To construct Kaplan-Meier estimates for S_{T_1}, S_{T_2}, and S_{T^*}, we used the `survfit` function from the `R` package `survival: Survival Analysis`.[8, 9] To fit cure models on S_{T^*}, we used the `flexsurvcure` function from the `R` package `flexsurvcure: Flexible Parametric Cure Models`.[5]

With the Exponential-Exponential-Gamma model, there are closed form solutions to each of the survival functions S_{T_1}, S_{T_2}, and S_{T^*}, regardless of whether or not we are performing independent deconvolution or correlated deconvolution. With the Weibull-Weibull-Gamma model, there is no closed form solution for any of the survival functions, with or without including correlation between T_1 and T_2. However, the equations 2.26 can be numerically integrated. To implement the numerical integration, we use the function `distrExIntegrate`

from the `R` package `distrEx: Extensions of Package 'distr'`.[10] When integrating with the lower bound of 0, we had to instead choose some small value $0 + \epsilon$ to approximate the lower bound of integration. For the validation and applications in chapters 4 and 5, we chose 10^{-3} for ϵ. Additionally, we chose to use 100 as the maximum number of subintervals in the numerical integration (implemented with the `subdivisions` argument) and set a relative tolerance of 10^{-2} (implemented with the `rel.tol` argument).

To minimize the objective function for both the Exponential-Exponential-Gamma model and the Weibull-Weibull-Gamma model, we used the optimizer `optimx` from the `R` package `optimx: Expanded Replacement and Extension of the 'optim' Function`.[11, 12] `optimx` allowed us to explore the use of various optimization algorithms, ultimately settling on the method `nlminb` which uses "unconstrained and box-constrained optimization using PORT routines."[13] We chose a relative tolerance for the `nlminb` method in `optimx` of 10^{-3} (implemented with the `rel.tol` argument).

Chapter 3

SIMULATION STUDY

To assess the validity of the deconvolution method, we perform a simulation study where we simulate data from independent and correlated distributions for T_1 and T_2, and then use the four methods described in section 2.4 to estimate T_1. In this section, we describe the methods of the simulation study and present summaries of the results. Further details about the results can be found in appendix B.

3.1 Simulation Study Design

To perform the simulation study, we first choose the parameters θ_{T_1}, θ_{T_2}, and θ_K. For the exponential case, the θ_{T_1} and θ_{T_2} parameters are both 1-dimensional. We refer to them as λ, and μ, respectively, and they represent the rate parameter of the exponential distribution. For the Weibull case, the θ_{T_1} and θ_{T_2} parameters are both 2-dimensional. We refer to their rate parameters as λ and μ, respectively, and their shape parameters as p and q, respectively. When we include correlation between T_1 and T_2, θ_K represents a 2-dimensional parameter for the Gamma distribution, but we fix the shape parameter of the Gamma distribution to be $\alpha = 1$, and refer to β as the rate parameter of the frailty distribution. For simplicity, we chose one set of parameter values from which to simulate data: we set $\lambda = 0.1$, $\mu = 0.075$, $p = 3$, $q = 4$, and $\beta = 1$. The four simulation setups are given in table 3.1. We did not simulate the data with a cure fraction for T_1, but note that this is a potential area for future work.

In a single simulation for one of these scenarios, we simulated 1000 individuals, each who had a T_1 and T_2 according to the distributions in table 3.1. If we were performing a correlated simulation, we first simulated k for each individual, and then simulated their T_1 and T_2 based

Table 3.1: Simulation setup for four scenarios with different specifications for T_1, T_2, and their shared frailty K

Distribution	Correlation	T_1	T_2	K
Exponential	Independent	Exponential(0.1)	Exponential(0.075)	–
Weibull	Independent	Weibull$(0.1, 3)$	Weibull$(0.075, 4)$	–
Exponential	Correlated	Exponential$(0.1 \cdot k)$	Exponential$(0.075 \cdot k)$	Gamma$(1, 1)$
Weibull	Correlated	Weibull$(0.1 \cdot k^{\frac{1}{3}})$	Weibull$(0.075 \cdot k^{\frac{1}{4}})$	Gamma$(1, 1)$

on their value of k. To get their total survival time T^*, we added their T_1 and T_2 together. We then used the Kaplan-Meier estimator to get $\tilde{S}_{T_2}$ and $\tilde{S}_{T^*}$. We repeated this process 50 times for each scenario, so that in total we had 1000 (individuals) $\times$ 50 (simulations) $\times$ 4 (scenarios). We then used each deconvolution model type described in section 2.4 for all of the scenarios, such that, for example, the Exponential-Exponential model is correctly specified for the first scenario, but is misspecified for the other scenarios. Likewise, the Weibull-Weibull model is correctly specified for the third scenario and the first scenario (because the Exponential-Exponential model is a special case of the Weibull-Weibull model), but misspecified for the second and fourth scenarios which include correlation. We expect that the Weibull-Weibull-Gamma model will be the most robust to misspecification across scenarios.

3.2 Results

Table 3.2.1 presents the results of our simulation study. We present the relative bias of the median estimate of T_1, T_2, and T^* for each of the 4 scenarios using each of the four deconvolution methods. We highlight the relative bias for the median estimate of survival for T_1, as this is the primary quantity of interest from the deconvolution.

The independent exponential deconvolution is described in section 2.4.1. The relative bias in the median survival for T_1 was small (0.98%) when the model was correctly specified,

but at least -18% for the three misspecification scenarios. The independent Weibull deconvolution, described in section 2.4.3, performed similarly under misspecification: the relative bias for T_1 was -1.21% under correct specification, but at least -9% under misspecification. Under the misspecification scenarios presented here, these independent deconvolution models consistently *underestimate* the median survival time for T_1.

The correlated exponential deconvolution is described in section 2.4.2. It had small bias in T_1 (0.77%) and modest bias under both the independent exponential and Weibull simulations (-28.36% and -24.32%). The bias was about twice as large for the correlated Weibull simulation (-42.76%). The correlated Weibull deconvolution had the lowest bias across simulations, but interestingly had the highest relative bias for T_1 when the model was correctly specified compared to other deconvolution models under correct model specification (-2.84%). This could be due to noise: we only did 50 simulations and the correlated Weibull deconvolution has the most parameters to estimate and thus likely the greatest variance in its estimates. Finally, the correlated Weibull deconvolution had very low bias for the independent Weibull simulation (-0.80%) and the correlated exponential simulation (0.83%), but higher bias for the independent exponential simulation (27.58%).

3.2.1 Discussion

Overall, when the models were correctly specified, the deconvolution method had low relative bias for the median survival for T_1. The degree of bias in the presence of misspecification varied by deconvolution model type. Typically, the deconvolution underestimated the time to the intermediate endpoint (T_1). Generally, the amount of bias in T_1 was directly related to large biases for T_2 and T^*. Thus, if we are able to more accurately capture the true T_2 and T^*, we might expect less bias in T_1.

There are several limitations of this simulation study, including that we fixed the sample size for each simulation at 1000 individuals, and only performed 50 simulations. Additionally, we chose only one set of fixed parameters for λ, μ, p, q, and β. This strategy resulted in different median survival times when these fixed parameters were used with different

simulation specifications (e.g. Weibull versus exponential). Future work should explore the biases of these four deconvolution models with different parameter values for the simulated data. An additional limitation is that we did not simulate data with a cure fraction, though we use cure fractions in our application to melanoma (section 5.3). Finally, we only simulated three types of misspecification for each of the deconvolution models. The data that we encounter likely does not perfectly match any of these simulation specifications. It would be illuminating to simulate data under more scenarios to see how the deconvolution models perform under a variety of misspecifications.

Table 3.2: Simulation study results of the relative bias for the median estimate with four deconvolution model choices each with four simulation specifications.

Distribution	Correlation	Relative Bias T_1	Relative Bias T_2	Relative Bias T^*
Independent Exponential Deconvolution				
Exponential	No	0.98%	-0.29%	0.28%
Weibull	No	-26.93%	-19.03%	-6.85%
Exponential	Yes	-52.02%	82.71%	13.52%
Weibull	Yes	-18.60%	-22.10%	-4.23%
Independent Weibull Deconvolution				
Exponential	No	-9.22%	1.80%	-1.58%
Weibull	No	-1.21%	-0.23%	-0.18%
Exponential	Yes	-62.56%	3.09%	4.95%
Weibull	Yes	-26.37%	4.10%	-2.34%
Correlated Exponential Deconvolution				
Exponential	No	-28.36%	-39.81%	-34.84%
Weibull	No	-24.32%	-14.61%	-2.52%
Exponential	Yes	0.77%	-0.76%	-0.09%
Weibull	Yes	-42.76%	-46.11%	-33.39%
Correlated Weibull Deconvolution				
Exponential	No	27.58%	-1.41%	-5.43%
Weibull	No	-0.80%	-3.19%	-2.70%
Exponential	Yes	0.83%	-0.44%	0.00%
Weibull	Yes	-2.84%	-3.14%	-2.65%

We ran the four deconvolution models described in section 2.4 on 50 simulated data sets for each specification in 3.1, and each dataset had a sample size of 1000.

