Summer 2018

Statistical Analysis In Cancer Survivorship

zhiyun Ge
Southern Methodist University, zge@smu.edu

STATICTICAL ANALYSIS IN CANCER SURVIVORSHIP

Approved by:

_____
Dr. Daniel Heitjan
Professor of Biostatistics

_____
Dr. Hon Keung Ng
Professor of Statistics

_____
Dr. Sandi Pruitt
Assistant Professor of Clinical Sciences

_____
Dr. Song Zhang
Associate Professor of Statistics

STATISTICAL ANALYSIS IN CANCER SURVIVORSHIP

A Dissertation Presented to the Graduate Faculty of

Dedman College

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

with a

Major in Biostatistics

by

Zhiyun Ge

B.S., Biochemistry & Molecular Biology, Nanjing University
M.S., Biostatistics, Duke University

August 7, 2018

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Dr. Heitjan. He is such an intelligent person that I remembered once with his suggestion, the computation time of my code was reduced from 20 hours to 20 seconds. He is also nice and patient. He spent a lot of time editing the manuscript that I draft. Without his guidance and encouragement, I couldn't complete my dissertation. I also sincerely thank Dr. Pruitt for her guidance and generous support for my PhD life. She always has great research insight and can always come up with brilliant ideas. I also would like to thank Dr. Ng and Dr. Zhang for their time and help for my dissertation. I want to thank all other teachers and students who helped me through my PhD life. Finally, I would like to thank my family, my husband, You, who is always supporting and encouraging me; my two kids, Grace and Alex, they make me become strong mom; my parents and parents in law, who help me take care of my kids so that I can focus on my dissertation.

Ge, Zhiyun B.S., Biochemistry & Molecular Biology, Nanjing University, 2010
M.S., Biostatistics, Duke University, 2014

Statistical Analysis in Cancer Survivorship

Advisor: Professor Daniel Heitjan
Doctor of Philosophy conferred August 7, 2018
Dissertation completed July 2, 2018

The advance in cancer early detection and cancer treatment have led to the rapid growth in the number of cancer survivors. It is good news that cancer is more survivable than ever, however, it also brings new challenges. Cancer survivors are exposed to the risk of a second primary cancer. In Chapter 1 and Chapter 3, we investigated the survival of lung cancer patients with a history of previous cancer. Another challenge that cancer survivors need to face is the follow-up care. Many survivors found that they are "lost in transition" from cancer patients to cancer survivors. In Chapter 2, we investigated the patterns of use and impact on emergency department utilization in a comprehensive cancer survivorship program.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Surprisingly, survival from a diagnosis of lung cancer has been found to be longer for those who experienced a previous cancer than for those with no previous cancer. A possible explanation is lead-time bias, which, by advancing the time of diagnosis, apparently extends survival among those with a previous cancer even when they enjoy no real clinical advantage. We propose a discrete parametric model to jointly describe survival in a no-previous-cancer group (where, by definition, lead-time bias cannot exist) and in a previous-cancer group (where lead-time bias is possible). We model the lead time with a negative binomial distribution and the post–lead-time survival with a linear spline on the logit hazard scale, which allows for survival to differ between groups even in the absence of bias; we denote our model Logit-Spline/Negative Binomial. We fit Logit-Spline/Negative Binomial to a propensity-score matched subset of the Surveillance, Epidemiology, and End Results–Medicare linked data set, con-ducting sensitivity analyses to assess the effects of key assumptions. With lung cancer–specific death as the end point, the estimated mean lead time is roughly 11 months for stage I&II patients; with overall survival, it is roughly 3.4 months in stage I&II. For patients with higher-stage lung cancers, the mean lead time is 1 month or less for both outcomes. Accounting for lead-time bias reduces the survival advantage of the previous-cancer group when one exists, but it does not nullify it in all cases.

1.1 Introduction

Lung cancer, with 5-year survival less than 20%, is the leading cause of cancer-related death in the United States.[1] It mainly affects older people, many of whom have experienced previous cancers and other chronic diseases. Indeed, in 1992-2009 linked Surveillance, Epidemiology, and End Results (SEER) – Medicare data, the proportion of lung cancer patients who were survivors of another cancer at the time of their lung cancer diagnosis ranged from 14% to 21%, depending on stage.[2-4] Because a previous diagnosis of cancer is thought to adversely affect clinical outcomes, it is a common exclusion criterion in lung cancer clinical trials, blocking up to 18% of otherwise eligible patients from participation.[5] Yet surprisingly, several studies have reported that among newly diagnosed lung cancer patients aged 66 and older, those with a previous cancer do not have worse survival than those with no previous diagnosis; indeed, they often do better.[2,3,6-8] For example, Laccetti et al[2] observed that among patients with a newly diagnosed stage IV lung cancer, those with a previous cancer diagnosis had longer all-cause survival and lung cancer–specific survival than similar patients who had not had a previous cancer.

A possible explanation is lead-time bias. Lead time is the length of time between the moment a disease becomes detectable (that is, by tests applied to an asymptomatic person) and the moment it becomes clinically manifest. If a lead time advances the date of diagnosis, the survival time will appear to be longer, even if earlier detection offers no clinical benefit. It is plausible that a lead-time bias could exist in the case of a cancer survivor, likely as a result of enhanced surveillance or the patient's seeking prompt evaluation of symptoms that could represent a subsequent tumor.

Statisticians have long recognized the potential biasing effects of early detection on apparent cancer survival[9-11]; consequently, statistical models of lead-time bias largely assume the

background of a cancer screening program.[12-14] In this paper, we propose a parametric method for estimating lead-time bias that has arisen not from formal screening but instead from whatever additional surveillance that patients and their doctors have implemented following a previous cancer diagnosis. We suppose that one has data from newly diagnosed lung cancer patients, some with a history of cancer (the *Previous* group), and some without (the *No-Previous* group). We assume that only the Previous group is subject to the bias, which takes the form of a random lead time that is added to the latent survival that the patient would have experienced under usual care. In the No-Previous group, we see the natural survival only, untainted by bias. To model these variables, we describe the logit of the post–lead-time death hazard by a spline, allowed to differ between Previous and No-Previous groups, and the lead time as an independent negative binomial (NB); we denote our model LS/NB, for Logit-Spline/Negative Binomial. These assumptions give us the means to construct parsimonious models. We apply the method to new lung cancer diagnoses from a large national database.

1.2 The Lung Cancer Data

We extracted our data from the linked SEER-Medicare database. We included patients 66 years or older with primary lung cancer diagnosed between 2000 and 2011, an interval that represents the most recent data available and produces a large sample size. All patients had full coverage of Medicare Parts A and B from 1 year before to 1 year after the lung cancer diagnosis. We included only patients with either non–small cell (NSCLC) or small cell (SCLC) lung cancer histology. To ensure complete claims data, we excluded patients who participated in health maintenance organizations and those with only autopsy or death certificate records. We also omitted patients with incomplete diagnosis or death dates or discrepancies in SEER and Medicare birth dates of a year or more.

To preserve patient anonymity, SEER-Medicare death and diagnosis data include only the month and year of these events. Thus, survival is measured as the interval, in integer months, between the month of diagnosis and the month of death, and the survival times are effectively discrete. This creates the possibility of survival times of 0 months.

We conducted analyses stratified by the stage of the diagnosed lung cancer, for 2 reasons: First, survival varies greatly by stage, and thus, the strata represent clinically distinct groups. Second, symptoms and tumor aggressiveness differ by stage, in that earlier stages are less likely to be symptomatic and therefore more susceptible to lead-time bias. We classified patients by the American Joint Committee on Cancer criteria into stages I&II, III, and IV. We combined stages I and II because they are more similar to each other than to higher stages, and they represent a relatively small proportion of lung cancer (in our data, stage II is only around 3% of all cases).[4] We excluded the heterogeneous "unstaged" stratum.

We used propensity-score matching to reduce confounding from differences in baseline mortality risk between the Previous and No-Previous groups. We computed a propensity score predicting previous cancer status from available covariates: age, sex (F, M), race/ethnicity (white, black, Hispanic, other), marital status (married, separated/divorced/widowed, single, unknown), histology (SCLC, NSCLC-adenocarcinoma, NSCLC-squamous, NSCLC-other), Charlson comorbidity score (0, 1, 2+, not available), Medicaid status (Y, N), and lung cancer treatment (surgery only, chemotherapy only, radiation only, ≥2 treatments, no surgery/chemo/radiation). As there were fewer patients in the Previous group, we paired a single Non-Previous patient with each Previous patient by nearest-neighbors matching.

The Institutional Review Board of the University of Texas Southwestern Medical Center approved our study.

1.3 Methods

1.3.1 The LS/NB Model

In the No-Previous group, we take the observed survival time to represent the actual survival time from clinical diagnosis, which we denote the *post–lead-time survival*; we label this variable X_N. In the Previous group, we assume that the observable survival time Z is the sum of 2 independent, latent components: the lead time T and the post–lead-time survival X_P; that is, $Z = T + X_P$.[15] We assume moreover that X_N, X_P, and T take values in the nonnegative integers. Our strategy is to assume flexible models for X_N and X_P that differ by at most a single parameter. Because X_N is fully observed (except for censoring), we can use the hypothesized similarity of X_N and X_P as a lever to extract information on the distribution of lead times.

We first consider the distribution of post–lead-time survival in the No-Previous group, labeled X_N. We denote the probability mass function of X_N as $f_N(x) = \Pr[X_N = x]$, its survival function as $S_N(x) = \Pr[X_N \geq x] = \sum_{j=x}^{\infty} f_N(j)$, and its hazard function as $h_N(x) = \Pr[X_N = x | X_N \geq x] = f_N(x)/S_N(x)$, for $x \in (0, 1, 2, \cdots)$.

Attempts to model survival with standard discrete and continuous distributions revealed substantial lack of fit in this large database. A purely nonparametric model was also unsuccessful (see the discussion below). Thus, there was a need for an intermediate approach — a survival model that offers reasonable flexibility with a modest number of parameters. Plots of the logit of the empirical hazard against time revealed that this function is amenable to description with a linear spline having a modest number of knots.[16] Thus, for the No-Previous group, we assume the logit hazard is of the form

$$\lambda_N(x;\beta) \equiv ln\frac{h_N(x)}{1-h_N(x)} = \beta_0 + \beta_1 x + \sum_{j=2}^{m+1} \beta_j \left(x - k_{j-1}\right)_+, x \in \{0,1,2,\cdots\},$$

where $0 < k_1 < \cdots < k_m$ are preselected knots and $(u)_+ = \max(0, u$We assume moreover that the post–lead-time survival in the Previous group differs from that in the No-Previous group according to a proportional odds model on the hazard function. That is, we take the logit hazard for the Previous group $\lambda_P(x)$ to be

$$\lambda_P(x;\beta,\gamma) \equiv ln\frac{h_P(x)}{1-h_P(x)} = \lambda_N(x;\beta) + \gamma, \gamma \in (-\infty,\infty).$$

The odds ratio (OR) of hazards comparing the Previous group with the No-Previous group is therefore OR $= \exp(\gamma)$.

In computations, we can begin with the logit hazard $\lambda(x)$ and compute the hazard as $h(x) = 1/[1 + \exp(-\lambda(x))]$, the survival function as $S(x) = \prod_{j<x}[1 - h(j)]$, and the probability mass function as $f(x) = h(x)S(x)$.

As indicated above, we take survival Z in the Previous group to be the sum of a lead time T and the post–lead-time survival X_P. Because T is a latent variable, it is convenient to model it with a low-parameter discrete distribution. We chose the NB, as it has only 2 parameters but can present unimodal shapes and imposes a less strict functional relationship between mean and variance than the Poisson. Specifically, we assume that lead time follows the NB distribution $T \sim \text{NB}(\rho, \sigma)$, with probability mass function parameterized as

$$f_T(t;\rho,\sigma) = \frac{\Gamma(t+\rho)}{\Gamma(\rho)\Gamma(t+1)}\sigma^\rho(1-\sigma)^t, t \in (0,1,2,\cdots)$$

for $\rho > 0, 0 < \sigma \leq 1$. The probability mass function for Z is the convolution

$$f_Z(z) = \sum_{t=0}^{z} f_T(t) f_P(z-t)\,, z \in (0,1,2,\cdots),$$

where $f_T(\cdot)$ and $f_P(\cdot)$ are the probability mass functions of $T$and X_P, respectively.

With this large data set of discrete event times, we can hasten computations by structuring the data in a frequency table. We categorize data by group (Previous or No-Previous), duration of survival, and event status (censored or dead), as shown in Table 1-1. The index x represents the possible survival times and runs from 0 to M; $n_x^{(A)}$represents the number of subjects alive going into time x in group A; $d_x^{(A)}$ represents the numbers of subjects dying at time x in group A; and $c_x^{(A)}$is the numbers of subjects censored at time x in group A. Thus, $n_x^{(A)} = n_{x-1}^{(A)} - d_{x-1}^{(A)} - c_{x-1}^{(A)}, x = 1, \cdots, M$. An empirical estimate of the hazard at time x in the No-Previous group is the fraction of deaths among those at risk:

$$\hat{h}_N(x) = \frac{d_x^{(N)}}{n_x^{(N)}}, x = 0, \cdots, M.$$

Table 1-1. Tabular representation of the Surveillance, Epidemiology, and End Results–Medicare lung cancer data

Time	Previous			No-Previous		
	At Risk	Died	Censored	At Risk	Died	Censored
0	$n_0^{(P)}$	$d_0^{(P)}$	$c_0^{(P)}$	$n_0^{(N)}$	$d_0^{(N)}$	$c_0^{(N)}$
1	$n_1^{(P)}$	$d_1^{(P)}$	$c_1^{(P)}$	$n_1^{(N)}$	$d_1^{(N)}$	$c_1^{(N)}$
2	$n_2^{(P)}$	$d_2^{(P)}$	$c_2^{(P)}$	$n_2^{(N)}$	$d_2^{(N)}$	$c_2^{(N)}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
x	$n_x^{(P)}$	$d_x^{(P)}$	$c_x^{(P)}$	$n_x^{(N)}$	$d_x^{(N)}$	$c_x^{(N)}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
M	$n_M^{(P)}$	$d_M^{(P)}$	$c_M^{(P)}$	$n_M^{(N)}$	$d_M^{(N)}$	$c_M^{(N)}$

1.3.2 Estimation

Denote the probability mass function and survival function at time x in the No-Previous group as $f_X(x;\beta)$ and $S_X(x;\beta)$, respectively. Using our tabular notation, the loglikelihood for β in the No-Previous group is

$$\ln L_N(\beta) = \sum_{x=0}^{M} [d_x^{(N)} \ln f_X(x;\beta) + c_x^{(N)} \ln S_X(x;\beta)].$$

Similarly, let $f_Z(z;\beta,\gamma,\rho,\sigma)$ and $S_Z(z;\beta,\gamma,\rho,\sigma)$ be the probability mass and survival functions, respectively, for the Previous group, derived form Equation 1. The loglikelihood contribution for β,γ (the log OR for the post-lead-time survival) and ρ,σ (the parameters of the lead-time distribution) from the Previous group is then

$$\ln L_P(\beta,\gamma,\rho,\sigma) = \sum_{z=0}^{M} [d_z^{(P)} \ln f_Z(z;\beta,\gamma,\rho,\sigma) + c_z^{(P)} \ln S_Z(z;\beta,\gamma,\rho,\sigma)].$$

Combining these expressions, the loglikelihood form the entire data set is

$$\ln L(\beta,\gamma,\rho,\sigma) = \ln L_N(\beta) + \ln L_P(\beta,\gamma,\rho,\sigma).$$

We obtain the maximum likelihood estimate (MLE) $\hat{\beta},\hat{\gamma},\hat{\rho},\hat{\sigma}$ by maximizing Equation 3 numerically using the limited-memory Broyden-Fletcher-Goldfard-Shanno method with box constraints, a quasi-Newton algorithm implemented in R function optim().[17] We estimated the mean lead time $E(T)$ as $\hat{\rho}(1-\hat{\sigma})/\hat{\sigma}$ and the OR as $\exp(\hat{\gamma})$, and we construct confidence intervals (CIs) for these parameters by the delta method. R code is available from the first author.

1.3.3 Sensitivity Analysis

We conducted a set of sensitivity analyses to evaluate robustness of results to key model assumptions.

The first analysis assessed the effect of varying assumptions about each part of the post-lead-time survival model on the parameters of the other part. We first assumed that there is no lead time ($T \equiv 0$) and estimated the corresponding difference in survival between the Previous and No-Previous groups (now completely described by the OR). Next, we assumed fixed values of γ (the log OR for the survival difference) and obtained the corresponding estimates of the remaining parameters $\beta, \widehat{\rho, \sigma}(\gamma)$.

The second analysis assessed robustness to the assumed independence of X_P and T. Using our MLE as the truth, we simulated survival time X_N in the No-Previous group through its spline model. For the Previous group, we assumed the MLE marginal distributions and simulated X_P and T from a bivariate normal copula with underlying correlation θ, generating $Z = X_P + T$. After simulating data for both groups, we estimated LS/NB assuming independence of lead time and survival, comparing the estimated mean lead time under the varying assigned correlations.

The third analysis examined the effect of the assumed distribution of lead time T. In addition to the NB, we evaluated a range of models including the geometric, Poisson, zero-inflated Poisson, zero-inflated NB, and a nonparametric (multinomial) distribution that assumes support on a small number of integers but is otherwise unrestricted. We calculated MLEs of $E(T)$ and OR under each model and compared fits via the Akaike information criterion (AIC).

In a final sensitivity analysis, we compared results from the propensity-score-matched sample with an unmatched analysis that applied the model to the entire data set.

1.4 Results

1.4.1 Preliminary Analyses

We identified 215 718 SEER-Medicare lung cancer diagnoses, of whom 22% were stage I&II, 24% stage III, and 39% stage IV; the remaining 15% were unstaged. Roughly 20% had a previous cancer (Table 1-2).

We first analyzed the data by computing mean survival (restricted to 160 month), estimating a proportional hazards model with Previous group status as the sole covariate, and comparing the groups by a logrank test; results appear in Table 1-3. For lung cancer mortality, mean survival is greater in the Previous group in each stage, with hazard ratios ranging from 0.82 to 0.78. For all-cause mortality, mean survival is shorter in the Previous group in stage I&II (HR = 1.05) but longer in the higher stages (HR = 0.94 in stage III and HR = 0.90 in stage IV). Kaplan-Meier survival plots appear in Figure 1-1.

Figure 1-2 displays estimated hazard functions for lung cancer and all-cause death in the No-Previous group, stratified by lung cancer stage. We computed the "raw" estimates by Equation 2 and the "LS/NB" estimates by fitting the model to the entire matched data set. We placed knots at$k_1 = 1, k_2 = 5, k_3 = 50$, and $k_1 = 100$, where visual inspection suggested a possible change in the slope of the logit hazard. Evidently the model offers a good fit. Table 1-4, which compares Kaplan-Meier estimates of the survival function with estimates under the spline model for the No-Previous group, again shows that the model fits well.

Table 1-2. Surveillance, Epidemiology, and End Results–Medicare data: count of patients by stage and previous cancer diagnosis

Stage	Previous, %	No-Previous, %	Total
Original data			
I&II	10 187 (22)	36 402 (78)	46 589
III	8474 (16)	43 841 (84)	52 315
IV	12 716 (15)	70 852 (85)	83 568
Matched data			
I&II	10 187 (50)	10 187 (50)	20 374
III	8473* (50)	8473 (50)	16 946
IV	12 715* (50)	12 715 (50)	25 430

*Omitting subjects who had missing marital status.

Table 1-3. Surveillance, Epidemiology, and End Results–Medicare lung cancer data: summary of mortality in the matched data

Stage	Mean Survival in Months*			
	Previous	No-Previous†	HR (95% CI)	Logrank P
Death from lung cancer				
I&II	90.4	83.0	0.82 (0.79-0.86)	< .0001
III	33.8	26.3	0.80 (0.77-0.83)	< .0001
IV	16.5	10.8	0.78 (0.76-0.80)	< .0001
Death from any cause				
I&II	52.9	56.4	1.05 (1.02-1.09)	0.0033
III	19.1	18.1	0.94 (0.91-0.97)	< .0001
IV	9.4	8.1	0.90 (0.88-0.93)	< .0001

Abbreviations: CI, confidence interval; HR, hazard ratio from a Cox model.
*Restricted mean with upper limit = 160 months
†Reference group

1.4.2 LS/NB Model Estimates

We applied LS/NB to the matched data set, estimating simultaneously the spline coefficients, OR, and the lead-time parameters. The estimated LS/NB survival curves in Figure 1-3 agree well with the superimposed Kaplan-Meier curves; the divergence of the empirical and estimated curves in the right tail partly reflects plotting survival on the log scale, which magnifies differences at small values, and partly the reduced precision in this range.

Maximum likelihood estimates of $E(T)$ and OR appear in the top panel of Table 1-5. For lung cancer mortality, the estimated $E(T)$ for patients with stage I&II lung cancer in the Previous group is 11.3 months; estimated mean lead times in stages III and IV are roughly 1 month and 1 week, respectively. Even allowing for a potential lead-time bias, the ORs are less than 1 (significantly so in stages III and IV); thus, accounting for lead-time bias does not nullify the beneficial effect of having had a previous cancer on lung cancer mortality. As mean lead time declines with advancing stage, the effect of surviving a previous cancer increases, with the greatest survival advantage (OR = 0.79; 95% CI, 0.77-0.82) appearing in stage IV.

For all-cause death (Figure 1-3, right panel; Table 1-5, top right), the estimated mean lead times are 3.4, 1.1, and 1.1 months for patients in stages I&II, III, and IV, respectively. In stage I&II, accounting for lead-time bias accentuates an already statistically significant overall survival advantage for the No-Previous group. For stages III and IV, incorporating lead time in the model renders the OR indistinguishable from unity. Thus, for all-cause death, the apparent survival advantage in the Previous group with stages III and IV cancer may well reflect a modest lead-time bias. In stage I&II, the survival advantage in the No-Previous group is larger than the estimated hazard ratio of 1.05 from the Cox model.

As demonstrated in Figure 1-3 and Table 1-5, the survival advantage of the Previous group is only apparent when one censors non-lung cancer deaths; for overall survival, the No-Previous group does slightly better in stage I&II and roughly the same in stages III and IV, even after accounting for lead time. Figure 1-4, which displays cumulative incidence curves[18] of death from cancer and other causes in stage I&II and stage IV, explains this observation (we omit stage III, which is similar to stage IV). The curves, unadjusted for lead-time bias, show that subjects in the Previous group at every stage have a lower rate of death from lung cancer but a higher rate of

death from other causes. In stage I&II, the risk of lung cancer death is low in both groups and similar to the risk of death from other causes; thus, overall death rates slightly favor the No-Previous group. In stage IV, the risk of lung cancer mortality is the dominant hazard component; thus, there is a modest advantage for the Previous group in overall mortality, mirroring the findings in Table 1-3, which also does not account for lead time.

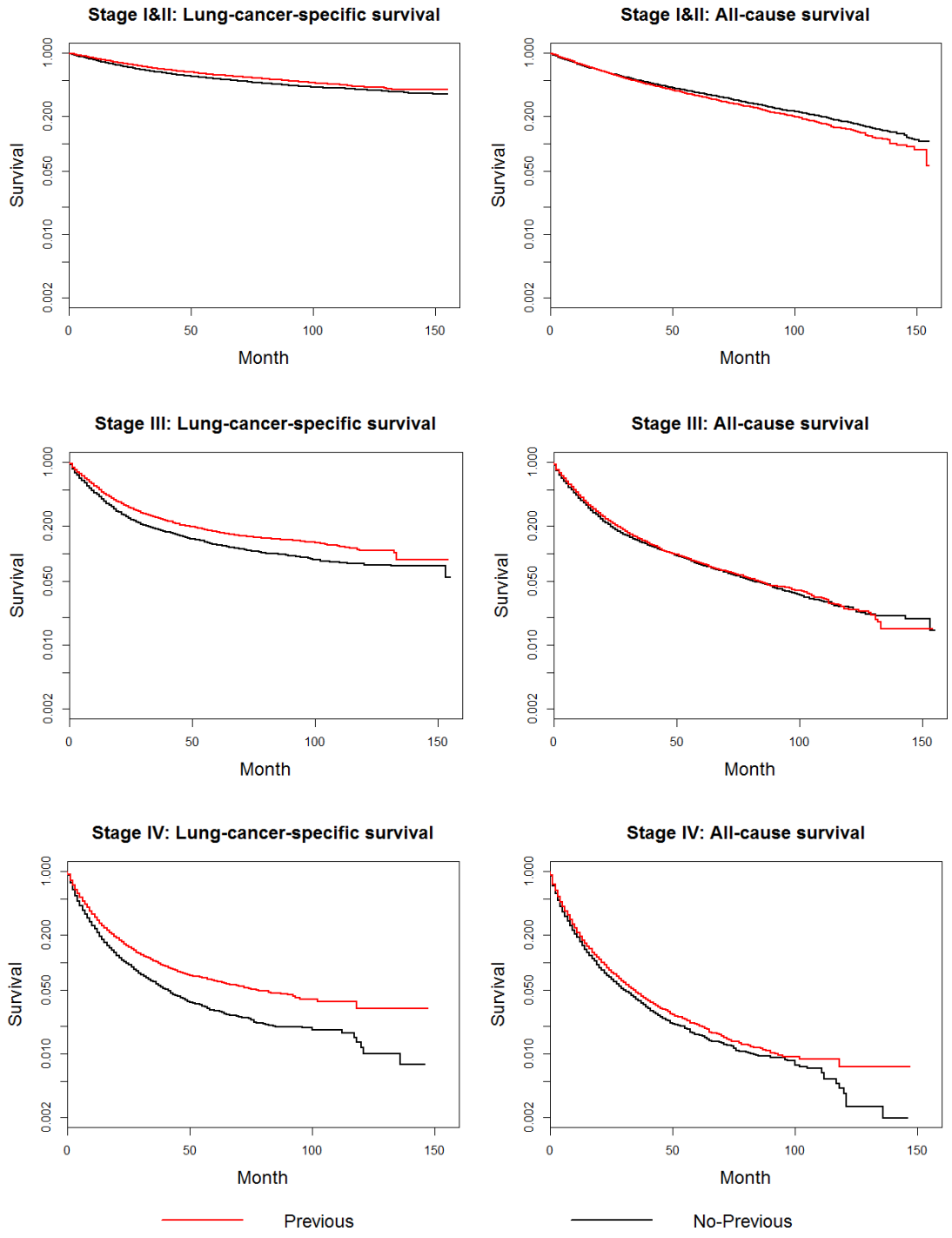

Figure 1-1. Kaplan-Meier survival curves (on the log scale) for newly diagnosed lung cancer patients, stratified by stage

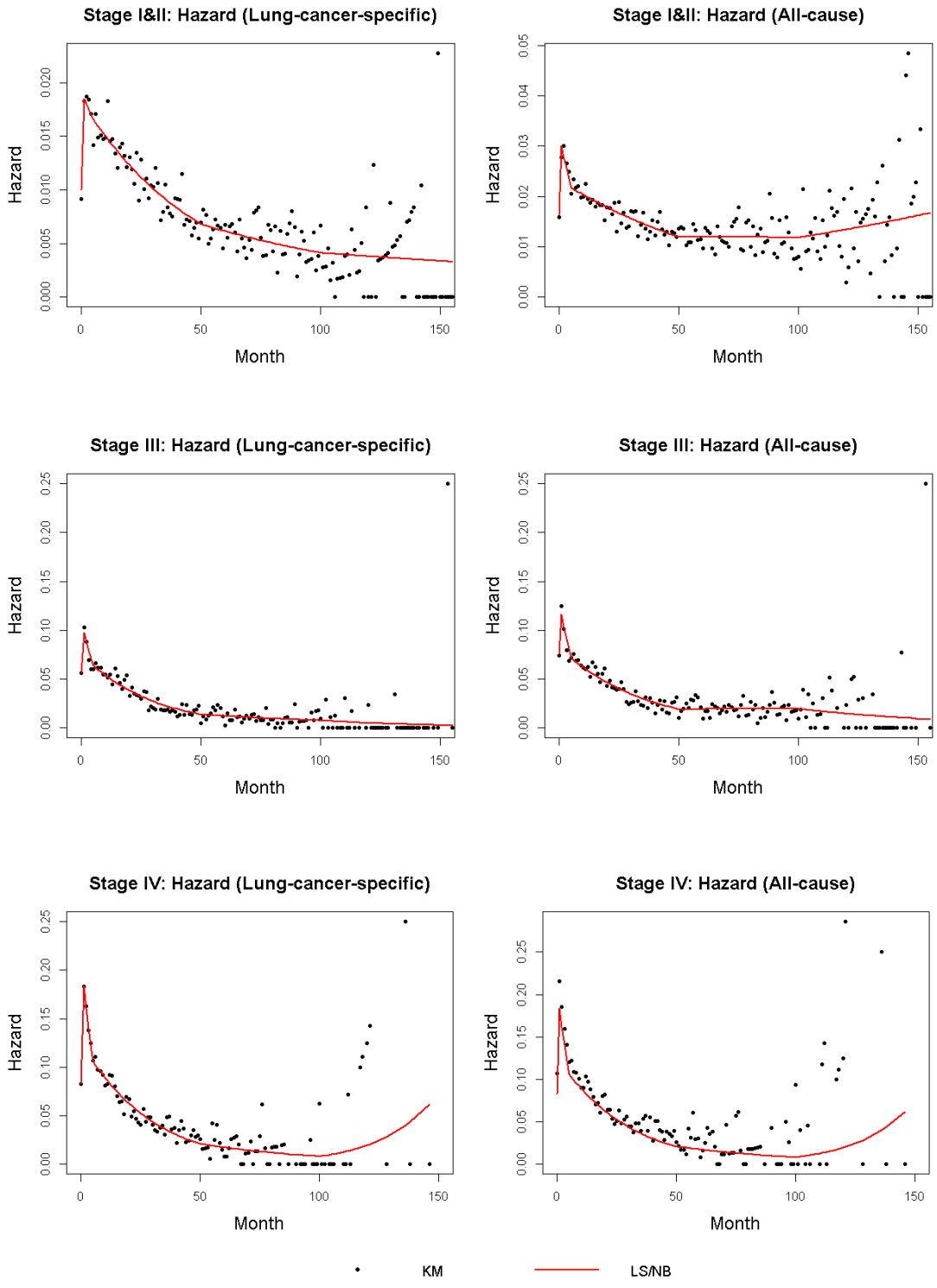

Figure 1-2. Raw (KM) and estimated hazard functions of lung cancer–specific and overall survival in the No-Previous group, stratified by stage. KM, Kaplan-Meier; LS/NB, Logit-Spline/Negative Binomial

Table 1-4. Estimated lung cancer–specific and all-cause survival in months (%) for the No-Previous group

	Lung Cancer-Specific Survival, %						All-Cause Survival, %					
Stage	I&II		III		IV		I&II		III		IV	
Time	KM	LS	KM	LS	KM	LS	KM	LS	KM	LS	KM	LS
0	99.1	99.0	94.4	94.4	91.7	91.7	98.4	98.4	92.6	92.4	89.3	89.3
1	97.3	97.2	84.7	85.2	74.9	74.8	95.7	95.4	81.0	81.7	70.1	70.1
3	93.7	93.8	71.9	71.5	54.0	54.0	90.3	90.4	67.0	66.6	48.0	47.9
5	90.8	90.6	63.5	62.2	42.3	42.4	86.3	86.3	57.9	56.8	36.2	36.5
40	59.8	59.7	17.2	16.7	5.0	4.8	47.3	46.9	11.9	11.5	3.2	3.0
80	46.4	46.5	10.1	10.1	2.1	2.3	28.6	28.8	5.1	5.2	1.0	1.1
100	42.1	42.5	8.6	8.6	1.8	1.9	22.7	22.7	3.6	3.5	0.8	0.8
120	39.7	39.2	7.6	7.6	1.2	1.5	17.6	17.6	2.6	2.5	0.4	0.5
140	36.2	36.5	7.3	7.0	0.8	0.8	13.5	13.2	2.1	1.9	0.2	0.2

1.4.3 Sensitivity Analyses

Our initial sensitivity analysis estimated the OR assuming no lead time ($T \equiv 0$) and estimated $E[T]$ while holding OR fixed at a range of likely values. Results appear in Table 1-6. For lung cancer death, with no lead-time bias, the estimated OR is 0.82, 0.80, and 0.77 for stages I&II, III, and IV, respectively (note the similarity to the hazard ratios in Table 1-3). As we allow the OR to increase to 1.2, the mean lead time rises to as high as 44 months for stage I&II and 7.7 months for stage IV (Table 1-6 and Figure 1-5). This is to be expected, because the larger the OR, the greater must be the lead-time bias to compensate for it. If the entire lung cancer survival difference is explained by lead time, that is, if OR = 1, then the mean lead time is estimated to be 15.4, 8.5, and 5.1 months for stages I&II, III, and IV, respectively. Thus, a substantial mean lead time—more than 1 year in stage I&II — is needed to nullify any apparent positive effect of previous cancer on lung cancer survival. For all-cause death with no lead time, we estimate the OR to be 1.05, 0.94, and 0.90 for stages I&II, III, and IV, respectively. The estimated $E[T]$ also increases as OR increases (Figure 1-5, right panel) but less dramatically than when lung cancer-specific death is the end point. A recurring theme of the analysis is that $\hat{E}[T]$ declines as the

stage increases. This is reasonable, as we expect higher-stage tumors to progress more rapidly from being just detectable to manifesting clinical signs and symptoms.