Chapter 4

VALIDATION STUDY USING SPCG4 CLINICAL TRIAL DATA

4.1 Introduction

In an ideal setting, deconvolution would not be necessary because we would have a well-defined cohort followed over time, where each of the intermediate events along an individual's natural history of cancer is noted. In that case, we would have Kaplan-Meier estimates for the time from diagnosis to intermediate endpoint (T_1), time from intermediate endpoint to death (T_2), and their sum T^*. We often do not have these ideal data sets, which creates the need for the deconvolution method. However, when we do have these types of data sets, they provide an opportunity to validate the deconvolution on real data rather than simulated data. In this section, we perform a validation study on the Scandanavian Prostate Cancer Group Study Number 4 (SPCG-4) clinical trial.

4.2 Data Source

From 1989 - 1999, 695 men in Scandanavia were enrolled in a clinical trial to investigate if there was any benefit on metastasis and long-term survival from a radical prostatectomy (RP) versus the standard of care of watchful waiting (WW).[14] All men were initially diagnosed with localized disease and were followed up over time, with intermediate events like metastasis being noted in the study data collection.[14].

In table 4.1, we display descriptive statistics for the participants stratified by whether or not they metastasized. About a third of participants (227 out of 695) went on to metastasize during the follow-up period. The participants that did and did not metastasize were similar in age at diagnosis (63 years and 65 years, respectively). Participants that metastasized had

higher baseline prostate-specific antigen (PSA) (15.8 compared to 11.50) and biopsy Gleason score at diagnosis (6.3 compared to 5.7). About 7% of the participants that did not metastasize were detected by screening for prostate cancer, compared to only 2% of the participants that did metastasize. Over 55% percent of those that did not metastasize had received a radical prostatectomy, whereas 39% of those that did metastasize had received a radical prostatectomy. None of the participants that did not metastasize died from prostate cancer, but 50% died from another cause and the other 50% were lost to follow-up. Alternatively 71% of those that metastasized died from prostate cancer, and 21% died from another cause. This differential implies that there may be a competing risk problem where participants who have a longer time to metastasis, but that ultimately would have died from prostate cancer, are dying from another cause before we can observe their metastasis.

4.3 Methods

We first analyze these data for all men enrolled, regardless of treatment assignment to radical prostatectomy versus the standard of care. We estimate Kaplan-Meier curves for the time from diagnosis with localized disease to metastasis (T_1), time from metastasis to death due to prostate cancer (T_2), and time from diagnosis with localized diseases to death due to prostate cancer (T^*). We define T_1 as the difference between the age at metastasis and the age at diagnosis of localized disease, T_2 as the difference between the age at death and the age at metastasis, and T^* as the difference between the age at death and the age at diagnosis of localized disease.

At metastasis, we considered participants as having the event of interest if they had metastatic cancer. We censored their T_1 if they died or were lost to follow-up before their cancer metastasized. At death, we considered participants as having the event of interest if they died due to prostate cancer. Those that were already censored at metastasis were censored with a follow-up time of zero from metastasis to death, those that were censored after metastasis were censored at their follow-up time, and those that died from some cause other than prostate cancer were censored at their event time. As previously mentioned in

Table 4.1: SPCG-4 Clinical Trial Descriptive Statistics by Metastases Status

	Did not Metastasize (N=468)	Metastasized (N=227)
Patient Age at Diagnosis	65.0 (5.0)	63.7 (5.2)
AJCC TNM Staging		
T1	131 (28.0%)	33 (14.5%)
T2	335 (71.6%)	194 (85.5%)
Unknown	2 (0.4%)	0 (0.0%)
Baseline PSA	11.50 (9.3)	15.8 (12.2)
Biopsy Gleason Score	5.7 (1.1)	6.3 (1.2)
Screening Detected	32 (6.8%)	4 (1.8%)
Randomized to Treatment	258 (55.1%)	89 (39.2%)
Status at Metastasis		
Metastasized	0 (0.0%)	227 (100.0%)
Dead from Other Cause	237 (50.6%)	0 (0.0%)
Lost to Follow-up	231 (49.4%)	0 (0.0%)
Status at Death		
Dead from Prostate Cancer	0 (0.0%)	162 (71.4%)
Dead from Other Cause	237 (50.6%)	48 (21.1%)
Lost to Follow-Up	231 (49.4%)	17 (7.5%)

section 4.2, these participants have competing risks. The Kaplan-Meier estimator does not account for competing risks, and we do not account for competing risks in the deconvolution framework. We discuss the implications of this in section 6.1.

Among the model specifications that we simulated and the deconvolution models that we tried, the simulation results (table 3.2.1) showed that the Weibull-Weibull-Gamma decon-

volution model was most robust to misspecification in terms of relative bias in the median for T_1. As such, we use the Weibull-Weibull-Gamma deconvolution model on Kaplan-Meier estimates for T_2 and T^*, and do not include a cure model. We independently sample the data for T_2 and T^* 500 times to provide bootstrap estimates of the deconvoluted fit. We independently sample T_2 and T^* rather than sampling the same individuals because when we perform the deconvolution on data from separate populations (as is whenever we apply it to non-clinical trial populations) the sampling must necessarily be done independently. We present 90%, 80%, and 70% confidence intervals for each survival curve for T_1, T_2, and T^*.

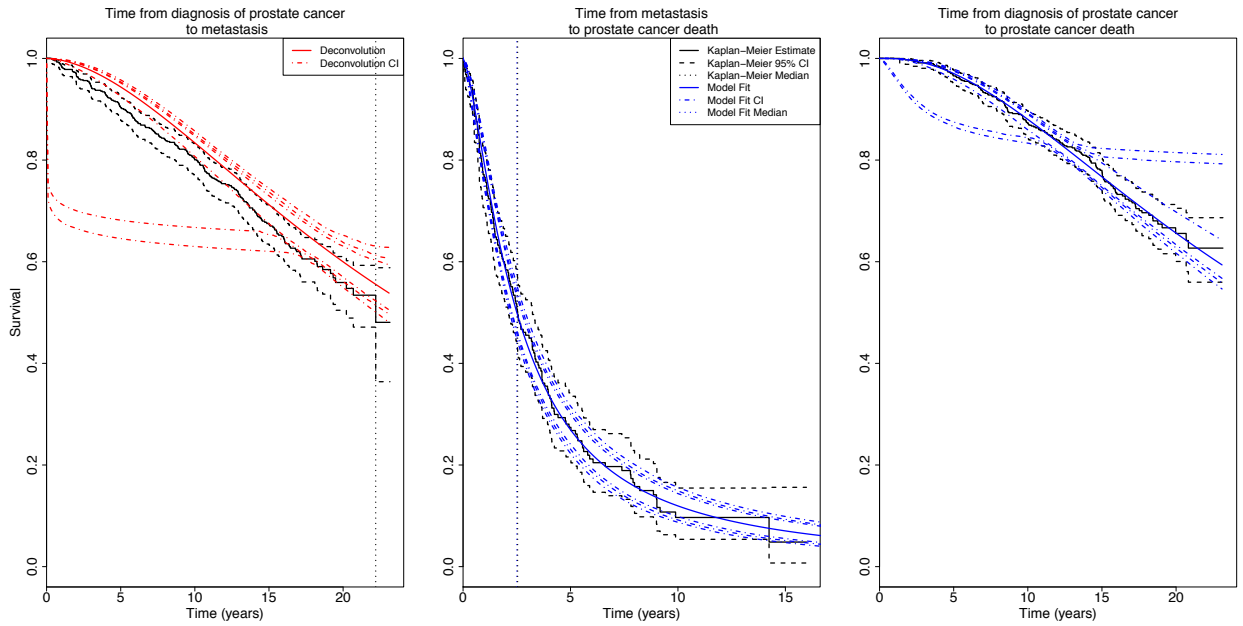
In the SPCG-4 trial, the participants were randomized to either radical prostatectomy (RP) or watchful waiting (WW). We additionally performed validation of the deconvolution method on the subgroups defined by their randomization status. We chose not to fit a cure model for the WW group or the RP group because there was not sufficient follow-up time to be able to estimate a cure fraction (as is visible from the Kaplan-Meier estimates in figures 4.2 and 4.3).

4.4 Results

We present the results of the deconvolution for T_1 and its comparison to the observed KM estimates for T_1 in the left panel of figures 4.1, 4.2, and 4.3. We also include the model fit for T_2 and T^* to evaluate how well the deconvolution model was able to fit to their KM estimates.

From figure 4.1, it is noteworthy that the deconvoluted estimate for T_1 is slightly more conservative than the Kaplan-Meier estimate, though the mean estimate for the deconvoluted T_1 survival curve falls approximately along the upper 95% confidence interval for the Kaplan-Meier estimate of the time from diagnosis of prostate cancer to metastatic recurrence. This is a promising validation result, especially when considering that the confidence intervals for the deconvoluted estimate and the Kaplan-Meier estimate always overlap, and they overlap considerably in the later time points. The 80% and 90% confidence intervals for T_1 are relatively unstable, especially the lower confidence interval. The instability in the confidence

Figure 4.1: SPCG-4 Kaplan-Meier estimates of time from diagnosis to recurrence, time from recurrence to death, and time from diagnosis to death, and the model fit and predictions for the Weibull-Weibull-Gamma deconvolution



interval is caused by a handful of the bootstrap replicates, and is also seen in a handful of simulations shown in B.4. Future work may explore different optimization routines, as this result could be due to the optimizer getting stuck in a local minimum.