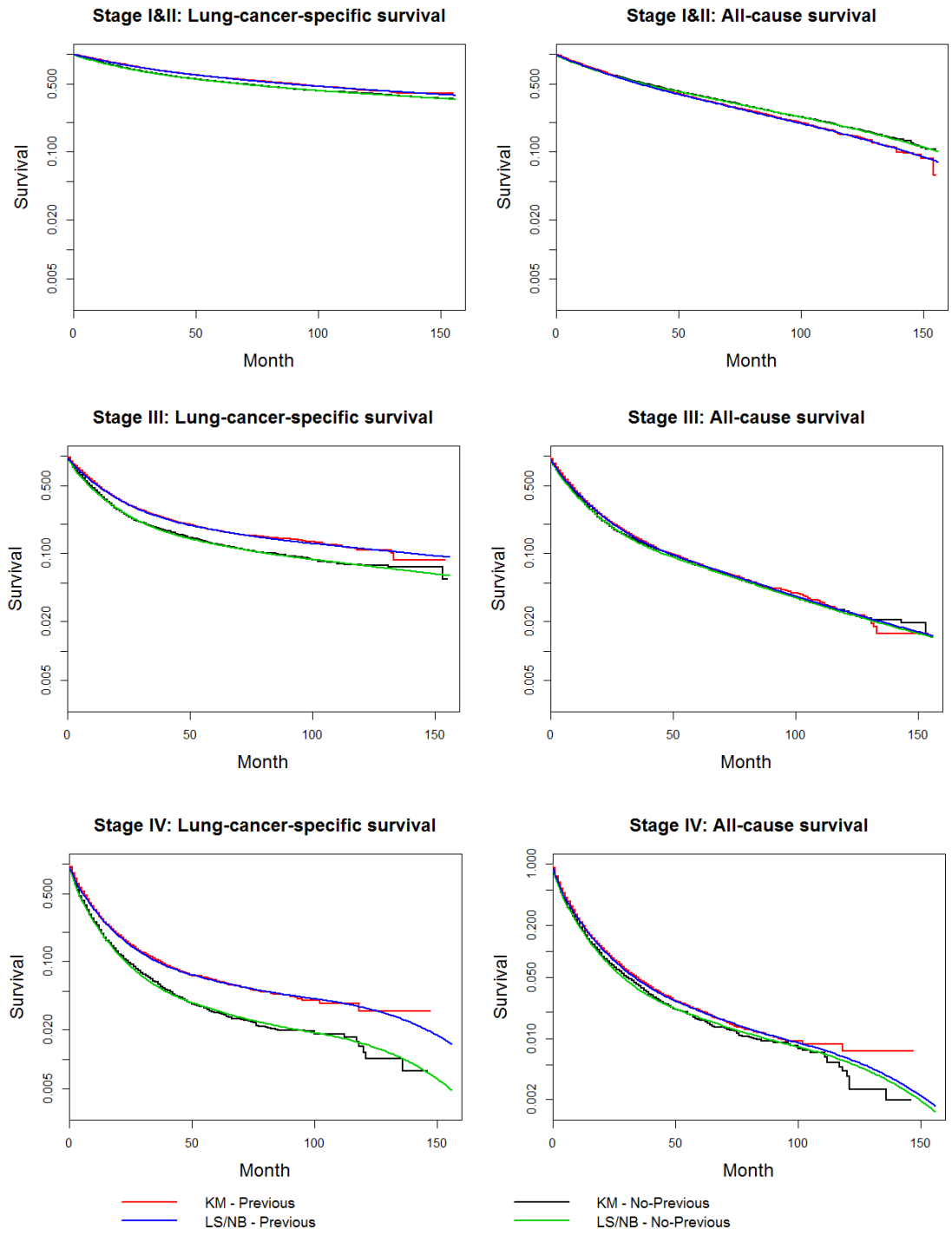

Figure 1-3. Estimated survival functions (on the log scale) by Kaplan-Meier (KM) and the Logit-Spline/Negative Binomial (LS/NB) method, stratified by stage

The second analysis evaluated sensitivity to the assumed independence between T and X_P by simulating data with marginal distributions similar to those in our lung cancer death data but with correlation of T and X_P induced by a normal copula. Results appear in Figure 1-6. With data generated under independence, $\hat{E}[T]$ is 9.97 (95% CI, 7.00-12.94), 1.22 (95% CI, 0.80-1.63), and 0.40 (95% CI, 0.15-0.65) months for stages I&II, III, and IV, respectively. Failure to account for correlation induces a negative bias when the correlation is positive and vice versa. The trend is most evident in stage I&II, where lead time is longest. Because positive correlation of T and X_P is the more plausible alternative to independence, a faulty assumption of independence will likely lead to underestimation of $E[T]$. Xu et al observed a similar tendency in a nonparametric model of breast screening.[19] For stage I&II, $\hat{E}[T]$ lies in the 95% CI (red dash-dotted line) under independence if the correlation is in the range (-0.2, 0.1). For stage IV, even if the correlation is as large as 0.4, $\hat{E}[T]$ is still within the 95% CI under independence. Thus, an incorrect assumption of independence can affect results, most likely leading to a negative bias, but the correlation must be substantial for this to occur. One can avoid this bias by conditioning on factors that confound the relationship between X_P and T. One can also attempt to model the correlation, but as only the sum of T and X_P is ever observed, it seems unlikely that it will be possible to estimate such a model robustly.

Third, we evaluated sensitivity to the assumed lead-time distribution by estimating parameters under a range of models for T: NB, geometric, Poisson, zero-inflated Poisson, zero-inflated NB, and a nonparametric distribution with mass at $(0, 1, \cdots, 15)$ for stage I&II or $(0, 1, \cdots, 5)$ for stage III and stage IV. In Table 1-7, we present for each model $\hat{E}[T]$, $\widehat{\mathrm{OR}}$, the first few values of the probability mass function of T, and the AIC. Among the parametric distributions, the geometric, Poisson, and zero-inflated Poisson never fit well; they give larger

AIC and usually underestimate both $E[T]$ and OR. Zero-inflated NB gives results similar to NB, although the latter has lower AIC. All estimated mass functions assign highest probability to $T = 0$; evidently, the important difference is that NB permits a longer tail and therefore a potentially higher mean. The mean lead time is sensitive to model assumptions, the OR less so. Thus, it appears that NB is a satisfactory model for reasons of flexibility and parsimony, although users must anticipate some sensitivity in the estimated $E[T]$.

Finally, we estimated the model on the entire data set; see the bottom panel of Table 1-5. Compared with the matched analysis, estimates of OR change by no more than about 1%. Estimates of $E[T]$ are less robust, possibly changing by up to a half-month but never leading to a qualitative difference in interpretation. Confidence intervals are narrower thanks to the larger sample size.

Table 1-5. Estimated mean lead time E[T] (month) and OR, by cause of death and stage

	Lung Cancer-Specific Mortality		All-cause Mortality	
Stage	$\hat{E}[T]$ (95% CI)	$\widehat{\text{OR}}$ (95% CI)	$\hat{E}[T]$ (95% CI)	$\widehat{\text{OR}}$ (95% CI)
		Matched sample		
I&II	11.3 (3.8-33.3)	0.96 (0.86-1.08)	3.4 (1.1-5.8)	1.14 (1.08-1.21)
III	1.1 (0.5-1.7)	0.85 (0.82-0.88)	1.1 (0.5-1.7)	1.00 (0.96-1.03)
IV	0.3 (0.02-0.5)	0.79 (0.77-0.82)	1.1 (0.4-1.7)	1.00 (0.96-1.04)
		Unmatched sample		
I&II	11.2 (2.4-20.1)	0.96 (0.89-1.05)	3.5 (1.9-5.1)	1.14 (1.10-1.18)
III	0.7 (0.2-1.1)	0.84 (0.82-0.86)	0.9 (0.2-1.6)	1.00 (0.96-1.04)
IV	0.4 (0.2-0.6)	0.80 (0.78-0.82)	0.9 (0.3-1.5)	0.99 (0.94-1.05)

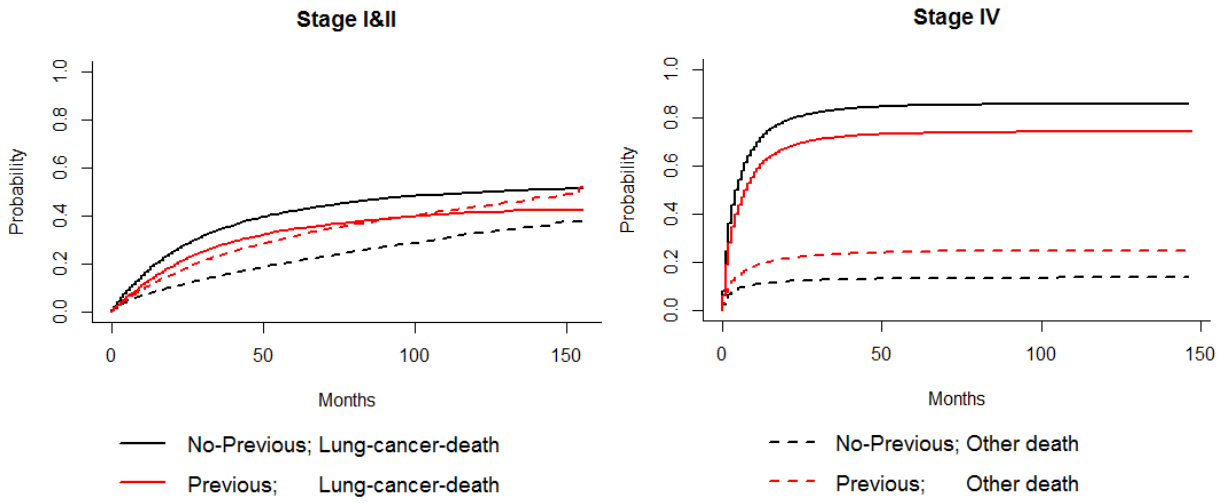

Figure 1-4. Cumulative incidence curves for stage I&II and stage IV (stage III is similar to stage IV)

Table 1-6. Sensitivity analysis of estimates of OR (assuming $T \equiv 0$) and mean lead time $E(T)$ (for fixed OR)

	Lung Cancer Mortality		All-Cause Mortality	
Stage	OR (SE)	$\hat{E}[T]$ (SE)	OR (SE)	$\hat{E}[T]$ (SE)
I&II	0.82 (0.013)	0	1.05 (0.013)	0
	0.90	5.0 (1.1)	--	--
	1.00	15.4 (2.6)	--	--
	1.10	28.6 (3.2)	1.10	2.4 (0.4)
	1.20	44.4 (3.2)	1.20	5.9 (0.8)
III	0.80 (0.011)	0	0.94 (0.011)	0
	0.90	2.2 (0.5)	--	--
	1.00	8.5 (1.2)	1.00	1.1 (0.2)
	1.10	14.4 (1.4)	1.10	3.2 (0.4)
	1.20	18.8 (1.4)	1.20	5.5 (0.4)
IV	0.77 (0.0082)	0	0.90 (0.0087)	0
	0.80	0.3 (0.1)	--	--
	0.90	2.9 (0.4)	--	--
	1.00	5.1 (0.4)	1.00	1.1 (0.1)
	1.10	6.6 (0.4)	1.10	2.1 (0.1)
	1.20	7.7 (0.4)	1.20	3.0 ()0.1

Abbreviations: OR, odds ratio; SE, standard error. Figures in *italic* are fixed in the sensitivity analysis.

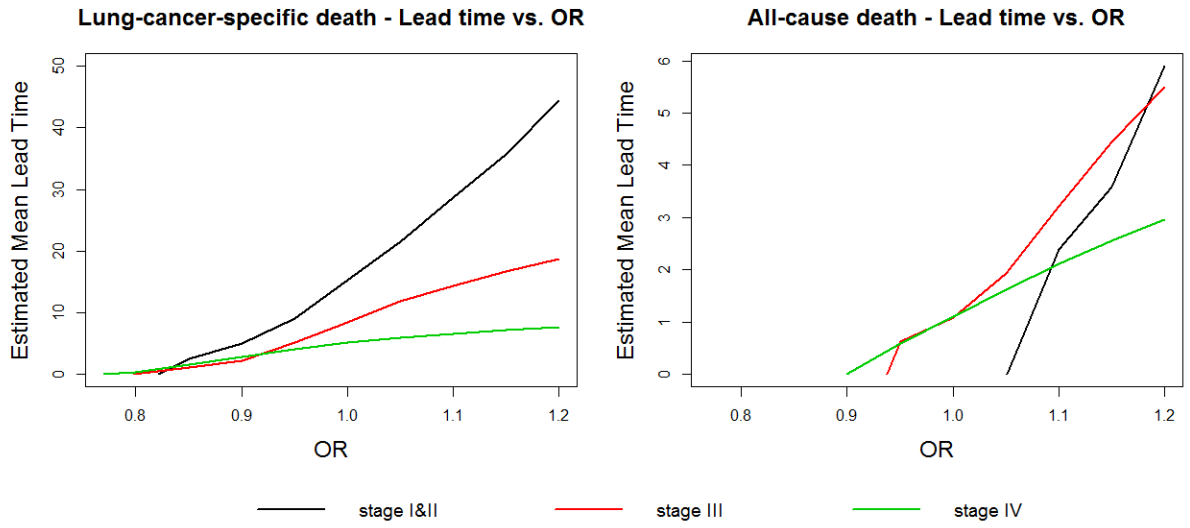

Figure 1-5. Estimated mean lead time (in month) as a function of odds ratio (OR), by stage

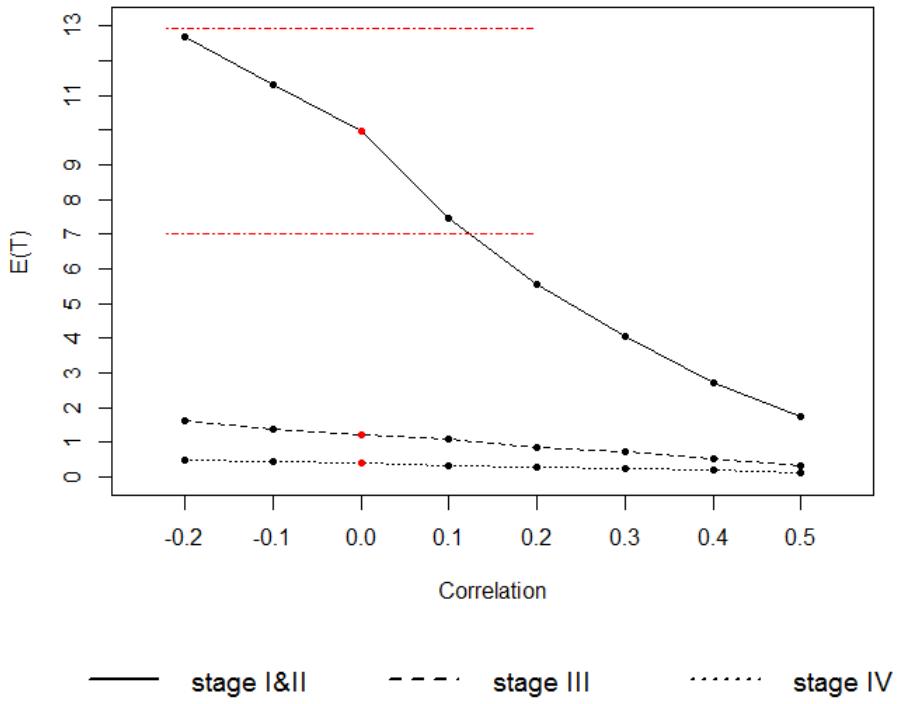

Figure 1-6. Analysis of sensitivity of $\hat{E}[T]$ (in months) to assumed independence of T and X_P

Table 1-7. Sensitivity to the assumed lead-time distribution

Stage	Parameter	NB	Geometric	Poisson	ZIP	ZINB	NP
I&II	$\hat{E}[T]$	11.3	1.42	0.35	0.40	8.39	2.24
	$\widehat{\mathrm{OR}}$	0.96	0.84	0.82	0.82	0.94	0.87
	$\Pr[T=0]$	0.640	0.704	0.704	0.712	0.677	0.741
	$\Pr[T=1]$	0.059	0.208	0.247	0.195	0.053	0.043
	$\Pr[T=2]$	0.032	0.062	0.043	0.071	0.029	0.023
	$\Pr[T=3]$	0.022	0.018	0.0051	0.017	0.020	0.016
	$\Pr[T=4]$	0.017	0.0054	4.4e-4	0.0031	0.016	0.012
	$\Pr[T=5]$	0.014	0.0016	3.1e-5	4.6e-4	0.013	0.010
	AIC	<u>85142.4</u>	85159.0	85160.0	85161.3	85143.8	85175.7
III	$\hat{E}[T]$	1.11	1.02	0.0064	0.019	0.83	0.57
	$\widehat{\mathrm{OR}}$	0.85	0.80	0.80	0.80	0.63	0.83
	$\Pr[T=0]$	0.828	0.983	0.994	0.981	0.860	0.829
	$\Pr[T=1]$	0.051	0.016	0.0063	0.018	0.047	0.048
	$\Pr[T=2]$	0.026	2.7e-4	2.0e-5	3.3e-4	0.024	0.025
	$\Pr[T=3]$	0.017	4.4e-6	4.2e-8	3.8e-6	0.015	0.0062
	$\Pr[T=4]$	0.012	7.2e-8	6.7e-11	3.4e-8	0.010	0.0046
	$\Pr[T=5]$	0.0092	1.2e-9	8.6e-14	2.4e-10	0.0078	0.087
	AIC	94836.9	94851.2	94850.7	94853.5	94839.8	<u>94834.8</u>
IV	$\hat{E}[T]$	0.28	1.02	0.013	0.011	0.25	0.20
	$\widehat{\mathrm{OR}}$	0.79	0.77	0.77	0.77	0.79	0.79
	$\Pr[T=0]$	0.925	0.984	0.986	0.989	0.927	0.924
	$\Pr[T=1]$	0.029	0.016	0.013	0.011	0.030	0.029
	$\Pr[T=2]$	0.013	2.6e-4	9.0e-5	8.5e-5	0.014	0.013
	$\Pr[T=3]$	0.0081	4.3e-6	4.0e-7	4.5e-7	0.0081	0.0082
	$\Pr[T=4]$	0.0054	7.1e-8	1.4e-9	1.8e-9	0.0054	0.0055
	$\Pr[T=5]$	0.0039	1.2e-9	3.7e-12	1.8e-9	0.0038	0.019
	AIC	<u>128990.1</u>	128993.7	128993.7	128995.6	128992.0	128993.5

Abbreviations: AIC, Akaike information criterion; NB, negative binomial; NP, nonparametric with support at (0,1,…,15) for stage I&II or (0,1,…,5) for stage III and stage IV; ZINB, zero-inflated NB; ZIP, zero-inflated Poisson.
Underlined value is the smallest for models in that stratum.

1.5 Discussion

1.5.1 Summary and Context

Our proposed LS/NB model allows estimation of the mean lead time in cancer patients who have a previous diagnosis of another cancer. Applying it to SEER-Medicare data with lung cancer–specific survival as the outcome, estimated mean lead times are roughly 11 months for stage I&II lung cancer and around 1 month or less for higher stages. For death from any cause, the estimated mean lead times are roughly 3 months for stage I&II and 1 month for higher stages.

Even after accounting for lead time, the Previous group has a lower lung cancer mortality hazard in all stages, statistically significantly so in stages III and IV. For all-cause mortality, accounting for lead time leaves survival slightly worse in the Previous group for stage I&II and practically equal to No-Previous survival in more advanced stages.

Most discussion of lead-time bias assumes a context of cancer screening; to our knowledge, this is the first analysis of lead time as it may arise from idiosyncratically enhanced surveillance in cancer survivors. Walter and Stitt proposed modeling the survival of screen-detected cases by the hazard function; their analysis requires specification of the duration of the detectable preclinical phase and assumptions of independent, exponential distributions for the lead time and the total survival time after diagnosis.[15] Xu and Prorok assumed that the lead time is exponential but used a nonparametric method to estimate post–lead-time survival.[20] Duffy et al assumed an exponential distribution of the lead time, adjusting the survival times of the screen-detected cases by subtracting an estimated conditional mean lead time.[21]

1.5.2 Modeling Issues

The LS/NB model departs from common practice in taking survival times to be discrete. This approach does not primarily reflect an impulse to model the data as they are, although SEER-Medicare survival times are in fact rounded to the nearest month. With the wide range of survival times, this is actually a fine rounding grid, and analyzing the data as though they are continuous should cause little bias.[22] An advantage of assuming discreteness is that it simplifies calculation, as one can compute all needed quantities — probabilities, means, and likelihood terms — by direct summation.

Because standard discrete distributions fit poorly to post-lead-time survival, we eschewed them in favor of flexible, easily estimated spline models on the logit hazard. Modeling survival

time in the Previous group is more challenging, because if one specifies both the lead time and the post–lead-time survival using nonparametric or overly flexible parametric forms, their convolution can be nonidentifiable. Therefore, for the lead-time distribution, we settled on the 2-parameter NB, which is more flexible than the Poisson and geometric distributions but retains an easily computed mean function. Analysis under a range of alternative models showed substantial sensitivity of estimates of $E(T)$ and modest sensitivity of estimates of OR. Supporting our initial intuition, the NB appeared to offer a good fit when evaluated by the AIC.

Our analysis is made possible by the availability of the No-Previous group — that is, a sample whose survival times are free of lead-time bias. Assuming that survival in the No-Previous group is the same as post–lead-time survival in the Previous group, possibly up to an OR parameter, we can readily identify the lead-time distribution. An approach that we tried initially was to estimate a common $S_N(x) = S_P(x)$ nonparametrically from the No-Previous data and solve for $f_T(t)$ by inverting the convolution equations (1). Unfortunately, the solution yielded probabilities outside [0, 1], even when we constrained the support of T to include only the first few nonnegative integers. Thus, despite the large sample size, some smoothing is necessary.

A key assumption is that T and X_P are independent. A plausible departure from this assumption is that the association is positive, in which case tumors that arise with shorter lead time are also more rapidly fatal.[19] As both lead time and post–lead-time survival in the Previous group are latent, one cannot test this hypothesis robustly. In a sensitivity analysis, we demonstrated that failure to account for correlation could induce bias when the true correlation is moderate. One could reduce this bias by adjusting for potential confounders of the relationship between lead time and post–lead-time survival.

We considered correction methods like those proposed by Duffy et al[21] but found them to have several shortcomings: First, the adjustments to the observed z values use only information on an assumed exponential distribution of T and ignore the model for X_P; that is, they adjust using the incorrect conditioning set $T \leq z$ rather than $T + X_P \leq z$. Second, the method requires that one possess estimates of the parameters of T from previous data; such estimates may be available in special cases, but in general, they are elusive. And finally, even if one could perform the adjustments correctly by subtracting the conditional mean of T, the method would be analogous to single imputation of the predicted mean lead time and therefore would understate uncertainty. Our attempts to implement these analyses (not shown) demonstrated that adjustment formulas do not work as well as full estimation within the models from which they are derived. A multiple-imputation approach that involves taking repeated draws from the predictive distribution of the latent X_P given (z, d) would address this concern.[23]

As indicated above, the survival advantage of the Previous group is only apparent for deaths from lung cancer, in an analysis that censors subjects at the time of death from other causes. Such an analysis implicitly assumes independence of times to death from cancer and other causes, a hypothesis whose validity is by no means certain and that one cannot test robustly. Moreover, because SEER cause-of-death data are not adjudicated, this outcome is subject to errors of misclassification. A possible enhancement of the method would be to jointly model lead time, mortality from cancer, and mortality from other causes, thereby creating valid estimates of cause-specific hazard functions.

Assuming that the Previous and No-Previous survival curves are identical except for a lead-time bias, one can estimate $E[T]$ by simply taking the difference between estimates of $E[Z]$ and $E[X_N]$, the first 2 columns in Table 1-3. Estimates computed in this way are similar to the model-

based estimates from the sensitivity analysis with OR = 1.0 in Table 1-6, at least for stages III and IV. The fact that estimates of $E[T]$ vary by end point suggests an inadequacy in the model, because the putative lead-time bias should be identical for survival end points measured from the same diagnosis time. Jointly modeling the 2 types of death would resolve this ambiguity.

1.5.3 Clinical Implications

Because the SEER-Medicare database contains only persons who are eligible for Medicare, we restricted our analysis to subjects aged 66 or older; thus, our findings may not be relevant to the entire lung cancer population. We note, however, that the median age at diagnosis of lung cancer is 70, and 69% of US lung cancer diagnoses occur at ages >65;[24] therefore, our data represent the majority of US lung cancer patients. Moreover, we recently demonstrated that, among lung cancer patients in SEER (2009-2013), 8.6% of those <65 years and 18.7% of those ≥65 years are survivors of a previous, non–lung cancer.[25] Thus, our study, while limited to older adults, represents the majority of all lung cancer patients and of lung cancer patients who have survived a previous cancer.

Our findings suggest that lead-time bias is one possible cause of the observed, modest, positive effect of a previous cancer diagnosis on lung cancer survival time. Other factors that underlie the observed differences are unknown but may include physiologic and health care delivery effects, misclassification, and residual confounding. Because SEER does not conduct active follow-up, it cannot provide validated data about metastatic disease occurring after initial cancer diagnosis. Nor does SEER measure smoking status, which is a potentially powerful confounder. Further studies with prospective, comprehensive data collection would help resolve these questions.

CHAPTER 2

The University of Texas Southwestern Medical Center Moncrief Cancer Institute has established a survivorship program to enhance the quality of life for cancer survivors, focusing on their mental and physical health. The program includes specialized exercise and nutrition training, as well as group and individual education and counseling. Benefits of participation in this program are unknown. We combined tumor registry and electronic medical record data for the safety-net healthcare system in Tarrant County, TX with participation data from the survivorship program. We identified patterns of participation through statistical clustering. We used regression models to measure the effect of participation on behaviors and on the frequency of Emergency Department (ED) visits. Among 467 program participants, we identified four clusters representing distinct patterns of participation. Our results demonstrated that participation in the survivorship program was associated with a 37% lower rate of ED visits ($p < 0.0001$). The study findings could further shape delivery of the survivorship services in our institution and similarly situated organizations across the country. In addition, these findings will provide insurers and policy makers with information to make evidence-based decisions regarding reimbursement for cancer survivorship programs.

2.1 Introduction

Cancer detection and treatment advances have led to rapid growth in the number of cancer survivors. An estimated 15.5 million survivors represented 4.8% of the United States population

in 2016 and projected to increase by 31% to 20.3 million by 2026.[26] Data from the Surveillance, Epidemiology, and End Results (SEER) Program[27] suggest about 67% of people diagnosed with cancer survive 5 years or more. Although cancer is more survivable than ever, many survivors become "lost in transition" once systematic care and treatment is finished.[28] In addition, survivors often face devastating physical, psychosocial, and economic effects from the disease and treatment that affects quality of life. Thus, it has been proposed comprehensive survivorship programs can help cancer survivors address the likely physical, psychological, social and financial problems encountered in their next stage of life.[29]

A major barrier to survivorship care is cost.[30] The majority of cancer survivorship programs are associated with large cancer centers or academic medical centers and limited in scope because they are expensive and poorly reimbursed. Cancer survivors typically face the high treatment costs, lost work time and/or impairment in the ability to work[31], and loss of health insurance. As a results, many are unable to afford the additional, unreimburseable costs for cancer survivorship services.

Moncrief Cancer Institute (MCI), an affiliate of UT Southwestern Medical Center (UTSW), is a non-profit community-based cancer prevention and support center. MCI has established a survivorship program using a community-based model to provide patient-centered care. This survivorship program provides opportunity for cancer survivors to improve their health and quality of life by addressing any lingering medical and psychosocial effects of illness in addition to promoting healthy lifestyle changes. MCI offers cancer survivors multidisciplinary services regardless of diagnosis, stage, treatment provider, socioeconomic status, or insurance coverage.

In this paper, we analyze sociodemographic and clinical characteristics of cancer survivors who took part in the MCI program. We use statistical clustering methods to identify common

patterns of service utilization. We further estimate the association of program participation with frequency of emergency department (ED) visits, to understand the potential impact of survivorship programs at the health system level.

2.2 Materials and Methods

2.2.1 Study procedures

This retrospective analysis uses three data sources: The John Peter Smith Hospital (JPS) tumor registry database, JPS healthcare system electronic medical records (JPS EMR), and the UTSW-MCI Surviviorship database. JPS is a safety-net healthcare system providing care for low-income, under- and un-insured patients living in Tarrant County, TX. Subjects were patients 18 years and older and diagnosed with cancer between January 1, 2005 and December 31, 2015, identified from the JPS tumor registry. We divided patients into two groups: The intervention group are participants who had one or more visits to the UTSW-MCI survivorship program from November 1, 2012 through December 31, 2016. The control group consisted of patients who did not participate in the cancer survivorship program during that period.

After the groups were identified, a patient level database was created that contained: 1) patient sex, age, race/ethnicity, language, marital status, alcohol use, tobacco use, cancer case class, cancer type, cancer stage, cancer grade, cancer diagnosis year from the tumor registry, 2) ED visits from the EMR and 3) survivorship program services from the MCI delivery database.

The survivorship program offered eleven different types of program services as follows:

- RN Encounter - An Oncology Certified Nurse (OCN) conducts a physical needs assessment based on cancer treatment and health history, including cancer screening and surveillance adherence, creating a survivorship care plan where appropriate.

- SW Encounter - A licensed medical social worker (LMSW), completes a psycho-social evaluation to determine need for care coordination and/or financial assistance.
- 1:1 Exercise session - An American College of Sports Medicine (ACSM) certified Cancer Exercise Trainer designs an individualized safe physical post-treatment activity plan for the participant after reviewing the medical history, exercise history, and fitness goals. The routine provides guidelines for improvement in areas like cardiovascular endurance, muscle strength and endurance, flexibility, range of motion, and balance.
- Nutrition Counseling – A registered dietitian (RD) evaluates the participants' dietary behaviors and needs to provide assistance with making nutritious foods and lifestyles choices, particularly when food security is an issue.
- Midlevel Provider Encounter - Medical consultations provided by a Physician Assistant for the treatment of comorbid conditions, interval testing post treatment, etc.
- Psychology Encounter - A psychologist consults with participants and their families to address psychosocial distress, anxiety and depression.
- Genetic Counseling - A board certified genetic counselor (CGC) assesses the participant's family and personal history along with any screening results, to identify the risk level for cancer and provides genetic testing where appropriate, along with guidance for early detection and prevention measures.
- Group Exercise Session - A safe physical activity designed and coordinated for a group setting by an ACSM certified Cancer Exercise Trainer.
- Support Group Session - Groups led by a licensed medical social worker (LMSW) to provide counseling for specific issues (e.g. smoking cessation, caregiving) or cancer type (prostate, brain, head-and-neck).

- Group Education – RD - A RD provides instruction on healthy food preparation and meal planning in a group setting.
- Group Education – RN – Diagnosis specific education and survivorship care planning provided in a group setting by an OCN.

Each service was designed to help cancer survivors to develop a healthy lifestyle, reduce the risk of recurrence or secondary cancers, and address psychological and social problems as a result of their disease and its treatment. All participants were strongly encouraged to attend 1 RN and 1 SW visit; after which, they were encouraged to engage in program services aligning with the needs identified during the initial encounters.

2.2.2 Statistical analysis

We first compared patient characteristics of program participants to all possible non-participants and to the matched set of non-participant controls using descriptive statistics (number, percent) and chi-square statistics from univariate logistic model.

For each type of program service, we estimated a random-effects, zero-inflated Poisson (ZIP) mixed model to predict the count of visits to that type of program service (Model 1). We had 11 models separately for 11 type of services and for each participant, 11 random effects were estimated correspondingly. We then applied principal component analysis (PCA) to the standardized (subtracted by mean and divided by standard deviation) estimated random effects from ZIP models. We used the selected components obtained from PCA methods as input to a *k*-means clustering algorithm[32] that identified distinct clusters in program utilization.

$$Pr(y_{ij}) = \begin{cases} w_j + (1 - w_j)e^{-\mu_{ij}} & y_{ij} = 0 \\ (1 - w_j)\dfrac{e^{-\mu_{ij}}{\mu_{ij}}^{y_{ij}}}{y_{ij}!} & y_{ij} > 0 \end{cases} \qquad \text{Model 1}$$

$\ln(\mu_{ij}) = \alpha_j + e_{ij} + \ln(t_{ij})$

where $= 1, 2, \dots, n$: participant ID;

$j = 1, 2, \dots, 11$: service type indicator;

y_{ij} : the number of visits of participant i to service j;

w_j : the zero-inflation proportion for service j;

μ_{ij} : mean number of program visits;

α_j: an intercept parameter;

e_{ij}: the random individual effect for participant *i* on type *j* service;

t_{ij} : the offset representing the duration between the individual's starting date of type j service and last observed program date of any type of service. Individual's starting date is defined as follows: if individual i participated in type j service, his starting date is either his enrollment date or the very first observed program date of type j service among all participants (overall first program date of type j service), whichever comes later; if individual i didn't participate in type j service, his starting date is his enrollment data if his last observed program date is before overall first program date of type j service, otherwise his starting date is either his enrollment data or the overall first program date of type j service, whichever comes later.

To study the effect of sociodemographic and clinical factors on program participation, we fitted a multivariate logistic regression model that estimated the expected probability of participation, with all measured covariates. Given the considerable number of missing observations for alcohol use and tobacco use, we imputed missing values using logistic regression. To do so, we built a logistic regression with observed alcohol use as outcome and all other covariates except tobacco use as predictors. This model was then used to predict the missing alcohol use. A similar method was used to impute tobacco use. Because patients can

have more than one primary reportable neoplasm over their lifetime, we describe characteristics of the patients' most recent cancer. Sequence number reflects whether the selected tumor was the patient's first or only tumor, or a second or higher-order tumor.[33] Case class identifies the role of the reporting facility in the patient's diagnosis and treatment. Analytical cases are those diagnosed by or receiving part or all of the first course of treatment at the reporting facility; non-analytical cases were diagnosed and received all of the first course of treatment at another facility.[34] We used these propensity scores to match each program participant to 3 non-participants, allowing us to estimate effects of program participation with minimal confounding bias.

To study the effect of program participation on frequency of ED visits, we also applied the random-effects ZIP model to the dataset. We estimated two models. For the first (Model 2), we examined whether the count of ED visits differed by any program participation. We applied the model to the propensity score matched data, with count of ED visits as the outcome and participation status as the predictor. The participation status is always 0 for non-participants; for a participant, the status switches from 0 to 1 at the time of first participation in a program service. In this model, the participation status could affect whether the survivors has ED visits and also the rates of their ED visits if they had any.

$$Pr(y_i) = \begin{cases} w + (1-w)e^{-\mu_i} & y_i = 0 \\ (1-w)\dfrac{e^{-\mu_i}\mu_i{}^{y_i}}{y_i!} & y_i > 0 \end{cases} \qquad \text{Model 2}$$

$$\ln(\mu_i) = \alpha_1 + \beta_1 X_i + e_i + \ln(t_i)$$

$$\ln\left(\frac{w}{1-w}\right) = \alpha_2 + \beta_2 X_i$$

where $i = 1, 2, \cdots, n$: patient ID;

y_i : the number of ED visits of patient i;

w : the zero-inflation proportion;

μ_i : mean number of ED visits;

α_1: the intercept parameter to predict the mean number of ED visits;

β_1 : the fixed participation effect to predict the mean number of ED visits;

X_i : the participation status – Participants before-program and non-participants: 0; participants after-program: 1;

e_i: the random individual effect;

t_i : the offset: for non-participants, the offset is the duration between the first and last recorded ED visit dates; for participants, there are two stages: before-program and after-program. In the before-program stage, the patients have not yet initiated participation in the survivorship program; the offset for this stage is the duration between the first ED visit date and the first survivorship program date. In the after-program stage, the patients have started participating the program; the offset is the duration between the first program date and the last observed date, either the last program participation date or the last ED visit date, whichever comes later;

α_2: the intercept parameter to predict the zero-inflation proportion;

β_2 : the fixed participation effect to predict the zero-inflation proportion.

For the second model (Model 3), we further examined the program effect on ED utilization by membership in the program participation clusters. We applied this model only to survivorship program participants. It took outcome as the count of ED visits and an indicator of cluster membership as predictor, where each patient was classified into one cluster.

$$Pr(y_i) = \begin{cases} w + (1-w)e^{-\mu_i} & y_i = 0 \\ (1-w)\dfrac{e^{-\mu_i}\mu_i{}^{y_i}}{y_i!} & y_i > 0 \end{cases} \qquad \text{Model 3}$$

$$\ln(\mu_i) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + e_i + \ln(t_i)$$

where $= 1, 2, \cdots, n$: participant ID;

y_i : the number of ED visits of participant i;

w : the zero-inflation proportion;

μ_i : the mean number of ED visits;

α: the intercept parameter to predict the mean number of ED visits;

$\beta_1, \beta_2, \beta_3, \beta_4$: the fixed effects of cluster 1-4 to predict the mean number of ED visits;

$X_{i1}, X_{i2}, X_{i3}, X_{i4}$: the cluster status indicator; $X_{ij} = 1$, if participants is in cluster j ; otherwise, $X_{ij} = 0, j = 1, 2, 3, 4$.

e_i: the random individual effect;

t_i : before-program: the offset is the duration between the first ED visit date and the first survivorship program date; after-program: the offset is the duration between the first program date and the last observed date, either the last program participation date or the last ED visit date, whichever comes later.