The mean deconvolution estimate closely tracks the Kaplan-Meier estimate for T_2, the time from metastatic recurrence to prostate cancer death, and T^*, the time from diagnosis of prostate cancer to prostate cancer death. At 90%, 80%, and 70% confidence levels, the bootstrap confidence intervals for T_2 are stable. However, for T^*, we again see instability in the bootstrap confidence interval above about 70% confidence level, similar to the confidence intervals for T_1.

The deconvolution method performed well for both the WW and the RP subgroups within the SPCG-4 trial. It is clear that the WW group had poorer survival overall, but also poorer

metastasis-free survival than the RP group. From the Kaplan-Meier estimates in the RP group, we only observe the median time from metastasis to prostate cancer death, which is approximately 3.7 years. The deconvolution method produces a median time from diagnosis to metastasis of 31.0 years, metastasis to prostate cancer death of approximately 3.3 years, for an overall time from diagnosis to death of 35.0 years. The deconvoluted estimates for the RP group closely track the Kaplan-Meier estimates for T_2 and T^*, and fall mostly within the 95% confidence interval of the Kaplan-Meier estimates for T_1. For the WW group, we observe the Kaplan-Meier estimates for the time from diagnosis to metastasis and the time from metastasis to death of 18.6 and 2.1 years, respectively. The deconvolution method produces a median time from diagnosis to metastasis of 21.1 years, metastasis to prostate cancer death of approximately 2.1 years, for an overall time from diagnosis to death of 23.8 years. All of these time to event medians are notably lower than those in the RP group, indicating that the deconvolution method is able to pick up on those subgroup differences when they are separately analyzed.

Figure 4.2: SPCG-4 KM and deconvolution estimates for the WW group

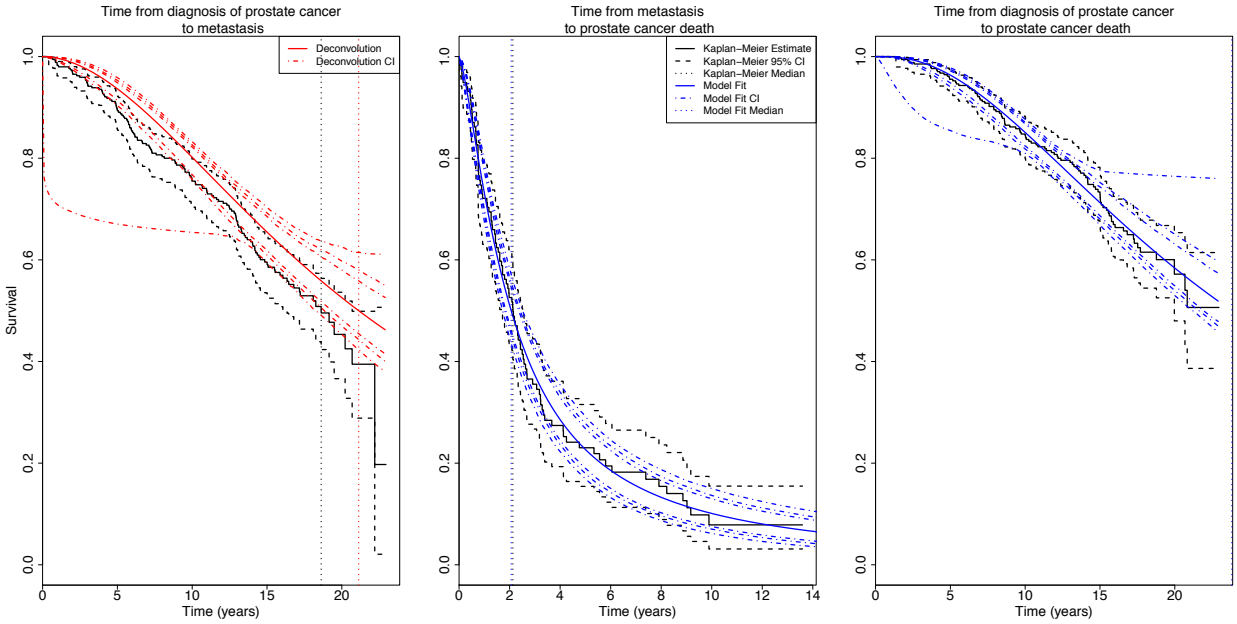
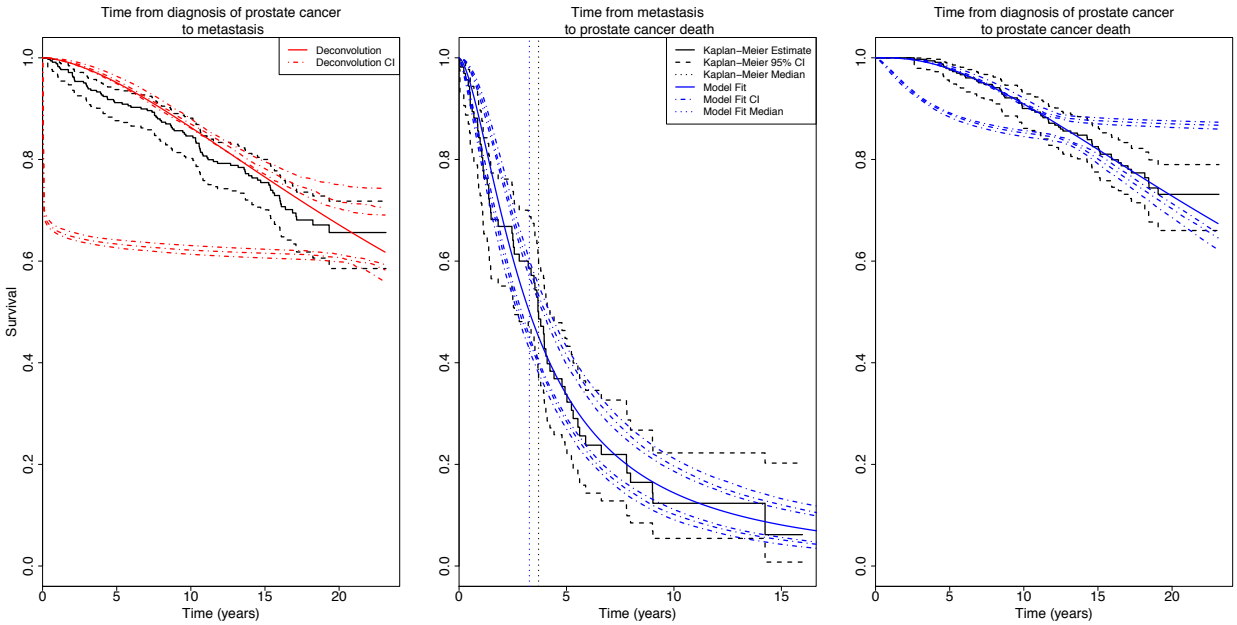


Figure 4.3: SPCG-4 KM and deconvolution estimates for the RR group



Chapter 5

APPLICATION TO MELANOMA DATA FROM SEER

5.1 Data Extraction

The Surveillance, Epidemiology, and End Results Program (SEER) is a national registry in the United States that tracks cancer diagnoses at various stages and links them with death records for the same individuals, noting whether or not they had a cause-specific death of interest (e.g. melanoma) or died of another cause. The SEER*Stat program allows us to extract individual-level data for melanoma diagnoses for individuals diagnosed with localized disease and for individuals diagnosed with distant disease.

We used the SEER 9 Regs Research Data, November 2018 Sub (1975-2016) [Katrina/Rita Population Adjustment][3]. The cases that we extracted all had disease that exhibited malignant behavior, the patients had known age, were diagnosed between 1985 - 2000, and had a maximum follow-up time of 180 months, or 15 years. We chose 2000 as the cutoff date so that we have a full 15 years of potential follow-up time for cases regardless of when they were diagnosed. SEER only captures initial diagnoses, and *not* recurrent cancers. Cases that had a missing or unknown cause of death were excluded. We used the SEER*Stat Site-Recode ICD-O-3/WHO 2008 Definition to extract only cases with melanoma.[1] To maximize the number of years of follow-up available and the comparability across years, we used the SEER historic stage A (1975 - 2015)[2], since follow-up for the cases that we extracted spans across 30 years from 1985 - 2015. The historic stage categories available were in situ, localized, regional, distant, un-staged, and missing. We grouped the cases by their SEER historic

[1]More information about the SEER Site-Recode ICD-O-3/WHO 2008 Definitions can be found at the SEER Website: `https://seer.cancer.gov/siterecode/icdo3_dwhoheme/index.html`

[2]Converted from Collaborative Stage (CS) for 2004+ and Extent of Disease (EOD) prior to that. For more information, see `https://seer.cancer.gov/seerstat/variables/seer/lrd-stage.`

stage value, and extracted separately those who were diagnosed with localized disease and those who were diagnosed with distant disease.

We extract data for T_2 on individuals that are initially diagnosed with distant melanoma, but as our goal is to estimate recurrence-free survival, we would ideally have T_2 that reflects those who are diagnosed with recurrent disease. This method is equivalent to how Mariotto and colleagues estimated the risk of recurrence of breast cancer[2]. The assumption underlying our methodology is that the hazard of death for individuals diagnosed initially with distant metastatic melanoma is the same as the hazard of death for individuals who have had a distant metastatic recurrence of melanoma after initially being diagnosed with localized disease. In the breast cancer application, Mariotto and colleagues used a hazard ratio from a clinical trial to adjust the hazard of death from those with "de novo" distant metastatic breast cancer to those with recurrent metastatic breast cancer, and then used the adjusted survival in the deconvolution[2]. In our application to melanoma, we do not apply a hazard ratio to make this adjustment. However the methodology to do so would be equivalent to that implemented by Mariotto and colleagues, and is an area for future work discussed in section 6.1.

5.2 Descriptive Statistics

Unlike in the SPCG-4 validation data set, the individuals diagnosed with localized and distant melanoma are not the same people tracked over time – which is the primary motivation for doing the population-level deconvolution. As such, we want to ensure that there are not significant demographic differences between individuals diagnosed with localized disease versus individuals diagnosed with distant disease. In Table 5.2, we present descriptive statistics comparing those diagnosed with localized disease to those diagnosed with regional disease on age of diagnosis, year of diagnosis (1985 - 1989, 1990 - 1994, and 1995 - 2000), sex, and race. We present the proportion of the cases that died, and the average number of years of follow-up.

There were many more individuals diagnosed with localized melanoma (N=41,106) than

distant melanoma (N=1,657). Individuals diagnosed with distant melanoma were slightly older on average than those with localized melanoma (59.4 years v. 52.7 years). The distributions of year of diagnosis (1985 - 1989, 1990 - 1994, and 1995 - 2000) and of race (defined as "white", "black", "other", or "unknown") were similar for individuals diagnosed with localized versus distant melanoma. A noteworthy difference is that nearly 50% of those diagnosed with localized melanoma were female, but only 33% of those diagnosed with distant melanoma were female. The mean follow-up time for those diagnosed with localized melanoma was 12.4 years with 9.9% dying from melanoma during the follow-up time, compared to a mean follow-up time of 2.34 years for those diagnosed with distant melanoma, of whom 83% died from melanoma during follow-up time.

5.3 Methods

For individuals diagnosed with localized disease and distant disease, we separately calculated their cause-specific survival using the Kaplan-Meier estimator for 180 months (15 years) at 1-month intervals. If individuals died due to a known other cause, or were alive at the end of follow-up, they were treated as censored. The Kaplan-Meier estimator does not account for competing risks, which is a limitation of our deconvolution method that we discuss further in section 6.1.

We first fit a cure model to the time from diagnosis of localized melanoma to death from distant melanoma, with a generalized Gamma distribution (details in section 2.12). We then used the survival for the non-cured population as T^* in the deconvolution and the time from distant melanoma diagnosis to death as T_2. We fit the deconvolution using the Weibull-Weibull-Gamma model described in section 2.4.4. The deconvolution estimate for T_1 is the time from diagnosis of localized melanoma to distant melanoma *for those that are un-cured*. In order to back-calculate the time from diagnosis of localized melanoma to distant melanoma for the mixed population that includes cured and un-cured individuals, we used equation 2.14. We used the bootstrap method described in section 2.3.2 with 500 replicates to calculate a 90% confidence interval. We present the median survival time where we have

Table 5.1: Descriptive Characteristics by Status at Diagnosis for Melanoma Cases in SEER Diagnosed between 1985 - 2000

	Localized Melanoma (N=40,106)	Distant Melanoma (N=1,657)
Age at Diagnosis	52.7 (16.9)	59.4 (16.5)
Year of Diagnosis		
1985 - 1989	10196 (25.4%)	436 (26.3%)
1990 - 1994	11667 (29.1%)	516 (31.1%)
1995 - 2000	18243 (45.5%)	705 (42.5%)
Female Sex	18998 (47.4%)	561 (33.9%)
Race		
White	38911 (97.0%)	1588 (95.8%)
Black	157 (0.4%)	34 (2.1%)
Other	294 (0.7%)	34 (2.1%)
Unknown	744 (1.9%)	1 (0.1%)
Dead	3961 (9.9%)	1383 (83.5%)
Follow-up Time (years)	12.4 (4.4)	2.34 (4.06)

sufficient follow-up to observe the median.

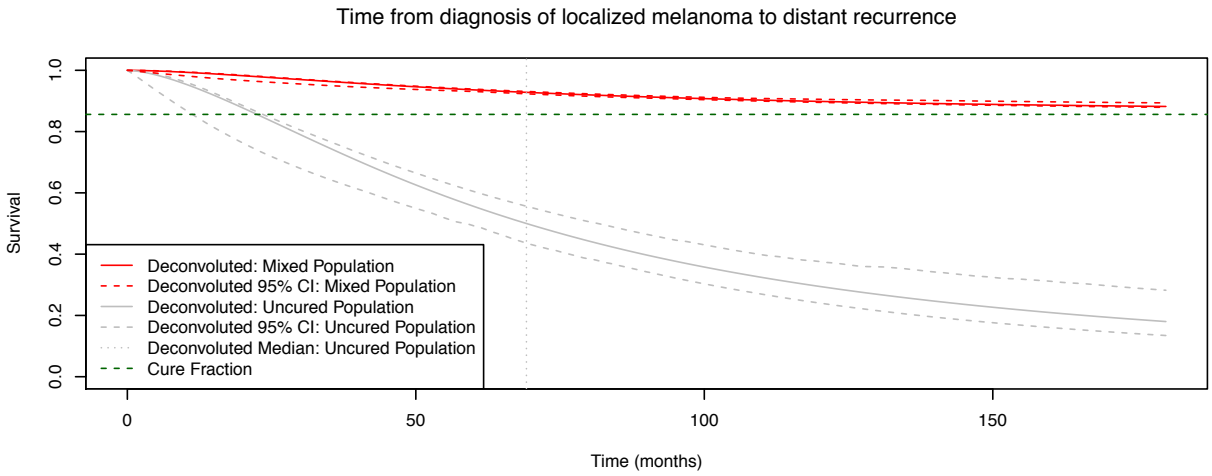
5.4 Results

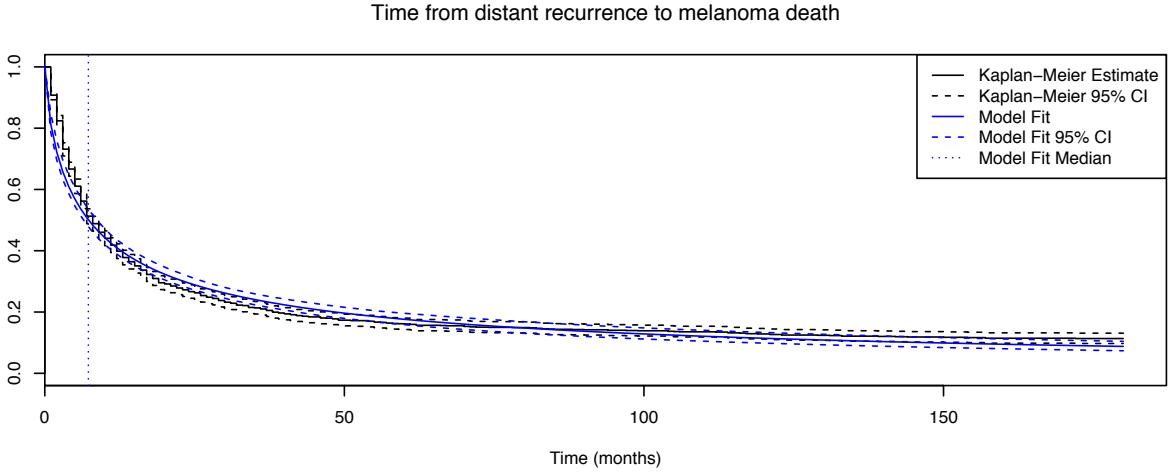
The results of the deconvolution are shown in figure 5.1. The deconvolution estimate of time from diagnosis of localized melanoma to distant recurrence for the mixed population of cured and un-cured individuals is shown in red in the first panel, and for the un-cured individuals is shown in grey in the first panel.

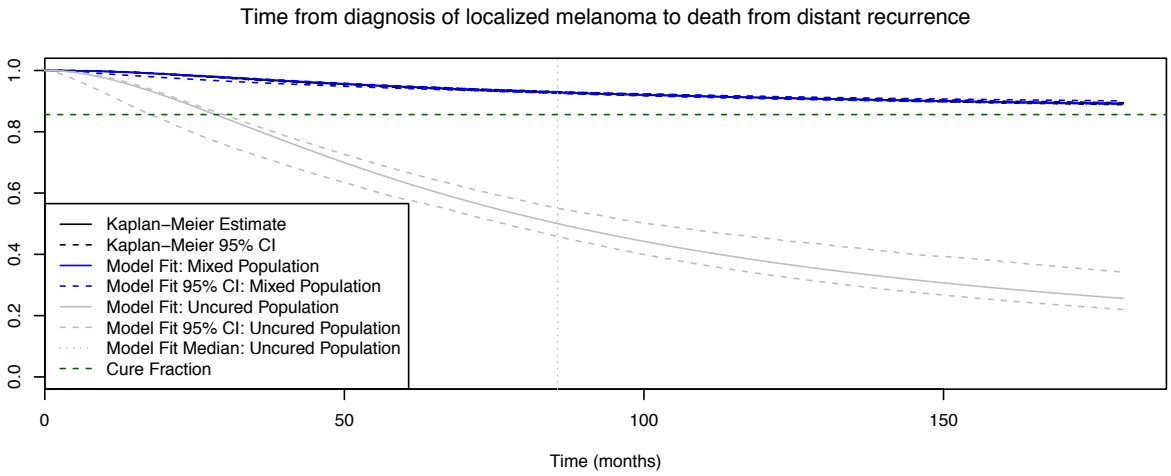
Based on the mixture cure model, we estimate that about 85% of individuals diagnosed with localized melanoma are cured (shown in green in the top and bottom plot in figure 5.1). We estimate that for the *uncured* individuals, the median time from diagnosis of localized melanoma to distant recurrence is 69.2 months, or 5.8 years, i.e. recurrence-free survival for un-cured individuals is 50% at approximately 70 months. In contrast, for the mixed population of cured and un-cured individuals, recurrence-free survival at 70 months is approximately 93.9%. The estimated median time to death for the un-cured population that is diagnosed with localized melanoma is 85.6 months, or 7.1 years. The estimated median time to death for individuals diagnosed with distant melanoma is 7.3 months.