2.3 Results

Among 8,435 cancer survivors, 467 (5.5%) participated in the survivorship program. The average age (interquartile range, IQR) is 51.9 (47-59) for participants and 54.8 (47-63) for non-participants, additional characteristics by participation status are in Table 2-1. Cancer survivors who are female, younger, Hispanic or black, have quit smoking, and those with certain cancer types are significantly more likely to participate in the survivorship program (Table 2-1). We

then matched each participant to three non-participants based on the propensity score obtained from the multivariate logistic regression model. The matching balanced the covariates between the participation groups ($p > 0.05$ for all covariates after matching). All remaining analyses compared participants to matched non-participants.

Table 2-1. Summary statistics of participants and non-participants

Covariates	Categories	Participants	Non-participants	Matched Non-participants	P value* before matching[a]	P value* after matching[b]
Total		467	7968	1401		
Sex	Female	331(70.9)	4296(53.9)	976(69.7)	<0.001	0.62
	Male	136(29.1)	3672(46.1)	425(30.3)		
Age	18-39	55(11.8)	996(12.5)	177(12.6)	<0.001	0.66
	40-54	213(45.6)	2576(32.3)	593(42.3)		
	55-64	167(35.8)	2887(36.2)	525(37.5)		
	65+	32(6.8)	1509(18.9)	106(7.6)		
Race	Hispanic	171(36.6)	1794(22.5)	498(35.5)	<0.001	0.98
	Non-Hisp White	134(28.7)	3740(46.9)	406(29.0)		
	Non-Hisp Black	149(31.9)	1962(24.6)	456(32.5)		
	Non-Hisp Other	13(2.8)	472(5.9)	41(2.9)		
Language	English	6663(83.6)	1085(77.4)	358(76.6)	<0.001	0.88
	Spanish	97(20.8)	916(11.5)	277(19.8)		
	Other	12(2.6)	389(4.9)	39(2.8)		
Marital Status	Single	2959(37.1)	499(35.6)	167(35.8)	0.044	0.95
	Married	174(37.2)	2818(35.4)	530(37.8)		
	Separated	22(4.7)	213(2.7)	63(4.5)		
	Divorced	71(15.2)	1166(14.6)	224(16.0)		
	Widow	24(5.1)	587(7.4)	66(4.7)		
	Unknown	9(1.9)	225(2.8)	19(1.4)		
Alcohol	Current	89(19.0)	1875(23.5)	272(19.4)	<0.001	0.99
	Previous	9(1.9)	504(6.3)	27(1.9)		
	Never	369(79.0)	5589(70.1)	1102(78.6)		
Tobacco	Current	115(24.6)	2928(36.7)	351(25.0)	<0.001	0.81
	Previous	80(17.1)	1470(18.4)	256(18.3)		
	Never	272(58.2)	3570(44.8)	794(56.7)		
Sequence number	First or only cancer	429(91.9)	6815(85.5)	1275(91.0)	<0.001	0.57
	Second or higher order cancer	38(8.1)	1153(14.5)	126(9.0)		

Case class	Analytic	417(89.3)	6615(83.0)	1246(88.9)	<0.001	0.83
	Non-analytic	50(10.7)	1353(17.0)	155(11.1)		
Cancer type	Breast	143(30.6)	970(12.2)	401(28.6)	<0.001	1.0
	Colon and Rectum	42(9.0)	759(9.5)	122(8.7)		
	Corpus and Uterus	29(6.2)	253(3.2)	90(6.4)		
	Kidney and Renal	24(5.1)	301(3.8)	73(5.2)		
	Leukemia	10(2.1)	95(1.2)	29(2.1)		
	Liver	10(2.1)	326(4.1)	28(2.0)		
	Lung and Bronchus	28(6.0)	994(12.5)	87(6.2)		
	Lymphoma-NHL	20(4.3)	208(2.6)	62(4.4)		
	Myeloma	11(2.4)	119(1.5)	31(2.2)		
	Oral Cavity Pharynx	10(2.1)	343(4.3)	32(2.3)		
	Prostate	30(6.4)	376(4.7)	106(7.6)		
	Vagina, Vulva, Ovary	14(3.0)	222(2.8)	40(2.8)		
	Other	96(20.6)	3002(37.7)	300(21.4)		
Stage	*In situ*	31(6.6)	481(6.0)	91(6.5)	<0.001	0.97
	Localized	167(35.8)	2315(29.0)	525(37.5)		
	Regional	135(28.9)	1684(21.1)	392(28.0)		
	Distant	93(19.9)	2144(26.9)	277(19.8)		
	Other	41(8.8)	1344(16.9)	116(8.3)		
Grade	Poor	101(21.6)	1265(15.9)	298(21.3)	<0.001	0.97
	Moderate	137(29.3)	1775(22.3)	409(29.2)		
	Well	60(12.8)	731(9.2)	192(13.7)		
	Other	169(36.2)	4197(52.7)	502(35.8)		
Cancer Diagnosis Year	Continuous				<0.001	0.85

***** Two-sided P value calculated from univariate logistic model by Wald Chi-squared test; [a]Comparing participants to all non-participants; [b]Comparing participants to nonparticipants (1:3) matched on propensity scores of all measured covariates.

We investigated the participation pattern of patients according to their service type and frequency. We applied PCA to the estimated patient random effects from the ZIP model. The first four components explained roughly 80% of the variability in types and frequency of program services received, so we applied *k*-means clustering on these components, identifying four clusters as the optimal solution. The clusters —of sizes 93 (20%), 130 (28%), 198 (42%), and 46(10%) — appear in Figure 2-1. We calculated the proportion of participants attending the different type of service as well as the average number of visits across clusters (Figure 2-2).

Among all participants, the approximate proportion of participation in the 1:1 exercise is 50%, 20% for different group-activity sessions and psychology encounters, 45% for nutrition counseling, over 80% for RN or SW encounter, and less than 5% in mid-level provider encounter and genetic counseling. The average number of visits of any type of service is 10.1, respectively 5 for 1:1 exercise session, 1 for nutrition counseling, psychology encounter, RN or SW encounter, and less than 1 for the other types of services. On average, participants spent 161 days on the program.

The clustering identifies different participation patterns: Patients in Cluster 1 received services related to exercise and diet lifestyle behaviors; 97% of the participants in Cluster 1 participated in a 1:1 exercise session where the average session count exceded 11; 86% participated in nutrition counseling with an average participation count of 2.6. Cluster 2 engaged in multiple types of sessions: 61% participated in 1:1 exercise with an average visit count of 7.4; 55% of them attended nutritional counseling, and around 40% participated in group-activity sessions. Cluster 3 opted for sessions involving interaction with nurses (97%) and social workers (90%); the frequencies for other types of service were all less than average. Cluster 4 gravitated toward group-activity sessions; rather than individual RN or SW encounters (around 25%), they preferred group-activity sessions (over 80% in all different group-activity session). The average number of visits of any type of service is 19.7, 15.3, 3.3 and 5.5 for Cluster 1, 2, 3, 4 respectively with average participation period as 165, 405, 32 and 18 days long respectively.

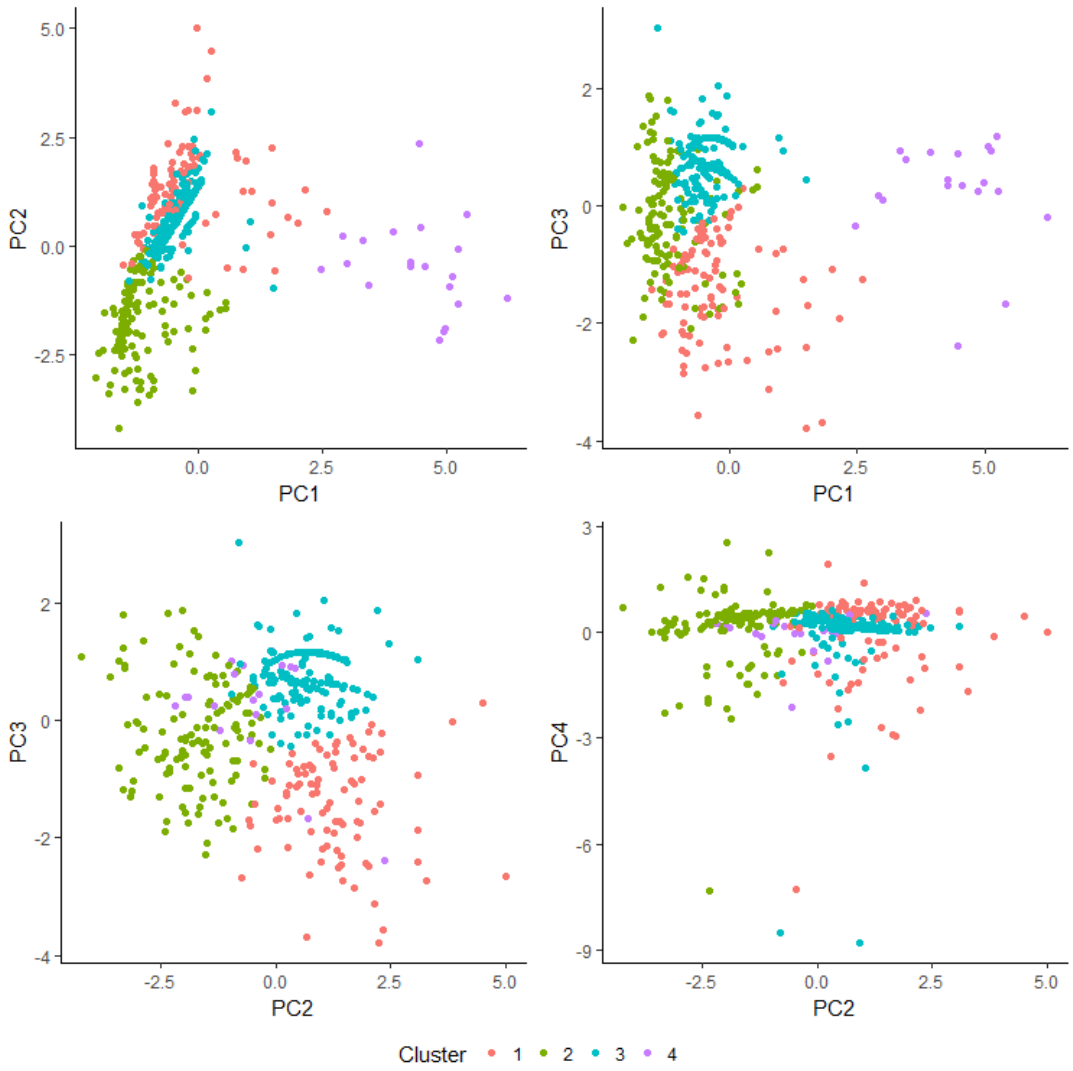

Figure 2-1. k-means clusters visualized in principal components coordinates. (PC1-PC4: the first four principal components)

Next, we analyzed ED visits. Most survivors (85%) had no ED visits, including non-participants (85%), participants before program (84%) and participants after the program (86%). Corresponding to non-participants, participants before the program and participants after program, the proportion of survivors who had 1 to 4 ED visits is 9.3%, 11.9% and 10.1%; the proportion of survivors who had more than 4 ED visits is 4.7%, 3.9% and 3.8%, respectively.

We used the mixed ZIP model to examine the impact of the survivorship program on ED visits. We first estimated the effect of any program participation. The ratio of ED counts of

participants in the program to that of participants before the program and non-participants was 0.63 (95% CI: 0.48 – 0.81), which means participation was associated with a 37% reduction in the number of ED visits (Figure 2-3, left panel). These models kept the zero-inflation proportion the same regardless of participation status; allowing it to differ by participation status gave similar results (not shown). We then examined, among participants, whether cluster membership had an effect on ED visits. The number of ED visits after program to that before program is 0.44 (95% CI: 0.28 – 0.72), 0.51 (95% CI: 0.36 – 0.74), 1.26 (95% CI: 0.88 – 1.82) and 0.88 (95% CI: 0.48 – 1.59) respectively for Clusters 1, 2, 3 and 4 (Figure 2-3, right panel). These results demonstrate that those who intensively participated in 1:1 exercise and nutrition counseling sessions (Cluster 1), as well as those who participated in a multiple, mixed types of sessions (Cluster 2) had significantly lower ED visit rate after vs. before program participation.

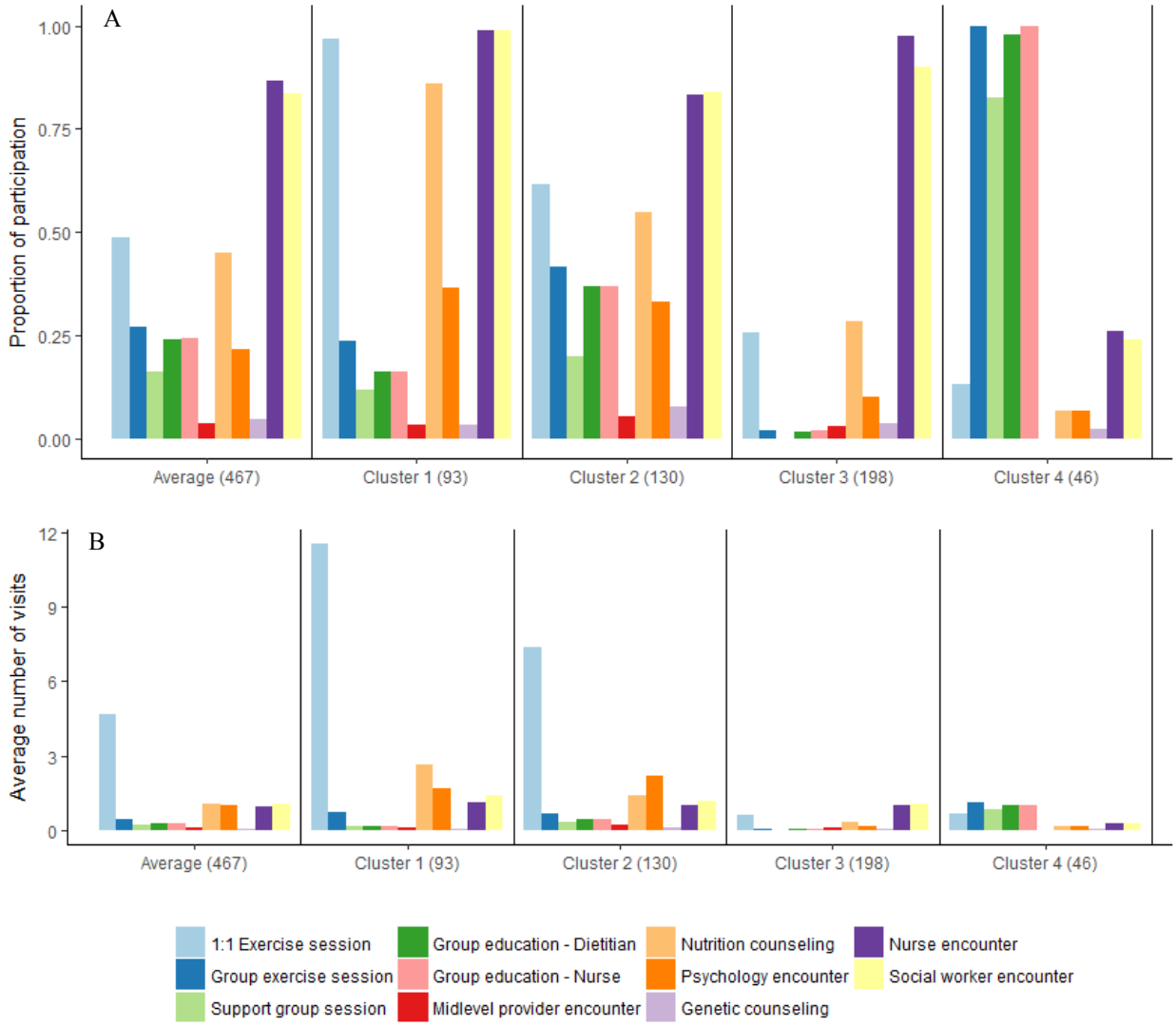

Figure 2-2. Participation in different type of survivorship services by cluster type. Panel A shows the proportion of all visits by visit type and Panel B shows the average number of visits by visit type.