As discussed in section 5.1, we base this deconvolution off of the assumption that the hazard of death for individuals diagnosed initially with distant metastatic melanoma is the same as the hazard of death for individuals who have had a distant metastatic recurrence of melanoma after initially being diagnosed with localized disease. If we believe that individuals diagnosed with recurrent disease have a greater hazard of death than those diagnosed initially with distant disease, then the survival for T_2 is conservative, i.e. the time from distant recurrence to melanoma death is shorter than what we observe in 5.1. Since the sum of T_1 and T_2 has to stay the same, the deconvoluted estimate that we present in 5.1 would then be too pessimistic, i.e. the median time to recurrence is longer than we estimate. Alternatively, if we believe that individuals diagnosed with recurrent disease have a lesser hazard of death than those diagnosed initially with distant disease, then the deconvoluted estimate that we present is too optimistic, i.e. the median time to recurrence is shorter than we estimate.

Figure 5.1: SEER Melanoma KM estimates and deconvolution estimates using the Weibull-Weibull-Gamma deconvolution with a cure fraction







Chapter 6

DISCUSSION AND CONCLUSIONS

6.1 Discussion

In this thesis, we used deconvolution methods to parse total time-to-death data into two parts: time from initial diagnosis to an intermediate endpoint and time from an intermediate endpoint to death. We described the methodological framework for deconvolution, allowing for individuals to be cured before reaching the intermediate endpoint, and allowing for correlation between the time from initial diagnosis to the intermediate endpoint and the time from the intermediate endpoint to death. We implemented the framework in four deconvolution models where the two time to event components were simulated as independent exponential random variables, correlated exponential random variables, independent Weibull random variables, and correlated Weibull random variables.

We then used our deconvolution method in three main applications. First, we performed a simulation study where we simulated individuals from four scenarios with known distributions of T_1, T_2, and T^*, and then used each of the four deconvolution models. The simulation study allowed us to characterize the performance of the models when the model is correctly specified and misspecified. Although we reported the bias of the deconvolution models in estimating T_1, T_2, and T^*, we primarily focused on the bias in T_1, the deconvoluted time-to-event outcome. We then validated the deconvolution method using the correlated Weibull model on data collected in a clinical trial on prostate cancer metastasis and survival for men in Scandanavia randomized to radical prostatectomy (RP) versus watchful waiting (WW). For this validation, we had the "gold-standard" time from diagnosis to intermediate endpoint to which we compared our deconvoluted estimates. We performed deconvolution on all individuals from the clinical trial together, and then split by treatment subgroup.

Finally, we extracted data on localized and distant melanoma diagnoses in the Surveillance, Epidemiology, and End Results (SEER) program database between 1985 - 2000 and applied the deconvolution method to estimate the time from localized diagnosis to distant metastatic recurrence, accounting for individuals with localized disease that are cured.

The U.S. SEER database is one of several population-based cancer registries that only capture first occurrence of cancer and thus do not have data to track cancer recurrences and metastases, including the Danish Cancer Registry [15] and the Australian Cancer Database [16]. Others have attempted to track recurrence by using algorithms that look for treatment codes that are temporally indicative of having a recurrent cancer, for example that a cancer-related treatment occurs well after treatment for the first cancer.[17, 18] Our method provides an innovative way to utilize the existing registry data to describe the time to progression, metastasis, or recurrence at the population level for any cancer by using survival data from diagnosis, and survival data from a well-defined intermediate endpoint. Our work builds off of the work from Mariotto and colleagues [2] in that allows for correlation between the two time-to-event quantities in the deconvolution by way of a shared frailty term.

There are a number of limitations to the deconvolution method. First, most data sources that could be used in the deconvolution are subject to left-censoring due to detection bias. We are not able to capture all diagnoses with, for example, localized disease, at the time when the individual first has cancer. Instead, their T_1 only captures when we were able to detect their cancer. This issue is further complicated when we consider that individuals diagnosed with late stage disease are probably more likely to be detected than individuals diagnosed with early stage disease. Whether or not individuals are detected at an early or late stage may also be affected by the speed of their cancer progression. When we use data like those from the SPCG-4 data set in the validation, there is the additional issue of interval censoring. Even if we were able to catch individuals when they initially develop early stage cancer, we do not know precisely when they metastasize unless they are continually monitored for metastasis. An additional limitation is that we do not account for competing risks in our estimates of survival, as discussed in the application to both SPCG-4 data (section 4.3) and

melanoma data from SEER (section 5.3). The implication of this is that we are not able to capture all potential events of interest in our analyses because individuals have died due to other causes and are thus censored. Future work should explore incorporating competing risks into the deconvolution framework.

Finally, when we apply the deconvolution method to data collected from SEER, we make the assumption that individuals diagnosed initially with late stage disease have the same hazard of death as individuals who were diagnosed at an early stage and are now at a late stage disease. Likewise, we assume that someone diagnosed initially with metastatic disease has the same hazard of death as someone who has a metastatic recurrence. We discussed the implications of this assumption in our application to estimating melanoma recurrence (section 5.4). In practice, we could account for differences in the hazard of death by applying a hazard ratio (HR) that represents the relative hazard of death that someone initially diagnosed with metastatic disease has compared to someone who has recurrent metastatic disease. This HR could be applied directly to the Kaplan-Meier estimates of survival and then the adjusted data could be used in the deconvolution.

Our deconvolution method, building off of the work from Mariotto and colleagues, has several promising potential applications. For cancers that do not have routine screening implemented in the United States, including liver cancer, stomach cancer, esophageal cancer, and pancreatic cancer, we could use deconvolution to estimate the benefit of early detection and screening. In this case, the T_2 data would be time to death from diagnosis at a later stage, e.g. Stage IV, and the T^* data would be time to death from diagnosis at an early stage, e.g. Stage I. The deconvolution method would estimate T_1, the time from diagnosis of Stage I to progression to Stage IV. Consider a cancer that when detected at Stage IV, patients have a median survival of 1 year. Now consider an early detection program that catches and treats that same cancer at Stage I, resulting in a median survival of 3 years from diagnosis at Stage I. If the median time from diagnosis of Stage I to progression to Stage IV estimated from the deconvolution model is approximately 2 years, then there is no benefit to early detection. Rather, the apparent 2-year survival "benefit" from early screening is just

a lead time effect: the cancer was detected 2 years earlier and treated at Stage I, but that did not improve the overall survival time of 3 years. However, if the median time estimated from the deconvolution model is notably less than 2 years, that suggests that there could be a benefit to early detection. On the other hand, for cancers that are already screened for and treated effectively at an early stage, we could use deconvolution methods to characterize the natural history of that cancer. We could apply deconvolution to a cancer that has a known, effective treatment and remove the treatment effect from T_1 by applying a hazard ratio. The results can then be interpreted as a natural history of that particular cancer in the absence of standard treatment.

Future work should also explore a number of interesting methodological extensions to the deconvolution method. When we use the deconvolution method in the presence of a cured fraction, we have a two-part model. First we estimate the fraction of the population that is cured, and then we use deconvolution on only the non-cured fraction of the population. A modification to the deconvolution method could make this a one-part model by estimating the cure fraction at the same time as doing the deconvolution. In our implementation of the deconvolution, we grouped all individuals together, and allowed for unmeasured heterogeneity within the population to be captured by a shared frailty term between T_1 and T_2. We could do the deconvolution separately by subgroups defined by covariates that we believe might influence T_1 and T_2. This was the strategy used by Mariotto and colleagues.[2] Alternatively, we could estimate the coefficients on covariates for T_1 and T_2 directly, so that we can use covariate information to actually inform the deconvolution. One potential way of doing so is through the hazard function similar to a Cox proportional hazards framework. We already are incorporating unmeasured heterogeneity by including k into the hazard function for the parametric distributions of T_1 and T_2. A simple extension would be to modify

$$k = \mathbf{X}\gamma + \mathbf{X_{T_1}}\gamma\mathbf{T_1} + \mathbf{X_{T_2}}\gamma\mathbf{T_2}$$

where $\mathbf{X}$, $\mathbf{X_{T_1}}$, and $\mathbf{X_{T_2}}$ are vectors of covariates of interest that influence both T_1 and T_2, only T_1, or only T_2, respectively.

6.2 Conclusion

In conclusion, deconvolution parses time-to-event data into the sum of two parts when one of those parts is unmeasured. Current applications of deconvolution described here and elsewhere allow us to estimate metastasis-free and recurrence-free survival at the population-level. Methodological extensions in deconvolution will allow us to more accurately estimate these quantities. Future work will continue to explore the exciting applications of the deconvolution method to cancer epidemiology.