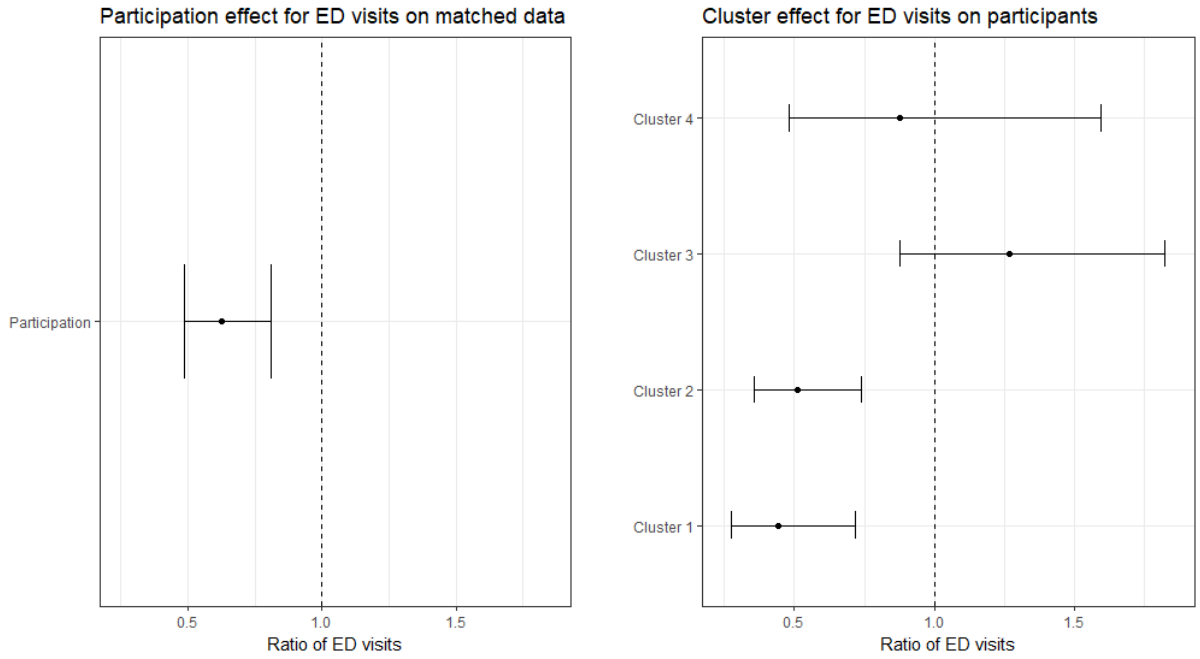

Figure 2-3. Effect of participation in the survivorship program on ratio (the number of ED visits in the intervention group to that in the reference group) of emergency department (ED) visits. Panel A shows the effect of participants in the program vs. participants before program and matched non-participants (reference group). Panel B shows the effect of different clusters of participation among participants only, comparing the ratio of ED visits after program participation to the ratio before program participation (reference group). A ratio of 1.0 indicates that the number of ED visits in the intervention group is the same as the reference group. A ratio less than 1.0 demonstrates fewer ED visits in the intervention vs. reference group.

2.4 Discussion

In 1986, the National Coalition for Cancer Survivorship (NCCS) was founded to establish programs to help cancer survivors deal with the long-term effects of their disease and its treatment.[35] The Institute of Medicine (IOM) has provided guidance for the implementation of comprehensive cancer survivorship care plans.[28,29,36] Researchers have studied cancer survivorship from a range of perspectives.[37] Wattchow *et al*.[38] found that colon cancer patients with follow-up led by surgeons or general practitioners experience similar outcomes regarding quality of life, anxiety and depression, and patient satisfaction. Knowles *et al*.[39] demonstrated

that a nurse-led model for colorectal cancer survivors was safe, efficient and cost-effective. In a randomized trial, Grunfeld *et al*.[40] found that receiving a survivorship care plan (SCP) did not improve breast cancer survivors' distress. Nevertheless, much remains unknown, and there is a pressing need for evidence-based guidance regarding the types and frequency of survivorship services, along with which models of survivorship care improve patient outcomes.

Prior studies investigated the effects of survivorship program participation and number of cancer follow-up providers on the occurrence of ED visits.[41,42] To our knowledge, there has been no study classifying survivorship program participants according to their patterns of use of the different types of service, not to mention analyzing the association between these patterns and the frequency of outcomes such as ED visits. Our study also goes beyond prior studies by focusing on low-income, under- and uninsured patients, including multiple disease sites, and by using propensity scores to account for differences between participants and non-participants.

We demonstrated participation in the MCI survivorship program is significantly associated with lower rates of ED visits. The exact mechanisms for this effect are not known. It is possible the services provided may reduce ED visits by lessening the severity of medical conditions needing urgent care. It is also possible participants are seeking advice or care from program personnel before their condition worsens to the point where an ED visit is necessary. Several studies have shown the frequency of ED visits among cancer survivors exceed those of the general population.[43-45] Moreover, Panattoni *et al*.[46] found 49.8% of ED visits in a commercially insured oncology population had a potentially preventable cancer-related diagnosis with a related median reimbursement of $1,029 per ED visit. Our findings suggest a survivorship program can provide an opportunity to prevent avoidable ED visits. The cost-effectiveness of this approach is unknown and deserves further study[47].

We also demonstrated distinct utilization patterns among survivors. We identified four clusters, perhaps reflecting differences in patient interests or preferences for services, or perhaps the availability and accessibility of different services. Participants in two of the four clusters had reductions in ED visit rates by roughly half after beginning participation. This suggests the most effective program components may be associated with these clusters: 1:1 exercise sessions, nutrition sessions, etc.

Our study had several strengths. Linking the three data sources allowed us to analyze patterns and effectiveness of cancer survivorship program use, to characterize and classify survivorship program usage, and to evaluate the effectiveness of program participation using ED visits. Our study also has several limitations. Because survivorship programs vary across institutions, our results may not be relevant to different programs. Also, ED visits were captured within a single healthcare system and survivors could have been seen elsewhere. However, because this population is low-income, under- or uninsured, and received all or most of their cancer diagnosis and treatment at JPS, and because JPS is the only integrated safety-net system in the county, this seems like an unlikely scenario. In addition, as an observational study, unobserved confounding may bias our results.

In conclusion, our study provides useful information for health care providers and cancer centers to guide development, implementation, and outcome evaluations for future cancer survivorship programs. These findings also provide evidence for insurers and payors to design benefits, policies, and reimbursement mechanisms to facilitate coverage. Future studies will evaluate the effect of survivorship program participation on other health service utilization, like the uptake of cancer screening and surveillance tests.

CHAPTER 3

Among newly diagnosed cancer patients, those who survived a previous cancer have been reported to live longer than those with no previous cancer. Possible explanations for this phenomenon are lead-time bias and true biological effects. We propose a discrete competing-risk model with adjustment for lead-time bias to describe the effect of a previous, non-lung cancer diagnosis on the cause-specific survival of patients with lung cancer. We assume that the observed survival for patients with previous cancer is the sum of lead time and post-lead-time survival. We describe the former with a negative binomial distribution, and the latter with a discrete cause-specific hazard, modeled as the inverse logit of a spline function on time. We assume that post-lead-time survival in patients with no previous cancer differs from that in cancer survivors by an odds ratio parameter. We applied our model to propensity score-matched linked SEER–Medicare data. We estimate mean lead time to be less than one month at all lung cancer stages; the effect of including lead time on estimates of group differences is modest. Patients with a previous cancer had significantly lower hazard for lung-cancer death and non-cancer death than patients without a previous cancer. Under a competing-risk model, lead-time bias is modest, and does not explain differentials in cause-specific survival between lung cancer patients with and without previous cancer. Patients with previous cancer have reduced hazards of both lung-cancer and non-cancer mortality, but many die of their original tumors.

3.1 Introduction

The number of cancer survivors is growing rapidly, leaving more of them at risk of a second primary cancer.[1,25] Lung and bronchus cancers are the leading cause of cancer-related death in the US.[1] Roughly 20% of patients aged 65 and over with a new diagnosis of a respiratory cancer have experienced a previous cancer of some other kind.[25]

Several studies have investigated the impact of a previous cancer diagnosis on survival in lung cancer and other diseases. Curiously, among lung cancer patients older than 65 years, those with a history of a previous cancer had similar or better all-cause survival, depending on the stage of lung cancer. Across all stages of lung cancer, those with a previous cancer had on average longer lung cancer-specific survival.[2-4]

A possible explanation for this observation is lead-time bias, which occurs when surveillance advances the date of diagnosis of a disease. In cancer, we typically define lead time to be the difference between the date of diagnosis when observed through screening and the (latent) date at which the diagnosis would have occurred without screening.[48-51] For lung cancer patients with a history of cancer, it is possible that a lead time could arise through enhanced surveillance. That is, cancer survivors who harbor an as yet undiagnosed lung tumor could have that tumor discovered early through additional testing they undergo as follow-up to their previous cancer.[52] Lead-time *bias* occurs when the diagnosis date is advanced such that mean survival time appears to be longer, even when no survival advantage exists.

In a previous article, we proposed a statistical model for lead-time bias and survivorship, and observed that even after accounting for lead time, time to death from lung cancer was significantly longer for patients with a previous cancer diagnosis.[52] Our method of analysis did not, however, account for competing risks of death, in that it treated all non-lung cancer deaths

as censoring events. This approach has several weaknesses: First, it implicitly assumes that the latent times of death from each cause are statistically independent, an assumption that is impossible to evaluate robustly.[53] Indeed, it is more plausible that latent death times are positively correlated, reflecting different levels of frailty, in which case estimates of survival assuming independence are biased upward.[54] Second, the survival curve assuming latent, independent death times estimates the survival curve that would occur if we could eliminate the competing causes of death. This is problematic because there is no reason to believe that eliminating non-lung cancer causes of death would leave lung cancer death times unaffected. Thus, not only the validity but the relevance of these estimates is questionable. Finally, we note that our estimates of mean lead time differed by survival outcome (death from any cause vs. death from lung cancer), possibly reflecting these biases and suggesting the need for a comprehensive approach to modeling lead time and cause-specific survival.[52]

As early as 1957, Cornfield[55] observed that the existence of competing risks complicates the interpretation of cause-specific mortality rates. Prentice *et al*.[56] proposed to study the interrelations among competing causes of failure through cause-specific hazard functions, which one can estimate without the need for unverifiable assumptions. Later, Fine and Gray[18] proposed a proportional hazards model to estimate the cumulative incidence of a competing risk. Yet Austin and Fine[57] observed that despite these advances, fewer than 20% of randomized trials in which competing risks data arise present a competing-risks analysis.

Although it is typical to treat survival data as though they are continuous, in fact they are, like all data, essentially discrete. With Surveillance, Epidemiology, & End Results (SEER)-Medicare cancer survival data, for example, diagnosis and survival dates are only accurate to the nearest month. Thus, many such observations may have equal, or "tied", survival times, a

circumstance that complicates the analysis of data that we assume to be continuous. Tutz[58] proposed a discrete competing-risks model that one can estimate in the framework of the generalized linear model. Ambrogi *et al*.[59] proposed to estimate cumulative incidences through multinomial logit regression analysis of discrete cause-specific hazards.

In this article, we propose a discrete competing-risk model to estimate the cause-specific hazard for lung cancer patients with a previous diagnosis of cancer. We adjust our results by assuming that lung cancer diagnosis time in survivors of a previous cancer is potentially subject to a lead-time bias. We model the cause-specific hazards as inverse logits of linear splines on time, assuming that a previous cancer diagnosis affects the cause-specific hazard through an odds ratio factor. We assume that the lead time follows a negative binomial distribution and is independent of the latent survival time. We estimate the mean lead time, the odds ratios of the cause-specific hazards, and the cause-specific cumulative incidence rates by maximum likelihood, applying our method to linked SEER-Medicare data on lung cancer patients.

3.2 Methods

3.2.1 Data Source

We used linked 1992–2011 National Cancer Institute SEER program files and 1991–2013 Medicare claims files from the Centers for Medicare & Medicaid Services. The Institutional Review Board of the University of Texas Southwestern Medical Center approved our study.

3.2.2 Study Population

We included patients older than 65 years with primary lung cancer diagnosed between 2000 and 2011. All patients had full coverage of Medicare Parts A and B from 1 year before to 1 year after the lung cancer diagnosis. We included only patients with either non–small cell (NSCLC) or small cell (SCLC) lung cancer histology. To ensure complete claims data, we excluded

patients who participated in health maintenance organizations and those with only autopsy or death certificate records. We omitted patients with incomplete diagnosis or death dates or discrepancies in SEER and Medicare birth dates of a year or more. We also excluded those who developed another cancer after the index lung cancer.

We divided the patients into two groups: Those with a history of cancer (the Previous group) and those without (the No-Previous group). The Previous group included those who had only one previous, invasive, primary cancer that was not a lung cancer. We stratified patients by American Joint Committee on Cancer lung cancer stage, grouping them into stages I&II, III, and IV, and excluding the heterogeneous "unstaged" stratum.

3.2.3 Measures

We assumed three possible competing causes of death: The previous cancer (possible for the Previous group only), the index lung cancer, and non-cancer causes. We measured survival as the interval in months between the lung-cancer diagnosis and the date of death derived from SEER.

To reduce confounding of previous cancer status with other potential correlates of mortality, we created a set of patients in the No-Previous group who were matched to the Previous group members. We matched on a propensity score that predicted previous cancer status from available potential confounders in the SEER-Medicare database: Age, sex (F,M), race/ethnicity (white, black, Hispanic, other), marital status (married, separated/divorced/widowed, single, unknown), histology (SCLC, NSCLC-adenocarcinoma, NSCLC-squamous, NSCLC-other), Charlson comorbidity score (0, 1, 2+, not available), Medicaid status (Y, N), and lung cancer treatment (surgery only, chemotherapy only, radiation only, ≥2 treatments, no surgery/chemo/radiation).

3.2.4 Statistical Analysis

We propose a discrete competing-risk model to describe the cause-specific hazards in the two groups defined by prior cancer status. We assume that there is a standard survival measure — the time from clinical diagnosis to death — that we denote *post-lead-time survival* and label X_N for subjects in the No-Previous group and X_P for subjects in the Previous group. Because there is, by definition, no possibility of lead-time bias in the No-Previous group, we observe X_N directly. In the Previous group, the observed survival is the sum of two independent components: The notional post-lead-time survival X_P and a random lead time $T \geq 0$ that is a consequence of additional surveillance that patients undergo as a result of the previous cancer. We see neither X_P nor T directly; rather we observe their sum, which we denote $Z = X_P + T$.

The cause-specific hazard for cause r in the No-Previous group is defined as $h_{Nr}(x) = \Pr(X = x, R = r | X \geq x)$. We model it with a linear spline on the logit scale:

$$h_{Nr}(x) = \frac{\exp[\eta_{Nr}(x)]}{1 + \sum_{r=1}^{R} \exp[\eta_{Nr}(x)]}, \qquad (1)$$

$$\eta_{Nr}(x) = \beta_{0r} + \beta_{1r} x + \sum_{j=2}^{m_r+1} \beta_{jr} \left(x - k_{r\,j-1}\right)_+, \qquad (2)$$

where m_r is the number of knots; β_{jr}, $j = 0, \ldots, m_r$ are the spline coefficients; $0 < k_{r_1} < \cdots < k_{r_{m_r}}$ are the spline knots; and $(u)_+ = \max(0, u)$. We model the cause-specific hazard in the Previous group as $\eta_{Pr}(x) = \eta_{Nr}(x) + \gamma_r$; with this specification, $\mathrm{OR}_r = \exp(\gamma_r)$ is the odds ratio of hazards comparing the Previous to the No-Previous group. The overall hazard at time x in the No-Previous group is $h_N(x) = \sum_{r=1}^{R} h_{Nr}(x)$; the overall survival function is $S_N(x) = \Pr[X_N \geq x] = \prod_{j<x} [1 - h_N(j)]$; the cause-specific probability mass function is $f_{Nr}(x) = h_{Nr}(x) S_N(x)$; and the cumulative incidence rate is $F_{Nr}(x) = \sum_{j<x} f_{Nr}(j)$.

We calculate the probability mass function of the latent X_P similarly. To derive a distribution for Z, the observable time from diagnosis to death in the Previous group, we first assume that lead time T follows the negative binomial distribution $T \sim NB(\rho, \sigma)$, with probability mass function parameterized as $f_T(t; \rho, \sigma) = \frac{\Gamma(t+\rho)}{\Gamma(\rho)\Gamma(t+1)} \sigma^\rho (1-\sigma)^t$ for $\rho>0$ and $0<\sigma\leq 1$. Then the probability mass function of Z is the convolution of the densities of X_P and T: $f_{Zr}(z) = \sum_{t=0}^{z} f_T(t) f_{Pr}(z-t)$. The cumulative incidence rate is therefore $F_{Zr}(z) = \sum_{j<z} f_{Zr}(j)$, and the overall survival is $S_Z(z) = 1 - \sum_{r=1}^{R} F_{Zr}(z)$.

The loglikelihood for the matched dataset is

$$\ln L(\beta, \gamma, \rho, \sigma) = \sum_{i=1}^{n_N} \left[\sum_{r=1}^{R} d_{iNr} \ln f_{Nr}(x_i; \beta) + (1 - \sum_{r=1}^{R} d_{iNr}) \ln S_N(x_i; \beta) \right] + \sum_{i=1}^{n_P} \left[\sum_{r=1}^{R} d_{iPr} \ln f_{Zr}(z_i; \beta, \gamma, \rho, \sigma) + (1 - \sum_{r=1}^{R} d_{iPr}) \ln S_Z(z_i; \beta, \gamma, \rho, \sigma) \right], \quad (3)$$

where n_N and n_P are the numbers of patients in No-Previous and Previous groups, respectively; and d_{iNr} and d_{iPr} are indicators of whether patient i died from cause r in the No-Previous group and Previous group, respectively.

We choose the spline knot locations by lasso variable selection. We assume for each cause of death that the knots are the same in the Previous and No-Previous groups. For lung-cancer death and other-cause death, we identify knots using data from the No-Previous group only. For previous-cancer death, we find knots using data from the Previous group only. Initially, we set knots at every fifth centile of the empirical distribution function of the survival data. Using Equation (2), we fit the linear spline on the empirical net hazard, using lasso variable selection to choose at most another two knots besides 1 and 5th centile.

We estimate the model parameters by using generic optimization functions in the R statistical language.[17] Having obtained maximum likelihood estimates of the parameters, we calculate estimates and confidence intervals for the odds ratios of the cause-specific mortality hazards, the mean lead time, and the cumulative incidence rate (CIR) for each event cause. The Online Supplement provides additional details.

We conducted a range of sensitivity analyses: The basic model assumes three causes of death: Lung cancer, prior cancer, and other. We also assumed a two-cause model in which we grouped together the prior cancer and other causes as a single cause of death. To examine the impact of lead-time bias in the two-cause model, we estimated the odds ratios assuming that there was no lead-time bias.