VITA

Marlena Bannick graduated from the University of Washington in 2016 with a Bachelor of Science in Public Health and a minor in Mathematics, and in 2019 with a Master of Science in Biostatistics. While pursuing her MS degree, she worked as a Post-Bachelor Fellow at the Institute for Health Metrics and Evaluation on various endeavors for the Global Burden of Disease Study. In her free time, she likes to read, cook, explore the Pacific Northwest, and spend time with her husband, Jonathan, and her cat, Peaches.

BIBLIOGRAPHY

[1] Lakdawalla DN, Shafrin J, Hou N, Peneva D, Vine S, Park J, et al. Predicting Real-World Effectiveness of Cancer Therapies Using Overall Survival and Progression-Free Survival from Clinical Trials: Empirical Evidence for the ASCO Value Framework. Value in Health. 2017;20(7):866–875.

[2] Mariotto AB, Zou Z, Zhang F, Howlader N, Kurian AW, Etzioni R. Can We Use Survival Data from Cancer Registries to Learn about Disease Recurrence? The Case of Breast Cancer. Cancer epidemiology, biomarkers prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2018;27(11).

[3] National Cancer Institute SRP DCCPS. Surveillance, Epidemiology, and End Results (SEER) Program SEER*Stat Database: Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2017 Sub (2000-2015) ¡Katrina/Rita Population Adjustment¿ - Linked To County Attributes - Total U.S., 1969-2016 Counties; April 2018. Data retrieved from SEER*Stat, `www.seer.cancer.gov`.

[4] Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association. 1958;53(282):457–481. Available from: `https://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501452`.

[5] Amdahl J. flexsurvcure: Flexible Parametric Cure Models; 2017. R package version 0.0.2. Available from: `https://cran.r-project.org/web/packages/flexsurvcure/index.html`.

[6] Molenberghs G, Verbeke G. On the Weibull-Gamma frailty model, its infinite moments, and its connection to generalized log-logistic, logistic, Cauchy, and extreme-value distributions. Journal of Statistical Planning and Inference. 2011;141(2):861 – 868. Available from: `http://www.sciencedirect.com/science/article/pii/S0378375810003782`.

[7] Bannick MS. survival-deconvolution. GitHub; 2019. `https://github.com/mbannick/deconvolution`.

[8] Therneau TM. A Package for Survival Analysis in S; 2015. Version 2.38. Available from: `https://CRAN.R-project.org/package=survival`.

[9] Terry M Therneau, Patricia M Grambsch. Modeling Survival Data: Extending the Cox Model. New York: Springer; 2000.

[10] Ruckdeschel P, Kohl M, Stabla T, Camphausen F. S4 Classes for Distributions. R News. 2006 May;6(2):2–6.

[11] Nash JC, Varadhan R. Unifying Optimization Algorithms to Aid Software System Users: optimx for R. Journal of Statistical Software. 2011;43(9):1–14. Available from: http://www.jstatsoft.org/v43/i09/.

[12] Nash JC. On Best Practice Optimization Methods in R. Journal of Statistical Software. 2014;60(2):1–14. Available from: http://www.jstatsoft.org/v60/i02/.

[13] Gay DM. Usage summary for selected optimization routines. Computing Science Technical Report 153, ATT Bell Laboratories, Murray Hill. 1990;.

[14] Holmberg L, Bill-Axelson A, Helgesen F, Salo JO, Folmerz P, Haggman M, et al. A randomized trial comparing radical prostatectomy with watchful waiting in early prostate cancer. N Engl J Med. 2002 Sep;347(11):781–789.

[15] Gjerstorff ML. The Danish Cancer Registry. Scandinavian Journal of Public Health. 2011;39(7_suppl):42–45. PMID: 21775350. Available from: https://doi.org/10.1177/1403494810393562.

[16] Australian Cancer Database [Dataset]; 2012. Available from: http://www.aihw.gov.au/datacat/index.cfm/action/showall/id/5018,https://researchdata.ands.org.au/australian-cancer-database.

[17] Rasmussen LA, Jensen H, Virgilsen LF, Falborg AZ, M?ller H, Vedsted P. Time from incident primary cancer until recurrence or second primary cancer: Risk factors and impact in general practice. Eur J Cancer Care (Engl). 2019 Jun;p. e13123.

[18] Warren JL, Mariotto A, Melbert D, Schrag D, Doria-Rose P, Penson D, et al. Sensitivity of Medicare Claims to Identify Cancer Recurrence in Elderly Colorectal and Breast Cancer Patients. Med Care. 2016 08;54(8):47–54.

[19] Pinsky PF. Estimation and Prediction for Cancer Screening Models Using Deconvolution and Smoothing. Biometrics. 2001 6;57(2):389–395. Available from: https://doi.org/10.1111/j.0006-341X.2001.00389.x.

[20] Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science (New York, NY). 2018;359(6378).

[21] Yuan P, Cao JL, Rustam A, Zhang C, Yuan XS, Bao FC, et al. Time-to-Progression of NSCLC from Early to Advanced Stages: An Analysis of data from SEER Registry and a Single Institute. Sci Rep. 2016;6(1):28477–28477.

Appendix A

MATHEMATICAL DERIVATIONS

A.1 Parametric Deconvolution with Independence

Assume that $T_1 \sim Exponential(\lambda)$ and $T_2 \sim Exponential(\mu)$ and that $T_1 \perp\!\!\!\perp T_2$. Then,

$$f_{T_1}(t) = \lambda e^{-\lambda t} \tag{A.1}$$

$$f_{T_2}(t) = \mu e^{-\mu t} \tag{A.2}$$

$$f_{T_1,T_2}(t_1, t_2) = \lambda e^{-\lambda t_1} \mu e^{-\mu t_2} \tag{A.3}$$

Now let $T^* = T_1 + T_2$, so $T_1 = T^* - T_2$ and $T_2 = T_2$. The Jacobian of this transformation is 1, and the joint density of T_2 and T^* is given by

$$f_{T^*,T_2}(t^*, t_2) = f_{T_1,T_2}(t^* - t_2, t_2) \tag{A.4}$$

$$= \lambda e^{-\lambda(t^* - t_2)} \mu e^{-\mu t_2} \tag{A.5}$$

We know that for an exponential random variable $T \sim Exponential(\lambda)$, the survival function is given by $S(t) := e^{-\lambda}$. To find the survival function for the convolution of the two exponential random variables, T^*, we need to marginalize f_{T^*,T_2} over T_2, and then integrate f_{T^*}. First, we find f_{T^*}:

$$f_{T^*}(t^*) = \int_0^{t^*} \lambda e^{-\lambda(t^*-t_2)} \mu e^{-\mu t_2} \, \mathrm{d}t_2 \tag{A.6}$$

$$= \lambda\mu e^{-\lambda t^*} \int_0^{t^*} e^{-t_2(\mu-\lambda)} \, \mathrm{d}t_2 \tag{A.7}$$

$$= \frac{\lambda\mu}{\lambda-\mu} e^{-\lambda t^*} e^{t_2(\mu-\lambda)} \Big|_0^{t^*} \tag{A.8}$$

$$= \frac{\lambda\mu}{\lambda-\mu} e^{-\lambda t^*} \left(e^{-t^*(\mu-\lambda)} - 1\right) \tag{A.9}$$

$$= \frac{\lambda\mu}{\lambda-\mu} \left(e^{-\mu t^*} - e^{-\lambda t^*}\right) \tag{A.10}$$

$$\tag{A.11}$$

Next, we integrate to get the survival function, $S_{T^*}(t)$:

$$S_{T^*}(t) = \int_t^{\infty} f_{T^*}(t^*) \, \mathrm{d}t^* \tag{A.12}$$

$$= \int_t^{\infty} \frac{\lambda\mu}{\lambda-\mu} \left(e^{-\mu t^*} - e^{-\lambda t^*}\right) \mathrm{d}t^* \tag{A.13}$$

$$= \frac{\lambda\mu}{\lambda-\mu} \left(\frac{e^{-\lambda t^*}}{\lambda} - \frac{e^{-\mu t^*}}{\mu}\right) \Big|_t^{\infty} \tag{A.14}$$

$$= \frac{\lambda\mu}{\lambda-\mu} \left(\frac{e^{-\mu t^*}}{\mu} - \frac{e^{-\lambda t^*}}{\lambda}\right) \tag{A.15}$$

$$= \frac{\lambda e^{-\mu t} - \mu e^{-\lambda t}}{\lambda-\mu} \tag{A.16}$$

$$\tag{A.17}$$

A.2 Parametric Deconvolution with Correlation

A.2.1 Marginal Distribution of an Exponential Random Variable with Gamma Frailty

For an exponential random variable, $T \sim Exponential(k\lambda)$ with $K \sim Gamma(\alpha, \beta)$, the marginal density function for T is given by:

$$f_T(t) := \int_0^\infty f_{T|K}(t|k) \cdot f_K(k)\,\mathrm{d}k \tag{A.18}$$

$$= \int_0^\infty \lambda k e^{-\lambda k t} k^{\alpha-1} e^{-k\beta} \frac{\beta^\alpha}{\Gamma(\alpha)}\,\mathrm{d}k \tag{A.19}$$

$$= \frac{\lambda\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty k^\alpha e^{-k(t\lambda+\beta)}\,\mathrm{d}k \tag{A.20}$$

$$= \frac{\lambda\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{(t\lambda+\beta)^{(\alpha+1)}} \qquad \text{recognizing Gamma}(\alpha+1, x\lambda+\beta) \text{ kernel}$$

$$= \frac{\lambda\beta^\alpha\alpha}{(t\lambda+\beta)^{(\alpha+1)}} \tag{A.21}$$

To find the survival function, we then integrate the density from t to ∞:

$$S_T(t) := \int_t^\infty \lambda\beta^\alpha\alpha(x\lambda+\beta)^{-(\alpha+1)}\,\mathrm{d}x \tag{A.22}$$

$$= \lambda\beta^\alpha\alpha \frac{\lambda x + \beta_{-\alpha}}{-\alpha\lambda}\bigg|_t^\infty \tag{A.23}$$

$$= -\left(\frac{\beta}{\lambda x+\beta}\right)^\alpha\bigg|_t^\infty \tag{A.24}$$

$$= \left(\frac{\beta}{\lambda t+\beta}\right)^\alpha - \lim_{x\to\infty}\left(\frac{\beta}{\lambda x+\beta}\right)^\alpha \qquad \text{using improper integration}$$

$$= \left(\frac{\beta}{\lambda t+\beta}\right)^\alpha \tag{A.25}$$

A.2.2 Marginal Distribution of the Sum of Exponential Random Variables with a Shared Gamma Frailty