3.3 Results

Among 173,635 eligible lung cancer patients, 42,994 (24.8%) were stage I&II; 50,084 (28.8%) were stage III; and 80,557 (46.4%) were stage IV. The proportions of lung cancer patients who had only one previous cancer were 15.3%, 12.5% and 12.0% for stages I&II, III, and IV, respectively. Before matching, previous cancer prevalence differed across measured sociodemographic and clinical covariates; it was higher ($P < 0.0001$, Tables 3-1, S1, S2) in lung cancer patients who were older, male and without Medicaid. A 1:1 propensity score matching eliminated these imbalances. The remaining analyses use this matched dataset.

Table 3-2 displays proportions of patients according to cause of death and stage. Combining the Previous and No-Previous groups, the proportion who died of any cause increased as stage increased: 65.9%, 91.6%, and 97.1% for stages I&II, III, and IV, respectively. For death from lung cancer the trend was similar: 39.0%, 71.6%, and 79.0% for stages I&II, III, and IV, respectively. As more patients died of lung cancer in the higher stages, the proportion who died

from non-cancer causes declined, 21.1%, 14.2%, and 10.6% for stages I&II, III, and IV, respectively. The proportions of overall, lung-cancer, and non-cancer deaths were all higher in the No-Previous group. In the Previous group, the proportion of patients who died from the previous cancer also increased as lung cancer stage increased: 11.5%, 11.6%, and 14.9% for stages I&II, III, and IV, respectively.

We computed estimates of the model parameters using the matched data. Figure 3-1 shows estimated CIRs from the three-cause model. It is clear that as stage advanced, the CIR for lung-cancer death (dashed line) increased, while the CIR for non-cancer death (dot dashed line) decreased. The No-Previous group (black lines) had higher CIR for both lung-cancer death and non-cancer death, compared to the Previous group (red lines). For the Previous group, in stage I&II, the CIR for previous-cancer death (dotted line) is lower than that for non-cancer death; in stage III, their difference decreased; in stage IV, the order is reversed as CIR for previous-cancer death is higher than non-cancer death. For the No-Previous group, the CIR for previous-cancer death is defined to be 0. We calculated the CIR for death from any cause by summing the cause-specific CIRs. In stage I&II and stage III, the CIR of overall death (solid line) for the Previous group is higher than that for the No-Previous group. In stage IV, the CIR of overall death for the Previous group is slightly less than that for the No-Previous group.

Table 3-1. Characteristics of patients with stage I&II lung cancer.

	n (%)			P value	
	Previous	No-Previous	Matched No-Previous	Unmatched	Matched
Total (*n*)	6594	36400	6594		
Age				<0.0001	0.88
<75	2811 (42.6)	17491 (48.1)	2785 (42.4)		
75-85	3093 (46.9)	15766 (43.3)	3107 (47.1)		
≥85	690 (10.5)	3143 (8.6)	702 (10.6)		
Sex				<0.0001	0.17
Female	2799 (42.4)	18220 (50.1)	2877 (43.6)		
Male	3795 (57.6)	18180 (49.9)	3717 (56.4)		

Race				0.052	0.99
White	5839 (88.6)	32006 (87.9)	5848 (88.7)		
Black	472 (7.2)	2541 (7.0)	468 (7.1)		
Hispanic	50 (0.8)	341 (0.9)	48 (0.7)		
Other	233 (3.5)	1512 (4.2)	230 (3.5)		
Marital Status				0.00032	0.77
Married	3635 (55.1)	19015 (52.2)	3587 (54.4)		
Sep/div/wid	2324 (35.2)	13655 (37.5)	2382 (36.1)		
Single	416 (6.3)	2456 (6.7)	409 (6.3)		
Unknown	219 (3.3)	1274 (3.5)	216 (3.3)		
Histology				<0.0001	0.98
Adenocarcinoma	3294 (50.0)	16394 (45.0)	3302 (50.1)		
Squamous	1927 (29.2)	11363 (31.2)	1911 (29.0)		
Small cell	230 (3.5)	1369 (3.8)	237 (3.6)		
NSCLS/other	1143 (17.3)	7274 (20.0)	1144 (17.3)		
Charlson Score				0.00083	0.72
0	2490 (37.8)	13047 (35.8)	2465 (37.4)		
1	2026 (30.7)	11494 (31.6)	2058 (31.2)		
2+	1895 (28.7)	10570 (29.0)	1906 (28.9)		
Not available	183 (2.8)	1289 (3.5)	165 (2.5)		
Medicaid				<0.0001	1.0
Yes	851 (12.9)	6151 (16.9)	851 (12.9)		
No	5743 (87.1)	30249 (83.1)	5743 (87.1)		
Lung cancer treatment				<0.0001	0.86
Surgery only	2507 (38.0)	15464 (42.5)	2518 (38.2)		
Chemo only	191 (2.9)	857 (2.4)	195 (3.0)		
Radiation only	1041 (15.8)	5261 (14.5)	1072 (16.3)		
≥2 treatments	1421 (21.5)	6980 (19.2)	1374 (20.8)		
No surg/chemo/rad	1434 (21.7)	7838 (21.5)	1435 (21.8)		

* Characteristics of patients with stage III and IV lung cancer appear in Tables S1 and S2.

Table 3-2. Mortality fraction by lung cancer stage and cause of death.

			Deaths by cause (%)			
Stage	Group	Patients	Overall	Lung cancer	Non-cancer	Previous cancer
I&II	Total	13188	8697 (65.9)	5150 (39.0)	2790 (21.1)	
	No-Previous	6594	4417 (67.0)	2824 (42.8)	1593 (24.2)	--
	Previous	6594	4280 (64.9)	2326 (35.3)	1197 (18.2)	757 (11.5)
III	Total	12500	11450 (91.6)	8955 (71.6)	1773 (14.2)	
	No-Previous	6250	5761 (92.2)	4746 (75.9)	1015 (16.2)	--
	Previous	6250	5689 (91.0)	4209 (67.3)	758 (12.1)	722 (11.6)
IV	Total	19430	18864 (97.1)	15354 (79.0)	2058 (10.6)	
	No-Previous	9715	9472 (97.5)	8153 (83.9)	1319 (13.6)	--
	Previous	9715	9392 (96.7)	7201 (74.1)	739 (7.6)	1452 (14.9)

Table 3-3 presents estimates of the mean lead time and the odds ratios of the cause-specific mortality hazards comparing the Previous group to the No-Previous group; we omit the OR for risk of death from previous cancer, which is by definition infinity. The estimated mean lead time

is less than 1 month: 0.53, 0.96, 0.48 months for stages I&II, III and IV respectively. It is interesting that the longest lead time is found in stage III, possibly because by the time stage I&II lung cancer becomes clinically detectable, it has already progressed to stage III.[60] After adjustment for lead-time bias, the odds ratios are significantly less than 1 for lung-cancer death (OR_l) and non-cancer death (OR_{nc}). This suggests that patients in the Previous group are relatively resistant to mortality from the subsequent lung cancer and to non-cancer causes, experiencing their greatest risk of death from previous tumors. Comparing the top and bottom panels of the table, it is apparent that estimated cause-specific odds ratios are robust to inclusion of lead time in the model.

The estimated cumulative incidence rate, mean lead time and odds ratios for the two-cause data appear in Figure S1 and Table S3 in the Online Supplement. In the two-cause data, estimates of mean lead time are modest and are similar to those in the three-cause data, as are values of OR_l. Combining previous-cancer and non-cancer death, odds ratios for other-cause death for the previous group versus non-previous group are significantly larger than one at all stages of lung cancer. This suggests that the competing risk from a previous cancer accounted for the reduction of the hazard of lung cancer.

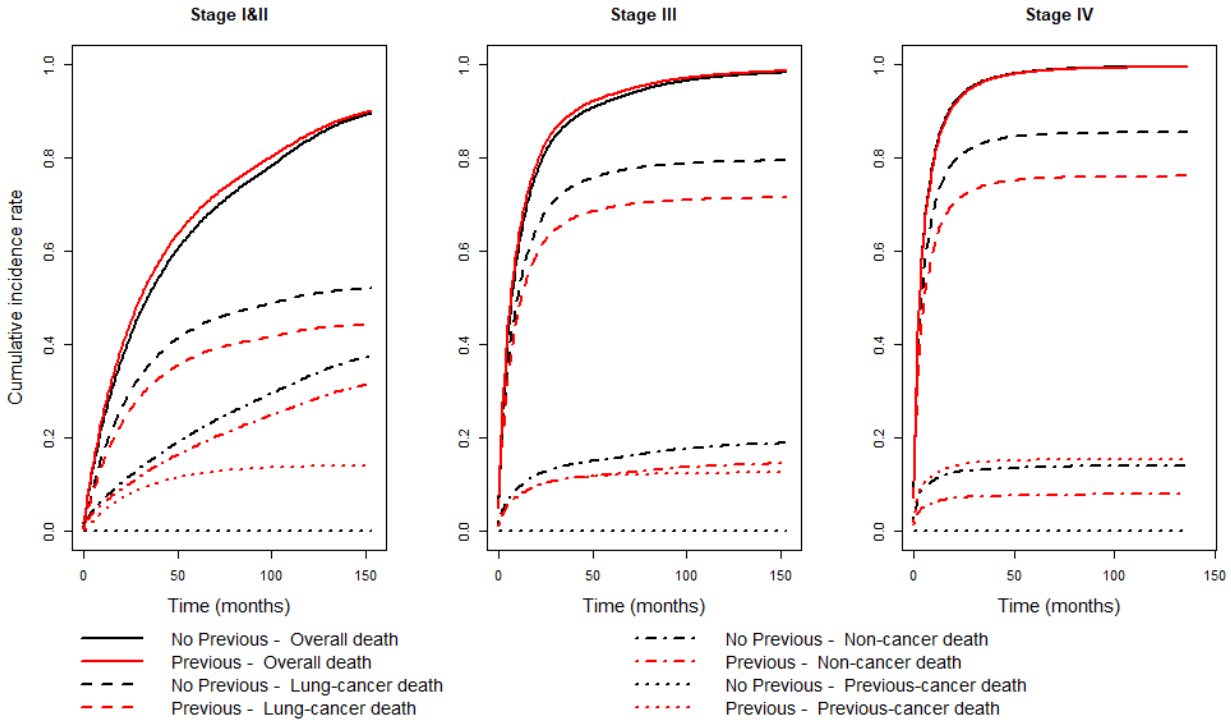

Figure 3-1. Cumulative incidence rate of death by stage and cause of death in the three-cause data for lung cancer patients with and without previous cancer.

Table 3-3. Estimated mean lead time (months) and odds ratios by cause of death and stage in the three-cause competing risk data.

Lung cancer stage	Mean lead time (95% CI)	Lung cancer mortality OR_l (95% CI)	Non-cancer mortality OR_{nc} (95% CI)
	Assuming lead time		
I&II	0.53 (0.35, 0.72)	0.89 (0.86, 0.93)	0.89 (0.84, 0.94)
III	0.96 (0.40, 1.53)	0.95 (0.91, 0.99)	0.83 (0.77, 0.89)
IV	0.48 (0.03, 0.94)	0.88 (0.83, 0.92)	0.56 (0.51, 0.60)
	Assuming no lead time		
I&II	0	0.88 (0.85, 0.92)	0.86 (0.82, 0.91)
III	0	0.89 (0.87, 0.92)	0.84 (0.78, 0.89)
IV	0	0.82 (0.80, 0.84)	0.66 (0.62, 0.70)

OR_l: Odds ratio of the lung cancer mortality hazard of the Previous group relative to the No-Previous group.
OR_{nc}: Odds ratio of the non-cancer mortality hazard of the Previous group relative to the No-Previous group.

3.4 Discussion

Applying our flexible, discrete-data competing-risk model to SEER-Medicare lung cancer data, we observed that estimated mean lead times are modest — less than one month in all

stages. Regardless of adjustment for lead-time bias, the results in the three-cause data demonstrated that the hazards of both lung cancer death and non-cancer death are moderately less among patients with previous cancer compared to those with no previous cancer. The two-cause data further revealed that the competing risk of other causes accounted for the reduction of hazard of lung-cancer death. Failure to adjust for lead-time bias results in modest underestimation of odds ratios, but as the mean lead time is small, so is the effect of estimating it on the OR parameters.

The estimates of mean lead time differ from those in our previous analysis,[52] which ignored the competing-risks aspect of the data. Our earlier estimates of the mean lead time for lung cancer survival were 11.3, 1.1, and 0.3 months for stages I&II, III and IV, respectively, whereas under the three-cause model the corresponding estimates are 0.53, 0.96 and 0.48 months. We conjecture that these discrepancies reflect a bias from ignoring competing risks. The largest discrepancy is in stage I&II, where the percent of censoring from competing risks is highest, and therefore there is the greatest opportunity for bias. Another possible explanation is that in the data set for our competing-risks analysis we excluded patients with multiple previous tumors and those who developed a second tumor after the index lung cancer. These exclusions were necessary to ensure the accuracy of measuring previous-cancer death. The excluded patients, especially those with multiple previous tumors, could have had long lead times that would have strongly influenced estimates in the original analysis. The large difference between these results indicates that competing risks are a likely source of bias in future studies about the survival of patients with multiple cancers.

We estimated the lung cancer lead time to be short. In coming years, as lung cancer screening becomes more routine, lead time for lung cancer in patients with previous cancers may

increase. In patients with previous cancer who go on to develop other cancer types with recommended early detection methods, such as breast, cervical, and colorectal cancer, lead times may be larger. As our proposed model handles bias from both competing risks and lead time, we advocate its application in future studies about the prognosis of cancer patients with multiple cancers.

Several studies have applied the Fine-Gray[18] competing-risk model to analyze the mortality of cancer patients with a history of previous cancer.[61,62] To our knowledge, none has proposed a discrete competing-risk model and none has further adjusted for lead-time bias.

Our study has some limitations. First, SEER records a limited set of baseline variables, and the absence of possible confounders may bias our estimates of the odds ratios on the cause-specific hazards. Most prominently, SEER does not include smoking status, which is associated with lung cancer, other cancers, and survival. Second, to reduce confounding, we applied our model to matched data. This simplified computations but prevented us from assessing the effects of these variables on mortality. Future studies could consider hazard models that include these predictors, both as a way to better describe mortality and to exploit all available observations. Third, we did not differentiate the types of previous cancers. Clearly, the previous cancer type is a powerful predictor of survival; early-stage breast cancer has a far better prognosis than, say, advanced pancreatic cancer.[62,63] Finally, our analysis relies on assumptions regarding the form of the joint distribution of the latent lead time and post-lead time survival. Previous analyses with a similar model, however, suggested that conclusions are only moderately sensitive to these untestable assumptions.[52]

In conclusion, under a discrete competing-risk model, the estimated mean lead time is less than one month for all stages of lung cancer. Both with and without adjustment for lead time, the

Previous group had a significantly lower hazard for both lung cancer and non-cancer mortality. Evidently, the different survival outcomes seen in the Previous and No-Previous groups represent true differences in mortality, and not a lead-time bias.

The number of cancer survivors in the U.S. is large and growing; as life expectancy for this population increases,[63] the U.S. will face a rising number of patients diagnosed with multiple primary cancers. Careful consideration of the prevalence and impact of multiple primary cancers on cancer outcomes is needed to ensure accurate estimation of mortality in descriptive cancer epidemiology. More importantly exclusion criteria in lung cancer clinical trials, which frequently prevent participation of patients with previous cancer,[2] should be carefully reconsidered in light of the observed survival advantage for lung cancer and non-cancer cause of death for this large, growing population of newly diagnosed patients with previous cancers.

BIBOLOGRAPHY

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. *CA: a cancer journal for clinicians.* 2018;68(1):7-30.

2. Laccetti AL, Pruitt SL, Xuan L, Halm EA, Gerber DE. Effect of prior cancer on outcomes in advanced lung cancer: implications for clinical trial eligibility and accrual. *JNCI: Journal of the National Cancer Institute.* 2015;107(4).

3. Laccetti AL, Pruitt SL, Xuan L, Halm EA, Gerber DE. Prior cancer does not adversely affect survival in locally advanced lung cancer: a national SEER-medicare analysis. *Lung Cancer.* 2016;98:106-113.

4. Pruitt SL, Laccetti AL, Xuan L, Halm EA, Gerber DE. Revisiting a longstanding clinical trial exclusion criterion: impact of prior cancer in early-stage lung cancer. *British journal of cancer.* 2017;116(6):717.

5. Gerber DE, Laccetti AL, Xuan L, Halm EA, Pruitt SL. Impact of prior cancer on eligibility for lung cancer clinical trials. *Journal of the National Cancer Institute.* 2014;106(11):dju302.

6. Aguiló R, Macià F, Porta M, Casamitjana M, Minguella J, Novoa AM. Multiple independent primary cancers do not adversely affect survival of the lung cancer patient. *European Journal of Cardio-Thoracic Surgery.* 2008;34(5):1075-1080.

7. Duchateau CS, Stokkel MP. Second primary tumors involving non-small cell lung cancer: prevalence and its influence on survival. *Chest.* 2005;127(4):1152-1158.

8. Utsumi T, Fujii Y, Takeda S-i, et al. Clinical study on lung cancer as a second primary cancer. *Surgery today.* 1998;28(5):487-491.

9. Day NE, Walter SD. Simplified models of screening for chronic disease: estimation procedures from mass screening programmes. *Biometrics.* 1984:1-13.

10. Prorok PC. Bounded recurrence times and lead time in the design of a repetitive screening program. Journal of Applied Probability. 1982;19(1):10-19.

11. Straatman H, Peer PG, Verbeek AL. Estimating lead time and sensitivity in a screening program without estimating the incidence in the screened group. Biometrics. 1997:217-229.

12. Draisma G, Boer R, Otto SJ, et al. Lead times and overdetection due to prostate-specific antigen screening: estimates from the European Randomized Study of Screening for Prostate Cancer. Journal of the National Cancer Institute. 2003;95(12):868-878.