Now consider the scenario outlined in A.1, but where T_1 and T_2 have the shared frailty term, K. The survival functions of T_1 and T_2 are given by A.22, but we need to find the survival

function for their convolution, T^*. We know that

$$f_{T_1,T_2|K}(t_1,t_2|k) = \lambda k e^{-\lambda k t_1}\mu k e^{\mu k t_2} \tag{A.26}$$

$$f_{T^*,T_2|K}(t^*,t_2|k) = \lambda k e^{-\lambda k (t^*-t_2)}\mu k e^{\mu k t_2} \quad \text{from A.4} \tag{A.27}$$

and we can marginalize over joint distribution of T^*, T_2, and K by first K and then T_2 to get the density of the convoluted random variable, T^*. First,

$$f_{T^*,T_2,K}(t^*,t_2,k) = \lambda\mu k^2 e^{-\lambda k t^* - k t_2(\mu-\lambda)}\frac{\beta^\alpha}{\Gamma(\alpha)}k^{\alpha-1}e^{-k\beta} \tag{A.28}$$

$$\tag{A.29}$$

Marginalizing over K, we get

$$f_{T^*,T_2} = \frac{\lambda\mu\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty k^{\alpha+1}e^{-k(\lambda t^* + t_2(\mu-\lambda)+\beta)}\,\mathrm{d}k \tag{A.30}$$

$$= \frac{\lambda\mu\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+2)}{(\lambda t^* + t_2(\mu-\lambda)+\beta)^{(\alpha+2)}} \tag{A.31}$$

recognizing the kernel of a Gamma$(\alpha+2, \lambda t^* + t_2(\mu-\lambda)+\beta)$ random variable. Then the marginal distribution of T^* is given by

$$f_{T^*} = \lambda\mu\beta^\alpha\alpha(\alpha+1)\int_0^{t^*}(\lambda t^* + t_2(\mu-\lambda)+\beta)^{-(\alpha+2)}\,\mathrm{d}t_2 \tag{A.32}$$

$$= \lambda\mu\beta^\alpha\alpha(\alpha+1)\frac{(\lambda t^* + t_2(\mu-\lambda)+\beta)^{-(\alpha+1)}}{-(\alpha+1)(\mu-\lambda)}\Bigg|_0^{t^*} \tag{A.33}$$

$$= \frac{\lambda\mu\beta^\alpha\alpha}{\lambda-\mu}\left[(\lambda t^* + t^*(\mu-\lambda)+\beta)^{-(\alpha+1)} - (\lambda t^*+\beta)^{-(\alpha+1)}\right] \tag{A.34}$$

$$= \frac{\lambda\mu\beta^\alpha\alpha}{\lambda-\mu}\left[(\mu t^*+\beta)^{-(\alpha+1)} - (\lambda t^*+\beta)^{-(\alpha+1)}\right] \tag{A.35}$$

$$\tag{A.36}$$

To find the survival function for T^*, we use the fact that $S(t) := 1 - F(t)$:

$$S_{T^*}(t) = 1 - F_{T^*} \tag{A.37}$$

$$= 1 - \int_0^t \frac{\lambda\mu\beta^\alpha\alpha}{\lambda-\mu}\left[(\mu t^* + \beta)^{-(\alpha+1)} - (\lambda t^* + \beta)^{-(\alpha+1)}\right] \mathrm{d}t^* \tag{A.38}$$

$$= 1 - \frac{\lambda\mu\beta^\alpha\alpha}{\lambda-\mu}\left[\frac{-(\mu t^* + \beta)^{-\alpha}}{\mu\alpha} + \frac{-(\lambda t^* + \beta)^{-\alpha}}{\lambda\alpha}\right]\Bigg|_0^t \tag{A.39}$$

$$= 1 - \frac{\lambda\mu\beta^\alpha}{\lambda-\mu}\left[\frac{-(\mu t + \beta)^{-\alpha}}{\mu} + \frac{-(\lambda t + \beta)^{-\alpha}}{\lambda} + \frac{\beta^{-\alpha}}{\mu} - \frac{\beta^{-\alpha}}{\lambda}\right] \tag{A.40}$$

$$= 1 + \frac{\lambda\beta^{-\alpha}(\mu t + \beta)^{-\alpha}}{\lambda-\mu} - \frac{\mu\beta^{-\alpha}(\lambda t + \beta)^{-\alpha}}{\lambda-\mu} - \frac{\lambda}{\lambda-\mu} + \frac{\mu}{\lambda-\mu} \tag{A.41}$$

$$= \frac{\beta^\alpha}{\lambda-\mu}\left[\frac{\lambda}{(\mu t + \beta)^\alpha} - \frac{\mu}{(\lambda t + \beta)^\alpha}\right] \tag{A.42}$$

Appendix B

SUPPLEMENTAL SIMULATION RESULTS

The results presented in Table 3.2.1 summarize the information that is contained in the following four plots. Recall from section 3.1 that we simulated data under four scenarios, and for each of those four scenarios performed deconvolution with each of the model types presented in section 2.4. In Table 3.2.1 we reported the bias across 50 simulated data sets in the deconvolution estimate for the median and compared that with the true median of the survival functions from which the data were simulated.

The plots show in the left column the deconvoluted estimates for T_1 for each simulation compared with the true survival function for T_1 shown in blue. The middle column shows the estimates from the simulations compared with the true survival function for T_2, and the right column shows the estimates from the simulations compared with the true survival function for T^*. Each of the four rows correspond to the four scenarios described in 3.1. Figures B.1, B.2, B.3, and B.4 present the simulation results using the Exponential-Exponential, Exponential-Exponential-Gamma, Weibull-Weibull, and Weibull-Weibull-Gamma deconvolution models, respectively.

Figure B.1: Results for the Exponential-Exponential Model where data have been simulated from 4 scenarios, with 50 simulations each