13. Shapiro S, Goldberg JD, Hutchison GB. Lead time in breast cancer detection and implications for periodicity of screening. American Journal of Epidemiology. 1974;100(5):357-366.

14. Wu D, Erwin D, Rosner GL. Sojourn time and lead time projection in lung cancer screening. *Lung Cancer.* 2011;72(3):322-326.

15. Walter S, Stitt L. Evaluating the survival of cancer cases detected by screening. *Statistics in Medicine.* 1987;6(8):885-900.

16. Efron B. Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American statistical Association.* 1988;83(402):414-425.

17. Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing.* 1995;16(5):1190-1208.

18. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association.* 1999;94(446):496-509.

19. Xu JL, Fagerstrom RM, Prorok PC. Estimation of post-lead-time survival under dependence between lead-time and post-lead-time survival. *Statistics in medicine.* 1999;18(2):155-162.

20. Xu JL, Prorok PC. Non-parametric estimation of the post-lead-time survival distribution of screen-detected cancer cases. *Statistics in medicine.* 1995;14(24):2715-2725.

21. Duffy SW, Nagtegaal ID, Wallis M, et al. Correcting for lead time and length bias in estimating the effect of screen detection on cancer survival. *American journal of epidemiology.* 2008;168(1):98-104.

22. Heitjan DF. Inference from grouped continuous data: a review. *Statistical science.* 1989:164-179.

23. Rubin DB. *Multiple imputation for nonresponse in surveys.* Vol 81: John Wiley & Sons; 2004.

24. Institute NC. SEER cancer stats facts: lung and bronchus cancer. https://seer.cancer.gov/statfacts/html/lungb.html. Accessed October 30, 2017.

25. Murphy CC, Gerber DE, Pruitt SL. Prevalence of prior cancer among persons newly diagnosed with cancer: an initial report from the Surveillance, Epidemiology, and End Results Program. *JAMA oncology.* 2017.

26. Bluethmann SM, Mariotto AB, Rowland JH. Anticipating the "Silver Tsunami": Prevalence Trajectories and Comorbidity Burden among Older Cancer Survivors in the United States. *Cancer Epidemiology Biomarkers & Prevention.* 2016;25(7):1029-1036.

27. Howlader N, Noone A, Krapcho M, Miller D, Bishop K, Altekruse S. SEER Cancer Statistics Review, 1975-2015, National Cancer Institute. Bethesda, MD: National Cancer Institute; 2015. 2015.

28. Oncology ASoC. Institute of Medicine. From cancer patient to cancer survivor: lost in transition. The National Academies Press; 2005.

29. Page AE, Adler NE. *Cancer care for the whole patient: Meeting psychosocial health needs.* National Academies Press; 2008.

30. Center NCSR. Policy and advocacy recommendation. https://www.cancer.org/health-care-professionals/national-cancer-survivorship-resource-center.html.

31. Mehnert A, de Boer A, Feuerstein M. Employment challenges for cancer survivors. *Cancer.* 2013;119(S11):2151-2159.

32. Hartigan JA, Wong MA. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society Series C (Applied Statistics).* 1979;28(1):100-108.

33. Health NIo. National Cancer Institute SEER Training Modules. 2016.

34. Services TDoSH. 2016 cancer reporting handbook. https://www.dshs.texas.gov/tcr/training/2016-handbook.aspx.

35. Morgan MA. Cancer survivorship: history, quality-of-life issues, and the evolving multidisciplinary approach to implementation of cancer survivorship care plans. Paper presented at: Oncology Nursing Forum2009.

36. Hewitt M, Ganz P. Implementing cancer survivorship care planning: workshop summary. 2007.

37. Halpern MT, Viswanathan M, Evans TS, Birken SA, Basch E, Mayer DK. Models of cancer survivorship care: overview and summary of current evidence. *Journal of oncology practice.* 2014;11(1):e19-e27.

38. Wattchow DA, Weller DP, Esterman A, et al. General practice vs surgical-based follow-up for patients with colon cancer: randomised controlled trial. *British journal of cancer.* 2006;94(8):1116.

39. Knowles G, Sherwood L, Dunlop MG, et al. Developing and piloting a nurse-led model of follow-up in the multidisciplinary management of colorectal cancer. *European Journal of Oncology Nursing.* 2007;11(3):212-223.

40. Grunfeld E, Julian JA, Pond G, et al. Evaluating survivorship care plans: results of a randomized, clinical trial of patients with breast cancer. *Journal of Clinical Oncology.* 2011;29(36):4755-4762.

41. Sutradhar R, Agha M, Pole JD, et al. Specialized survivor clinic attendance is associated with decreased rates of emergency department visits in adult survivors of childhood cancer. *Cancer.* 2015;121(24):4389-4397.

42. Cannon AJ, Darrington DL, McIlvain HE, et al. Association of number of follow-up providers with outcomes in survivors of hematologic malignancies. *Leukemia & lymphoma.* 2010;51(10):1862-1869.

43. Lash RS, Bell JF, Reed MSC, et al. A systematic review of emergency department use among cancer patients. *Cancer nursing.* 2017;40(2):135.

44. Hsu J, Donnelly JP, Moore JX, Meneses K, Williams G, Wang HE. National characteristics of Emergency Department visits by patients with cancer in the United States. *The American journal of emergency medicine.* 2018.

45. Rivera DR, Gallicchio L, Brown J, Liu B, Kyriacou DN, Shelburne N. Trends in Adult Cancer–Related Emergency Department Utilization: An Analysis of Data From the Nationwide Emergency Department Sample. *JAMA oncology.* 2017;3(10):e172450-e172450.

46. Panattoni L, Fedorenko C, Greenwood-Hickman MA, et al. Characterizing Potentially Preventable Cancer-and Chronic Disease–Related Emergency Department Use in the Year After Treatment Initiation: A Regional Study. *Journal of oncology practice.* 2018;14(3):e176-e185.

47. Kokko R, Hakama M, Holli K. Follow-up cost of breast cancer patients with localized disease after primary treatment: a randomized trial. *Breast cancer research and treatment.* 2005;93(3):255-260.

48. Draisma G, Etzioni R, Tsodikov A, et al. Lead Time and Overdiagnosis in Prostate-Specific Antigen Screening: Importance of Methods and Context. *JNCI: Journal of the National Cancer Institute.* 2009;101(6):374-383.

49. Hutchison GB, Shapiro S. Lead Time Gained by Diagnostic Screening for Breast Cancer23. *JNCI: Journal of the National Cancer Institute.* 1968;41(3):665-681.

50. Telesca D, Etzioni R, Gulati R. Estimating lead time and overdiagnosis associated with PSA screening from prostate cancer incidence trends. *Biometrics.* 2008;64(1):10-19.

51. Törnblom M, Eriksson H, FRANZen S, et al. Lead time associated with screening for prostate cancer. *International journal of cancer.* 2004;108(1):122-129.

52. Ge Z, Heitjan DF, Gerber DE, Xuan L, Pruitt SL. Estimating lead-time bias in lung cancer diagnosis of patients with previous cancers. *Statistics in medicine.* 2018:1-14.

53. Tsiatis A. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences.* 1975;72(1):20-22.

54. Austin PC, Lee DS, Fine JP. Introduction to the analysis of survival data in the presence of competing risks. *Circulation.* 2016;133(6):601-609.

55. Cornfield J. Estimation of the probability of developing a disease in the presence of competing risks. *American Journal of Public Health and the Nations Health.* 1957;47(5):601-607.

56. Prentice RL, Kalbfleisch JD, Peterson Jr AV, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. *Biometrics.* 1978:541-554.

57. Austin PC, Fine JP. Accounting for competing risks in randomized controlled trials: a review and recommendations for improvement. *Statistics in medicine.* 2017;36(8):1203-1209.

58. Tutz G. Competing risks models in discrete time with nominal or ordinal categories of response. *Quality and Quantity.* 1995;29(4):405-420.

59. Ambrogi F, Biganzoli E, Boracchi P. Estimating crude cumulative incidences through multinomial logit regression on discrete cause-specific hazards. *Computational Statistics & Data Analysis.* 2009;53(7):2767-2779.

60. Quekel LG, Kessels AG, Goei R, van Engelshoven JM. Miss rate of lung cancer on the chest radiograph in clinical practice. *Chest.* 1999;115(3):720-724.

61. Mirabeau-Beale K, Chen M-H, D'Amico AV. Prior-cancer diagnosis in men with nonmetastatic prostate cancer and the risk of prostate-cancer-specific and all-cause mortality. *ISRN oncology.* 2014;2014.

62. Dinh KT, Mahal BA, Ziehr DR, et al. Risk of prostate cancer mortality in men with a history of prior cancer. *BJU international.* 2016;117(6B).

63. Van Hemelrijck M, Drevin L, Holmberg L, Garmo H, Adolfsson J, Stattin P. Primary cancers before and after prostate cancer diagnosis. *Cancer.* 2012;118(24):6207-6216.

APPENDIX

Table S1. Characteristics of patients with stage III lung cancer.

	n (%)			P value	
	Previous	No-Previous	Matched No-Previous	Unmatched	Matched
Total (*n*)	6250	43834	6250		
Age				<0.0001	0.65
<75	2383 (38.1)	20128 (45.9)	2391 (38.3)		
75–85	2996 (47.9)	18667 (42.6)	3023 (48.4)		
≥85	871 (13.9)	5039 (11.5)	836 (13.4)		
Sex				<0.0001	0.93
Female	2424 (38.8)	20586 (47.0)	2429 (38.9)		
Male	3826 (61.2)	23248 (53.0)	3821 (61.1)		
Race				0.00084	0.76
White	5415 (86.6)	37360 (85.2)	5455 (87.3)		
Black	550 (8.8)	3975 (9.1)	520 (8.3)		
Hispanic	44 (0.7)	487 (1.1)	42 (0.7)		
Other	241 (3.9)	2012 (4.6)	233 (3.7)		
Marital Status				<0.0001	0.92
Married	3371 (53.9)	21471 (49.0)	3377 (54.0)		
Sep/div/wid	2278 (36.4)	17626 (40.2)	2257 (36.1)		
Single	406 (6.5)	3162 (7.2)	424 (6.8)		
Unknown	195 (3.1)	1575 (3.6)	192 (3.1)		
Histology				<0.0001	0.56
Adenocarcinoma	2264 (36.2)	13731 (31.3)	2244 (35.9)		
Squamous	1535 (24.6)	11095 (25.3)	1484 (23.7)		
Small cell	785 (12.6)	6288 (14.3)	819 (13.1)		
NSCLS/other	1666 (26.7)	12720 (29.0)	1703 (27.2)		
Charlson Score				<0.0001	0.66
0	2369 (37.9)	16408 (37.4)	2405 (38.5)		
1	1851 (29.6)	12864 (29.3)	1849 (29.6)		
2+	1813 (29.0)	11957 (27.3)	1802 (28.8)		
Unavailable	217 (3.5)	2605 (5.9)	194 (3.1)		
Medicaid				<0.0001	0.40
Yes	877 (14.0)	8646 (19.7)	910 (14.6)		
No	5373 (86.0)	35188 (80.3)	5340 (85.4)		

Lung cancer treatment				0.028	0.88
Surgery only	326 (5.2)	2184 (5.0)	340 (5.4)		
Chemotherapy only	710 (11.4)	4451 (10.2)	702 (11.2)		
Radiation only	984 (15.7)	6782 (15.5)	949 (15.2)		
≥2 treatments	2170 (34.7)	15580 (35.5)	2170 (34.7)		
No surg/chemo/rad	2060 (33.0)	14837 (33.8)	2089 (33.4)		

Table S2. Characteristics of patients with stage IV lung cancer.

	n (%)			P value	
	Previous	No-Previous	Matched No-Previous	Unmatched	Matched
Total (*n*)	9715	70842	9715		
Age				<0.0001	0.88
<75	3822 (39.3)	33932 (47.9)	3820 (39.3)		
75–85	4651 (47.9)	29571 (41.7)	4649 (47.9)		
≥85	1242 (12.8)	7339 (10.4)	1246 (12.8)		
Sex				<0.0001	0.17
Female	3622 (37.3)	33233 (46.9)	3591 (37.0)		
Male	6093 (62.7)	37609 (53.1)	6124 (63.0)		
Race				0.00028	0.99
White	8387 (86.3)	60596 (85.5)	8410 (86.6)		
Black	859 (8.8)	6055 (8.5)	830 (8.5)		
Hispanic	94 (1.0)	846 (1.2)	94 (1.0)		
Other	375 (3.9)	3345 (4.7)	381 (3.9)		
Marital Status				<0.0001	0.77
Married	5284 (54.4)	34619 (48.9)	5290 (54.5)		
Sep/div/wid	3454 (35.6)	27864 (39.3)	3453 (35.5)		
Single	656 (6.8)	5814 (8.2)	665 (6.8)		
Unknown	321 (3.3)	2545 (3.6)	307 (3.2)		
Histology				<0.0001	0.98
Adenocarcinoma	3336 (34.3)	22742 (32.1)	3296 (33.9)		
Squamous	1530 (15.7)	10239 (14.5)	1532 (15.8)		
Small cell	1668 (17.2)	13212 (18.6)	1641 (16.9)		
NSCLS/other	3181 (32.7)	24649 (34.8)	3246 (33.4)		
Charlson Score				<0.0001	0.72
0	3946 (40.6)	28555 (40.3)	3991 (41.1)		
1	2772 (28.5)	19864 (28.0)	2763 (28.4)		
2+	2649 (27.3)	17371 (24.5)	2592 (26.7)		
Unavailable	348 (3.6)	5052 (7.1)	369 (3.8)		
Medicaid				<0.0001	1.0
Yes	1372 (14.1)	13365 (18.9)	1378 (14.2)		
No	8343 (85.9)	57477 (81.1)	8337 (85.8)		

Lung cancer treatment				<0.0001	0.86
Surgery only	150 (1.5)	909 (1.3)	151 (1.6)		
Chemotherapy only	1502 (15.5)	9948 (14.0)	1445 (14.9)		
Radiation only	2058 (21.2)	15142 (21.4)	2093 (21.5)		
≥2 treatments	2261 (23.3)	16143 (22.8)	2275 (23.4)		
No surg/chemo/rad	3744 (38.5)	28700 (40.5)	3751 (38.6)		

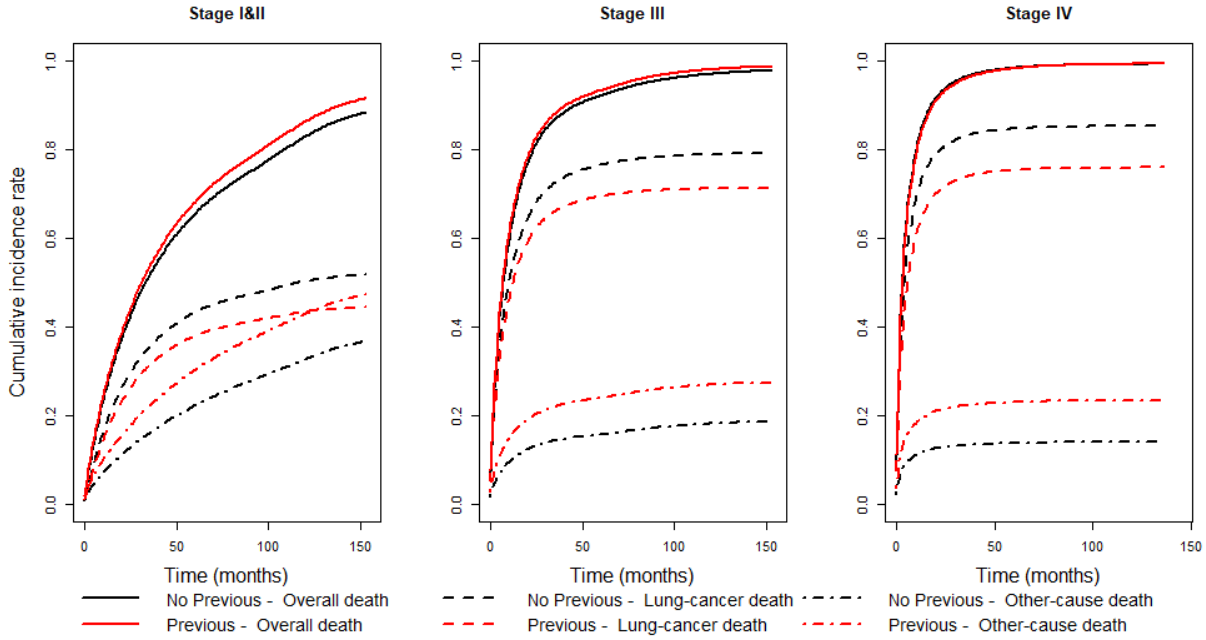

Figure S1. Cumulative incidence rate by stage and cause of death in the two-cause data for lung cancer patients with and without previous cancer.

Table S3. Estimated mean lead time (months) and odds ratios by cause of death and stage in the two-cause competing risk data.

Lung cancer stage	Mean lead time (95% CI)	OR_l (95% CI)	OR_o (95% CI)
	Assuming lead time		
I&II	0.60 (0.11, 1.10)	0.90 (0.87, 0.94)	1.41 (1.34, 1.48)
III	0.76 (0, 1.55)	0.94 (0.89, 0.99)	1.59 (1.48, 1.71)
IV	0.43 (0.09, 0.77)	0.87 (0.83, 0.91)	1.64 (1.54, 1.74)
	Assuming no lead time		
I&II	0	0.88 (0.85, 0.92)	1.39 (1.33, 1.46)
III	0	0.89 (0.86, 0.92)	1.51 (1.43, 1.60)
IV	0	0.82 (0.80, 0.84)	1.56 (1.49, 1.64)

OR_l: Odds ratio of the lung cancer mortality hazard of the Previous group relative to the No-Previous group.

OR_o: Odds ratio of the other (non-cancer and previous cancer) mortality hazard of the Previous group relative to the No-Previous group.