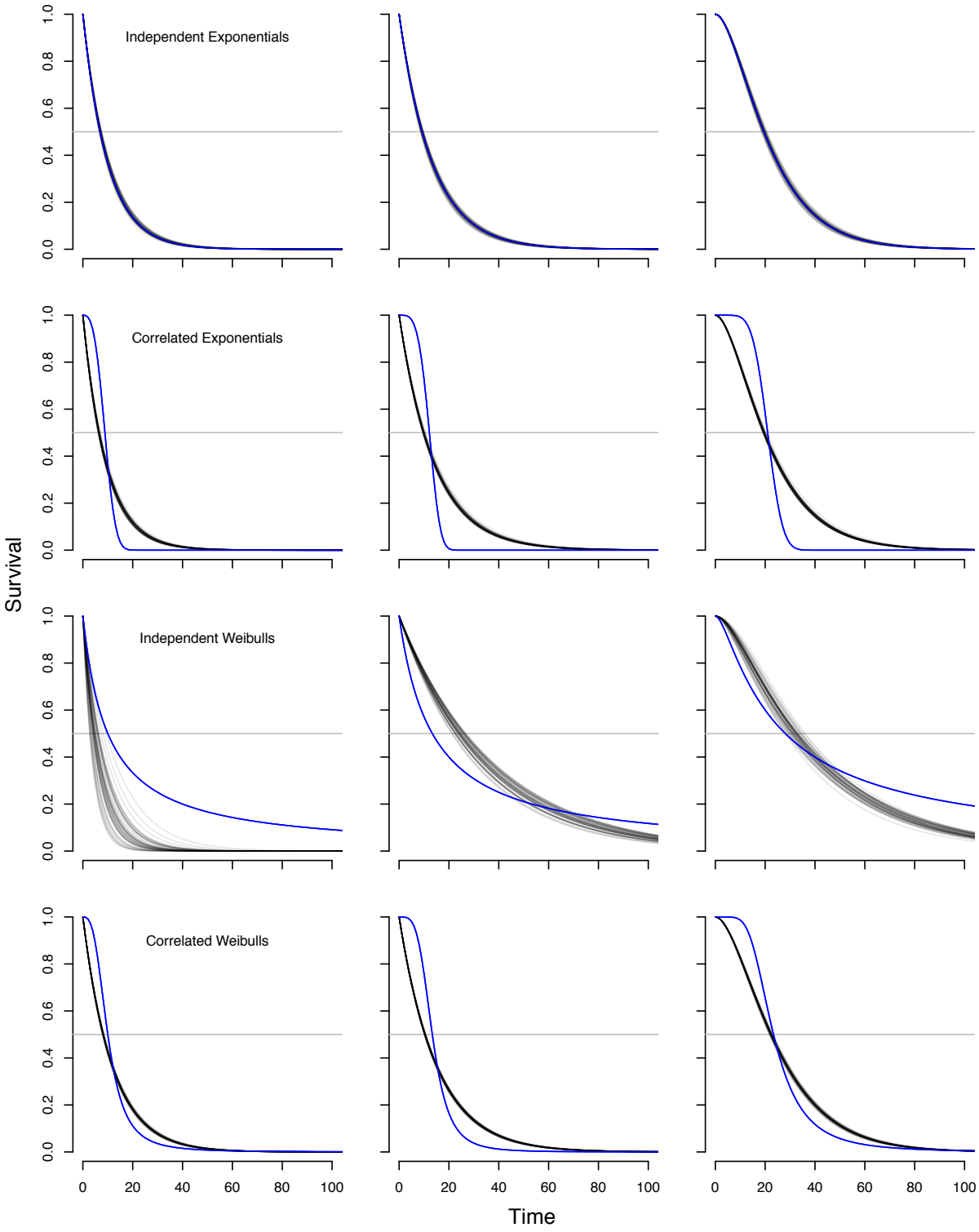


Figure B.2: Results for the Exponential-Exponential-Gamma Model where data have been simulated from 4 scenarios, with 50 simulations each

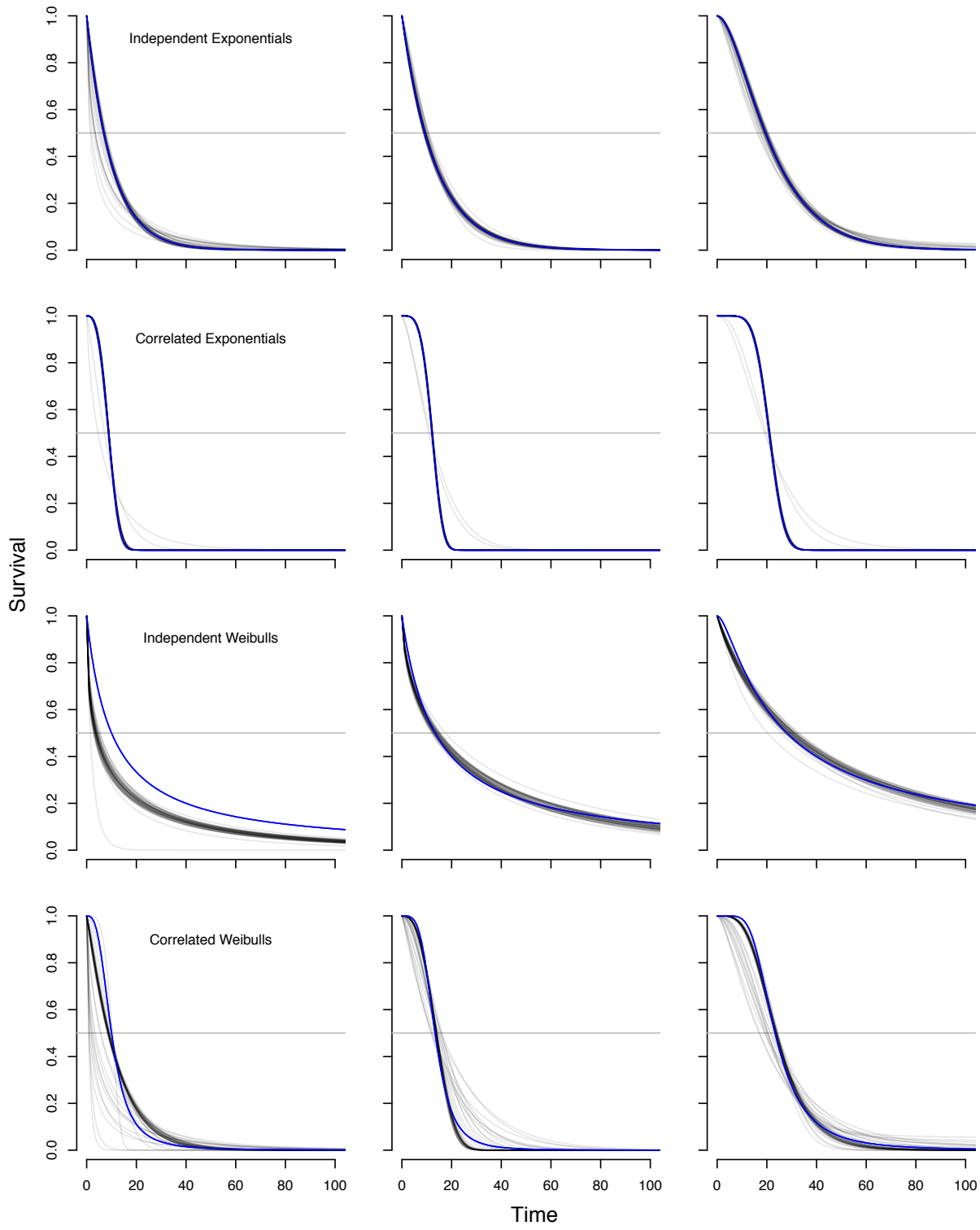


Figure B.3: Results for the Weibull-Weibull Model where data have been simulated from 4 scenarios, with 50 simulations each

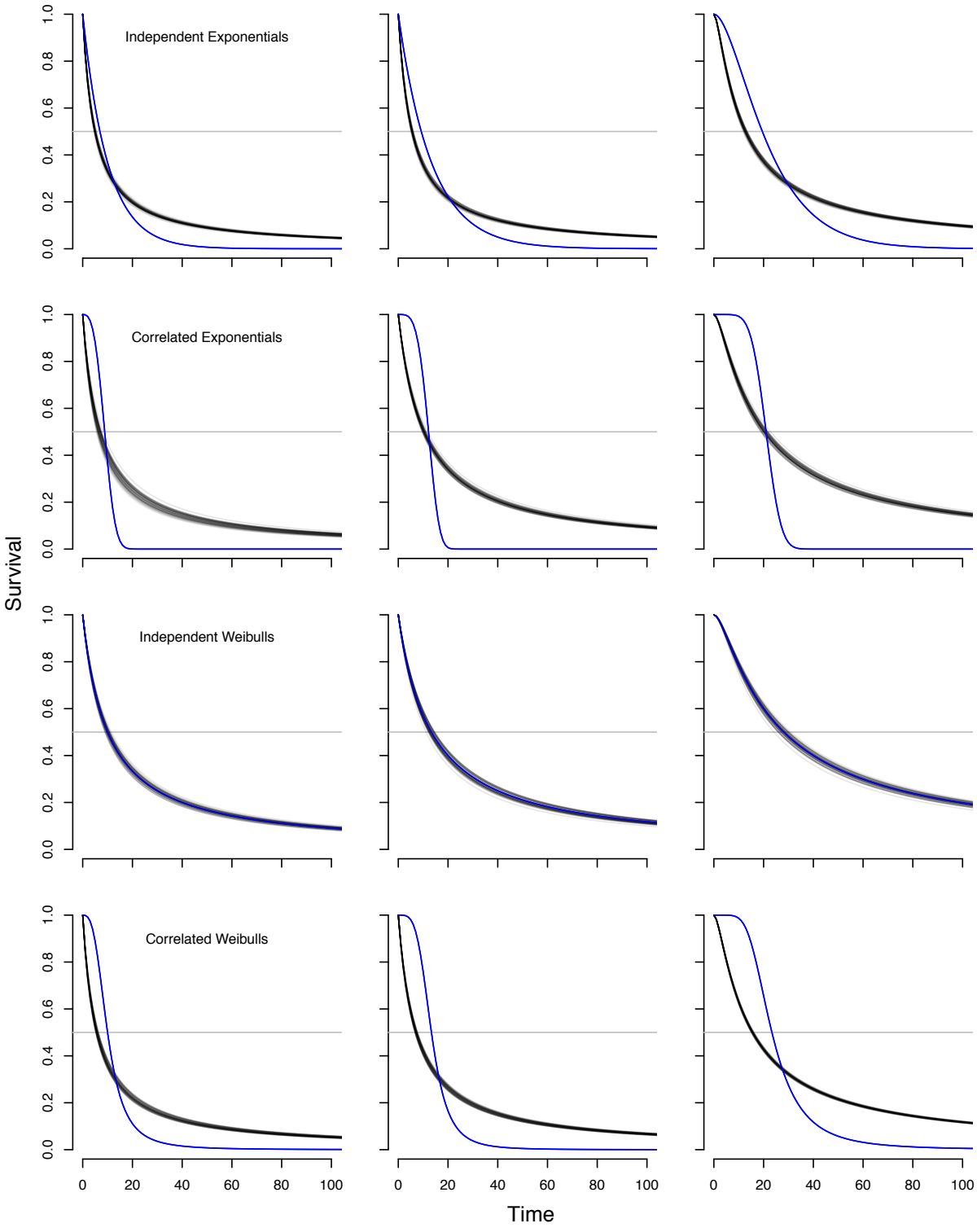


Figure B.4: Results for the Weibull-Weibull-Gamma Model where data have been simulated from 4 scenarios, with 50 simulations each